Identifying Cellular Niches in Spatial Transcriptomics: An Investigation into the Capabilities of Large Language Models

Huanhuan Wei[♡], **Xiao Luo**[♠], **Hongyi Yu**[♡], **Jinping Liang**[♡], **Luning Yang**[♡], **Lixing Lin**[♡], **Alexandra Popa**[♢], **Xiting Yan**[♡♣†]

[♡] Section of Pulmonary, Critical Care and Sleep Medicine,Yale School of Medicine, New Haven, CT, USA [♣] Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA [♠] University of California, Los Angeles, CA, USA
[♢] Boehringer Ingelheim RCV GmbH & Co KG, Doktor-Boehringer-Gasse 5-11, Vienna 1120, Austria.

xiting.yan@yale.edu

Abstract

Spatial transcriptomic technologies enable measuring gene expression profile and spatial information of cells in tissues simultaneously. Clustering of captured cells/spots in the spatial transcriptomic data is crucial for understanding tissue niches and uncovering disease-related changes. Current methods to cluster spatial transcriptomic data encounter obstacles, including inefficiency in handling multi-replicate data, lack of prior knowledge incorporation, and producing uninterpretable cluster labels. We introduce a novel approach, LLMiniST[1], to identify spatial niche using a zero-shot large language models (LLMs) by transforming spatial transcriptomic data into spatial context prompts, leveraging gene expression of neighboring cells/spots, cell type composition, tissue information, and external knowledge. The model was further enhanced using a two-stage fine-tuning strategy for improved generalizability. We also develop a user-friendly annotation tool[2] to accelerate the creation of well-annotated spatial dataset for fine-tuning. Comprehensive method performance evaluations showed that both zero-shot and fine-tunned LLMiniST had superior performance than current non-LLM methods in many circumstances. Notably, the two-stage fine-tuning strategy facilitated substantial cross-subject transferability. The results demonstrate the feasibility of LLMs for tissue niche identification using spatial transcriptomic data and the potential of LLMs as a scalable solution to efficiently integrate minimal human guidance for improved performance across large-scale datasets.

1 Introduction

Spatial transcriptomic technologies have enabled us to profile gene expression and preserve the spatial location of cells/spots within intact tissues (Moffitt et al., 2022) simultaneously. A common and critical analytical task for such data is to identify spatial niches that are higher-order tissue structures. This task, often termed spatial clustering, is fundamental to construct spatial atlases (Zeng et al., 2023) and plays a pivotal role in visualizing tissue anatomy, inferring spatial continuity, detecting niche-specific marker genes (Cable et al., 2022), uncovering spatial signatures of development and disease (Elhanani et al., 2023), and identifying molecular regulatory networks within distinct niches (Vandereyken et al., 2023).

Current spatial clustering techniques perform unsupervised clustering on spatial transcriptomic data to delineate distinct spatial niches within tissues. A prevalent strategy uses graph neural networks (GNNs) (Wu et al., 2021) to aggregate spatial gene expression profiles. This process yields low-dimensional embeddings that capture both gene expression and spatial context, which are then subjected to clustering algorithms (Dong and Zhang, 2022; Long et al., 2023; Ren et al., 2022; Zong et al., 2022). However, existing inductive spatial clustering methods face several limitations. First, they typically require training a new model from scratch for each dataset, relying solely on data without incorporating prior knowledge (Liu et al., 2024a). Second, they often struggle with data with replicates (Yuan et al., 2024), necessitating analysis for each replicate separately and thus hindering the identification of consistent niches across different subjects. Finally, current methods typically output uninterpretable cluster labels, requiring further analysis to determine their biological relevance.

In this paper, we introduce Large Language Model for identifying niches in Spatial

[1]The source code of our implementation can be found at https://github.com/wJDKnight/LLMiniST.

[2]The source code of the software can be found at https://github.com/wJDKnight/draw_spatial.

[†] Corresponding authors.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9275–9289

Transcriptomics (LLMiniST), a novel approach that identifies spatial niches using LLMs. We converts spatial transcriptomic data into spatial context prompts that effectively encode spatial information for the LLM to interpret. These prompts integrate gene expression profiles from spatially neighboring cells and cell type composition, along with basic tissue section information and external knowledge about established niches within the given tissue type. We explore the application of LLMs in two ways: zero-shot prompting and a dedicated two-stage fine-tuning approach. The zero-shot prompting directly applied trained LLMs while the two-stage fine-tuning approach fine-tunes the LLMs using human guidance to generalize to new, unseen data. The latter approach enhances the robustness and applicability of LLMiniST to a wider range of spatial transcriptomic datasets. LLMiniST offers a new paradigm for spatial niche identification, moving beyond purely data-driven approaches towards a more context-aware and knowledge-informed analysis.

In comparison to non-LLM clustering methods, applying large language models (LLMs) to identify spatial niche types in spatial transcriptomics data presents unique challenges. First, methodology for converting complex spatial transcriptomic information into effective LLM prompt is not well established (Sahoo et al., 2024). This involves translating high-dimensional gene expression patterns and spatial coordinates into a format that leverages the language processing capabilities of the LLM. Second, LLMs rely heavily on the knowledge for pre-training, which may be insufficient or inconsistent with the specific biological context of a given spatial transcriptomic dataset. The inherent noise and variability in gene expression data further complicate this issue (Vandereyken et al., 2023). Finally, the inherent heterogeneity across biological samples poses a significant hurdle, potentially causing the biological "truth" derived from the LLM's knowledge not universally applicable and leading to inaccurate or incomplete niche type identification in subjects different from those in the knowledge data for pre-training.

This study embarks on a comprehensive evaluation of LLMiniST to determine its efficacy in identifying spatial niches using diverse spatial transcriptomic datasets. First, we find the pre-trained knowledge within general LLMs well aligns with the biological principles governing spatial niche formation and organization. Second, by leveraging a limited set of pathologist-annotated samples from a single subject, we can fine-tune an LLM, creating a specialized model capable of accurately delineating spatial niches across the remaining unannotated tissue sections from the same subject. Third, a fine-tuned LLM generalize well to data from other subjects that have similar tissue architectures, providing broadly applicable models for spatial niche identification.

The contributions of this work are summarized as follows: (1) We are the first to address the use of LLMs for niche identification in spatial transcriptomics. (2) We demonstrate the potential of LLMs in this area, achieving comparable or superior performance in some instances, while also highlighting limitations when pre-trained knowledge conflicts with real data and providing a two-stage fine-tuning solution for it. (3) To address the data scarcity bottleneck and promote the widespread adoption of our fine-tuning approach, we have developed a user-friendly graphic software to empower researchers to efficiently generate high-quality, ground-truth annotated datasets. We believe this work can pave the way for future advancements in applying LLMs to spatial transcriptomic analysis.

2 Related Work

2.1 Large Language Models

LLMs have demonstrated remarkable performance across various topics, leading to increasing efforts in exploring their potential in specialized fields such as single-cell transcriptomics (Bian et al., 2024; Chen and Zou, 2024; Hou and Ji, 2024). Prompt engineering has emerged as a crucial technique for extending LLMs' capabilities (Sahoo et al., 2024). Parameter Efficient Fine-Tuning (PEFT) offers solutions for adapting these models to specific tasks while minimizing computational overhead (Han et al., 2024; Mao et al., 2024). LLMs also show a promising potential on graph-related task (Wang et al., 2024; Li et al., 2024). Motivated by such advancements, we explored the potential of LLMs to identify spatial niches using spatial transcriptomic data.

2.2 Spatial Clustering Methods

Numerous spatial clustering methods have been developed to identify spatial niches in spatial transcriptomic data. These methods exploit the spatial locations of cells in various manners to enhance

clustering accuracy (Pham et al., 2023; Zhao et al., 2021). A particularly popular approach is to represent the spatial transcriptomic data using graphs. Graph neural networks (GNNs) (Liu et al., 2024b; Li et al., 2022; Ren et al., 2022; Zong et al., 2022; Long et al., 2023) and its variants like graph convolutional networks (GCNs) (Hu et al., 2021) and graph attention networks (GAT) (Dong and Zhang, 2022), can be adopted for learning cell representations from graphs. These learned representations are then utilized in downstream clustering tasks. Given the success of GNNs in effectively encoding spatial information for representation learning, we adopt a similar paradigm to construct spatially-aware prompts. .

3 Method: LLMiniST

3.1 Problem Definition

In this work, our objective is to infer the niche label of each cell/spot [3] in a spatial transcriptomic dataset, leveraging gene expression, cell type information (if available), and spatial location. Suppose the spatial transcriptomic dataset $\mathcal{D}$ comprises data from P subjects. We denote the data for subject p as $\mathcal{D}_p$, thus $\mathcal{D} = \{\mathcal{D}_p\}_{p=1}^{P}$. For each subject p, the data $\mathcal{D}_p$ consists of M_p tissue sections (replicates), represented as $\mathcal{D}_p = \{\mathbf{R}_{p,m}\}_{m=1}^{M_p}$, where $\mathbf{R}_{p,m}$ denotes the m-th replicate of subject p. Each replicate $\mathbf{R}_{p,m}$ comprises the following components:

1) Gene Expression Matrix: $\mathbf{X}_{p,m} \in \mathbb{R}^{N_{p,m} \times G}$, where $N_{p,m}$ is the number of cells and G represents the number of genes. We assume a common set of genes is measured across all replicates and subjects for simplicity.

2) Spatial Coordinates: $\mathbf{S}_{p,m} \in \mathbb{R}^{N_{p,m} \times D}$, where D is the dimensionality of the spatial coordinates (typically $D \in \{2, 3\}$).

3) Cell Type Annotations (except for Visium): $\mathbf{c}_{p,m} \in \{1, 2, \ldots, C_p\}^{N_{p,m}}$, where C_p is the number of cell types of subject p.

4) Niche Annotations (Optional): $\mathcal{A}_{p,m} \subseteq \{1, 2, \ldots, N_{p,m}\}$ denoting the indices of cells with such annotations. The niche labels are given by $\mathbf{n}_{p,m} \in \{1, 2, \ldots, K\}^{|\mathcal{A}_{p,m}|}$, where K is the number of distinct niches. Each element $n_{p,m,i}$ indicates the niche label assigned to cell $i \in \mathcal{A}_{p,m}$. If no niche annotations are available, then $\mathcal{A}_{p,m} = \emptyset$ and $\mathbf{n}_{p,m}$ are undefined.

Given a replicate $\mathbf{R}_{p,m} = (\mathbf{X}_{p,m}, \mathbf{S}_{p,m}, \mathbf{c}_{p,m}, (\mathcal{A}_{p,m}, \mathbf{n}_{p,m}))$, we aim to infer the niche label for each cell in this replicate or the whole spatial transcriptomic dataset.

3.2 Spatial Context Prompt Engineering

We assessed two approaches for leveraging LLMs to identify niches using spatial transcriptomic data: the zero-shot approach (LLMiniST-Z) and the fine-tuning approach (LLMiniST-F). An overview of this framework is presented in Figure 1.

LLMiniST-Z uses $\mathbf{R}_{p,m} = (\mathbf{X}_{p,m}, \mathbf{S}_{p,m}, \mathbf{c}_{p,m})$ to construct a spatial context prompt for cell i, which is directly input into an LLM to predict its niche type n_i[4]. The spatial context prompt employed in this approach comprises three components: (i) task description, (ii) spatial profile, and (iii) response format.

- **Task Description:** This component establishes the objective, specifies the tissue region and potential niches for all cells, and defines the input format.
- **Spatial Profile:** The spatial profile for cell i is constructed by first defining its neighborhood $\mathcal{N}_i$. A cell j is considered a neighbor of cell i (i.e., $j \in \mathcal{N}_i$) if their spatial distance $d(\mathbf{s}_i, \mathbf{s}_j)$ is less than a predefined threshold δ. For each cell i and its neighborhood $\mathcal{N}_i$, the spatial profile is characterized by two ordered lists:
  1. **Neighbor Cell Type List:** Ordered by the frequency of each cell type t in neighborhood $\mathcal{N}_i$ defined as: $f_{i,t} = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \mathbb{I}(c_j = t)$, where $\mathbb{I}(\cdot)$ is the indicator function, equal to 1 if the condition is true and 0 otherwise. The cell type list is sorted in descending order of $f_{i,t}$.
  2. **Neighbor Marker Gene List:** Ordered by the average expression level of each marker gene g in neighborhood $\mathcal{N}_i$ calculated as: $\text{expr}_{i,g} = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} x_{j,g}$. The list of marker genes is sorted in descending order of $\text{expr}_{i,g}$.
- **Response Format:** To facilitate convenient downstream processing, we require LLMs to output only the most probable niche, without explanation. However, users can optionally

[3] Hereafter, we use "cells" to refer to both cells and spots, unless there is a need to distinguish between them.

[4] For clarity and simplicity, we omit the subscripts p and m when the context makes the meaning unambiguous.

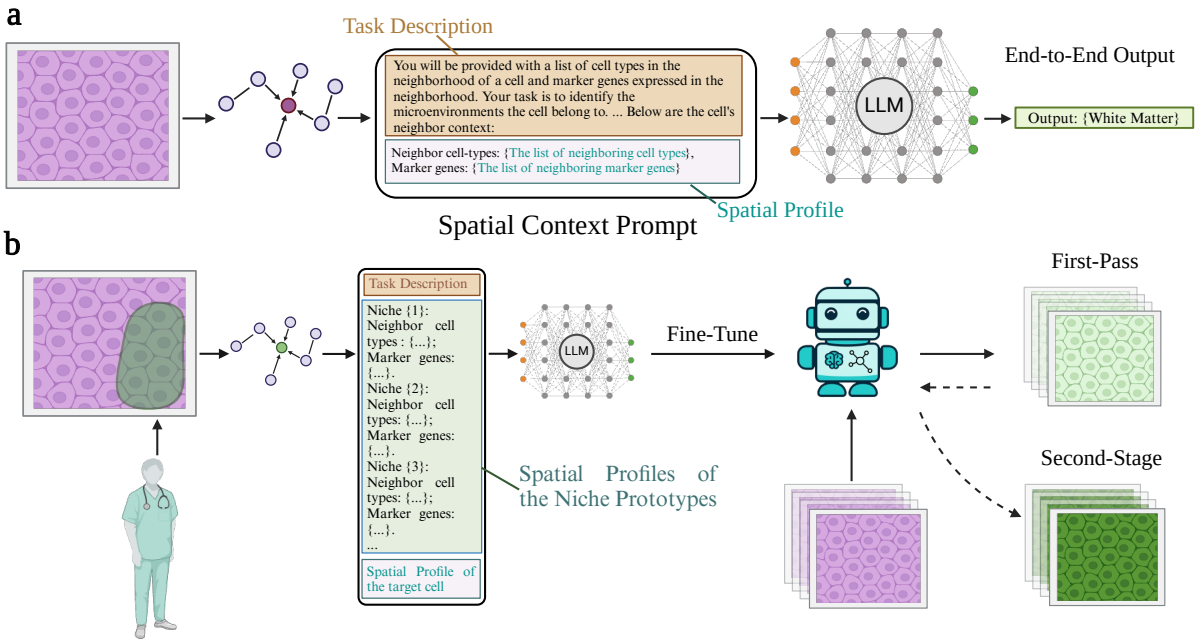

Figure 1: An overview of the pipeline of (**a**) zero-shot approach and (**b**) fine-tuning approach.

request detailed reasoning steps for improved interpretability (Examples in Appendix A).

LLMiniST-F leverages cells with label to construct example prompts and responses to fine-tune the LLM. While the prompts share the same structural components as those in LLMiniST-Z, the task description within LLMiniST-F is augmented to include the spatial profiles of niche prototypes generated as follows. The mean cell type frequency and the mean marker gene expression for niche e is calculated by averaging the $f_{p,m,i,t}$ and $\text{expr}_{p,m,i,g}$ of the cells belongs to niche e. The spatial profile of niche prototype is generated similarly to that of a cell. The augmented prompts of the cells with labels are used to fine-tune the LLM. When identifying cells from different samples or subjects, we proposed a two-stage strategy to adjust the difference between samples or subjects (referred as LLMiniST-Fs). In the first stage, an initial fine-tuned model generates predictions for the test data. Cells with majority of their neighbors sharing their predicted niche are identified as high-confidence predictions. We update the spatial profiles of the predicted niches for these high-confidence cells. Niches lacking high-confidence cells keep their original prototypes. The updated spatial profiles and the initial model together generate improved predictions for the remaining low-confidence cells. The examples of prompts for both approaches can be found in Appendix A.

Labeling Tool is a user-friendly software for biologists to easily annotate regions of interest for ground truth, promoting the use of our fine-tuning approach in more datasets. This tool allows for the creation of new high-quality data sets, which is critical for ongoing model improvement and benchmarking, especially when working with rare tissue types. A brief description of the usage of the software can be found in Appendix B.

3.3 Compared Methods

For the zero-shot approach (LLMiniST-Z), we evaluate three LLMs: GPT-4o mini (gpt-4o-mini-2024-07-18) (OpenAI, 2024), GPT-4o (gpt-4o-2024-08-06) (OpenAI, 2024), and Gemini 1.5 Pro (gemini-1.5-pro) (Team, 2024), which are closed-source and pre-trained models developed by OpenAI and Google, respectively. While we are aware of biomedical-specific LLMs (Wu et al., 2024; Bolton et al., 2024; Labrak et al., 2024), we choose GPT-4o and Gemini 1.5 because these biomedical LLMs do not outperform general-purpose LLMs in biomedical question-resolution tasks according to their benchmarking. In addition, none of the models are trained for spatial transcriptomics. We are continuously updating and will use advanced LLMs as soon as they become available. For the fine-tuning approach (LLMiniST-F), we employ GPT-4o mini as the base model. We access their functionalities and generated outputs through their respective

APIs. We aslo benchmark LLMiniST against 12 state-of-the-art non-LLM spatial clustering methods and 2 non-spatial clustering methods (Yuan et al., 2024).

3.4 Datasets

We selected three distinct types of spatial transcriptomic datasets with manual annotations: STARmap (Wang et al., 2018), MERFISH (Moffitt et al., 2018), and Visium (Maynard et al., 2021). Statistics of the datasets can be found in Appendix C. All datasets and corresponding ground-truth annotations are downloaded from `https://figshare.com/projects/SDMBench/163942` (Yuan et al., 2024).

4 Experiment

4.1 Experimental Settings

Validation and Testing We evaluated the zero-shot approach by executing it 3 times for each replicate of each subject, except for MERFISH dataset. Given the suboptimal performance of the zero-shot approach on MERFISH data, we conducted only a single trial for this dataset. For the evaluation of fine-tuning approaches, both LLMiniST-F and LLMiniST-Fs were run 3 times on each replicate of each subject. After fine-tuning the LLM, we define the application of LLMiniST-F to unlabeled cells within the same replicate as *supervised validation*. Applying the fine-tuned LLM to different replicates of the same subject is termed *intra-subject testing*. Conversely, testing on data from other subjects is designated as *cross-subject testing*. For clarity, we denote the LLMiniST-F fine-tuned with spots from Subject 1 ($\mathcal{D}_1$) as LLMiniST-D1-F, and its second-stage results are referred to as LLMiniST-D1-Fs. The LLMiniST fine-tuned with spots from $\mathcal{D}_2$ and $\mathcal{D}_3$ follow analogous naming conventions. The fine-tuning replicates were excluded from testing.

Evaluation Metrics We employ three clustering evaluation metrics to assess the accuracy of predicted cluster labeling using ground truth: Normalized Mutual Information (NMI), Homogeneity score (HOM), and Completeness score (COM) (Pedregosa et al., 2012). On the other hand, to evaluate the spatial continuity of the predicted segmentations, we employ three metrics: CHAOS (Yuan et al., 2024), the Percentage of Allowed outliers for Segmentation (PAS) (Shang and Zhou, 2022), and the Average Silhouette Width (ASW) (Yuan et al., 2024). Benchmarking of non-LLM methods on those metrics was effectively conducted by Yuan et al. (2024), and results of non-LLM methods are derived from the work. The detailed explanation for those metrics is given in Appendix E.

Spatial Neighborhood Definition Two cells or spots are considered neighbors if the spatial distance between them is below a specified threshold δ, which was set to 72 (700 pixels), 100 (100 pixels) and 344 (600 pixels) for STARmap, MERFISH, and Visium, respectively (representative examples are illustrated in the Appendix F). These thresholds were selected such that the resulting neighborhood represents the minimal functional unit of the tissue niches, encompassing the characteristic scale of its structural complexity. Spatial profiles are constructed from all cells residing within the neighborhood. Dataset-specific preprocessing procedures are applied, the specifics of which are elaborated upon in the Appendix G.

4.2 Performance Comparison

We first report the results of intra-subject testing. The accuracy and continuity performance of the zero-shot approach on STARmap is shown in Figure 2. The NMI of zero-shot and fine-tuning approaches in all datasets are in Table 1.

Observation 1. General LLMs possess intrinsic knowledge and capability to discern spatial niches. All three large language models (LLMs) demonstrated superior performance compared to current non-LLM spatial clustering methods as illustrated in the Figure 2. Specifically, the zero-shot approach with Gemini 1.5 Pro consistently achieved the highest rank across all evaluated metrics, excelling in both accuracy-related and continuity-related assessments. This is uncharacteristic of non-LLM methods, which may perform well in accuracy but fall short in continuity. Notably, even the GPT-4o mini, which is a smaller LLM, achieved a higher average rank than the best model-based method across all considered metrics, whether in terms of accuracy or continuity, as depicted in the Figure 2. We also tested replacing cell type names by abstract labels. This change led to a NMI decrease from 0.72 to 0.19 under the zero-shot settings, which underscores the importance of biological knowledge inside the cell type names for LLM performance in spatial niche identification.

Observation 2. A larger general LLM tends to exhibit superior performance. Comparing the results between GPT-4o mini and GPT-4o suggests that the model with a larger parameter count gen-

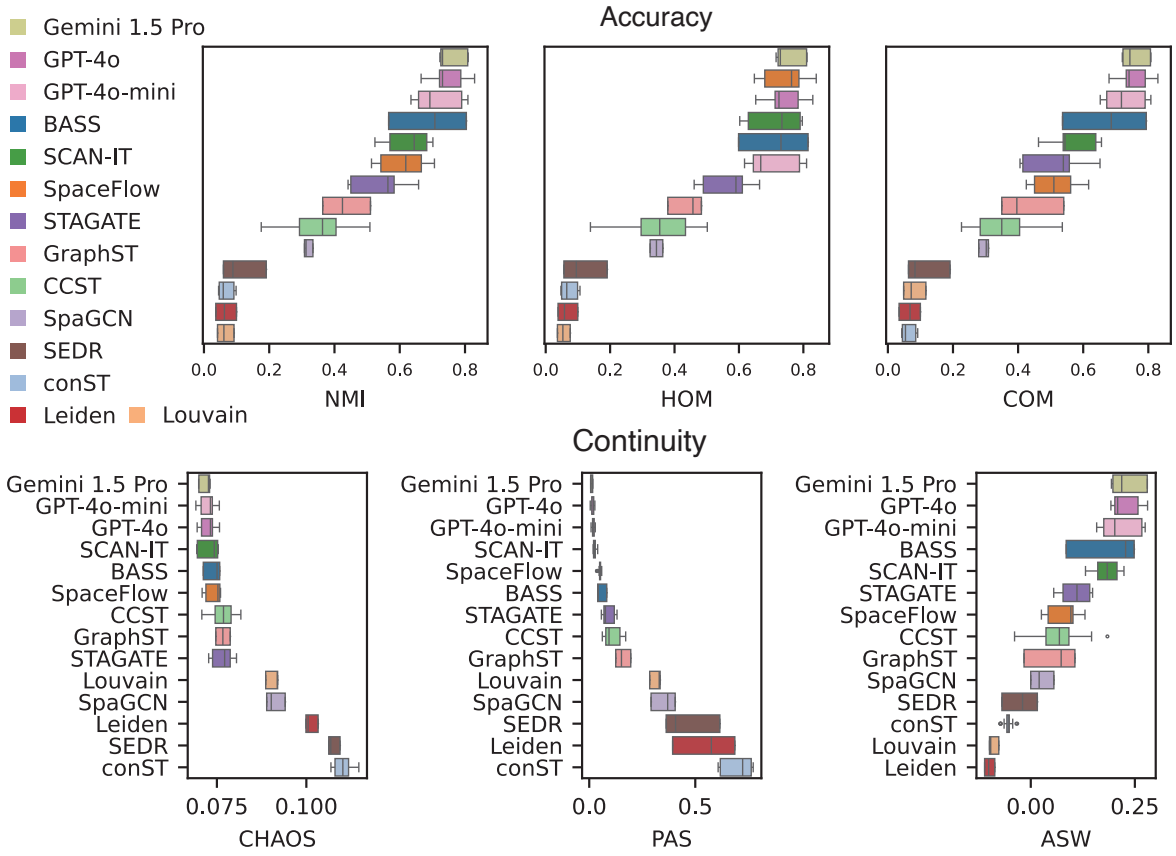

Figure 2: Assessment of the zero-shot approach on STARmap in terms of accuracy and continuity. The LLMiniST-Z with GPT-4o mini, GPT-4o, or Gemini 1.5 Pro is denoted as the respective LLM name. For NMI, HOM, COM, and ASW, higher values are better. For CHAOS and PAS, lower values are better.

erally demonstrates enhanced performance (Figure 2). This observation underscores the potential advantages conferred by the scale of an LLM in the zero-shot task. Considering that Gemini 1.5 Pro performs the best, it will represent the zero-shot approach (LLMiniST-Z) in subsequent comparisons.

Observation 3. End-to-end spatial niche identification is achievable. Our method not only accurately classifies cells but also automatically generates meaningful labels for each identified niche category, surpassing the capabilities of traditional clustering methods that only offer generic cluster tags. This end-to-end approach is more interpretable and readily applicable in practice.

Observation 4. The complexity of spatial transcriptomic data limits the feasibility of the zero-shot approach. Suboptimal results were observed in MERFISH and Visium datasets (Table 1), particularly in the MERFISH dataset, where the average NMI of the zero-shot approach (with Gemini 1.5 Pro as a representative) was lower than that of baseline non-spatial clustering methods. This is likely attributed to coarse-grained cell type annotations and the intricate spatial niche structures present in the MERFISH dataset. In Visium data, the mixed-cell spots necessitate deconvolution to infer cell-type composition, a process susceptible to inaccuracies. We also tested a domain-specific LLM specifically designed for single-cell analysis, ChatCell (Fang et al., 2024), in addition to general-purpose LLMs (GPT4o and Gemini). On a Visium sample, ChatCell achieved an NMI of 0.45, comparable to Gemini 1.5 Pro (NMI 0.44). These results show that current LLMs are not adapted for complex spatial transcriptomic data, highlighting an urgent need for tissue-specific tuning in this area.

Observation 5. The fine-tuned approach demonstrates superior performance and generalizability across replicates of the same subject. The fine-tuned model not only achieved the highest accuracy in the supervised validation, but also outperformed other non-LLM methods in intra-subject testing (Table 1). Specifically, the results imply that annotating a portion of cells in a single replicate provides sufficient information for LLMiniST-F to generalize and accurately identify cells across all other replicates. Besides, the superior performance of LLMiniST-F over Gemini 1.5 Pro highlights

Method Type	Method	STARmap	Visium	MERFISH	Avg.	Rank
LLM-based	LLMiniST-F (supervised)	0.811±0.003	0.760±0.002	0.795±0.006	-	-
	LLMiniST-Fs	**0.752±0.013**	**0.695±0.036**	**0.610±0.027**	0.686	1.7
	LLMiniST-F	**0.753±0.011**	**0.678±0.051**	0.581±0.020	0.67	2.0
	LLMiniST-Z	**0.755±0.040**	0.471±0.090	0.068±0.031	0.431	9.7
Non-LLM based	BASS	**0.693±0.100**	0.581±0.021	0.519±0.053	0.598	4.3
	SCAN-IT	0.630±0.055	0.546±0.047	0.578±0.045	0.585	5.0
	BayesSpace	-	0.565±0.087	-	0.565	5.0
	GraphST	0.433±0.061	0.592±0.049	0.317±0.056	0.448	6.0
	stLearn	-	0.552±0.014	-	0.552	6.0
	SpaceFlow	0.606±0.061	0.433±0.042	0.535±0.077	0.525	8.3
	CCST	0.353±0.083	0.507±0.022	0.468±0.031	0.443	8.7
	SpaGCN	0.318±0.011	0.513±0.047	0.214±0.015	0.348	9.0
	STAGATE	0.538±0.079	0.507±0.042	0.204±0.085	0.417	9.3
	SEDR	0.113±0.057	0.532±0.030	0.142±0.045	0.263	10.3
	conST	0.067±0.021	0.511±0.084	0.107±0.012	0.228	11.7
	SpaGCN(HE)	-	0.475±0.043	-	0.475	13.0
Non-Spatial	Leiden	0.066±0.026	0.329±0.009	0.177±0.004	0.191	13.0
	Louvain	0.065±0.021	0.336±0.014	0.169±0.009	0.19	13.3

Table 1: Comparison of Normalized Mutual Information (NMI) for Different Methods on Three Datasets. LLMiniST-F (supervised) is the supervised validation results, which is not included in comparison. The highest NMI values and those statistically indistinguishable from the highest (t-test, p >0.05) are boldfaced. (Mean ± Standard Deviation)

the effectiveness of fine-tuning in addressing the limitations of the zero-shot approach when applied to complex spatial transcriptomic data. Except for NMI, LLMiniST-Fs also perform well in terms of other matrics (Appendix H). We also conduct noise resilience experiments for LLMiniST-F on MERFISH dataset validating our method's stability under data perturbations (Appendix I).

Then, we evaluate how LLMiniST-F generalize to data from three different subjects of Visium. (Figure 3). We also compared the performance of LLMiniST-D1-F with LLMiniST-D2-F and LLMiniST-D3-F in Subjects 2 and 3, respectively (Figure 4**a**).

Observation 6. The fine-tuned model exhibits good zero-shot generalizability. Although the spots used to fine-tune LLMiniST-D1-Fs are not from Subject 2 or 3, it still achieved the highest average rank across the three accuracy matrices (Figure 3a). This demonstrates the LLM is efficient in domain-adaption, and the fine-tunning approach can facilitate zero-shot generalization on completely unseen data. While LLMiniST-D1-Fs does not achieve the best performance in terms of continuity, this may be attributed to the lack of considering the overall spatial context.

Observation 7. The two-stage strategy enhances generalizability. Both intra-subject (Table 1) and cross-subject (Figure 3) evaluations

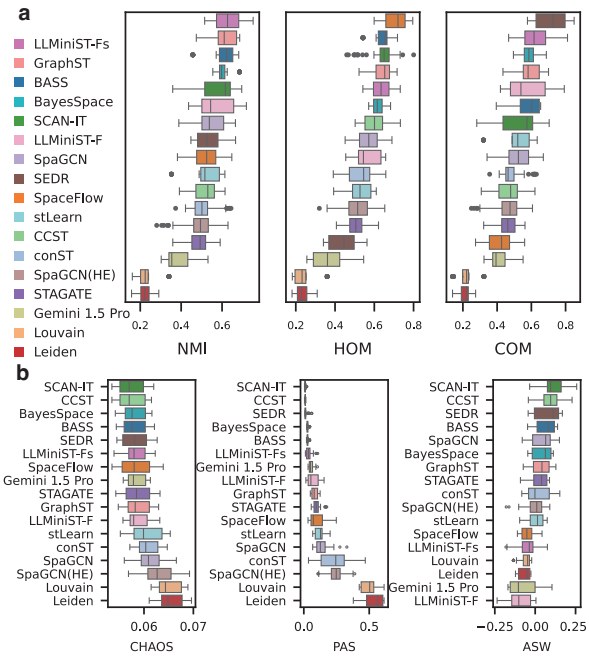

Figure 3: The performance of LLMiniST on the Subjects 2 and 3 of the Visium dataset. **a**, Accuracy-related metrics; **b**, Continuity related metrics. LLMiniST-F is fine-tuned with Subject 1. For NMI, HOM, COM, and ASW, higher values are better. For CHAOS and PAS, lower values are better.

demonstrate the effectiveness of the two-stage strategy, as evidenced by the superior performance of LLMiniST-Fs compared to LLMiniST-F. Furthermore, the two-stage approach enables

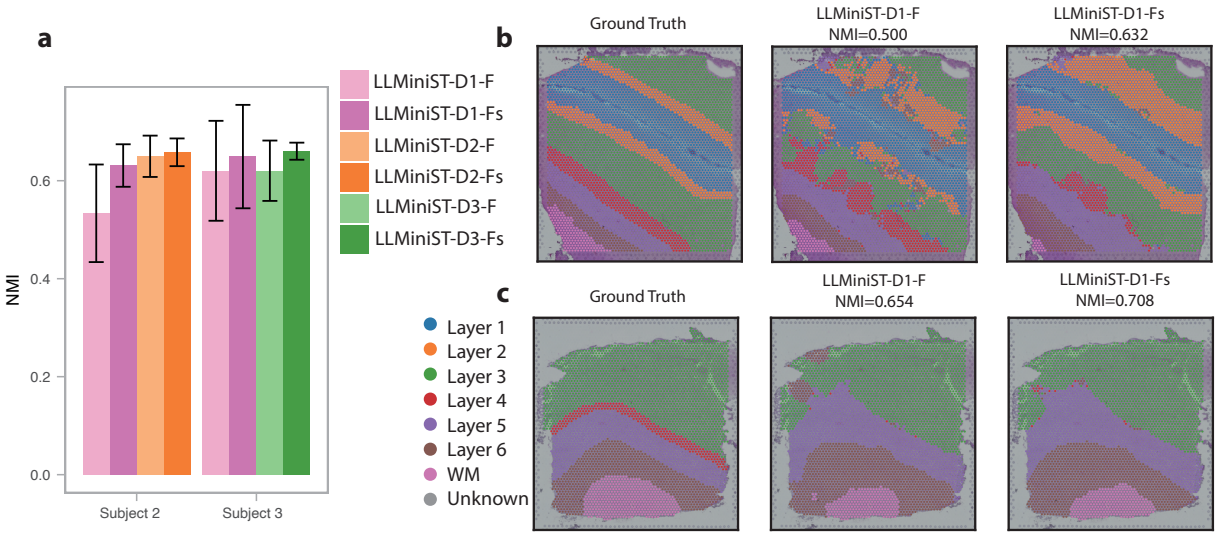

Figure 4: NMI improvement achieved by the proposed two-stage stategy on the Subjects 2 and 3 of the Visium dataset. (a) Comparison of LLMiniST fine-tuned using spots from different subjects, with and without two-stage strategy. (b) Identification results for a replicate of Subject 2. (c) Identification results for a replicate of Subject 3.

LLMiniST-D1-Fs to achieve an accuracy comparable to or even surpassing that of LLMiniST-D2-F and LLMiniST-D3-F within their respective subjects (Figure 4 a). Notably, as illustrated in Figure 4 b and c, the second stage successfully rectifies misidentifications in low-confidence instances.

Observation 8. Fine-tuned models effectively handle mismatched labels. Unlike other non-LLM methods that require a pre-defined number of clusters, k, LLMiniST-F operates without this constraint, offering a more robust and unbiased solution. Despite the misalignment of niche components between Subject 3 and the other subjects, particularly concerning the absent niches of Layers 1 and 2, both LLMiniST-D1-F and LLMiniST-D1-Fs proficiently circumvent the misclassification of spots into these non-existent niches (Figure 4c).

4.3 Proportions of Examples for Fine-tuning

We examined the impact of varying the proportion of labeled cells on fine-tuning performance using the MERFISH dataset (Figure 5). Specifically, we fine-tuned LLMiniST with subsets ranging from 5% to 70% of labeled cells from a single replicate.

Observation 9. Fine-tuning requires only a small fraction of labeled cells for high accuracy. Fine-tuning with just 5% of the cells already surpasses the accuracy of other non-LLM methods (Figure 5). Furthermore, the benefits of increasing the proportion of labeled cells appear to saturate beyond 30%. This suggests that labeling a relatively small portion of a sample is sufficient to adapt general LLMs into specialized models with high accuracy. The fine-tuning approach offers a scalable solution where minimal human input can yield substantial results, even with massive datasets.

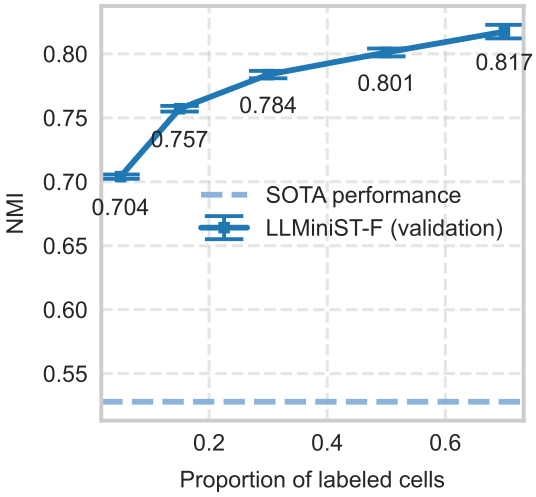

Figure 5: Impact of labeled cell proportion on the performance of LLMiniST-F, evaluated on a supervised validation set from the MERFISH dataset. Error bars indicate the standard deviation.

5 Conclusion

This work compared the performance of LLMs with non-LLM methods for niche identification in spatial transcriptomic data. We demonstrate the feasibility of zero-shot LLMiniST. Furthermore, by incorporating a two-stage fine-tuning strategy, LLMiniST exhibits a strong generalizability to samples from either the same or different subjects. We further demonstrate that fine-tuning with just a sub-

set of labeled cells is sufficient to boost LLMs beyond state-of-the-art performance, and we have designed a user-friendly software tool to facilitate the efficient annotation of such training data.

6 Limitations

Recent advances in spatial transcriptomics have introduced a suite of new technologies that require thorough assessment. In addition, the availability of datasets with reliable ground-truth information is currently a limiting factor. We will continuously promote our annotation software among biologists to obtain more ground-truth data from a broader range of tissue regions. To improve annotation efficiency, future work will investigate methods for selecting a minimal subset of cells that provides maximal information for fine-tuning, as opposed to the current random selection approach.

With the emergence of new open-source local LLMs, a promising future strategy involves applying knowledge distillation to these smaller architectures and integrating this with fine-tuning. This combined approach aims to improve operational speed while preserving high levels of performance.

7 Ethics Statement

This research adheres to the ethical guidelines outlined in the ACL Code of Ethics[5]. All datasets employed in this study are publicly available resources. Similarly, all language models used are publicly accessible and utilized in accordance with their respective terms of use, as specified by OpenAI's Terms of Use[6] and Gemini API Terms of Service[7]. To ensure the reproducibility and transparency of our findings, we commit to releasing the code and associated materials upon publication.

8 Acknowledgment

This work was partially supported by *NIH R01LM014087* and *Yale University and Boehringer Ingelheim Biomedical Data Science Fellowship Program*.

References

Haiyang Bian, Yixin Chen, Erpai Luo, Xinze Wu, Minsheng Hao, Lei Wei, and Xuegong Zhang. 2024. General-purpose pre-trained large cellular models for single-cell transcriptomics. National Science Review, 11(11):nwae340.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. Biomedlm: A 2.7b parameter language model trained on biomedical text. Preprint, arXiv:2403.18421.

Dylan M Cable, Evan Murray, Vignesh Shanmugam, Simon Zhang, Luli S Zou, Michael Diao, Haiqi Chen, Evan Z Macosko, Rafael A Irizarry, and Fei Chen. 2022. Cell type-specific inference of differential expression in spatial transcriptomics. Nat Methods, 19(9):1076–1087.

Zixuan Cang, Xinyi Ning, Annika Nie, Min Xu, and Jing Zhang. 2021. SCAN-IT: Domain segmentation of spatial transcriptomics images by graph neural network. BMVC, 32.

Yiqun Chen and James Zou. 2024. Simple and effective embedding model for single-cell biology built from chatgpt. Nature Biomedical Engineering.

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

Ofer Elhanani, Raz Ben-Uri, and Leeat Keren. 2023. Spatial profiling technologies illuminate the tumor microenvironment. Cancer Cell, 41(3):404–420.

Yin Fang, Kangwei Liu, Ningyu Zhang, Xinle Deng, Penghui Yang, Zhuo Chen, Xiangru Tang, Mark Gerstein, Xiaohui Fan, and Huajun Chen. 2024. Chatcell: Facilitating single-cell analysis with natural language. Preprint, arXiv:2402.08303.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. Preprint, arXiv:2403.14608.

Yuhan Hao, Tim Stuart, Madeline H Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, and Rahul Satija. 2023. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature Biotechnology.

Wenpin Hou and Zhicheng Ji. 2024. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. Nature Methods, 21(8):1462–1465.

[5] https://www.aclweb.org/portal/content/acl-code-ethics

[6] https://openai.com/policies/row-terms-of-use/

[7] https://ai.google.dev/gemini-api/terms

Jian Hu, Xiangjie Li, Kyle Coleman, et al. 2021. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nature Methods, 18:1342–1351.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. Preprint, arXiv:2402.10373.

Jiachen Li, Siheng Chen, Xiaoyong Pan, Ye Yuan, and Hong-Bin Shen. 2022. Cell clustering for spatial transcriptomics data with graph neural networks. Nature Computational Science, 2:399–408.

Rui Li, Jiwei Li, Jiawei Han, and Guoyin Wang. 2024. Similarity-based neighbor selection for graph llms. Preprint, arXiv:2402.03720.

Zheng Li and Xiang Zhou. 2022. Bass: Multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. Genome Biology, 23:168.

Teng Liu, Zhao-Yu Fang, Zongbo Zhang, Yongxiang Yu, Min Li, and Ming-Zhu Yin. 2024a. A comprehensive overview of graph neural network-based approaches to clustering for spatial transcriptomics. Comput Struct Biotechnol J, 23:106–128.

Yunqing Liu, Ningshan Li, Ji Qi, Gang Xu, Jiayi Zhao, Nating Wang, Xiayuan Huang, Wenhao Jiang, Huanhuan Wei, Aurélien Justet, Taylor S Adams, Robert Homer, Amei Amei, Ivan O Rosas, Naftali Kaminski, Zuoheng Wang, and Xiting Yan. 2024b. SDePER: a hybrid machine learning and regression method for cell-type deconvolution of spatial barcoding-based transcriptomic data. Genome Biol., 25(1):271.

Yahui Long, Kok Siong Ang, Mengwei Li, et al. 2023. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. Nature Communications, 14:1155.

Yuren Mao, Yuhang Ge, Yijiang Fan, Wenyi Xu, Yu Mi, Zhonghao Hu, and Yunjun Gao. 2024. A survey on lora of large language models. Frontiers of Computer Science, 19(7).

Kristen R Maynard, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, 2nd, Matthew N Tran, Zachary Besich, Madhavi Tippani, Jennifer Chew, Yifeng Yin, Joel E Kleinman, Thomas M Hyde, Nikhil Rao, Stephanie C Hicks, Keri Martinowich, and Andrew E Jaffe. 2021. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nat. Neurosci., 24(3):425–436.

Jeffrey R Moffitt, Dhananjay Bambah-Mukku, Stephen W Eichhorn, Eric Vaughn, Karthik Shekhar, Julio D Perez, Nimrod D Rubinstein, Junjie Hao, Aviv Regev, Catherine Dulac, and Xiaowei Zhuang. 2018. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science, 362(6416):eaau5324.

Jeffrey R Moffitt, Emma Lundberg, and Holger Heyn. 2022. The emerging landscape of spatial profiling technologies. Nat Rev Genet, 23(12):741–759.

OpenAI. 2024. Gpt-4o system card. Preprint, arXiv:2410.21276.

Fabian Pedregosa, Ga"el Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2012. Scikit-learn: Machine learning in python. CoRR, abs/1201.0490.

Duy Pham, Xiao Tan, Blake Balderson, et al. 2023. Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues. Nature Communications, 14:7739.

Honglei Ren, Benjamin L. Walker, Zixuan Cang, and Nie Qing. 2022. Identifying multicellular spatiotemporal organization of cells with spaceflow. Nature Communications, 13:4076.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. Preprint, arXiv:2402.07927.

Lulu Shang and Xiang Zhou. 2022. Spatially aware dimension reduction for spatial transcriptomics. Nature Communications, 13(1):7203.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Preprint, arXiv:2403.05530.

V. A. Traag, L. Waltman, and N. J. van Eck. 2019. From louvain to leiden: guaranteeing well-connected communities. Scientific Reports, 9(1):5233.

Katy Vandereyken, Alejandro Sifrim, Bernard Thienpont, and Thierry Voet. 2023. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet, 24(8):494–515.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. 2024. Can language models solve graph problems in natural language? Preprint, arXiv:2305.10037.

Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P Nolan, Felice-Alessio Bava, and Karl Deisseroth. 2018. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science, 361(6400).

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. Journal of the American Medical Informatics Association, 31(9):1833–1843.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2021. A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1):4–24.

Hang Xu, Huazhu Fu, Yahui Long, and Kok Siong Ang. 2024. Unsupervised spatially embedded deep representation of spatial transcriptomics. Genome Medicine, 16:12.

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

Hu Zeng, Jiahao Huang, Jingyi Ren, Connie Kangni Wang, Zefang Tang, Haowen Zhou, Yiming Zhou, Hailing Shi, Abhishek Aditham, Xin Sui, Hongyu Chen, Jennifer A Lo, and Xiao Wang. 2023. Spatially resolved single-cell translatomics at molecular resolution. Science, 380(6652):eadd3067.

Edward Zhao, Matthew R. Stone, Xing Ren, et al. 2021. Spatial transcriptomics at subspot resolution with bayesspace. Nature Biotechnology, 39:1375–1384.

Yongshuo Zong, Tingyang Yu, Xuesong Wang, Yixuan Wang, Zhihang Hu, and Yu Li. 2022. const: an interpretable multi-modal contrastive learning framework for spatial transcriptomics. bioRxiv.

A Examples of prompt

We provide examples of spatial context prompts for LLMiniST-Z and LLMiniST-F (Figure 6). Note that the prompt format of LLMiniST-F is identical to that of LLMiniST-Fs.

B Software for Annotations

We develop as software (Figure 7) that provides a comprehensive set of tools for visualizing, annotating, and performing basic and AI-powered analysis on spatial transcriptomics data. It combines data loading, interactive plotting, multiple selection methods, annotation management, and advanced analytical capabilities within a user-friendly graphical interface. The detailed usage and source code of the software can be found at `https://github.com/wJDKnight/draw_spatial`.

	STARmap	Visium	MERFISH
Number of subjects	3	3	1
Number of replicates for each subject	1	4	5
Total number of cells/spots	3,268	47,681	28,317

Table 2: Statistics of datasets.

C Statistics of Datasets

STARmap and MERFISH both have single-cell resolution and therefore provide cell type annotations. In contrast, Visium is a spatial transcriptomics platform that generates data at the resolution of spots, each encompassing 1-10 cells, without providing individual cell type annotations. Table 2 shows statistics of datasets.

D Full List of Compared Methods

We compare LLMiniST against spatial clustering methods, including BASS (Li and Zhou, 2022), conST (Zong et al., 2022), SpaceFlow (Ren et al., 2022), SCAN-IT (Cang et al., 2021), CCST (Li et al., 2022), GraphST (Long et al., 2023), STAGATE (Dong and Zhang, 2022), SpaGCN (Hu et al., 2021), SpaGCN(HE) (Hu et al., 2021), SEDR (Xu et al., 2024), stLearn (Pham et al., 2023), and BayesSpace (Zhao et al., 2021). Notably, BayesSpace, stLearn, and SpaGCN(HE) rely on the presence of histological images, so these methods could not be evaluated on the STARmap and MERFISH datasets due to the absence of image data. To provide a baseline for comparison, we also incorporate the Leiden (Traag et al., 2019) and Louvain (Blondel et al., 2008) algorithms, which do not use spatial information.

E Explanation of Metrics

All of the accuracy-related metrics, NMI, HOM, and COM, range from 0 to 1, with higher values indicating better agreement between the predicted and true cluster assignments. To provide a comprehensive comparison of overall accuracy across different methods, we calculate the average rank based on the average across the three metrics. For continuity-related metircs, lower values of CHAOS and PAS indicate higher spatial continuity. In contrast, ASW is rescaled to 0 - 1 by Yuan et al. (2024), with higher values corresponding to greater spatial coherence of the predicted segments.

F Neighborhood Examples

The spatial scope of our neighborhood analysis is visualized in Figure 8. Each sub-figure represents

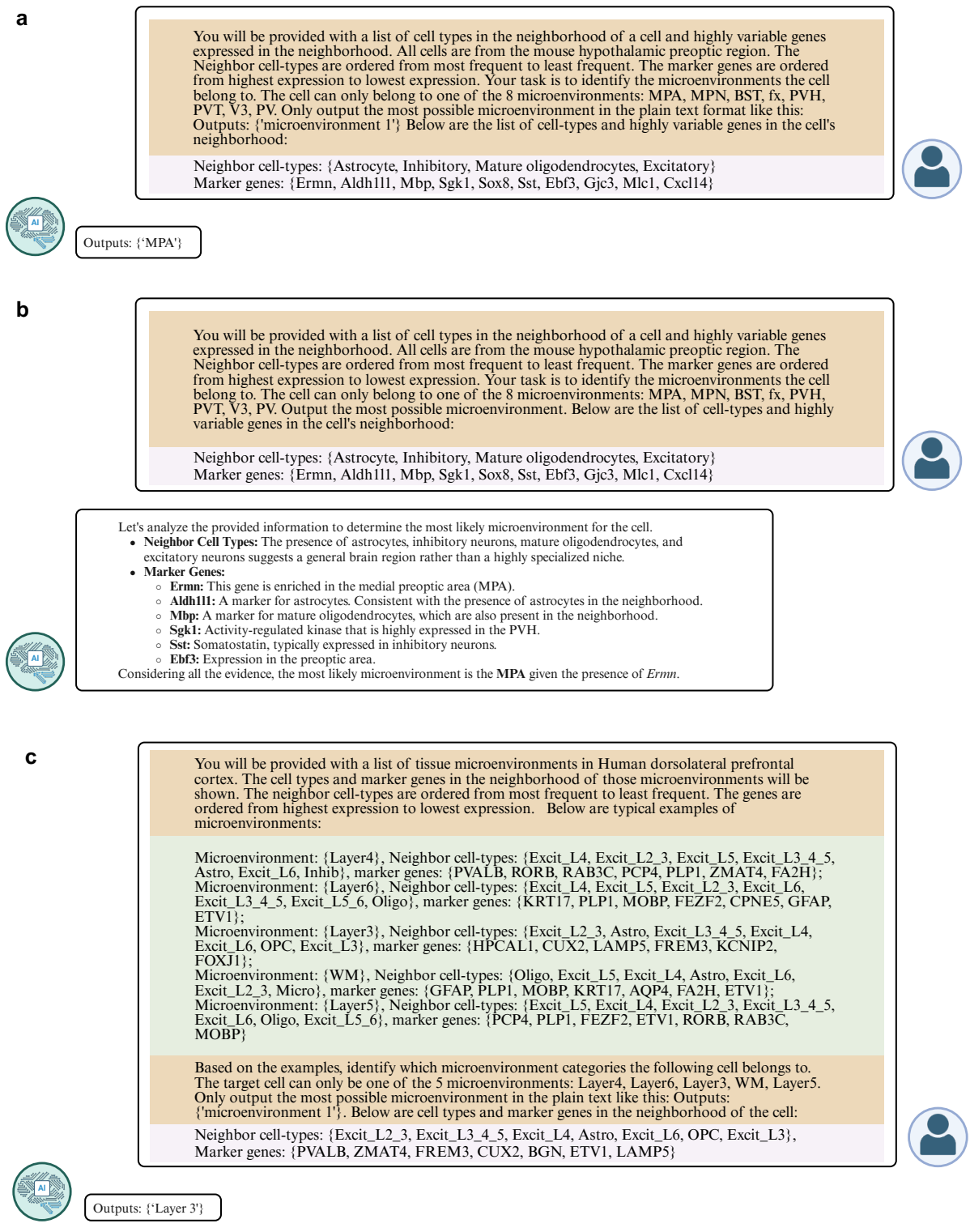

Figure 6: Examples of spatial context prompts. **a**, a zero-shot prompt for a cell in MERFISH dataset; **b**, a zero-shot prompt without limitation for output format; **c**, a fine-tuning prompt for a spot in Visium dataset.

a different dataset, and within each, a red circle delineates the perimeter of the neighborhood that was taken into account for constructing spatial profile.

G Data Preprocessing for Spatial Context Prompt

Spatial Context Prompt for STARmap The STARmap dataset features clear cell type anno-

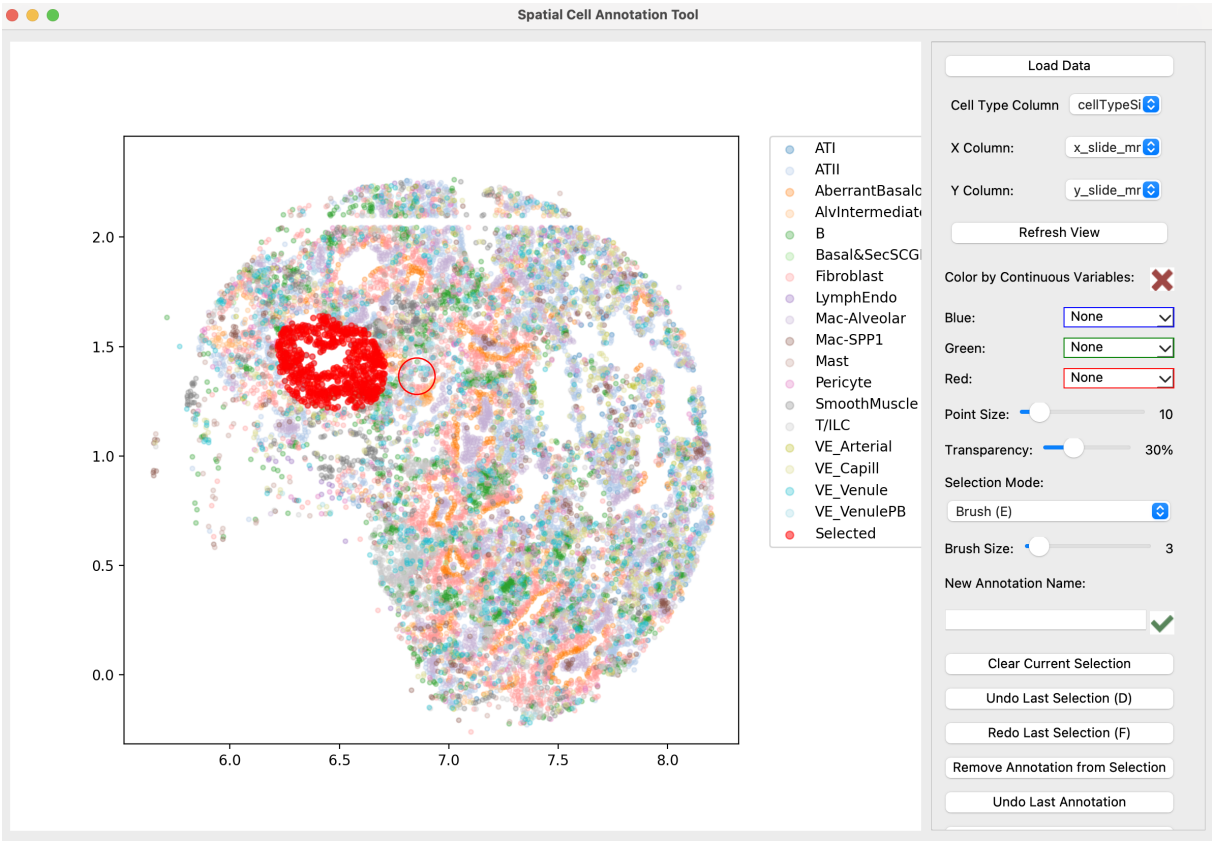

Figure 7: Screenshot of the spatial annotation software.

tations and a simple layered structure (Wang et al., 2018). We only used neighborhood cell type composition for spatial context prompt. In addition, we identified distally located cell types from the target cell and used the 3 farthest cell types as negative examples in prompt engineering.

Spatial Context Prompt for MERFISH MERFISH data provides a less detailed cell type resolution, making it challenging to distinguish cells across spatial niches. To enhance our analysis, we incorporated expression data from adjacent genes 3. Our gene selection process involves combining all five samples, normalizing each cell by total gene counts, and scaling genes to have unit variance and zero mean. Utilizing 30 principal components, we conducted CCAIntegration with the Seurat package (Hao et al., 2023) and clustered the integrated data with FindClusters() at a resolution of 0.1. Finally, we identified the top five marker genes per cluster using FindConservedMarkers().

Spatial Context Prompt for Visium Visium data does not inherently offer cell type annotations, as each spot encompasses the gene expression profile of multiple cells. To mitigate this limitation, we employed spatial deconvolution techniques, SDePER (Liu et al., 2024b), to estimate the proportions of cell types within each spot. Subsequently, the cell type composition for a specific spot was determined as the mean of the cell type compositions of its adjacent spots. Acknowledging that the precision of these estimated cell type proportions can be influenced by the efficacy of the deconvolution algorithm, we incorporated neighboring gene expression data into our analysis. For Visium, 22 established marker genes (Yuan et al., 2024) of the dorsolateral prefrontal cortex were selected.

Post-refinement Assuming that the spatial niche should be smooth across the spatial, we add a refinement step for the identification result like SpaGCN and GraphST (Hu et al., 2021; Long et al., 2023). In this step, we examine the assignment of the niche of each cell and its surrounding areas. For a given cell, if more than half of its surrounding cells are assigned to a different niche, this cell will be relabeled to the same niche as the major label of its surrounding cells. For the zero-shot approach, we add this step in the predictions. For the fine-tuning approach, this step is added after both LLMiniST-F and LLMiniST-Fs, except for identifying high-confidence cells. The high-confident

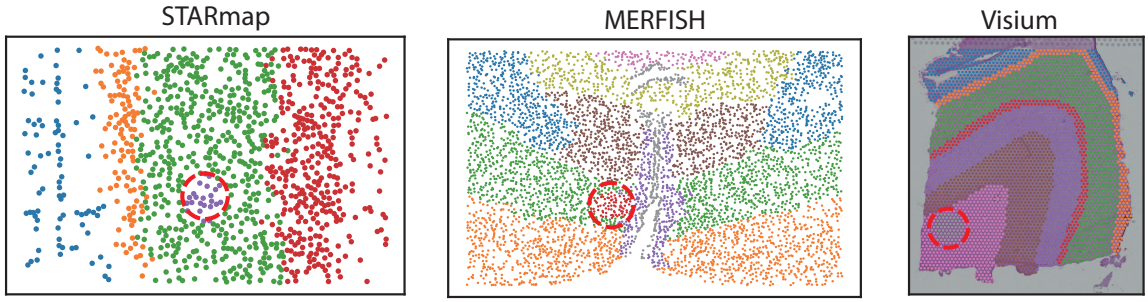

Figure 8: Depiction of the neighborhood size across three different datasets.

	MERFISH	STARmap	Visium	Avg.	Rank
LLMiniST-Fs	0.613±0.021	0.757±0.020	0.683±0.036	0.684	1.0
LLMiniST-F	0.580±0.017	0.755±0.018	0.662±0.044	0.666	2.3
SCAN-IT	0.598±0.048	0.716±0.077	0.561±0.054	0.625	4.3
BayesSpace	-	-	0.565±0.079	0.565	5.0
BASS	0.432±0.043	0.715±0.090	0.585±0.020	0.577	5.3
GraphST	0.219±0.042	0.440±0.045	0.591±0.054	0.416	6.0
SpaceFlow	0.483±0.080	0.753±0.066	0.529±0.048	0.588	6.0
CCST	0.536±0.035	0.360±0.091	0.534±0.028	0.477	7.0
stLearn	-	-	0.543±0.009	0.543	7.0
STAGATE	0.207±0.087	0.564±0.069	0.509±0.037	0.426	9.3
SpaGCN	0.210±0.013	0.344±0.016	0.520±0.045	0.358	9.3
Gemini 1.5 Pro	0.045±0.025	0.753±0.043	0.411±0.079	0.403	10.7
conST	0.110±0.012	0.072±0.022	0.515±0.087	0.232	12.0
SEDR	0.116±0.036	0.114±0.058	0.455±0.067	0.228	12.3
SpaGCN(HE)	-	-	0.483±0.043	0.483	13.0
Leiden	0.167±0.007	0.065±0.026	0.334±0.008	0.189	13.3
Louvain	0.157±0.007	0.055±0.017	0.335±0.014	0.182	13.7

Table 3: Intra-subject HOM: Mean ± SD

cells are selected based on the unrefined LLMiniST-F's results.

H Results of Intra-Subject Testing

The following results present the performance of LLMiniST in terms of HOM, COM, CHAOS, PAS, and ASW, evaluated using intra-subject testing across all datasets (Table 3 - 7).

I Noise resilience experiments

To demonstrate robustness to noise, we simulated non-capture noise by randomly zeroing gene expression values on MERFISH data. We counted the number of genes changed in the top-10 marker gene rankings before and after randomly zeroing. We observed minimal changes in the top-10 marker gene rankings (Figure 9). Furthermore, the fine-tuned LLMiniST showed resilient performance on MERFISH data with increasing noise (Figure 9). Even with 50% of the values set to 0, the NMI drops by only 12.5%.

	MERFISH	STARmap	Visium	Avg.	Rank
LLMiniST-Fs	0.608±0.033	0.748±0.006	0.707±0.040	0.688	2.0
LLMiniST-F	0.582±0.025	0.750±0.006	0.695±0.060	0.676	3.0
BASS	0.651±0.070	0.672±0.107	0.577±0.022	0.633	3.3
GraphST	0.593±0.129	0.429±0.083	0.594±0.044	0.538	5.3
BayesSpace	-	-	0.566±0.096	0.566	6.0
SCAN-IT	0.560±0.045	0.566±0.057	0.533±0.045	0.553	6.7
stLearn	-	-	0.562±0.022	0.562	7.0
Gemini 1.5 Pro	0.153±0.036	0.757±0.038	0.554±0.110	0.488	7.3
SpaceFlow	0.603±0.077	0.508±0.059	0.367±0.037	0.493	8.3
SEDR	0.187±0.063	0.113±0.057	0.659±0.047	0.32	8.3
STAGATE	0.201±0.083	0.516±0.087	0.506±0.047	0.408	9.0
SpaGCN	0.218±0.018	0.297±0.013	0.506±0.050	0.34	9.7
CCST	0.415±0.030	0.350±0.082	0.483±0.021	0.416	9.7
conST	0.104±0.012	0.063±0.019	0.507±0.081	0.224	12.7
Louvain	0.183±0.011	0.079±0.029	0.337±0.016	0.2	13.3
Leiden	0.190±0.004	0.067±0.027	0.325±0.011	0.194	13.3
SpaGCN(HE)	-	-	0.468±0.042	0.468	14.0

Table 4: Intra-subject COM: Mean ± SD

	MERFISH	STARmap	Visium	Avg.	Rank
SCAN-IT	0.029±0.000	0.073±0.003	0.061±0.001	0.054	2.3
CCST	0.028±0.000	0.077±0.003	0.061±0.001	0.055	3.0
BASS	0.029±0.001	0.074±0.002	0.061±0.001	0.055	3.7
Gemini 1.5 Pro	0.030±0.001	0.072±0.001	0.062±0.001	0.054	4.0
LLMiniST-Fs	0.030±0.001	0.073±0.001	0.062±0.002	0.055	4.3
LLMiniST-F	0.030±0.001	0.074±0.001	0.062±0.001	0.055	5.7
SpaceFlow	0.029±0.001	0.074±0.002	0.064±0.002	0.056	6.7
GraphST	0.030±0.001	0.077±0.002	0.063±0.002	0.056	8.3
STAGATE	0.048±0.004	0.077±0.003	0.062±0.001	0.062	9.0
SEDR	0.046±0.004	0.107±0.001	0.062±0.001	0.072	9.7
BayesSpace	-	-	0.063±0.002	0.063	10.0
stLearn	-	-	0.064±0.001	0.064	11.0
SpaGCN	0.049±0.001	0.091±0.002	0.065±0.002	0.068	12.0
Louvain	0.055±0.001	0.091±0.002	0.068±0.003	0.071	12.7
conST	0.060±0.002	0.110±0.002	0.065±0.003	0.078	13.7
Leiden	0.055±0.002	0.101±0.002	0.069±0.002	0.075	14.0
SpaGCN(HE)	-	-	0.067±0.002	0.067	15.0

Table 5: Intra-subject CHAOS: Mean ± SD

	MERFISH	STARmap	Visium	Avg.	Rank
SCAN-IT	0.027±0.003	0.025±0.006	0.015±0.003	0.022	3.0
CCST	0.005±0.001	0.111±0.036	0.011±0.003	0.042	3.3
BASS	0.026±0.006	0.055±0.020	0.029±0.000	0.037	3.7
LLMiniST-Fs	0.041±0.008	0.014±0.001	0.048±0.011	0.034	4.0
Gemini 1.5 Pro	0.068±0.039	0.012±0.004	0.057±0.033	0.046	5.3
LLMiniST-F	0.053±0.009	0.022±0.006	0.062±0.017	0.046	5.7
BayesSpace	-	-	0.053±0.013	0.053	6.0
SpaceFlow	0.028±0.005	0.050±0.006	0.199±0.056	0.092	7.3
SEDR	0.392±0.089	0.462±0.112	0.038±0.011	0.298	8.3
GraphST	0.064±0.028	0.158±0.030	0.118±0.014	0.113	8.7
STAGATE	0.589±0.100	0.089±0.025	0.084±0.029	0.254	9.3
stLearn	-	-	0.126±0.012	0.126	11.0
Louvain	0.568±0.042	0.316±0.023	0.392±0.074	0.426	12.0
SpaGCN	0.590±0.039	0.356±0.048	0.133±0.028	0.36	12.0
Leiden	0.579±0.036	0.552±0.123	0.442±0.041	0.524	13.7
conST	0.847±0.023	0.700±0.065	0.202±0.148	0.583	14.0
SpaGCN(HE)	-	-	0.228±0.053	0.228	15.0

Table 6: Intra-subject PAS: Mean ± SD

	MERFISH	STARmap	Visium	Avg.	Rank
BASS	-0.017±0.020	0.187±0.074	0.087±0.022	0.086	3.3
CCST	0.292±0.018	0.064±0.052	0.170±0.078	0.175	3.3
SCAN-IT	-0.018±0.056	0.184±0.032	0.162±0.080	0.109	3.3
BayesSpace	-	-	0.085±0.070	0.085	5.0
LLMiniST-Fs	-0.117±0.034	0.226±0.003	0.010±0.034	0.04	6.3
SEDR	-0.121±0.067	-0.024±0.035	0.105±0.070	-0.013	7.0
stLearn	-	-	0.044±0.010	0.044	8.0
STAGATE	-0.195±0.044	0.111±0.030	0.060±0.023	-0.008	8.7
GraphST	-0.126±0.035	0.054±0.052	0.040±0.012	-0.011	8.7
LLMiniST-F	-0.147±0.044	0.226±0.005	0.001±0.046	0.026	9.0
SpaGCN	-0.172±0.011	0.025±0.023	0.082±0.043	-0.022	9.3
SpaceFlow	-0.029±0.059	0.080±0.036	-0.066±0.029	-0.005	9.3
SpaGCN(HE)	-	-	0.021±0.035	0.021	10.0
Gemini 1.5 Pro	-0.225±0.096	0.231±0.037	-0.045±0.087	-0.013	10.3
conST	-0.101±0.015	-0.056±0.009	-0.001±0.024	-0.052	10.7
Louvain	-0.143±0.011	-0.091±0.010	0.010±0.019	-0.075	11.3
Leiden	-0.168±0.025	-0.100±0.010	0.001±0.010	-0.089	12.7

Table 7: Intra-subject ASW: Mean ± SD

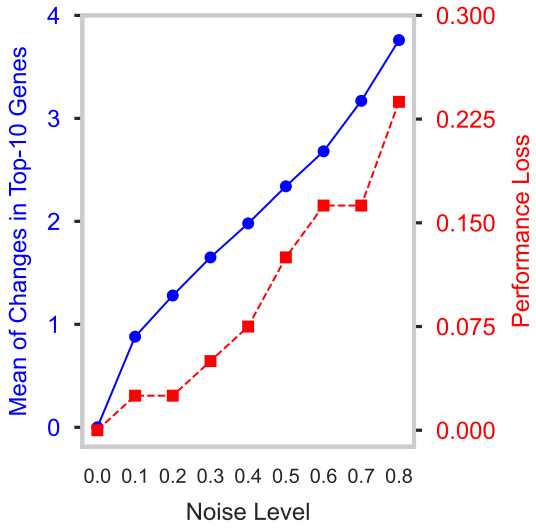

Figure 9: Noise Resilience Experiments on MERFISH dataset. The changes of NMI is calculated as the performance loss.

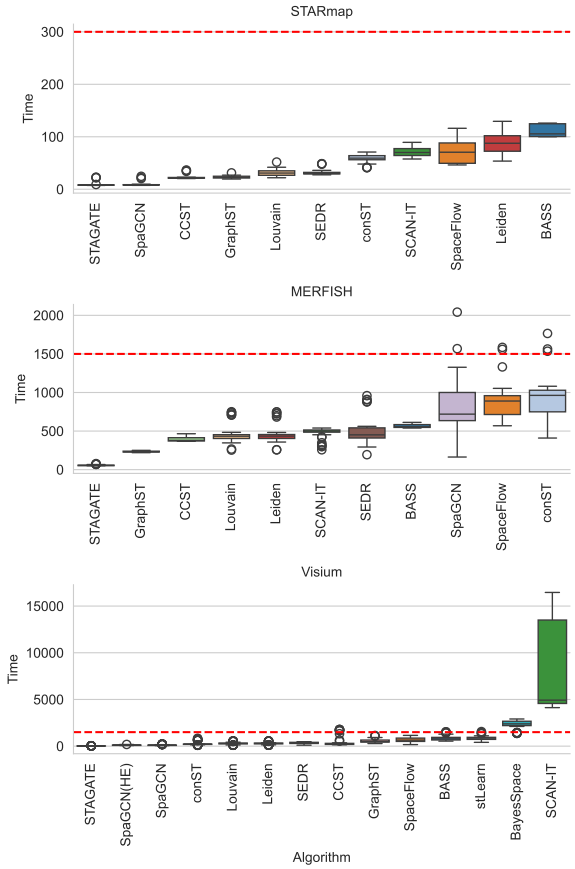

Figure 10: Runtime of non-spatial methods in three datasets. The red auxiliary line marks the estimated time of our method.

J Running Time

We used the official APIs for both GPT and Gemini. Our zero-shot method takes approximately 0.2–0.5 seconds per cell, resulting in a total runtime of around 1,500 seconds for each Visium or MERFISH sample, and roughly 300 seconds for each STARmap sample. The runtime of existing non-LLM methods is benchmarked by Yuan et al. (2024). We summarized the runtime of different methods in different datasets in Figure 10. Our approach offers two advantages: (1) it avoids out-of-memory errors, and (2) it scales linearly with the number of cells. In comparison, GNN-based methods are often limited by graph size.