

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

Anonymous ACL submission

## 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, visualizing tissue anatomy, 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, LLMiST<sup>1</sup>, 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. Comprehensive method performance evaluations showed that both zero-shot and fine-tuned LLMiST had superior performance than current non-LLM methods in many circumstances. Notably, the two-stage fine-tuning strategy facilitated substantial cross-subject generalizability. 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 in 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 Transcriptomics (LLMiST), 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

<sup>1</sup>The source code of our implementation can be found at <https://anonymous.4open.science/r/LLMiST-7000>.

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 align 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, it is feasible for a fine-tuned LLM generalize to data from other subjects that have similar tissue archi-

ectures, providing broadly applicable models for spatial niche identification.

The contributions of this work are summarized as follows: (1) We bridges the gap between LLMs and spatial transcriptomics, opening new avenues for spatial clustering analysis. (2) We performed extensive benchmarking using a variety of spatial transcriptomic datasets, comparing the performance of LLMs with that of traditional non-LLM methods. (3) Through this rigorous evaluation, we not only demonstrate the effectiveness of our apporach, LLMiniST, but also uncover key insights into the capabilities and limitations of LLMs in this specific application. 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). Prompt engineering has emerged as a crucial technique for extending LLMs' capabilities (Sahoo et al., 2024). For example, Chen and Zou (2024) demonstrated the ability to encode single-cell biological knowledge by leveraging LLM embeddings of gene descriptions from literatures. 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). Hou and Ji (2024) achieved cell type annotation in single-cell transcriptomic data using LLMs. 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. Then, 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<sup>2</sup> 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):  $\mathbf{c}_{p,m} \in \{1, 2, \dots, C_p\}^{N_{p,m}}$ , where  $C_p$  is the number of cell types of subject  $p$ .

4) Niche Annotations (Optional):  $\mathcal{A}_{p,m} \subseteq \{1, 2, \dots, N_{p,m}\}$  denoting the indices of cells with such annotations. The niche labels are given by  $\mathbf{n}_{p,m} \in \{1, 2, \dots, K\}^{|\mathcal{A}_{p,m}|}$ , 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:

<sup>2</sup>Hereafter, we use "cells" to refer to both cells and spots, unless there is a need to distinguish between them.

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  $n_i$ <sup>3</sup>. 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  $\delta$ . 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:** We require the LLM to output the most possible niche without any explanation to facilitate downstream processing.

**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

<sup>3</sup>For clarity and simplicity, we omit the subscripts  $p$  and  $m$  when the context makes the meaning unambiguous.

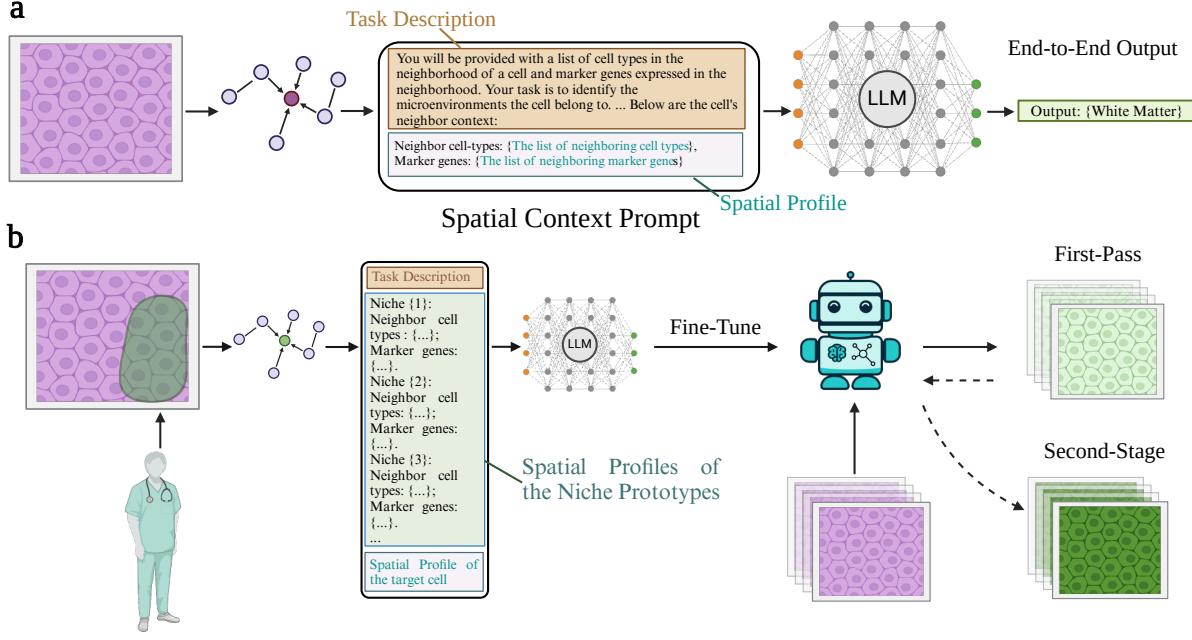


Figure 1: A overview of the pipeline of zero-shot approach (a) and fine-tuning approach (b).

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.

When identifying cells from different samples or subjects, we proposed a two-stage strategy to adjust the difference between samples or subjects. 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 results of LLMinist-F with two-stage strategy is referred as LLMinist-Fs. The examples of prompts for both approaches can be found in Appendix A.

### 3.3 LLM Selection

For the zero-shot approach (LLMinist-Z), we evaluate three widely-used large language models (LLMs): GPT-4o mini (OpenAI, 2024), GPT-4o (OpenAI, 2024), and Gemini 1.5 Pro (Team, 2024), which are closed-source and pre-trained models developed by OpenAI and Google, respectively. 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. For clarity and simplicity, the LLMinist-Z with GPT-4o mini, GPT-4o, or Gemini 1.5 Pro is denoted as the respective LLM name. The LLMinist-F with GPT-4o mini is referred to as LLMinist-F.

### 3.4 Compared Methods

We benchmark LLMinist against a comprehensive set of state-of-the-art non-LLM 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.

### 3.5 Datasets

We selected three distinct types of spatial transcriptomic datasets with manual annotations:

275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294

304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331

332 STARmap (Wang et al., 2018), MERFISH (Moffitt  
333 et al., 2018), and Visium (Maynard et al., 2021).  
334 STARmap and MERFISH both have single-cell  
335 resolution and therefore provide cell type annotations.  
336 In contrast, Visium is a spatial transcriptomics  
337 platform that generates data at the resolution  
338 of spots, each encompassing 1-10 cells, without  
339 providing individual cell type annotations. The  
340 STARmap dataset comprises one subject with three  
341 replicates, while the MERFISH dataset contains  
342 one subject with five replicates. The Visium dataset  
343 encompasses three subjects with each having four  
344 replicates. All datasets and corresponding ground-  
345 truth annotations are downloaded from <https://figshare.com/projects/SDMBench/163942>.  
346

## 347 4 Experiment

### 348 4.1 Experimental Settings

349 **Validation and Testing** We evaluated the zero-shot  
350 approach by executing it 3 times for each repli-  
351 cate of each subject, except for MERFISH dataset.  
352 Given the suboptimal performance of zero-shot  
353 approach on MERFISH data, we conducted only  
354 a single trial for this dataset. For the evaluation  
355 of fine-tuning approaches, both LLMiST-F and  
356 LLMiST-Fs were run 3 times on each replicate  
357 of each subject. After fine-tuning the LLM, we  
358 define the application of LLMiST-F to unlabeled  
359 cells within the same replicate as *supervised val-*  
360 *idation*. Applying the fine-tuned LLM to differ-  
361 ent replicates of the same subject is termed *intra-*  
362 *subject testing*. Conversely, testing on data from  
363 other subjects is designated as *cross-subject testing*.  
364 For clarity, we denote the LLMiST-F fine-tuned  
365 with spots from Subject 1 ( $\mathcal{D}_1$ ) as LLMiST-D1-  
366 F, and its second-stage results are referred to as  
367 LLMiST-D1-Fs. The LLMiST fine-tuned with  
368 spots from  $\mathcal{D}_2$  and  $\mathcal{D}_3$  follow analogous naming  
369 conventions. Importantly, the fine-tuning replicates  
370 were excluded from testing, ensuring that perfor-  
371 mance comparisons were conducted solely within  
372 the testing replicates.

373 **Evaluation Metrics** We employ three clustering  
374 evaluation metrics to assess the accuracy of pre-  
375 dicted cluster labeling using ground truth: Nor-  
376 malized Mutual Information (NMI), Homogeneity  
377 score (HOM), and Completeness score (COM) (Pe-  
378 dredosa et al., 2012). On the other hand, to evaluate  
379 the spatial continuity of the predicted segmenta-  
380 tions, we employ three metrics: CHAOS (Yuan  
381 et al., 2024), the Percentage of Allowed outliers for

382 Segmentation (PAS) (Shang and Zhou, 2022), and  
383 the Average Silhouette Width (ASW) (Yuan et al.,  
384 2024). Benchmarking of non-LLM methods on  
385 those metrics was effectively conducted by Yuan  
386 et al. (2024), and results of non-LLM methods are  
387 derived from the work. The detailed explanation  
388 for those metrics is given in Appendix B.

389 **Spatial Neighborhood Definition** Two cells or  
390 spots are considered neighbors if the spatial dis-  
391 tance between them is below a specified thresh-  
392 old  $\delta$ , which was set to 72 (700 pixels), 100 (100 pix-  
393 els) and 344 (600 pixels) for STARmap, MERFISH,  
394 and Visium, respectively (representative examples  
395 are illustrated in the Appendix C). These thresh-  
396 olds were selected such that the resulting neighbor-  
397 hood represents the minimal functional unit of the tissue  
398 niches, encompassing the characteristic scale of  
399 its structural complexity. Spatial profiles are con-  
400 structed from all cells residing within the neighbor-  
401 hood. Dataset-specific preprocessing procedures  
402 are applied, the specifics of which are elaborated  
403 upon in the Appendix D.

### 404 4.2 Performance Comparison

405 We first report the results of intra-subject testing.  
406 The accuracy and continuity performance of the  
407 zero-shot approach on STARmap is shown in Fig-  
408 ure 2. The NMI of zero-shot and fine-tuning ap-  
409 proaches in all datasets are in Table 1.

410 **Observation 1. General LLMs possess intrinsic  
411 knowledge and capability to discern spatial niches.**  
412 All three large language models (LLMs) demon-  
413 strated superior performance compared to current  
414 non-LLM spatial clustering methods as illus-  
415 trated in the Figure 2. Specifically, the zero-shot approach  
416 with Gemini 1.5 Pro consistently achieved the high-  
417 est rank across all evaluated metrics, excelling in  
418 both accuracy-related and continuity-related as-  
419 sessments. This is uncharacteristic of non-LLM meth-  
420 ods, which may perform well in accuracy but fall  
421 short in continuity. Notably, even the GPT-4o mini,  
422 which is a smaller LLM, achieved a higher average  
423 rank than the best model-based method across all  
424 considered metrics, whether in terms of accuracy  
425 or continuity, as depicted in the Figure 2.

426 **Observation 2. A larger general LLM tends  
427 to exhibit superior performance.** Comparing the  
428 results between GPT-4o mini and GPT-4o suggests  
429 that the model with a larger parameter count gen-  
430 erally demonstrates enhanced performance (Fig-  
431 ure 2). This observation underscores the potential  
432 advantages conferred by the scale of an LLM in the

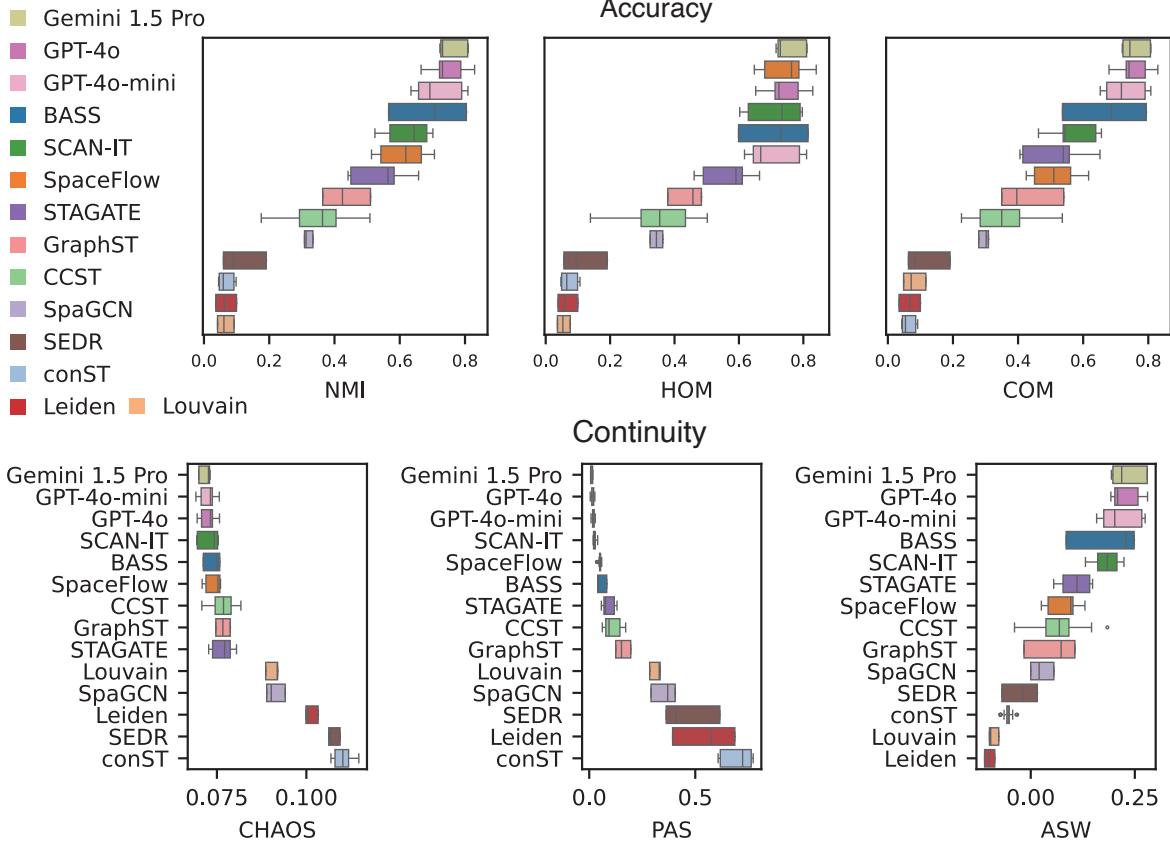


Figure 2: Assessment of the zero-shot approach on STARmap in terms of accuracy and continuity.

zero-shot task. In subsequent comparisons, Gemini 1.5 Pro will represent the zero-shot approach.

**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) 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. Further investigation revealed instances where the knowledge encoded in LLMs conflicted with the

observed data. These findings underscore the limitations of the zero-shot approach when applied to complex spatial transcriptomic data.

**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 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 metrics (Appendix E).

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

Method Type	Method	MERFISH	STARmap	Visium	Avg.	Rank
LLM-based	LLMiniST-F (supervised)	0.795±0.006	0.811±0.003	0.760±0.002	-	-
	LLMiniST-Fs	<b>0.610±0.027</b>	<b>0.752±0.013</b>	<b>0.695±0.036</b>	0.686	1.7
	LLMiniST-F	0.581±0.020	<b>0.753±0.011</b>	<b>0.678±0.051</b>	0.67	2.0
	Gemini 1.5 Pro	0.068±0.031	<b>0.755±0.040</b>	0.471±0.090	0.431	9.7
Non-LLM based	BASS	0.519±0.053	<b>0.693±0.100</b>	0.581±0.021	0.598	4.3
	SCAN-IT	0.578±0.045	0.630±0.055	<b>0.546±0.047</b>	0.585	5.0
	BayesSpace	-	-	0.565±0.087	0.565	5.0
	GraphST	0.317±0.056	0.433±0.061	0.592±0.049	0.448	6.0
	stLearn	-	-	0.552±0.014	0.552	6.0
	SpaceFlow	0.535±0.077	0.606±0.061	0.433±0.042	0.525	8.3
	CCST	0.468±0.031	0.353±0.083	0.507±0.022	0.443	8.7
	SpaGCN	0.214±0.015	0.318±0.011	0.513±0.047	0.348	9.0
	STAGATE	0.204±0.085	0.538±0.079	0.507±0.042	0.417	9.3
	SEDR	0.142±0.045	0.113±0.057	0.532±0.030	0.263	10.3
	conST	0.107±0.012	0.067±0.021	0.511±0.084	0.228	11.7
Non-Spatial	SpaGCN(HE)	-	-	0.475±0.043	0.475	13.0
	Leiden	0.177±0.004	0.066±0.026	0.329±0.009	0.191	13.0
	Louvain	0.169±0.009	0.065±0.021	0.336±0.014	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  $\pm$  Standard Deviation)

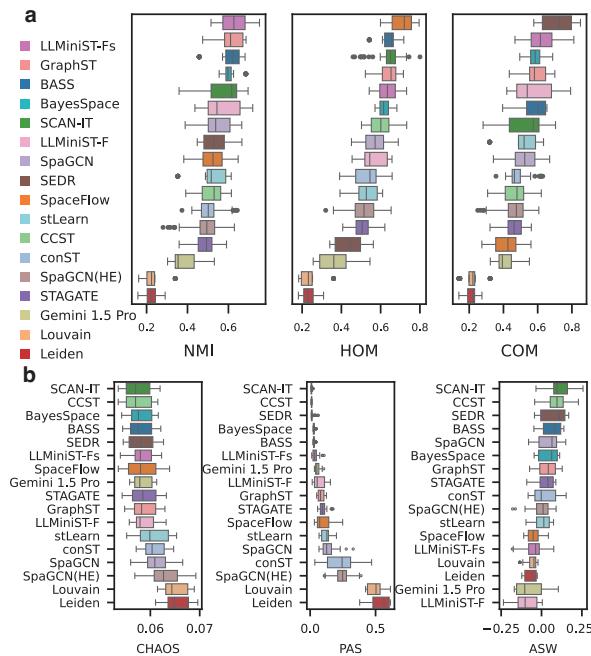


Figure 3: The performance of LLMiniST on the Subjects 2 and 3 of the Visium dataset. **a**, Accuracy-related metrics; **b**, Continuity related metrics. LLMiniST-F is fine-tuned with Subject 1.

**Observation 6. The fine-tuning approach exhibits high tolerance to subject-level variance, demonstrating good cross-subject generalizability.** We compared the performance of LLMiniST-D1-Fs with other methods on replicates from Subjects 2 and 3. Notably, even though the spots used

to fine-tune LLMiniST originated from Subject 1, LLMiniST-D1-Fs still achieved the highest average rank across the three accuracy matrices (Figure 3). This demonstrates the generalizability of the fine-tuning approach to replicates from different subjects. While LLMiniST-D1-Fs does not achieve the best performance in terms of continuity, this may be attributed to the lack of strong spatial constraints within the LLMiniST framework. LLMiniST processes each spot individually, without considering the overall spatial context.

**Observation 7. The two-stage strategy enhances generalizability.** Both intra-subject (Table 1) and cross-subject (Figure 3) evaluations 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 LLMiniST-D1-Fs to achieve a accuracy comparable to or even surpassing that of LLMiniST-D2-F and LLMiniST-D3-F within their respective subjects (Figure 3a). Notably, as illustrated in Figure 3b 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 so-

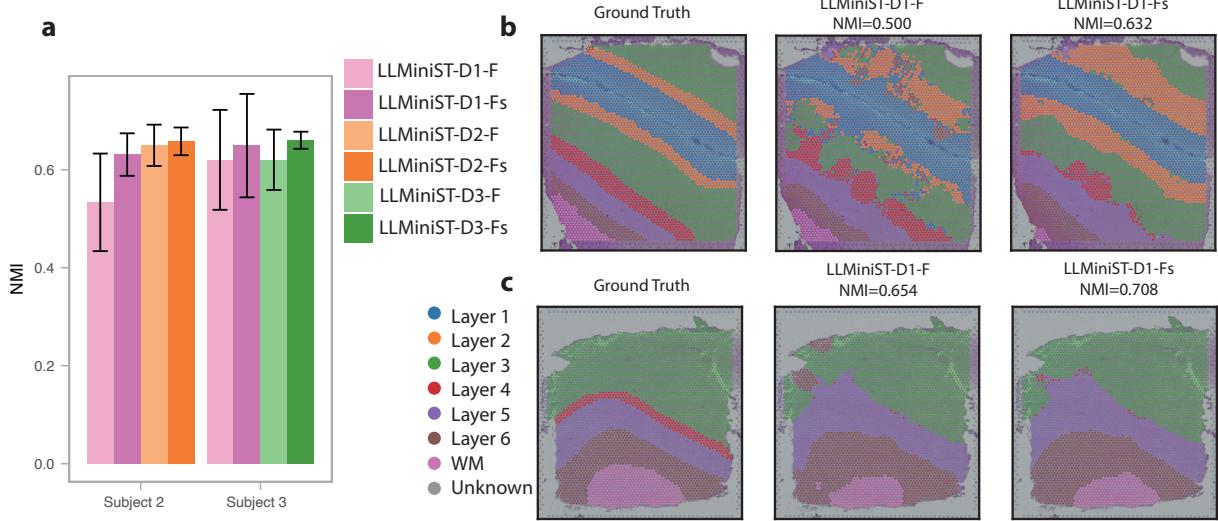


Figure 4: NMI improvement achieved by the proposed two-stage strategy on the Subjects 2 and 3 of the Visium dataset. (a) Comparison of LLMiST fine-tuned using spots from different subjects, with and without two-stage strategy. (b) Identification results for a replicate of Subject 2. (c) Identification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

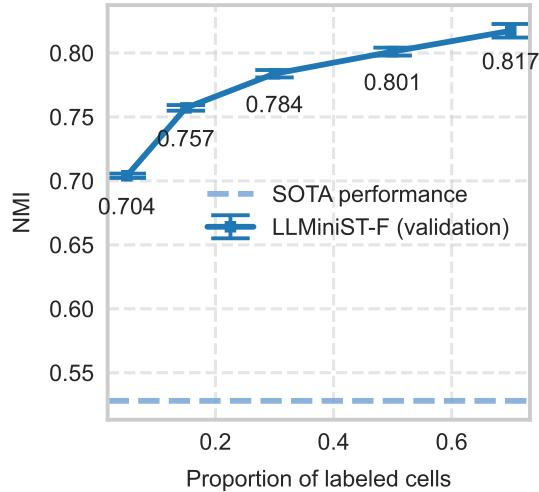


Figure 5: Impact of labeled cell proportion on the performance of LLMiST-F, evaluated on a supervised validation set from the MERFISH dataset. Error bars indicate the standard deviation.

lution. Despite the misalignment of niche components between Subject 3 and the other subjects, particularly concerning the absent niches of Layers 1 and 2, both LLMiST-D1-F and LLMiST-D1-Fs proficiently circumvent the misclassification of spots into these non-existent niches (Figure 3c).

#### 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 LLMiST 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.

## 5 Conclusion

In this work, we comprehensively compared the performance of LLMs with traditional non-LLM methods for niche identification in spatial transcriptomic data. Our proposed spatial context prompt successfully leverage the rich knowledge embedded within pre-trained LLMs to identify spatial niches without requiring data-specific training. We demonstrate the feasibility of zero-shot LLMiST, particularly when clear cell type annotations and simple niche structures are available. Futhurmore, by incorporating a two-stage fine-tuning strategy, LLMiST exhibits strong generalizability to samples from either the same or different subjects, highlighting its potential for application to large-scale datasets. The strength of LLMiST lies in its simplicity, effectiveness, and broad utility.

518  
519  
520  
521  
522  
523

524  
525  
526  
527  
528

529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558

## 559 6 Limitations

560 Recent advancements in spatial transcriptomics  
561 have introduced a suite of new technologies that  
562 require thorough assessment. Besides, the avail-  
563 ability of datasets with reliable ground truth in-  
564 formation is currently a limiting factor. Further  
565 development and validation necessitate the acqui-  
566 sition of ground truth data from a greater variety  
567 of tissue regions. To improve annotation efficiency,  
568 future work will investigate methods for selecting  
569 a minimal subset of cells that provides maximal in-  
570 formation for fine-tuning, as opposed to the current  
571 random selection approach.

## 572 7 Ethics Statement

573 This research adheres to the ethical guidelines out-  
574 lined in the ACL Code of Ethics<sup>4</sup>. All datasets  
575 employed in this study are publicly available re-  
576 sources. Similarly, all language models used are  
577 publicly accessible and utilized in accordance with  
578 their respective terms of use, as specified by Ope-  
579 nAI's Terms of Use<sup>5</sup> and Gemini API Terms of  
580 Service<sup>6</sup>. To ensure the reproducibility and trans-  
581 parency of our findings, we commit to releasing the  
582 code and associated materials upon publication.

## 583 References

- 584 Haiyang Bian, Yixin Chen, Erpai Luo, Xinze Wu, Min-  
585 sheng Hao, Lei Wei, and Xuegong Zhang. 2024.  
586 General-purpose pre-trained large cellular models for  
587 single-cell transcriptomics. *National Science Review*,  
588 11(11):nwae340.
- 589 Vincent D Blondel, Jean-Loup Guillaume, Renaud Lam-  
590 biotte, and Etienne Lefebvre. 2008. Fast unfold-  
591 ing of communities in large networks. *Journal*  
592 of Statistical Mechanics: Theory and Experiment,  
593 2008(10):P10008.
- 594 Dylan M Cable, Evan Murray, Vignesh Shanmugam, Si-  
595 mon Zhang, Luli S Zou, Michael Diao, Haiqi Chen,  
596 Evan Z Macosko, Rafael A Irizarry, and Fei Chen.  
597 2022. Cell type-specific inference of differential ex-  
598 pression in spatial transcriptomics. *Nat Methods*,  
599 19(9):1076–1087.
- 600 Zixuan Cang, Xinyi Ning, Annika Nie, Min Xu, and  
601 Jing Zhang. 2021. SCAN-IT: Domain segmentation  
602 of spatial transcriptomics images by graph neural  
603 network. *BMVC*, 32.

<sup>4</sup><https://www.aclweb.org/portal/content/acl-code-ethics>

<sup>5</sup><https://openai.com/policies/row-terms-of-use/>

<sup>6</sup><https://ai.google.dev/gemini-api/terms>

Yiqun Chen and James Zou. 2024. Simple and effective embedding model for single-cell biology built from chatgpt. <i>Nature Biomedical Engineering</i> .	604
	605
	606
Kangning Dong and Shihua Zhang. 2022. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. <i>Nature Communications</i> , 13:1739.	607
	608
	609
	610
Ofer Elhanani, Raz Ben-Uri, and Leeat Keren. 2023. Spatial profiling technologies illuminate the tumor microenvironment. <i>Cancer Cell</i> , 41(3):404–420.	611
	612
	613
Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. <i>Preprint</i> , arXiv:2403.14608.	614
	615
	616
	617
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. <i>Nature Biotechnology</i> .	618
	619
	620
	621
	622
	623
Wenpin Hou and Zhicheng Ji. 2024. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. <i>Nature Methods</i> , 21(8):1462–1465.	624
	625
	626
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. <i>Nature Methods</i> , 18:1342–1351.	627
	628
	629
	630
	631
Jiachen Li, Siheng Chen, Xiaoyong Pan, Ye Yuan, and Hong-Bin Shen. 2022. Cell clustering for spatial transcriptomics data with graph neural networks. <i>Nature Computational Science</i> , 2:399–408.	632
	633
	634
	635
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. <i>Genome Biology</i> , 23:168.	636
	637
	638
	639
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. <i>Comput Struct Biotechnol J</i> , 23:106–128.	640
	641
	642
	643
	644
Yunqing Liu, Ningshan Li, Ji Qi, Gang Xu, Jiayi Zhao, Nating Wang, Xiayuan Huang, Wenhao Jiang, Huan-huan 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. <i>Genome Biol.</i> , 25(1):271.	645
	646
	647
	648
	649
	650
	651
	652
Yahui Long, Kok Siong Ang, Mengwei Li, et al. 2023. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. <i>Nature Communications</i> , 14:1155.	653
	654
	655
	656

657	Yuren Mao, Yuhang Ge, Yijiang Fan, Wenyi Xu, Yu Mi, Zhonghao Hu, and Yunjun Gao. 2024. <b>A survey on lora of large language models.</b> <i>Frontiers of Computer Science</i> , 19(7).	711
658		712
659		713
660		714
661	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. <b>Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex.</b> <i>Nat. Neurosci.</i> , 24(3):425–436.	715
662		716
663		717
664		718
665		719
666		720
667		721
668		
669		
670	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. <i>Science</i> , 362(6416):eaau5324.	722
671		723
672		724
673		725
674		726
675		
676		
677	Jeffrey R Moffitt, Emma Lundberg, and Holger Heyn. 2022. <b>The emerging landscape of spatial profiling technologies.</b> <i>Nat Rev Genet</i> , 23(12):741–759.	727
678		728
679		729
680	OpenAI. 2024. <b>Gpt-4o system card.</b> <i>Preprint</i> , arXiv:2410.21276.	730
681		
682	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. <b>Scikit-learn: Machine learning in python.</b> <i>CoRR</i> , abs/1201.0490.	731
683		732
684		733
685		734
686		735
687		
688		
689	Duy Pham, Xiao Tan, Blake Balderson, et al. 2023. <b>Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues.</b> <i>Nature Communications</i> , 14:7739.	736
690		737
691		738
692		739
693	Honglei Ren, Benjamin L. Walker, Zixuan Cang, and Nie Qing. 2022. <b>Identifying multicellular spatiotemporal organization of cells with spaceflow.</b> <i>Nature Communications</i> , 13:4076.	740
694		741
695		
696		
697	Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. <b>A systematic survey of prompt engineering in large language models: Techniques and applications.</b> <i>Preprint</i> , arXiv:2402.07927.	742
698		743
699		744
700		
701		
702	Lulu Shang and Xiang Zhou. 2022. <b>Spatially aware dimension reduction for spatial transcriptomics.</b> <i>Nature Communications</i> , 13(1):7203.	745
703		746
704		747
705	Gemini Team. 2024. <b>Gemini 1.5: Unlocking multi-modal understanding across millions of tokens of context.</b> <i>Preprint</i> , arXiv:2403.05530.	748
706		
707		
708	V. A. Traag, L. Waltman, and N. J. van Eck. 2019. <b>From louvain to leiden: guaranteeing well-connected communities.</b> <i>Scientific Reports</i> , 9(1):5233.	750
709		751
710		752
711		753
712		
713		
714		
715	Katy Vandereyken, Alejandro Sifrim, Bernard Thienpont, and Thierry Voet. 2023. <b>Methods and applications for single-cell and spatial multi-omics.</b> <i>Nat Rev Genet</i> , 24(8):494–515.	715
716		716
717		717
718		718
719		719
720		720
721		721
722	Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2021. <b>A comprehensive survey on graph neural networks.</b> <i>IEEE Transactions on Neural Networks and Learning Systems</i> , 32(1):4–24.	722
723		723
724		724
725		725
726		726
727	Hang Xu, Huazhu Fu, Yahui Long, and Kok Siong Ang. 2024. <b>Unsupervised spatially embedded deep representation of spatial transcriptomics.</b> <i>Genome Medicine</i> , 16:12.	727
728		728
729		729
730		730
731	Zhiyuan Yuan, Fangyuan Zhao, Senlin Lin, Yu Zhao, Jianhua Yao, Yan Cui, Xiao-Yong Zhang, and Yi Zhao. 2024. <b>Benchmarking spatial clustering methods with spatially resolved transcriptomics data.</b> <i>Nature Methods</i> , 21(4):712–722.	731
732		732
733		733
734		734
735		735
736	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. <b>Spatially resolved single-cell transcriptomics at molecular resolution.</b> <i>Science</i> , 380(6652):eadd3067.	736
737		737
738		738
739		739
740		740
741		741
742	Edward Zhao, Matthew R. Stone, Xing Ren, et al. 2021. <b>Spatial transcriptomics at subspot resolution with bayesspace.</b> <i>Nature Biotechnology</i> , 39:1375–1384.	742
743		743
744		744
745	Yongshuo Zong, Tingyang Yu, Xuesong Wang, Yixuan Wang, Zhihang Hu, and Yu Li. 2022. <b>const: an interpretable multi-modal contrastive learning framework for spatial transcriptomics.</b> <i>bioRxiv</i> .	745
746		746
747		747
748		748
749	<b>A Examples of prompt</b>	749
750	We provide examples of spatail 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.	750
751		751
752		752
753		753
754	<b>B Explanation of Metrics</b>	754
755	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	755
756		756
757		757
758		758
759		759
760		760
761		761
762		762

**a**

Outputs: {'fx'}

You will be provided with a list of cell types in the neighborhood of a cell and highly variable genes expressed in the neighborhood. All cells are from the mouse hypothalamic preoptic region. The Neighbor cell-types are ordered from most frequent to least frequent. The marker genes are ordered from highest expression to lowest expression. Your task is to identify the microenvironments the cell belong to. The cell can only belong to one of the 8 microenvironments: MPA, MPN, BST, fx, PVH, PVT, V3, PV. Only output the most possible microenvironment in the plain text format like this:

Outputs: {'microenvironment 1'} Below are the list of cell-types and highly variable genes in the cell's neighborhood:

Neighbor cell-types: {Astrocyte, Inhibitory, Mature oligodendrocytes, Excitatory}  
Marker genes: {Ermn, Aldh1l1, Mbp, Sgk1, Sox8, Sst, Ebf3, Gjc3, Mlc1, Cxcl14}

**b**

Outputs: {'Layer 3'}

You will be provided with a list of tissue microenvironments in Human dorsolateral prefrontal cortex. The cell types and marker genes in the neighborhood of those microenvironments will be shown. The neighbor cell-types are ordered from most frequent to least frequent. The genes are ordered from highest expression to lowest expression. Below are typical examples of microenvironments:

Microenvironment: {Layer4}, Neighbor cell-types: {Excit\_L4, Excit\_L2\_3, Excit\_L5, Excit\_L3\_4\_5, Astro, Excit\_L6, Inhib}, marker genes: {PVALB, RORB, RAB3C, PCP4, PLP1, ZMAT4, FA2H};  
Microenvironment: {Layer6}, Neighbor cell-types: {Excit\_L4, Excit\_L5, Excit\_L2\_3, Excit\_L6, Excit\_L3\_4\_5, Excit\_L5\_6, Oligo}, marker genes: {KRT17, PLP1, MOBP, FEZF2, CPNE5, GFAP, ETV1};

Microenvironment: {Layer3}, Neighbor cell-types: {Excit\_L2\_3, Astro, Excit\_L3\_4\_5, Excit\_L4, Excit\_L6, OPC, Excit\_L3}, marker genes: {HPCAL1, CUX2, LAMP5, FREM3, KCNIP2, FOXJ1};

Microenvironment: {WM}, Neighbor cell-types: {Oligo, Excit\_L5, Excit\_L4, Astro, Excit\_L6, Excit\_L2\_3, Micro}, marker genes: {GFAP, PLP1, MOBP, KRT17, AQP4, FA2H, ETV1};  
Microenvironment: {Layer5}, Neighbor cell-types: {Excit\_L5, Excit\_L4, Excit\_L2\_3, Excit\_L3\_4\_5, Excit\_L6, Oligo, Excit\_L5\_6}, marker genes: {PCP4, PLP1, FEZF2, ETV1, RORB, RAB3C, MOBP}

Based on the examples, identify which microenvironment categories the following cell belongs to. The target cell can only be one of the 5 microenvironments: Layer4, Layer6, Layer3, WM, Layer5. Only output the most possible microenvironment in the plain text like this: Outputs: {'microenvironment 1'}. Below are cell types and marker genes in the neighborhood of the cell:

Neighbor cell-types: {Excit\_L2\_3, Excit\_L3\_4\_5, Excit\_L4, Astro, Excit\_L6, OPC, Excit\_L3},  
Marker genes: {PVALB, ZMAT4, FREM3, CUX2, BGN, ETV1, LAMP5}



Figure 6: Examples of spatial context prompts. a, a zero-shot prompt for a cell in MERFISH dataset; b, a fine-tuning prompt for a spot in Visium dataset.

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.

## C Neighborhood Examples

The spatial scope of our neighborhood analysis is visualized in Figure 7. Each sub-figure represents a different dataset, and within each, a red circle delineates the perimeter of the neighborhood that was taken into account for constructing spatial profile.

## D Data Preprocessing for Spatial Context Prompt

**Spatial Context Prompt for STARmap** The STARmap dataset features clear cell type annotations 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, SDe-

763  
764  
765  
766

767

768  
769  
770  
771  
772

773  
774

775  
776  
777  
778  
779  
780

781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801

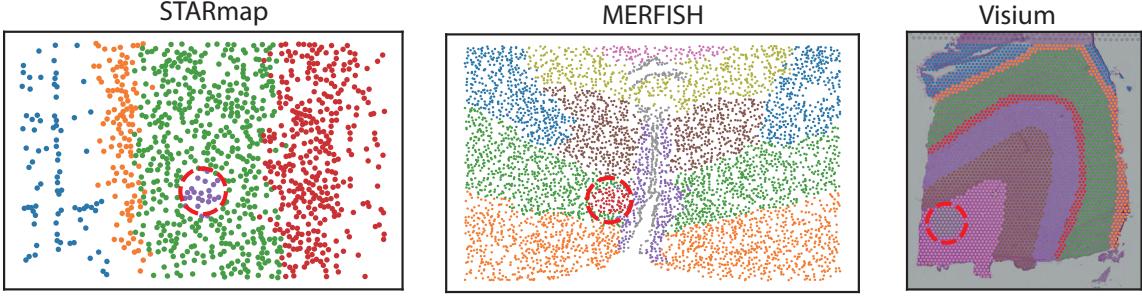


Figure 7: Depiction of the neighborhood size across three different datasets.

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

## E 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 2 - 6).

	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 2: Intra-subject HOM: Mean ± SD

	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 3: 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 4: 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 5: 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 6: Intra-subject ASW: Mean ± SD