



## OPEN ACCESS

## EDITED BY

Martin Hemberg,  
Harvard Medical School, United States

## REVIEWED BY

Hyun Jung Park,  
University of Pittsburgh, United States  
Alexandro Trevino,  
Enable Medicine, United States

## \*CORRESPONDENCE

Ziheng Duan,  
✉ zihend1@uci.edu

RECEIVED 26 September 2025

REVISED 27 December 2025

ACCEPTED 30 December 2025

PUBLISHED 12 January 2026

## CITATION

Zou Z and Duan Z (2026) SpaLLM: a general  
framework for spatial domain identification  
with large language models.  
*Front. Bioinform.* 5:1713975.  
doi: 10.3389/fbinf.2025.1713975

## COPYRIGHT

© 2026 Zou and Duan. This is an open-access  
article distributed under the terms of the  
[Creative Commons Attribution License \(CC  
BY\)](#). The use, distribution or reproduction in  
other forums is permitted, provided the  
original author(s) and the copyright owner(s)  
are credited and that the original publication  
in this journal is cited, in accordance with  
accepted academic practice. No use,  
distribution or reproduction is permitted  
which does not comply with these terms.

# SpaLLM: a general framework for spatial domain identification with large language models

Zeyu Zou and Ziheng Duan\*

Department of Computer Science, University of California, Irvine, Irvine, CA, United States

Spatial transcriptomics (ST) technologies enable the profiling of gene expression while preserving spatial context, offering unprecedented insights into tissue organization. However, traditional spatial domain identification methods primarily rely on gene expression matrices and spatial coordinates while overlooking the rich biological knowledge encoded in gene functional descriptions. Here, we propose SpaLLM, a general framework that integrates large language model (LLM) embeddings of gene descriptions with conventional spatial transcriptomics analysis. Our approach leverages pre-computed GenePT embeddings from NCBI gene summaries to create biologically-informed gene representations. SpaLLM combines these LLM-derived gene features with cell-gene expression matrices through matrix multiplication, generating enriched cell representations that capture both expression patterns and functional knowledge. These enriched features are then integrated with existing graph-based spatial analysis methods for improved spatial domain identification. Extensive validation on 12 sequencing-based Visium sections and an independent imaging-based osmFISH dataset demonstrate that SpaLLM consistently enhances spatial domain identification. Our modular framework can be seamlessly integrated with existing spatial analysis pipelines, making it broadly applicable to diverse research scenarios.

## KEYWORDS

graph neural networks, large language models, multimodality, spatial domain identification, spatial transcriptomics

## 1 Introduction

Spatial transcriptomics (ST) technologies have revolutionized our understanding of tissue architecture by enabling simultaneous measurement of gene expression and spatial location information (Rodrigues et al., 2019; Emani et al., 2024; Ruzicka et al., 2024). A fundamental task in ST analysis is spatial domain identification, which aims to partition tissue sections into distinct regions based on similar gene expression patterns and spatial proximity. These spatial domains often correspond to anatomical structures, functional units, or pathological states, making their accurate identification crucial for understanding tissue organization and disease mechanisms (Chen et al., 2022).

Current spatial domain identification methods predominantly follow a graph-based approach, where spots or cells are represented as nodes in a spatial graph, and edges encode spatial proximity relationships (Hu et al., 2021; Zhao et al., 2021; Duan et al., 2024a; Duan et al., 2024b; Duan et al., 2025b). These methods typically employ graph neural networks (GNNs) or graph autoencoders to learn latent representations from cell-by-gene expression matrices and spatial coordinates, followed by clustering algorithms to identify

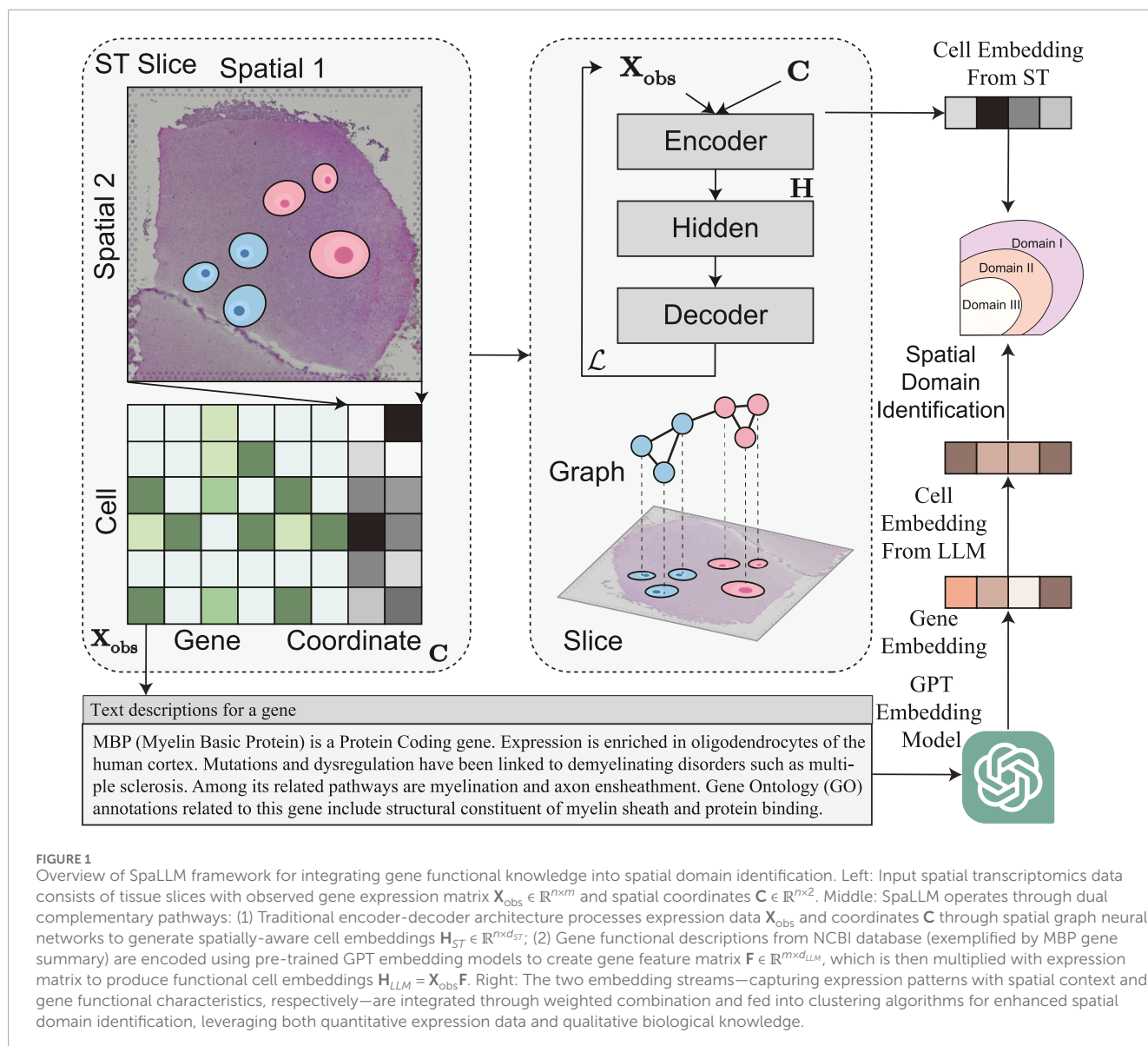


FIGURE 1

Overview of SpaLLM framework for integrating gene functional knowledge into spatial domain identification. Left: Input spatial transcriptomics data consists of tissue slices with observed gene expression matrix  $\mathbf{X}_{\text{obs}} \in \mathbb{R}^{n \times m}$  and spatial coordinates  $\mathbf{C} \in \mathbb{R}^{n \times 2}$ . Middle: SpaLLM operates through dual complementary pathways: (1) Traditional encoder-decoder architecture processes expression data  $\mathbf{X}_{\text{obs}}$  and coordinates  $\mathbf{C}$  through spatial graph neural networks to generate spatially-aware cell embeddings  $\mathbf{H}_{\text{ST}} \in \mathbb{R}^{n \times d_{\text{ST}}}$ ; (2) Gene functional descriptions from NCBI database (exemplified by MBP gene summary) are encoded using pre-trained GPT embedding models to create gene feature matrix  $\mathbf{F} \in \mathbb{R}^{m \times d_{\text{LLM}}}$ , which is then multiplied with expression matrix to produce functional cell embeddings  $\mathbf{H}_{\text{LLM}} = \mathbf{X}_{\text{obs}} \mathbf{F}$ . Right: The two embedding streams—capturing expression patterns with spatial context and gene functional characteristics, respectively—are integrated through weighted combination and fed into clustering algorithms for enhanced spatial domain identification, leveraging both quantitative expression data and qualitative biological knowledge.

spatial domains (Dong and Zhang, 2022; Long et al., 2023). While effective, these approaches have a fundamental limitation: they treat genes merely as numerical features without leveraging the extensive biological knowledge accumulated about gene functions, pathways, and interactions.

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in understanding biological text. The GenePT framework has shown that LLM embeddings of gene descriptions from NCBI can effectively capture biological relationships and improve downstream tasks in single-cell analysis (Chen and Zou, 2024). Specifically, GenePT uses pre-computed OpenAI text embeddings on NCBI gene summaries, demonstrating that these embeddings often outperform expression-based methods for various biological tasks.

Motivated by these observations, we propose SpaLLM, as shown in Figure 1, a general framework that integrates LLM-derived gene functional features with traditional spatial transcriptomics analysis. Our key insight is that gene functional

descriptions contain rich biological knowledge that can inform spatial domain identification beyond what expression patterns alone can reveal. By leveraging pre-trained language models to encode gene descriptions from biological databases, we create biologically-informed gene representations that capture functional relationships, pathway memberships, and molecular mechanisms.

The SpaLLM framework introduces a novel approach that enhances traditional spatial transcriptomics analysis by integrating gene functional knowledge. Following the standard encoder-decoder paradigm, we first obtain cell embeddings from ST data using existing graph-based methods. Simultaneously, we derive functional cell embeddings by multiplying the cell-by-gene expression matrix with LLM-derived gene embeddings ( $\mathbf{H} = \mathbf{X}_{\text{obs}} \mathbf{F}$ ), where the gene feature matrix  $\mathbf{F} \in \mathbb{R}^{m \times d}$  encodes biological functions extracted from NCBI descriptions. We then combine these two complementary cell representations—one capturing expression patterns and spatial context, the other encoding functional

characteristics—and feed the integrated features into clustering algorithms for improved spatial domain identification.

Our main contributions are as follows:

- We introduce the first systematic framework to integrate LLM embeddings of gene functional descriptions into spatial transcriptomics analysis;
- We demonstrate that incorporating biological knowledge through gene text descriptions significantly improves spatial domain identification accuracy, particularly for low-quality data;
- We provide a modular framework compatible with existing spatial analysis methods, enabling broad adoption across different research pipelines;
- We conduct comprehensive experiments across multiple datasets with varying quality levels, showing consistent 3.8%–400% improvements over state-of-the-art methods.

## 2 Methods

### 2.1 Problem formulation and SpaLLM architecture

We formulate the problem of spatial domain identification as a clustering task on a graph. Let  $\mathbf{X} \in \mathbb{R}^{n \times m}$  be the raw gene expression matrix, where  $n$  is the number of spatial spots and  $m$  is the number of genes. The spatial coordinates of the spots are given by the matrix  $\mathbf{C} \in \mathbb{R}^{n \times 2}$ . The goal is to learn a mapping  $f: (\mathbf{X}, \mathbf{C}) \rightarrow \mathbf{Z}$  that transforms the input data into a latent feature space  $\mathbf{Z} \in \mathbb{R}^{n \times d_{\text{final}}}$ , where  $d_{\text{final}}$  is the final embedding dimension. Subsequently, a clustering algorithm  $\mathcal{C}$  is applied to  $\mathbf{Z}$  to identify the spatial domains, i.e.,  $\mathbf{D} = \mathcal{C}(\mathbf{Z})$ .

The SpaLLM framework enhances this process by incorporating biological knowledge from gene functional descriptions. We introduce a dual-stream encoding architecture that combines a conventional spatial encoder with a novel LLM-based functional encoder, as illustrated in a figure. The outputs of these two encoders are fused to produce the final enriched embeddings  $\mathbf{Z}$ , which are then used for clustering.

### 2.2 Gene functional encoding

To capture the biological meaning of genes, we leverage pre-computed embeddings from a large language model trained on a corpus of gene functional descriptions (e.g., NCBI gene summaries). Let  $\mathbf{F} \in \mathbb{R}^{m \times d_{\text{LLM}}}$  be the pre-computed gene feature matrix, where each row  $\mathbf{f}_j \in \mathbb{R}^{d_{\text{LLM}}}$  is the vector representation of the  $j$ -th gene's description. This feature matrix serves as a look-up table for the functional representation of each gene. The functional encoder transforms the gene expression data into a functional feature space. We define the functional spot embeddings  $\mathbf{H}_{\text{LLM}} \in \mathbb{R}^{n \times d_{\text{LLM}}}$  as:

$$\mathbf{H}_{\text{LLM}} = \mathcal{E}_{\text{func}}(\mathbf{X}) = \mathbf{X}\mathbf{F}$$

Here,  $\mathcal{E}_{\text{func}}$  is the functional encoder. This operation implicitly weighs the contribution of each functional dimension based on the expression level of corresponding genes in each spot.

### 2.3 Spatial encoder

The spatial encoder captures both gene expression patterns and spatial context. We construct a spatial graph  $G = (V, E)$  where the set of vertices  $V$  corresponds to the  $n$  spots. The edge weights are defined by the adjacency matrix  $\mathbf{A}$ , where  $A_{ij}$  represents the spatial proximity between spot  $i$  and spot  $j$ . A common choice for computing  $\mathbf{A}$  is a Gaussian kernel on the spatial coordinates:

$$A_{ij} = \exp\left(-\frac{\|\mathbf{c}_i - \mathbf{c}_j\|^2}{2\sigma^2}\right) \quad \text{if } \|\mathbf{c}_i - \mathbf{c}_j\| \leq \tau, \text{ otherwise } 0$$

Here,  $\sigma$  and  $\tau$  are hyperparameters.

The spatial encoder,  $\mathcal{E}_{\text{spat}}$ , is a Graph Neural Network (GNN) that learns spatially-aware spot embeddings  $\mathbf{H}_{\text{ST}} \in \mathbb{R}^{n \times d_{\text{ST}}}$  by propagating and aggregating features over the graph. The input to the GNN is the normalized gene expression matrix,  $\mathbf{X}_{\text{norm}}$ :

$$\mathbf{H}_{\text{ST}} = \mathcal{E}_{\text{spat}}(\mathbf{X}_{\text{norm}}, \mathbf{A})$$

The GNN layers are typically defined by a message passing scheme. For a multi-layer GNN, the update rule for the  $l$ -th layer is:

$$\mathbf{H}^{(l+1)} = \sigma(\widehat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)})$$

where  $\mathbf{H}^{(0)} = \mathbf{X}_{\text{norm}}$ ,  $\widehat{\mathbf{A}}$  is the normalized adjacency matrix,  $\mathbf{W}^{(l)}$  is a trainable weight matrix, and  $\sigma(\cdot)$  is a non-linear activation function.

### 2.4 Feature fusion and clustering

The embeddings from the two encoders,  $\mathbf{H}_{\text{ST}} \in \mathbb{R}^{n \times d_{\text{ST}}}$  and  $\mathbf{H}_{\text{LLM}} \in \mathbb{R}^{n \times d_{\text{LLM}}}$ , may have different dimensions ( $d_{\text{ST}} \neq d_{\text{LLM}}$ ). To ensure feature alignment before fusion, we first perform Principal Component Analysis (PCA) on the LLM embeddings to project them into the same dimension as the spatial embeddings. This yields the dimension-reduced functional embeddings,  $\mathbf{H}_{\text{LLM}}^{\text{PCA}} \in \mathbb{R}^{n \times d_{\text{ST}}}$ .

The final spot representation  $\mathbf{Z} \in \mathbb{R}^{n \times d_{\text{ST}}}$  is then obtained by a weighted combination of the spatial and dimension-reduced functional embeddings:

$$\mathbf{Z} = \mathcal{F}(\mathbf{H}_{\text{ST}}, \mathbf{H}_{\text{LLM}}^{\text{PCA}}) = \alpha \cdot \mathbf{H}_{\text{ST}} + \beta \cdot \mathbf{H}_{\text{LLM}}^{\text{PCA}}$$

The hyperparameters  $\alpha$  and  $\beta$  control the contribution of each embedding stream. The combined embeddings  $\mathbf{Z}$  capture both spatial proximity/expression patterns and biological functional knowledge.

Finally, the spatial domains are identified by applying a clustering algorithm, such as K-means, Louvain, or mclust, to the fused feature matrix  $\mathbf{Z}$ .

## 3 Experimental setup

### 3.1 Datasets and data quality simulation

We evaluate SpaLLM on the human dorsolateral prefrontal cortex (DLPFC) spatial transcriptomics datasets from

TABLE 1 Summary of datasets and quality simulation.

Slice ID	Dimensions (cells × genes)	Original density (%)	Quality level density (%)			
			Q1	Q2	Q3	Q4
151507	(4221, 3000)	12.75	6.38	3.19	1.60	0.80
151508	(4381, 3000)	11.96	5.98	2.99	1.50	0.75
151509	(4788, 3000)	12.92	6.46	3.23	1.62	0.81
151510	(4595, 3000)	12.88	6.44	3.22	1.61	0.81
151669	(3636, 3000)	12.71	6.36	3.18	1.59	0.80
151670	(3484, 3000)	12.72	6.36	3.18	1.59	0.80
151671	(4093, 3000)	13.70	6.85	3.43	1.72	0.86
151672	(3888, 3000)	13.41	6.70	3.35	1.68	0.84
151673	(3611, 3000)	15.18	7.59	3.80	1.90	0.95
151674	(3635, 3000)	16.81	8.41	4.20	2.10	1.05
151675	(3566, 3000)	13.60	6.80	3.40	1.70	0.85
151676	(3431, 3000)	14.36	7.18	3.59	1.80	0.90

Maynard et al. (2021), which consist of 12 tissue sections with manually annotated spatial domains. To assess the robustness of our framework against varying data quality, we adopt a systematic simulation strategy where data quality is reduced by introducing sparsity Duan et al. (2025a). The original, unaltered data serves as a baseline for comparison. We create four simulated quality levels by randomly masking a percentage of the non-zero gene expression values: Q1 (50% masked), Q2 (75% masked), Q3 (87.5% masked), and Q4 (93.75% masked). This process generates a comprehensive testbed of 48 simulated datasets (12 original sections × 4 quality levels), enabling a robust evaluation of SpaLLM’s performance under different conditions. The characteristics of these datasets are summarized in Table 1.

To validate cross-platform generalizability, we incorporated the osmFISH dataset (Codeluppi et al., 2018) of the mouse somatosensory cortex. This dataset utilizes imaging-based technology, which provides a higher spatial resolution but fewer genes compared to the sequencing-based Visium platform, offering a complementary modality for testing.

3.2 Implementation details

3.2.1 GenePT feature configuration

We use the pre-computed GenePT embeddings (Chen and Zou, 2024) with the following specifications:

- Embedding model: OpenAI *text-embedding-ada-002*.

- Feature dimension:  $d = 1536$ .
- Gene coverage: 33,000+ genes with NCBI annotations.

3.2.2 Model hyperparameters

We adopt four representative baselines: SpaceFlow (Ren et al., 2022), STAGATE (Dong and Zhang, 2022), GraphST (Long et al., 2023), and stCluster (Wang et al., 2024). The hyperparameters for each method are set following the configurations recommended in their original papers. For SpaLLM integration, we set the weighting parameters  $\alpha = 0.5$  and  $\beta = 0.5$  as default values to achieve balanced integration between spatial expression features and functional embeddings.

3.3 Evaluation metrics

We evaluate spatial domain identification using the Adjusted Rand Index (ARI), a robust metric for measuring the similarity between a predicted clustering and the ground truth. The ARI corrects for chance agreements and has a value of 1.0 for perfect clustering and 0 for random assignments. The formula for ARI is defined as:

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

Where  $RI$  is the Rand Index,  $E[RI]$  is the expected Rand Index for a random partition, and  $max(RI)$  is the maximum possible Rand Index.

TABLE 2 Spatial domain identification performance (ARI) on Donor 1 datasets. The best performance for each quality level and dataset is bolded.

Method	Quality	Donor 1			
		151507	151508	151509	151510
SpaceFlow	Q0	0.41 $\pm$ 0.02	0.38 $\pm$ 0.03	0.34 $\pm$ 0.04	0.36 $\pm$ 0.02
	Q1	0.39 $\pm$ 0.02	0.36 $\pm$ 0.04	0.32 $\pm$ 0.04	0.34 $\pm$ 0.02
	Q2	0.38 $\pm$ 0.04	0.30 $\pm$ 0.04	0.31 $\pm$ 0.04	0.32 $\pm$ 0.07
	Q3	0.17 $\pm$ 0.03	0.16 $\pm$ 0.02	0.25 $\pm$ 0.04	0.18 $\pm$ 0.02
	Q4	0.07 $\pm$ 0.01	0.10 $\pm$ 0.03	0.05 $\pm$ 0.07	0.06 $\pm$ 0.03
SpaceFlow + SpaLLM	Q0	0.43 $\pm$ 0.02	0.40 $\pm$ 0.03	0.36 $\pm$ 0.04	0.38 $\pm$ 0.02
	Q1	0.42 $\pm$ 0.02	0.39 $\pm$ 0.04	0.35 $\pm$ 0.04	0.37 $\pm$ 0.02
	Q2	0.41 $\pm$ 0.04	<b>0.33</b> $\pm$ 0.04	0.34 $\pm$ 0.04	0.35 $\pm$ 0.07
	Q3	0.20 $\pm$ 0.03	<b>0.19</b> $\pm$ 0.02	0.28 $\pm$ 0.04	0.21 $\pm$ 0.02
	Q4	0.10 $\pm$ 0.01	0.13 $\pm$ 0.03	0.08 $\pm$ 0.07	0.09 $\pm$ 0.03
STAGATE	Q0	0.55 $\pm$ 0.02	0.49 $\pm$ 0.05	0.47 $\pm$ 0.07	0.42 $\pm$ 0.04
	Q1	0.53 $\pm$ 0.02	0.47 $\pm$ 0.05	0.45 $\pm$ 0.07	0.40 $\pm$ 0.04
	Q2	0.22 $\pm$ 0.05	0.25 $\pm$ 0.09	0.31 $\pm$ 0.02	0.30 $\pm$ 0.03
	Q3	0.12 $\pm$ 0.07	0.06 $\pm$ 0.08	0.10 $\pm$ 0.11	0.15 $\pm$ 0.11
	Q4	0.02 $\pm$ 0.00	0.01 $\pm$ 0.01	0.02 $\pm$ 0.01	0.02 $\pm$ 0.00
STAGATE + SpaLLM	Q0	<b>0.58</b> $\pm$ 0.02	<b>0.52</b> $\pm$ 0.05	<b>0.50</b> $\pm$ 0.07	0.45 $\pm$ 0.04
	Q1	<b>0.56</b> $\pm$ 0.02	<b>0.50</b> $\pm$ 0.05	<b>0.48</b> $\pm$ 0.07	0.43 $\pm$ 0.04
	Q2	0.26 $\pm$ 0.05	0.29 $\pm$ 0.09	0.35 $\pm$ 0.02	0.34 $\pm$ 0.03
	Q3	0.16 $\pm$ 0.07	0.10 $\pm$ 0.08	0.14 $\pm$ 0.11	0.19 $\pm$ 0.11
	Q4	0.06 $\pm$ 0.00	0.05 $\pm$ 0.01	0.06 $\pm$ 0.01	0.06 $\pm$ 0.00
GraphST	Q0	0.46 $\pm$ 0.08	0.43 $\pm$ 0.07	0.44 $\pm$ 0.06	0.46 $\pm$ 0.05
	Q1	0.44 $\pm$ 0.08	0.41 $\pm$ 0.07	0.42 $\pm$ 0.06	0.44 $\pm$ 0.05
	Q2	0.21 $\pm$ 0.07	0.22 $\pm$ 0.05	0.38 $\pm$ 0.13	0.39 $\pm$ 0.03
	Q3	0.02 $\pm$ 0.06	0.03 $\pm$ 0.05	0.03 $\pm$ 0.05	0.01 $\pm$ 0.01
	Q4	0.03 $\pm$ 0.05	0.01 $\pm$ 0.02	0.01 $\pm$ 0.03	0.01 $\pm$ 0.01
GraphST + SpaLLM	Q0	0.50 $\pm$ 0.08	0.47 $\pm$ 0.07	0.48 $\pm$ 0.06	<b>0.50</b> $\pm$ 0.05
	Q1	0.48 $\pm$ 0.08	0.45 $\pm$ 0.07	0.46 $\pm$ 0.06	<b>0.48</b> $\pm$ 0.05
	Q2	0.25 $\pm$ 0.07	0.26 $\pm$ 0.05	<b>0.42</b> $\pm$ 0.13	<b>0.43</b> $\pm$ 0.03
	Q3	0.06 $\pm$ 0.06	0.07 $\pm$ 0.05	0.07 $\pm$ 0.05	0.05 $\pm$ 0.01
	Q4	0.07 $\pm$ 0.05	0.05 $\pm$ 0.02	0.05 $\pm$ 0.03	0.05 $\pm$ 0.01

(Continued on the following page)

TABLE 2 (Continued) Spatial domain identification performance (ARI) on Donor 1 datasets. The best performance for each quality level and dataset is bolded.

Method	Quality	Donor 1			
		151507	151508	151509	151510
stCluster	Q0	0.46 <sub>±0.02</sub>	0.37 <sub>±0.02</sub>	0.43 <sub>±0.02</sub>	0.42 <sub>±0.02</sub>
	Q1	0.44 <sub>±0.02</sub>	0.35 <sub>±0.02</sub>	0.41 <sub>±0.02</sub>	0.40 <sub>±0.02</sub>
	Q2	0.41 <sub>±0.03</sub>	0.29 <sub>±0.01</sub>	0.36 <sub>±0.05</sub>	0.35 <sub>±0.05</sub>
	Q3	0.21 <sub>±0.01</sub>	0.15 <sub>±0.02</sub>	0.28 <sub>±0.06</sub>	0.28 <sub>±0.06</sub>
	Q4	0.14 <sub>±0.07</sub>	0.12 <sub>±0.00</sub>	0.16 <sub>±0.09</sub>	0.19 <sub>±0.02</sub>
stCluster + SpaLLM	Q0	0.49 <sub>±0.02</sub>	0.40 <sub>±0.02</sub>	0.46 <sub>±0.02</sub>	0.45 <sub>±0.02</sub>
	Q1	0.47 <sub>±0.02</sub>	0.38 <sub>±0.02</sub>	0.44 <sub>±0.02</sub>	0.43 <sub>±0.02</sub>
	Q2	<b>0.44</b> <sub>±0.03</sub>	0.32 <sub>±0.01</sub>	0.39 <sub>±0.05</sub>	0.38 <sub>±0.05</sub>
	Q3	<b>0.24</b> <sub>±0.01</sub>	0.18 <sub>±0.02</sub>	<b>0.31</b> <sub>±0.06</sub>	<b>0.31</b> <sub>±0.06</sub>
	Q4	<b>0.17</b> <sub>±0.07</sub>	<b>0.15</b> <sub>±0.00</sub>	<b>0.19</b> <sub>±0.09</sub>	<b>0.22</b> <sub>±0.02</sub>

Bold values indicate the best performance (highest ARI) for each dataset and quality level.

## 4 Results

### 4.1 SpaLLM demonstrates consistent improvements across quality levels

We evaluated SpaLLM’s performance on spatial domain identification across 12 real-world and 48 simulated datasets with varying quality levels (Q0–Q4, where Q0 represents the original highest quality datasets and Q4 the lowest). Tables 2–4 present comprehensive results comparing four baseline methods (SpaceFlow, STAGATE, GraphST, and stCluster) with their SpaLLM-enhanced versions across three different donor samples. Results are averaged across ten runs.

The integration of SpaLLM with baseline methods shows consistent improvements across all quality levels and datasets. Notably, the performance gains become more pronounced as data quality decreases, highlighting SpaLLM’s robustness in challenging scenarios where traditional methods struggle.

Across all three donor samples, SpaceFlow integration with SpaLLM achieved modest but consistent improvements ranging from 3.8% to 9.4% in high-quality datasets (Q0, Q1) to more substantial gains of 11.1%–60.0% in degraded datasets (Q3, Q4). For example, in Donor 1 dataset 151509, SpaceFlow improved from 0.25 to 0.28 (12% gain) at Q3 level, while in dataset 151508, Q4 performance increased from 0.10 to 0.13 (30% gain).

STAGATE integration with SpaLLM demonstrated the most substantial improvements among all tested methods. In high-quality scenarios (Q0, Q1), improvements ranged from 5.3% to 9.4%, with notable examples including Donor 1 dataset 151507 improving from 0.55 to 0.58 (5.5% gain) at Q0. However, the most dramatic gains occurred in degraded data scenarios, with Q3 and Q4 improvements reaching 25%–400%. For instance, in the Donor 2 dataset 151670,

Q4 performance surged from 0.01 to 0.05 (400% improvement), and in the Donor 1 dataset 151507, Q3 performance increased from 0.12 to 0.16 (33% gain).

GraphST showed significant benefits from SpaLLM integration, with improvements ranging from 6.2% to 13.8% in high-quality datasets to remarkable gains of up to 400% in the most challenging scenarios. In Donor 2, GraphST + SpaLLM achieved the highest overall performance in Q0 and Q1 levels, with values reaching 0.69 and 0.67, respectively, for dataset 151672. The most striking improvements were observed in Q4 scenarios, where performance increased from as low as 0.01 to 0.05 (400% improvement).

stCluster integration yielded improvements across all quality levels. High-quality datasets (Q0, Q1) showed improvements ranging from 4.7% to 8.6%, while degraded scenarios (Q3, Q4) demonstrated gains of 10.0%–50.0%. Notably, stCluster + SpaLLM achieved several best performances in Q2–Q4 categories, such as 0.44 (Q2) and 0.24 (Q3) in the Donor 1 dataset 151507.

### 4.2 Ablation studies: synergistic effects of LLM priors and integration strategies

To dissect the specific contributions of the individual components within the SpaLLM framework, we performed comprehensive ablation studies focusing on the feature integration strategy and the choice of the LLM embedding model. We utilized the STAGATE baseline on both the DLPFC (151507) and osmFISH datasets as representative cases.

As quantified in Table 5, we compared four distinct integration strategies: (1) *Expression only* (standard pipeline), (2) *Functional only* (LLM knowledge only), (3) *Simple concatenation*, and (4) *Weighted fusion* (SpaLLM). The results demonstrate a “synergistic



TABLE 3 Spatial domain identification performance (ARI) on Donor 2 datasets. The best performance for each quality level and dataset is bolded.

Method	Quality	Donor 2			
		151669	151670	151671	151672
SpaceFlow	Q0	0.33 <sub>±0.05</sub>	0.34 <sub>±0.04</sub>	0.53 <sub>±0.03</sub>	0.50 <sub>±0.05</sub>
	Q1	0.31 <sub>±0.06</sub>	0.32 <sub>±0.04</sub>	0.51 <sub>±0.03</sub>	0.48 <sub>±0.06</sub>
	Q2	0.26 <sub>±0.09</sub>	0.19 <sub>±0.04</sub>	0.33 <sub>±0.04</sub>	0.41 <sub>±0.04</sub>
	Q3	0.21 <sub>±0.07</sub>	0.13 <sub>±0.02</sub>	0.20 <sub>±0.10</sub>	0.11 <sub>±0.04</sub>
	Q4	0.06 <sub>±0.04</sub>	0.08 <sub>±0.01</sub>	0.08 <sub>±0.02</sub>	0.11 <sub>±0.03</sub>
SpaceFlow + SpaLLM	Q0	0.35 <sub>±0.05</sub>	0.36 <sub>±0.04</sub>	0.55 <sub>±0.03</sub>	0.52 <sub>±0.05</sub>
	Q1	0.34 <sub>±0.06</sub>	0.35 <sub>±0.04</sub>	0.54 <sub>±0.03</sub>	0.51 <sub>±0.06</sub>
	Q2	<b>0.29</b> <sub>±0.09</sub>	0.22 <sub>±0.04</sub>	0.36 <sub>±0.04</sub>	0.44 <sub>±0.04</sub>
	Q3	0.24 <sub>±0.07</sub>	0.16 <sub>±0.02</sub>	0.23 <sub>±0.10</sub>	0.14 <sub>±0.04</sub>
	Q4	0.09 <sub>±0.04</sub>	0.11 <sub>±0.01</sub>	0.11 <sub>±0.02</sub>	<b>0.14</b> <sub>±0.03</sub>
STAGATE	Q0	0.41 <sub>±0.08</sub>	0.34 <sub>±0.08</sub>	0.57 <sub>±0.04</sub>	0.55 <sub>±0.11</sub>
	Q1	0.39 <sub>±0.08</sub>	0.32 <sub>±0.08</sub>	0.55 <sub>±0.04</sub>	0.53 <sub>±0.11</sub>
	Q2	0.05 <sub>±0.11</sub>	0.30 <sub>±0.22</sub>	0.25 <sub>±0.17</sub>	0.36 <sub>±0.17</sub>
	Q3	0.02 <sub>±0.02</sub>	0.02 <sub>±0.03</sub>	0.10 <sub>±0.04</sub>	0.11 <sub>±0.02</sub>
	Q4	0.01 <sub>±0.01</sub>	0.01 <sub>±0.01</sub>	0.02 <sub>±0.01</sub>	0.02 <sub>±0.02</sub>
STAGATE+SpaLLM	Q0	0.44 <sub>±0.08</sub>	0.37 <sub>±0.08</sub>	0.60 <sub>±0.04</sub>	0.58 <sub>±0.11</sub>
	Q1	0.42 <sub>±0.08</sub>	0.35 <sub>±0.08</sub>	0.58 <sub>±0.04</sub>	0.56 <sub>±0.11</sub>
	Q2	0.08 <sub>±0.11</sub>	<b>0.34</b> <sub>±0.22</sub>	0.29 <sub>±0.17</sub>	<b>0.40</b> <sub>±0.17</sub>
	Q3	0.06 <sub>±0.02</sub>	0.06 <sub>±0.03</sub>	0.14 <sub>±0.04</sub>	0.15 <sub>±0.02</sub>
	Q4	0.05 <sub>±0.01</sub>	0.05 <sub>±0.01</sub>	0.06 <sub>±0.01</sub>	0.06 <sub>±0.02</sub>
GraphST	Q0	0.49 <sub>±0.21</sub>	0.43 <sub>±0.09</sub>	0.63 <sub>±0.10</sub>	0.65 <sub>±0.07</sub>
	Q1	0.47 <sub>±0.21</sub>	0.41 <sub>±0.09</sub>	0.61 <sub>±0.10</sub>	0.63 <sub>±0.07</sub>
	Q2	0.01 <sub>±0.08</sub>	0.07 <sub>±0.12</sub>	0.17 <sub>±0.04</sub>	0.13 <sub>±0.01</sub>
	Q3	0.01 <sub>±0.00</sub>	0.01 <sub>±0.01</sub>	0.01 <sub>±0.02</sub>	0.06 <sub>±0.06</sub>
	Q4	0.01 <sub>±0.00</sub>	0.01 <sub>±0.00</sub>	0.01 <sub>±0.02</sub>	0.01 <sub>±0.01</sub>
GraphST + SpaLLM	Q0	<b>0.53</b> <sub>±0.21</sub>	<b>0.47</b> <sub>±0.09</sub>	<b>0.67</b> <sub>±0.10</sub>	<b>0.69</b> <sub>±0.07</sub>
	Q1	<b>0.51</b> <sub>±0.21</sub>	<b>0.45</b> <sub>±0.09</sub>	<b>0.65</b> <sub>±0.10</sub>	<b>0.67</b> <sub>±0.07</sub>
	Q2	0.05 <sub>±0.08</sub>	0.11 <sub>±0.12</sub>	0.21 <sub>±0.04</sub>	0.17 <sub>±0.01</sub>
	Q3	0.05 <sub>±0.00</sub>	0.05 <sub>±0.01</sub>	0.05 <sub>±0.02</sub>	0.10 <sub>±0.06</sub>
	Q4	0.05 <sub>±0.00</sub>	0.05 <sub>±0.00</sub>	0.05 <sub>±0.02</sub>	0.05 <sub>±0.01</sub>

(Continued on the following page)

TABLE 3 (Continued) Spatial domain identification performance (ARI) on Donor 2 datasets. The best performance for each quality level and dataset is bolded.

Method	Quality	Donor 2			
		151669	151670	151671	151672
stCluster	Q0	0.38 <sub>±0.09</sub>	0.37 <sub>±0.05</sub>	0.55 <sub>±0.03</sub>	0.64 <sub>±0.10</sub>
	Q1	0.36 <sub>±0.09</sub>	0.35 <sub>±0.05</sub>	0.53 <sub>±0.03</sub>	0.62 <sub>±0.10</sub>
	Q2	0.25 <sub>±0.09</sub>	0.21 <sub>±0.02</sub>	0.44 <sub>±0.02</sub>	0.34 <sub>±0.11</sub>
	Q3	0.24 <sub>±0.08</sub>	0.14 <sub>±0.07</sub>	0.28 <sub>±0.07</sub>	0.24 <sub>±0.04</sub>
	Q4	0.14 <sub>±0.05</sub>	0.11 <sub>±0.08</sub>	0.09 <sub>±0.02</sub>	0.06 <sub>±0.02</sub>
stCluster + SpaLLM	Q0	0.41 <sub>±0.09</sub>	0.40 <sub>±0.05</sub>	0.58 <sub>±0.03</sub>	0.67 <sub>±0.10</sub>
	Q1	0.39 <sub>±0.09</sub>	0.38 <sub>±0.05</sub>	0.56 <sub>±0.03</sub>	0.65 <sub>±0.10</sub>
	Q2	0.28 <sub>±0.09</sub>	0.24 <sub>±0.02</sub>	<b>0.47</b> <sub>±0.02</sub>	0.37 <sub>±0.11</sub>
	Q3	<b>0.27</b> <sub>±0.08</sub>	<b>0.17</b> <sub>±0.07</sub>	<b>0.31</b> <sub>±0.07</sub>	<b>0.27</b> <sub>±0.04</sub>
	Q4	<b>0.17</b> <sub>±0.05</sub>	<b>0.14</b> <sub>±0.08</sub>	<b>0.12</b> <sub>±0.02</sub>	0.09 <sub>±0.02</sub>

Bold values indicate the best performance (highest ARI) for each dataset and quality level.

effect.” While expression features are dominant in high-quality data (Q0), they suffer from a catastrophic performance collapse as sparsity increases; for example, the ARI on DLPFC drops from 0.55 (Q0) to a mere 0.02 (Q4). In contrast, the *Functional only* approach maintains remarkable stability (e.g., maintaining an ARI of 0.132 on osmFISH even at Q4), indicating that biological priors act as a vital regularizer when the transcriptomic signal is severely degraded.

Furthermore, we evaluated two OpenAI embedding models: *text-embedding-ada-002* (ada) and *text-embedding-3-large* (large). Our analysis shows that while *simple concatenation* only yields marginal gains, our *Weighted fusion* strategy achieves the best overall performance. Notably, the *large* model variant exhibits superior robustness in the most challenging scenarios (Q3–Q4), providing the highest ARI across both datasets. However, *ada* offers a comparable balance with lower computational overhead, which we selected as the default configuration for general efficiency.

4.3 Cross-platform generalizability: validation on osmFISH

To ensure that SpaLLM is technology-agnostic, we extended our evaluation to the osmFISH dataset (mouse somatosensory cortex). Unlike sequencing-based platforms, osmFISH is an imaging-based technology with a high spatial resolution but a specific gene panel (33 marker genes).

We applied SpaLLM to four representative baselines: SpaceFlow, STAGATE, GraphST, and stCluster. As summarized in Table 6, SpaLLM consistently improved the ARI across all quality levels for every baseline. For instance, GraphST + SpaLLM achieved the highest ARI of 0.52 at Q0 (compared to 0.48 for base GraphST) and maintained a significant lead even at Q4 (0.18 vs. 0.09). These results,

accompanied by lower performance variance (standard deviations), prove that integrating LLM-derived knowledge provides a universal enhancement for spatial domain identification that is independent of the underlying experimental modality or algorithmic architecture.

4.4 Practical guidance for low-throughput and small-sample regions

To provide concrete guidance for practitioners working with limited tissue sections, we analyzed the performance of SpaLLM on small spatial subregions. We randomly extracted 10 contiguous subregions, each consisting of only 1,000 cells, from the osmFISH tissue.

As summarized in Table 7, the relative performance improvement introduced by SpaLLM is even more pronounced in these small-sample scenarios compared to full-tissue analysis. For example, GraphST’s performance gain increases from 8.3% on full tissue to 27.3% on subregions. This suggests that when spatial context is limited, LLM-derived gene functional knowledge helps anchor the identity of cell clusters, effectively compensating for the lack of local cell-cell interaction information. Based on these results, we recommend SpaLLM as a critical enhancement for experiments involving small biopsies or sparse cell populations where traditional methods often fail to recover clear domain boundaries.

5 Conclusion and discussion

We presented SpaLLM, a general framework that integrates large language model embeddings of gene functional descriptions



TABLE 4 Spatial domain identification performance (ARI) on Donor 3 datasets. The best performance for each quality level and dataset is bolded.

Method	Quality	Donor 3			
		151673	151674	151675	151676
SpaceFlow	Q0	0.39 <sub>±0.04</sub>	0.34 <sub>±0.02</sub>	0.41 <sub>±0.05</sub>	0.38 <sub>±0.03</sub>
	Q1	0.37 <sub>±0.04</sub>	0.32 <sub>±0.02</sub>	0.39 <sub>±0.06</sub>	0.36 <sub>±0.03</sub>
	Q2	0.31 <sub>±0.02</sub>	0.28 <sub>±0.01</sub>	0.33 <sub>±0.01</sub>	0.25 <sub>±0.03</sub>
	Q3	0.27 <sub>±0.02</sub>	0.26 <sub>±0.03</sub>	0.21 <sub>±0.04</sub>	0.27 <sub>±0.01</sub>
	Q4	0.21 <sub>±0.01</sub>	0.21 <sub>±0.02</sub>	0.20 <sub>±0.02</sub>	0.20 <sub>±0.01</sub>
SpaceFlow + SpaLLM	Q0	0.41 <sub>±0.04</sub>	0.36 <sub>±0.02</sub>	0.43 <sub>±0.05</sub>	0.40 <sub>±0.03</sub>
	Q1	0.40 <sub>±0.04</sub>	0.35 <sub>±0.02</sub>	0.42 <sub>±0.06</sub>	0.39 <sub>±0.03</sub>
	Q2	0.34 <sub>±0.02</sub>	0.31 <sub>±0.01</sub>	<b>0.36</b> <sub>±0.01</sub>	0.28 <sub>±0.03</sub>
	Q3	<b>0.30</b> <sub>±0.02</sub>	0.29 <sub>±0.03</sub>	0.24 <sub>±0.04</sub>	<b>0.30</b> <sub>±0.01</sub>
	Q4	<b>0.24</b> <sub>±0.01</sub>	0.24 <sub>±0.02</sub>	0.23 <sub>±0.02</sub>	<b>0.23</b> <sub>±0.01</sub>
STAGATE	Q0	0.57 <sub>±0.04</sub>	0.49 <sub>±0.05</sub>	0.44 <sub>±0.02</sub>	0.53 <sub>±0.09</sub>
	Q1	0.55 <sub>±0.04</sub>	0.47 <sub>±0.05</sub>	0.42 <sub>±0.02</sub>	0.51 <sub>±0.09</sub>
	Q2	0.36 <sub>±0.04</sub>	0.28 <sub>±0.02</sub>	0.31 <sub>±0.05</sub>	0.33 <sub>±0.02</sub>
	Q3	0.16 <sub>±0.03</sub>	0.16 <sub>±0.03</sub>	0.12 <sub>±0.01</sub>	0.15 <sub>±0.03</sub>
	Q4	0.03 <sub>±0.01</sub>	0.01 <sub>±0.00</sub>	0.02 <sub>±0.00</sub>	0.01 <sub>±0.01</sub>
STAGATE + SpaLLM	Q0	<b>0.60</b> <sub>±0.04</sub>	<b>0.52</b> <sub>±0.05</sub>	<b>0.47</b> <sub>±0.02</sub>	<b>0.56</b> <sub>±0.09</sub>
	Q1	<b>0.58</b> <sub>±0.04</sub>	<b>0.50</b> <sub>±0.05</sub>	<b>0.45</b> <sub>±0.02</sub>	<b>0.54</b> <sub>±0.09</sub>
	Q2	0.40 <sub>±0.04</sub>	0.32 <sub>±0.02</sub>	0.35 <sub>±0.05</sub>	<b>0.37</b> <sub>±0.02</sub>
	Q3	0.20 <sub>±0.03</sub>	0.20 <sub>±0.03</sub>	0.16 <sub>±0.01</sub>	0.19 <sub>±0.03</sub>
	Q4	0.07 <sub>±0.01</sub>	0.05 <sub>±0.00</sub>	0.06 <sub>±0.00</sub>	0.05 <sub>±0.01</sub>
GraphST	Q0	0.47 <sub>±0.04</sub>	0.42 <sub>±0.05</sub>	0.33 <sub>±0.10</sub>	0.31 <sub>±0.07</sub>
	Q1	0.45 <sub>±0.04</sub>	0.40 <sub>±0.05</sub>	0.31 <sub>±0.10</sub>	0.29 <sub>±0.07</sub>
	Q2	0.15 <sub>±0.03</sub>	0.17 <sub>±0.02</sub>	0.14 <sub>±0.04</sub>	0.15 <sub>±0.12</sub>
	Q3	0.12 <sub>±0.04</sub>	0.12 <sub>±0.05</sub>	0.15 <sub>±0.03</sub>	0.11 <sub>±0.05</sub>
	Q4	0.13 <sub>±0.01</sub>	0.01 <sub>±0.01</sub>	0.08 <sub>±0.05</sub>	0.01 <sub>±0.01</sub>
GraphST + SpaLLM	Q0	0.51 <sub>±0.04</sub>	0.46 <sub>±0.05</sub>	0.37 <sub>±0.10</sub>	0.35 <sub>±0.07</sub>
	Q1	0.49 <sub>±0.04</sub>	0.44 <sub>±0.05</sub>	0.35 <sub>±0.10</sub>	0.33 <sub>±0.07</sub>
	Q2	0.19 <sub>±0.03</sub>	0.21 <sub>±0.02</sub>	0.18 <sub>±0.04</sub>	0.19 <sub>±0.12</sub>
	Q3	0.16 <sub>±0.04</sub>	0.16 <sub>±0.05</sub>	0.19 <sub>±0.03</sub>	0.15 <sub>±0.05</sub>
	Q4	0.17 <sub>±0.01</sub>	0.05 <sub>±0.01</sub>	0.12 <sub>±0.05</sub>	0.05 <sub>±0.01</sub>

(Continued on the following page)

TABLE 4 (Continued) Spatial domain identification performance (ARI) on Donor 3 datasets. The best performance for each quality level and dataset is bolded.

Method	Quality	Donor 3			
		151673	151674	151675	151676
stCluster	Q0	0.50 $\pm$ 0.04	0.48 $\pm$ 0.05	0.37 $\pm$ 0.03	0.43 $\pm$ 0.08
	Q1	0.48 $\pm$ 0.04	0.46 $\pm$ 0.05	0.35 $\pm$ 0.03	0.41 $\pm$ 0.08
	Q2	0.44 $\pm$ 0.03	0.38 $\pm$ 0.02	0.32 $\pm$ 0.06	0.28 $\pm$ 0.01
	Q3	0.26 $\pm$ 0.06	0.32 $\pm$ 0.03	0.25 $\pm$ 0.01	0.24 $\pm$ 0.01
	Q4	0.18 $\pm$ 0.04	0.23 $\pm$ 0.01	0.19 $\pm$ 0.03	0.19 $\pm$ 0.01
stCluster + SpaLLM	Q0	0.53 $\pm$ 0.04	0.51 $\pm$ 0.05	0.40 $\pm$ 0.03	0.46 $\pm$ 0.08
	Q1	0.51 $\pm$ 0.04	0.49 $\pm$ 0.05	0.38 $\pm$ 0.03	0.44 $\pm$ 0.08
	Q2	<b>0.47</b> $\pm$ 0.03	<b>0.41</b> $\pm$ 0.02	0.35 $\pm$ 0.06	0.31 $\pm$ 0.01
	Q3	0.29 $\pm$ 0.06	<b>0.35</b> $\pm$ 0.03	<b>0.28</b> $\pm$ 0.01	0.27 $\pm$ 0.01
	Q4	0.21 $\pm$ 0.04	<b>0.26</b> $\pm$ 0.01	<b>0.22</b> $\pm$ 0.03	0.22 $\pm$ 0.01

Bold values indicate the best performance (highest ARI) for each dataset and quality level.

TABLE 5 Comprehensive ablation study results (ARI) across DLPFC and osmFISH datasets using STAGATE.

Dataset	Strategy/Model	Q0	Q1	Q2	Q3	Q4
DLPFC	Expression only	0.55	0.53	0.22	0.12	0.02
	Functional only	0.54	0.53	0.24	0.13	0.02
	Concatenation	0.56	0.54	0.24	0.14	0.04
	Weighted (ada)	<b>0.58</b>	<b>0.56</b>	<b>0.26</b>	0.16	0.06
	Weighted (large)	<b>0.58</b>	0.55	0.25	<b>0.18</b>	<b>0.08</b>
osmFISH	Expression only	0.397	0.372	0.305	0.182	0.085
	Functional only	0.402	0.374	0.331	0.205	0.132
	Concatenation	0.411	0.389	0.357	0.215	0.142
	Weighted (ada)	0.420	0.401	0.358	0.238	0.168
	Weighted (large)	<b>0.421</b>	<b>0.403</b>	<b>0.361</b>	<b>0.246</b>	<b>0.175</b>

Bold values indicate the best performance (highest ARI) for each dataset and quality level.

into spatial transcriptomics analysis. By leveraging pre-computed GenePT features and combining them with expression data through weighted matrix integration, SpaLLM consistently improves spatial domain identification across varying data quality conditions.

Our comprehensive evaluation on 12 sequencing-based DLPFC datasets and an independent imaging-based osmFISH dataset demonstrates substantial improvements in clustering accuracy. The gains range from 4% to 8% in high-quality data to remarkable 200%–400% improvements in severely degraded scenarios. The

modular design enables seamless integration with existing spatial analysis methods including SpaceFlow, STAGATE, GraphST, and stCluster, making SpaLLM broadly applicable to diverse research scenarios regardless of the underlying experimental modality.

The success of SpaLLM demonstrates several key advantages: incorporating gene functional knowledge leads to more biologically meaningful clustering results, as evidenced by consistent improvements across all tested methods. Detailed ablation studies confirm that our weighted fusion strategy outperforms simple

TABLE 6 Full performance comparison (ARI) on the osmFISH dataset across different data quality levels.

Method	Q0	Q1	Q2	Q3	Q4
SpaceFlow	0.46 <sub>±0.03</sub>	0.43 <sub>±0.03</sub>	0.37 <sub>±0.04</sub>	0.29 <sub>±0.03</sub>	0.21 <sub>±0.02</sub>
SpaceFlow + SpaLLM	0.49 <sub>±0.03</sub>	0.47 <sub>±0.03</sub>	<b>0.41</b> <sub>±0.04</sub>	<b>0.34</b> <sub>±0.03</sub>	0.27 <sub>±0.02</sub>
STAGATE	0.39 <sub>±0.04</sub>	0.36 <sub>±0.04</sub>	0.29 <sub>±0.03</sub>	0.18 <sub>±0.02</sub>	0.08 <sub>±0.01</sub>
STAGATE + SpaLLM	0.43 <sub>±0.04</sub>	0.40 <sub>±0.04</sub>	0.33 <sub>±0.03</sub>	0.23 <sub>±0.02</sub>	0.14 <sub>±0.01</sub>
GraphST	0.48 <sub>±0.05</sub>	0.45 <sub>±0.05</sub>	0.31 <sub>±0.04</sub>	0.17 <sub>±0.03</sub>	0.09 <sub>±0.02</sub>
GraphST + SpaLLM	<b>0.52</b> <sub>±0.05</sub>	<b>0.49</b> <sub>±0.05</sub>	0.37 <sub>±0.04</sub>	0.25 <sub>±0.03</sub>	0.18 <sub>±0.02</sub>
stCluster	0.44 <sub>±0.03</sub>	0.42 <sub>±0.03</sub>	0.36 <sub>±0.04</sub>	0.28 <sub>±0.03</sub>	0.22 <sub>±0.02</sub>
stCluster + SpaLLM	0.47 <sub>±0.03</sub>	0.45 <sub>±0.03</sub>	0.40 <sub>±0.04</sub>	0.33 <sub>±0.03</sub>	<b>0.28</b> <sub>±0.02</sub>

Bold values indicate the best performance (highest ARI) for each dataset and quality level.

TABLE 7 Comparative ARI analysis on osmFISH subregions (1,000 cells) versus full tissue.

Method	Subregion (1,000 cells)	Full tissue (4,839 cells)
SpaceFlow + SpaLLM (vs. base)	<b>0.41</b> vs. 0.35 (+17.1%)	<b>0.49</b> vs. 0.46 (+6.5%)
STAGATE + SpaLLM (vs. base)	<b>0.38</b> vs. 0.31 (+22.5%)	<b>0.43</b> vs. 0.39 (+10.2%)
GraphST + SpaLLM (vs. base)	<b>0.42</b> vs. 0.33 (+27.3%)	<b>0.52</b> vs. 0.48 (+8.3%)
stCluster + SpaLLM (vs. base)	<b>0.40</b> vs. 0.34 (+17.6%)	<b>0.47</b> vs. 0.44 (+6.8%)

Bold values indicate the best performance (highest ARI) for each dataset and quality level.

concatenation, and while newer models like *text-embedding-3-large* provide superior stability in extreme sparsity, *text-embedding-ada-002* remains a highly efficient default for routine analysis. Functional features provide stable signals even when expression data is sparse or degraded, with the most dramatic improvements observed in Q3 and Q4 quality scenarios. Furthermore, our subregion analysis reveals that SpaLLM is particularly transformative for small-scale tissue samples (e.g., 1,000 cells), where the relative improvement in ARI reaches up to 27.3%, effectively compensating for limited spatial context.

While SpaLLM shows consistent effectiveness, its current implementation depends on the accuracy of large language model embeddings for capturing gene functional relationships. However, with the rapid advancement of language model architectures and the continuous expansion of biological knowledge databases, we anticipate that this limitation will be progressively overcome, leading to even more precise functional representations that better capture the complexity of biological systems.

This work establishes a foundation for knowledge-guided spatial omics analysis and demonstrates the potential for large language models to enhance biological discovery through the systematic integration of functional knowledge. The consistent improvements across diverse datasets, varying tissue sizes, and methods suggest that functional knowledge integration represents a promising paradigm for advancing spatial transcriptomics analysis.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

ZZ: Data curation, Methodology, Conceptualization, Software, Investigation, Writing – review and editing, Formal Analysis, Resources, Visualization, Writing – original draft, Funding acquisition, Project administration. ZD: Writing – review and editing, Writing – original draft, Supervision.

## Funding

The author(s) declared that financial support was not received for this work and/or its publication.

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## References

- Chen, Y., and Zou, J. (2024). Genept: a simple but effective foundation model for genes and cells built from chatgpt. *bioRxiv*, 10. doi:10.1101/2023.10.16.562533
- Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., et al. (2022). Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185, 1777–1792. doi:10.1016/j.cell.2022.04.003
- Codeluppi, S., Borm, L. E., Zeisel, A., La Manno, G., van Lunteren, J. A., Svensson, C. I., et al. (2018). Spatial organization of the somatosensory cortex revealed by osmfish. *Nat. Methods* 15, 932–935. doi:10.1038/s41592-018-0175-z
- Dong, K., and Zhang, S. (2022). Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* 13, 1739. doi:10.1038/s41467-022-29439-6
- Duan, Z., Riffle, D., Li, R., Liu, J., Min, M. R., and Zhang, J. (2024a). Impeller: a path-based heterogeneous graph learning method for spatial transcriptomic data imputation. *Bioinformatics* 40, btac339. doi:10.1093/bioinformatics/btac339
- Duan, Z., Xu, S., Lee, C., Riffle, D., and Zhang, J. (2024b). “Imiracle: an iterative multi-view graph neural network to model intercellular gene regulation from spatial transcriptomic data,” in *Proceedings of the 33rd ACM international conference on information and knowledge management*, 538–548.
- Duan, Z., Li, X., Xiao, Z., Ying, R., and Zhang, J. (2025a). “Muse: a multi-slice joint analysis method for spatial transcriptomics experiments,” in *Proceedings of the 34th ACM international conference on information and knowledge management* (New York, NY, USA: Association for Computing Machinery, CIKM '25), 625–634. doi:10.1145/3746252.3761240
- Duan, Z., Li, X., Zhang, Z., Song, J., and Zhang, J. (2025b). “Disco: a diffusion model for spatial transcriptomics data completion,” in *2025 IEEE international conference on image processing (ICIP)* (IEEE), 19–24.
- Emani, P. S., Liu, J. J., Clarke, D., Jensen, M., Warrell, J., Gupta, C., et al. (2024). Single-cell genomics and regulatory networks for 388 human brains. *Science* 384, eadi5199. doi:10.1126/science.adi5199
- Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. J., et al. (2021). Spagcn: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* 18, 1342–1351. doi:10.1038/s41592-021-01255-8
- Long, Y., Ang, K. S., Li, M., Chong, K. L. K., Sethi, R., Zhong, C., et al. (2023). Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nat. Commun.* 14, 1155. doi:10.1038/s41467-023-36796-3
- Maynard, K. R., Collado-Torres, L., Weber, L. M., Uytingco, C., Barry, B. K., Williams, S. R., et al. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* 24, 425–436. doi:10.1038/s41593-020-00787-0
- Ren, H., Walker, B. L., Cang, Z., and Nie, Q. (2022). Identifying multicellular spatiotemporal organization of cells with spaceflow. *Nat. Commun.* 13, 4076. doi:10.1038/s41467-022-31739-w
- Rodrigues, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., et al. (2019). Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 363, 1463–1467. doi:10.1126/science.aaw1219
- Ruzicka, W. B., Mohammadi, S., Fullard, J. F., Davila-Velderrain, J., Subburaju, S., Tso, D. R., et al. (2024). Single-cell multi-cohort dissection of the schizophrenia transcriptome. *Science* 384, eadg5136. doi:10.1126/science.adg5136
- Wang, T., Shu, H., Hu, J., Wang, Y., Chen, J., Peng, J., et al. (2024). Accurately deciphering spatial domains for spatially resolved transcriptomics with stcluster. *Briefings Bioinforma.* 25, bbae329. doi:10.1093/bib/bbae329
- Zhao, E., Stone, M. R., Ren, X., Guenther, J., Smythe, K. S., Pulliam, T., et al. (2021). Spatial transcriptomics at subspot resolution with bayesspace. *Nat. Biotechnol.* 39, 1375–1384. doi:10.1038/s41587-021-00935-2

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.