Multiscale Microenvironment Vector Space Projection for Uncovering Diverse Pathological Biomarker

Zixiu Ding[1,2†] Wenxuan Hao[1†] Liya Ding[1,3] Yuan Dai[1]
Jiazhen Yang[1] Shuwen Han[4] Lingxiang Jia[1] Zunlei Feng[1*]
[1]Zhejiang University
[2]Central South University
[3]Sir Run Run Shaw Hospital
[4]Tarim University
{hwx_vipa,daiyuan2025,jiazhenyang}@zju.edu.cn
{dingliya,lingxiangjia,zunleifeng}@zju.edu.cn
8209230227@csu.edu.cn 120250064@taru.edu.cn

Abstract

Due to the challenge of uncovering biologically meaningful spatial tissue patterns from standard H&E-stained whole-slide images (WSIs) without relying on multiplexed staining or fixed-scale modeling, a unified framework is proposed for spatial biomarker discovery that integrates cell classification, adaptive multi-scale microenvironment construction, and colocalization pattern analysis. Cell-level predictions assign discrete, semantically meaningful types that serve as the basis for microenvironment modeling, where physical adjacency initializes local ecological features and a combination of spatial competition-based seed selection with homogeneity-constrained recursive region growing allows microenvironments to expand from locally pure regions in a data-driven manner, capturing heterogeneity across scales without predefined spatial ranges. Microenvironment-level cell composition features are then used to construct a colocalization feature space, explicitly modeling spatial relationships among cell types, and differential biomarkers are identified via a distributional comparison framework that systematically detects group-specific colocalization patterns. Experiments demonstrate that this framework reliably discovers discriminative and biologically interpretable spatial biomarkers from H&E images alone, exhibiting robustness in tissues with complex structures and heterogeneous scales, and offering a flexible, extensible paradigm for spatial phenotypic biomarker discovery in standard histopathology.

1 Introduction

The exploration of spatial biomarkers plays a crucial role in analyzing the tumor microenvironment and constructing prognostic models. With the advent of digital pathology, computational approaches based on whole-slide images (WSIs) have opened new avenues for large-scale tissue architecture analysis [1]. However, most existing spatial biomarker analysis methods rely heavily on multiplex immunofluorescence or other multimodal data to obtain accurate cell-type annotations and expression profiles, which are often costly and challenging to implement in clinical practice. In contrast, hematoxylin and eosin (H&E) staining, the most routine pathological modality, is widely available, yet its potential for spatial phenotyping and biomarker discovery remains largely underexplored [2].

†Equal contribution.
*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Recent efforts in spatial analysis based on H&E images generally follow a pipeline of cell detection, cell classification, and spatial statistical analysis. Despite these advances, existing methods exhibit significant limitations in modeling the tumor microenvironment and identifying spatial biomarkers. Most mainstream approaches rely on clustering-based statistics, which typically construct microenvironmental clusters using fixed-radius neighborhoods or regular patch divisions, operating under the assumption that local cellular composition is representative at a predefined scale [3]. However, real tissues exhibit substantial heterogeneity: cell density, type distributions, and tissue architecture can vary dramatically across regions. Fixed-scale neighborhood partitioning fails to accommodate this heterogeneity, often mixing or averaging distinct spatial patterns, thereby obscuring fine-grained differences in local cellular composition. Consequently, microenvironmental boundaries become blurred, complex spatial structures are inadequately represented, and the ability of spatial biomarkers to capture the true characteristics of the tumor microenvironment is limited [4].

Some studies have also employed heatmap-based approaches or sharpness-estimation methods, which generate significance distributions or spatial pattern scores over entire slides or large cellular assemblies to identify potentially functional microenvironmental regions. While these approaches can reveal global trends in large-sample settings, they suffer from three major limitations. First, they rely on a substantial number of samples to obtain stable statistical distributions, making significance estimation prone to bias in small or highly heterogeneous tissues [5]. Second, the smoothing inherent in the statistical processing often attenuates fine-grained spatial co-occurrence among cells, preventing accurate capture of local microenvironmental features [6]. Third, these methods frequently require manual verification to confirm significant regions, increasing analytical cost and complexity while reducing reproducibility and automation [7].

To address these limitations, we propose a framework for spatial biomarker discovery from H&E images. It leverages H&E-predicted cell types to construct microenvironmental units via an adaptive, growth-based procedure that learns variable-scale neighborhoods, capturing higher-order intercellular interactions. A cell-level classifier first generates discrete, semantically meaningful cell distributions, which initialize microenvironment representations. High-purity seeds are selected through spatial competitive inhibition and expanded recursively via phenotype-homogeneity-driven region growing, producing data-driven clusters without predefined spatial scales. Hierarchical microenvironment representations are then projected into a spatial feature space to map cellular compositions into explicit distribution patterns. A spatially aware differential biomarker strategy models these distributions and performs distribution-level contrast analysis to identify significant inter-group shifts. By accommodating the non-uniform, adaptive microenvironmental units, this approach enhances the stability and interpretability of discovered spatial-level and tissue-level biomarkers, providing a systematic framework for H&E-based spatial phenotyping.

In summary, the main contributions of this study are as follows:

- We propose a unified framework for spatial biomarker discovery directly from H&E-stained whole slide images (WSIs), removing the reliance on multimodal imaging or molecular profiling and making large-scale, clinically feasible analyses more accessible.
- An adaptive, growth-based microenvironment modeling approach is introduced to dynamically learn variable-scale neighborhoods. Limitations of fixed-radius or patch-based clustering are overcome, and tissue heterogeneity is captured more faithfully across regions with diverse cellular compositions.
- A spatial composition feature space is constructed to capture high-order interactions among cell subsets. Distribution-level differential analysis is employed to systematically identify combinatorial biomarkers, by which statistical robustness and biological interpretability are enhanced. Tissue-level biomarkers are derived from the full cellular composition across microenvironment clusters, through which discriminative regions and hierarchical tissue-scale organization are revealed.
- We validate the framework on multiple public and private datasets covering hepatocellular carcinoma, bladder cancer, and breast cancer. Results show that it yields superior discriminative power in identifying prognostic and functional biomarkers, with improved interpretability over traditional clustering- or patch-based methods.

2 Related work

2.1 H&E-Based Spatial Phenotype Analysis

Computational pathology using H&E whole-slide images has advanced rapidly with deep learning. Early CNN and multiple-instance learning methods extract patch-level features for tumor diagnosis and prognosis but overlook spatial structure and global tissue architecture [8, 9]. Although such methods connect morphology to molecular phenotypes, patch-level representations restrict explicit modeling of cellular spatial organization and interpretability.

Modeling the tumor microenvironment has become a critical direction. Traditional methods quantify heterogeneity using cell density and local composition; Xu et al. [10] analyzed spatial patterns of mutation burden and immune infiltration from WSIs. Graph-based methods and graph neural networks model cells as nodes and spatial interactions as edges [11, 12, 13], and hierarchical graphs further enable multi-scale fusion of cellular and tissue contexts.

Despite progress, current approaches still struggle to model continuous spatial composition, high-order cell interactions, and intrinsic tissue heterogeneity accurately.

2.2 Spatial Biomarker Discovery

Spatial biomarker discovery aims to identify tissue patterns associated with patient prognosis, treatment response, or molecular subtypes, and is a key task in tumor microenvironment analysis. Existing approaches can be broadly categorized into three groups:

Clustering-based methods using multi-modal data. These approaches use mIF, spatial transcriptomics, or proteomics to model tissue architecture through precise cell typing and spatial coordinates. Common clustering includes k-means, hierarchical clustering, spectral clustering, and graph-based methods, often combined with DBSCAN to detect local cell aggregates such as tertiary lymphoid structures [14, 15, 16, 17].Hierarchical strategies further integrate multi-scale spatial units and downstream predictive models [18]. Recent graph-based and manifold learning methods have extended this paradigm: Squidpy [19] integrates spatial graphs with neighborhood enrichment analysis; MELD [20] smooths experimental labels across spatial manifolds to identify condition-associated regions; and diffcyt [21] combines high-resolution clustering with moderated statistical tests for differential abundance analysis. These approaches overcome traditional limitations such as reliance on predefined cluster numbers and difficulty capturing continuous spatial composition, providing adaptive, statistically rigorous frameworks for spatial biomarker discovery.

Fixed-scale clustering on H&E images. H&E images provide a widely available data source. Typical pipelines involve cell detection, classification, and clustering within fixed-radius neighborhoods or regular patches [3]. While computationally efficient, fixed-scale assumptions fail to adapt to heterogeneous cell densities and tissue structures, often blending distinct spatial patterns and blurring microenvironment boundaries [4].

Global pattern analysis via heatmaps or significance scoring. These methods compute slide-level spatial scores to identify global microenvironments. Yuan et al. [5] quantified cellular heterogeneity from pathology images but relies on global summary features. Saltz et al. [6] used deep learning to map lymphocytes globally, at the cost of fine-grained spatial details. Ali et al. [7] predicted therapy response using global lymphocyte density, lacking local structure resolution. Such approaches reveal overall trends but smooth fine-scale features and limit automation.

3 Tumor Microenvironment Biomarker Discovery Task Formulation

The tumor microenvironment (TME) comprises tumor, immune, and stromal cells, whose spatial composition and local interactions influence tumor progression and clinical outcomes. In H&E whole-slide images, cells form functionally specialized microenvironmental units rather than being independently distributed, with substantial regional heterogeneity. Traditional single-cell or patch-level analyses fail to capture fine-grained multi-cellular interactions, limiting comprehensive characterization of spatial heterogeneity and interpretability.

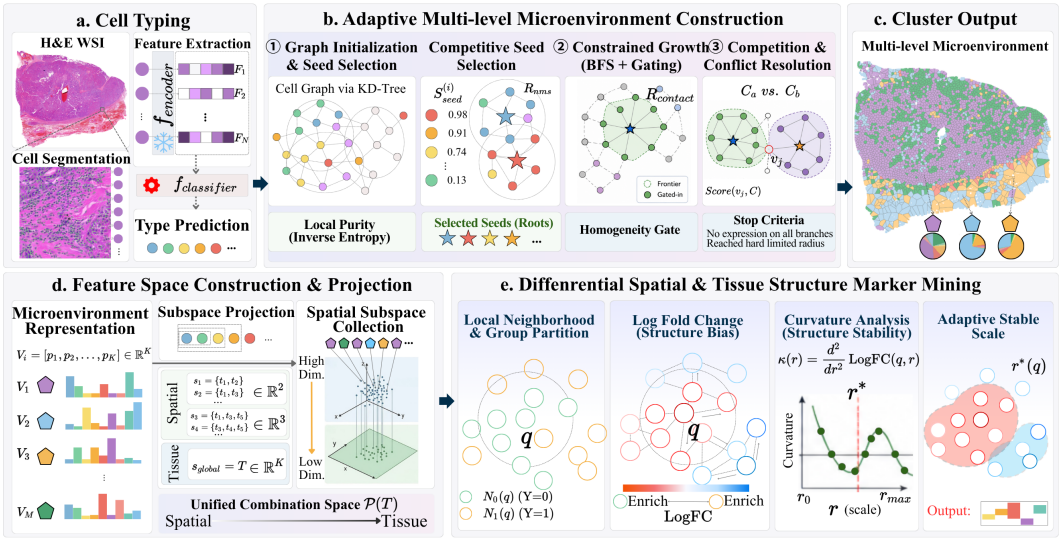

Figure 1: **Overview of the adaptive microenvironment framework.** The pipeline includes: **a. Cell Typing:** Segmenting H&E WSIs and encoding features to predict cell labels. **b. Adaptive Construction:** Building a cell graph via KD-Tree and expanding clusters from NMS-selected seeds using an adaptive BFS-gating mechanism. **c. Cluster Output:** Partitioning tissue into phenotypic units represented by composition vectors. **d. Subspace Projection:** Decomposing representations into multi-scale subspaces to capture cellular synergies. **e. Differential Mining:** Identifying stable biomarkers through cohort-specific LogFC analysis and curvature-based scale optimization.

To address this, we define a multi-scale spatial biomarker discovery task to systematically identify, characterize, and quantify hierarchical spatial patterns in the TME. Biomarkers capture local cellular enrichment or depletion, multi-cellular coordination, and tissue-level composition, including regional boundaries and global organization, spanning from local cellular configurations to tissue structures in a continuous, multi-resolution representation. The task involves adaptive construction of hierarchical microenvironment units without predefined scales, projection of multi-cellular configurations into spatial composition subspaces, and quantification of distributional shifts across groups, producing a comprehensive and interpretable framework for H&E-based spatial analysis.

4 Method

4.1 Adaptive Multi-level Microenvironment Construction

To construct phenotypic units with spatial consistency and interpretable structure from H&E whole-slide images (WSIs), we propose an adaptive microenvironment generation framework driven by cellular composition and constrained by physical space. This framework operates on a cell-level graph structure, forming multi-level microenvironmental partitions through a "seed selection – constrained growth – competitive allocation" procedure without requiring predefined scales.

Cells in WSIs are segmented using deep learning models. For each cell v_i, morphological features are extracted using encoder $F_i = f_{\text{encoder}}(v_i)$, and a pre-trained classifier assigns a type label:

$$T_i = f_{\text{classifier}}(F_i), \quad T_i \in \{T_1, \ldots, T_K\}, \tag{1}$$

where K denotes the number of cell types. These labels serve as the basis for microenvironment construction (Fig.1, a).

Each tissue slide is represented as a graph, with absolute physical coordinates establishing inter-cellular relationships. A KD-Tree is built to query neighboring cells within radius r, forming the local neighborhood G_i. The neighborhood composition is encoded as a K-dimensional vector $P_i = [p_1, p_2, \ldots, p_K]$, where p_k is the proportion of cell type k in G_i. Local purity scores $S^{(i)}_{\text{seed}}$ are

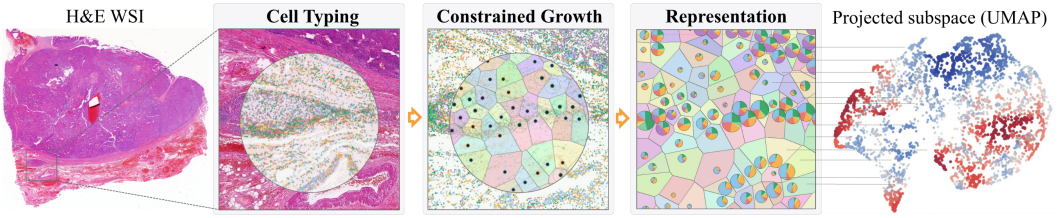

Figure 2: Adaptive microenvironment partition and representation.

computed as the inverse of Shannon entropy:

$$S_{\text{seed}}^{(i)} = \frac{1}{H(P_i)}, \; H(P_i) = -\sum_{k=1}^{K} p_k \log p_k. \tag{2}$$

Representative "tissue centers" are selected via competitive inhibition. Cells are ranked by S_{seed}, and non-maximum suppression (NMS) iteratively selects the highest-scoring unmasked cells v_{top} as cluster roots. Cluster centroids are initialized as $\bar{P}_{C,0} = P_{\text{top}}$, where P_{top} denotes the high-dimensional phenotypic vector of v_{top}. To ensure spatial distinctness, a spatial radius R_{nms} is applied to mask neighboring cells around each root, preventing the redundant selection of overlapping tissue units (Fig. 1, b(1)).

Clusters grow recursively via a breadth-first search (BFS) with an adaptive gating mechanism. A neighbor P_j is included in cluster C if its cosine similarity to the current n-sized centroid $\bar{P}_C^n$ exceeds a dynamic threshold $\mathcal{T}_C(n)$:

$$\cos(\theta) = \frac{P_j \cdot \bar{P}_C^n}{|P_j||\bar{P}_C^n|} > \mathcal{T}_C(n), \tag{3}$$

where the centroid updates incrementally as:

$$\bar{P}_C^{n+1} = \frac{n \cdot \bar{P}_C^n + P_j}{n+1}. \tag{4}$$

To prevent over-expansion into heterogeneous regions, the similarity threshold $\mathcal{T}_C(n)$ increases as cluster size n expands:

$$\mathcal{T}_C(n) = \mathcal{T}_{\text{base}} + \sigma \log(n), \tag{5}$$

where $\mathcal{T}_{\text{base}}$ is the initial stringency and σ is a scaling coefficient. Growth continues until no neighbors satisfy the dynamic similarity criterion or the maximum extent $R_{\text{hard_limit}}$ is reached (Fig.1, b(2)).

Cluster collisions are resolved using an arbitration mechanism. For a cell v_j at coordinates $\mathbf{xy}_j = (x_j, y_j)$, its score relative to a competing cluster C is computed as:

$$\text{Score}(v_j, C) = \omega D_M(v_j, C) + (1-\omega)\frac{||\mathbf{xy}_j - \mathbf{xy}_{\text{top},C}||_2}{R_{\text{current},C}}, \tag{6}$$

where $D_M(v_j, C)$ is the Mahalanobis distance in phenotypic space and $\mathbf{xy}_{\text{top},C}$ is the coordinate of the cluster root v_{top} initially selected via NMS. $R_{\text{current},C}$ denotes the current maximum spatial extent of cluster C, serving as a normalization factor for the Euclidean distance. Nodes are reassigned to the cluster that yields the minimum score, ensuring optimal spatial-phenotypic partitioning (Fig. 1, b(3)).

Growth continues until no eligible neighbors remain, the maximum hard radius $R_{\text{hard_limit}}$ is reached, or the expansion queue is empty. Each resulting microenvironment cluster is represented by a K-dimensional vector $V \in \mathbb{R}^K$, encoding the final composition of cell types (Fig.1, c). Overlapping initialization and local composition preservation ensure high spatial autocorrelation, providing a robust foundation for downstream differential biomarker discovery. The complete procedure, from initial cell-level identification through recursive constrained expansion to the final vectorized spatial partitioning, is visually summarized in Fig.2 and procedure is summarized in Appendix D.

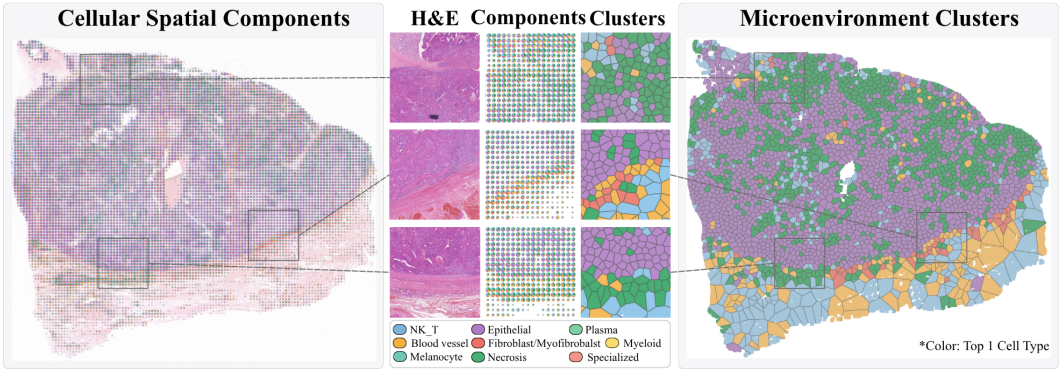

Figure 3: **Qualitative validation of adaptive microenvironment generation.** The workflow illustrates the transition from raw H&E morphology (left) to structured phenotypic units. High-magnification views (middle) showcase the alignment between cellular composition and resulting clusters, while the global partition (right) demonstrates the framework's ability to preserve spatial consistency and histological boundaries.

4.2 Construction and Projection of Spatial and Tissue Feature Space

Building upon the adaptive multi-level microenvironment construction, we further propose a hierarchical feature space construction and projection framework to stably extract structurally differential patterns across patient groups (Fig.1, d).

During microenvironment construction, the m-th identified microenvironment is represented by a K-dimensional cellular composition vector $V_m = [p_1, p_2, \ldots, p_K]$, where p_k denotes the proportion of cell type k within that specific microenvironment. This representation preserves the local spatial topology and structural boundaries derived from the adaptive growth process, while highlighting high-purity regions secured through competitive allocation.

To capture intercellular synergistic interaction patterns, we define a subspace projection operator on the cell type set $T = \{t_1, t_2, \ldots, t_K\}$. For any cell type subset $s \subseteq T$ representing a specific functional unit, the projection operator $\Pi_s : \mathbb{R}^K \to \mathbb{R}^{|s|}$ extracts the corresponding dimensions:

$$V_{m,s} = \Pi_s(V_m), \tag{7}$$

where $V_{m,s} \in \mathbb{R}^{|s|}$ encodes the local composition of cell types in s. By iterating over all subsets $s \in \mathcal{P}(T)$, the original high-dimensional microenvironment vector is decomposed into a set of low-dimensional spatial subspace representations

$$\mathcal{V}_m = \{V_{m,s} \mid s \in \mathcal{P}(T)\}, \tag{8}$$

explicitly capturing complex combinatorial relationships and local dependency patterns among heterogeneous cell type subsets.

Semantically, this framework unifies representations at multiple granularities. When s is a strict subset of T, $V_{m,s}$ corresponds to a spatial subspace describing local multi-cell interactions. When $s = T$, the projection reduces to the identity map $V_{m,s} = V_m$, reflecting the global composition of the tissue structure. Thus, $\mathcal{V}_m$ provides a continuous representation from local cellular co-occurrence to global tissue structure while maintaining a shared generative mechanism across all subspaces.

4.3 Differential Biomarker Discovery Based on Spatial and Tissue Structures

To avoid exhaustive search in the high-dimensional phenotypic space, we propose an instance-based density-aware retrieval strategy. For any query microenvironment q, we define its scale-dependent neighborhood as:

$$N(q, r) = \{z \mid \|z - q\|_2 \leq r\}, \tag{9}$$

where r denotes the phenotypic distance threshold in the latent representation space. This neighborhood is partitioned by prognostic labels $Y \in \{0, 1\}$ into sub-neighborhoods $N_y(q, r) = \{z \in$

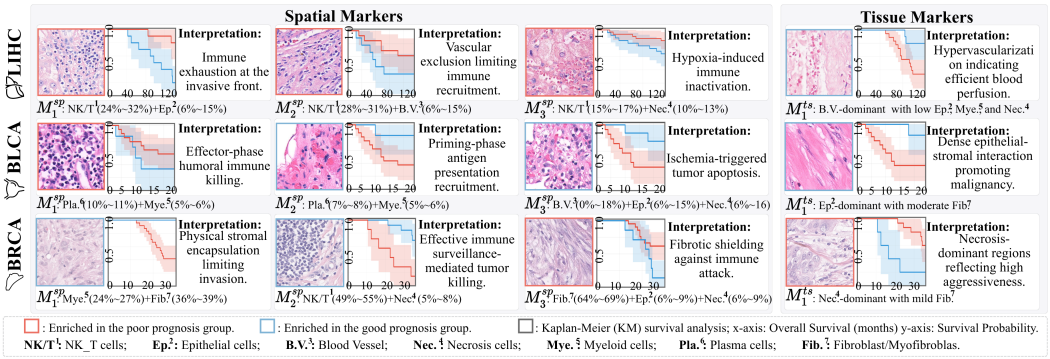

Figure 4: **Prognostic validation and pathological interpretation of spatial and tissue markers.** Representative histology and KM curves across LIHC, BLCA, and BRCA cohorts illustrate survival stratification for identified markers. Red and blue borders denote enrichment in poor and good prognosis groups, respectively, while concise interpretations suggest immune-spatial mechanisms. KM curves illustrate associations with overall survival.

$N(q,r) \mid Y_z = y\}$ for $y \in \{0,1\}$. To accurately capture systematic shifts of spatial structures across patient cohorts, we compute the differential trajectory $\mathrm{LogFC}(q,r)$:

$$\mathrm{LogFC}(q,r) = \log_2 \frac{|N_0(q,r)| + \epsilon}{|N_1(q,r)| + \epsilon}, \tag{10}$$

where ϵ is a small smoothing constant to ensure numerical stability and prevent division by zero. This trajectory encodes the accumulation of structural differences as the scale r increases.

To mitigate bias induced by arbitrary fixed-radius selection, we analyze the curvature of this trajectory to determine structural stability:

$$\frac{d^2}{dr^2}\mathrm{LogFC}(q,r). \tag{11}$$

A curvature approaching zero indicates a transition from a structurally consistent phenotypic core to a mixed region where the differential signal stabilizes. Accordingly, we define the adaptive optimal radius $r^*(q)$ as:

$$r^*(q) = \arg\min_r \left| \frac{d^2}{dr^2}\mathrm{LogFC}(q,r) \right|, \tag{12}$$

ensuring that the neighborhood $N(q, r^*(q))$ captures a stable differential interval that is both statistically coherent and maximally discriminative (Fig. 1, e).

5 Experiments

5.1 Experimental Settings

Datasets and Model Architectures. The cell-level classifier $f_{classfier}$ was trained on the sthelar dataset [22], containing approximately 11 million cells with spatial transcriptomics-based annotations. Microenvironment cluster construction and spatial biomarker discovery were performed on multiple H&E WSI datasets with complete clinical and prognostic information, including both public (TCGA [23]) and private institutional collections (reviewed and approved by the relevant institutional ethics committees Protocol ID: IIT20240720BR1). All data were preprocessed to ensure accurate alignment between WSIs and clinical labels, supporting high-resolution spatial analysis (details in Appendix A).

Parameter Settings. For formal marker mining, the input cell-level table includes coordinates (x, y), predicted type T_i, and cell-type probabilities P_i (cell scheme). Root nodes are selected via non-maximum suppression with $R_{\text{nms}} = 500$, and clusters grow by including neighbors within $R_{\text{contact}} = 80$, with maximum expansion $R_{\text{hard-limit}} = 1500$. Cluster growth uses a similarity threshold $\mathcal{T}_{\text{base}} = 0.90$ and decay $\sigma = 0.01$. Effect size and significance thresholds are $\min |\text{LogFC}| = 0.6$

and $p < 0.05$, with balanced group sampling and merge factor 1.2. Detailed rules are provided in Appendix B and information regarding computational costs is provided in Appendix E.

5.2 Biomarker Discovery

To assess cross-center and cross-cancer generalization, biomarkers were discovered on one dataset and validated on independent cohorts: LIHC biomarkers were mined from Private and evaluated on TCGA [23], while BRCA and BLCA biomarkers were identified and evaluated within their respective datasets. Local spatial Log Fold Change (LogFC) measures enrichment during discovery, while Mann-Whitney U tests assess marker-burden differences across patient groups; external survival tests are identified separately in Table 1 (see also Figure 4). The experimental results describe prognostic heterogeneity across cohorts and cancer types, motivating further independent validation.

5.3 Effectiveness Validation

Encoder Training and Evaluation. Several candidate classifiers were evaluated, including ConvNeXt Tiny, EfficientNet-B0, and ResNet50. Selection was based on overall Top-1 accuracy, macro F1, and the ability to distinguish fine-grained cell types relevant to tumor microenvironments. ConvNeXt Tiny was chosen as the primary backbone, with its hard-negative variant used for the frozen downstream model. Detailed parameter settings, the nine cell-type definitions, and example prediction visualizations are provided in Appendix C.

Microenvironment Cluster Generation. Visualizing phenotypic units on H&E WSIs (Fig.3) shows that the adaptive growth and competitive allocation mechanism effectively partitions tissues into spatially contiguous clusters that correspond to biological structures. Tumor nests, necrotic zones, and stromal boundaries are delineated with high local purity, demonstrating that the framework captures spatial heterogeneity and cellular synergy without relying on predefined scales.

Prognostic Evaluation. To assess the clinical relevance of the identified phenotypic units, we performed Kaplan-Meier survival analysis across multiple cancer cohorts to examine the discovered spatial and tissue markers for patient risk stratification. Furthermore, to assess the biological interpretability of these findings, we collaborated with clinical pathology experts to provide structured semantic definitions and qualitative interpretations for each marker, relating the algorithmic discoveries to established pathological mechanisms. This integration of survival analysis and expert-led pathological interpretation supports exploratory biomarker discovery (Fig. 4).

Paired molecular-reference validation. On five paired H&E–Xenium slides, we independently constructed microenvironment units from H&E predictions and molecular labels, then compared assignments on shared cells and matched-unit compositions. Across slides, mean ARI was 0.677, NMI 0.909, and composition cosine similarity 0.769; dominant-lineage agreement was 0.706. Matched IoU was 0.348 and boundary F1 was 0.315, showing that compositional correspondence was stronger than exact geometric agreement. Restricting the molecular reference to high-confidence A-tier labels gave NMI 0.879, composition cosine 0.776, and dominant-lineage agreement 0.748. This molecular-reference comparison complements the qualitative tissue examples by quantifying both semantic and boundary correspondence.

Sensitivity to cell-classification errors. We perturbed cell types using a prediction-conditioned confusion matrix, interpolating between identity and confusion-based reassignment at strengths $\alpha \in \{0.25, 0.50, 0.75, 1.00\}$, with five random seeds per setting. Specifically, a prediction a was reassigned to b with probability $(1-\alpha)\mathbf{1}[a = b]+\alpha M_{ab}$, where $M_{ab} = P(\text{true} = b \mid \text{predicted} = a)$ was estimated from classifier evaluation. Spatial geometry and candidate definitions were held fixed to isolate label sensitivity. For the three highlighted LIHC markers, patient-burden Spearman correlations ranged from 0.672 to 0.843 and risk-group agreement from 0.755 to 0.838. Global candidate-rank correlations were lower, at 0.085–0.230. Thus, selected patient-level summaries retained greater agreement than fine-grained candidate ordering; the perturbation strength represents a sensitivity setting rather than an estimated WSI error rate.

Direct fixed-grid comparison. On the same 28 LIHC slides, replacing adaptive units with 800-, 1,200-, or 1,600-pixel grids while keeping cell inputs and downstream analysis fixed yielded cell-

Table 1: **Summary of spatial cell-ecology configuration markers and tissue stratification markers across cancer types.** Each marker is shown as $M_i(\text{sign}, p\text{-value})$, where $+$ and $-$ denote enrichment in the favorable and poor prognosis groups. For each cohort, the top 3 spatial markers and the top 1 tissue marker are selected by nominal significance. External LIHC spatial p-values use log-rank for M_1^{sp} and univariable Cox tests for $M_{2,3}^{\text{sp}}$.

	LIHC		**BLCA**	**BRCA**
	Discovery set	External val. set		
Spatial	$M_1^{\text{sp}}(-,\ 4.77\times10^{-3})$	$(-,\ 7.46\times10^{-2})$	$M_1^{\text{sp}}(-,\ 1.33\times10^{-2})$	$M_1^{\text{sp}}(+,\ 1.53\times10^{-2})$
	$M_2^{\text{sp}}(-,\ 4.59\times10^{-6})$	$(-,\ 1.20\times10^{-1})$	$M_2^{\text{sp}}(+,\ 1.41\times10^{-2})$	$M_2^{\text{sp}}(+,\ 3.49\times10^{-2})$
	$M_3^{\text{sp}}(-,\ 2.61\times10^{-2})$	$(-,\ 1.32\times10^{-1})$	$M_3^{\text{sp}}(+,\ 3.66\times10^{-4})$	$M_3^{\text{sp}}(-,\ 1.75\times10^{-2})$
Tissue	$M_1^{\text{ts}}(+,\ 1.70\times10^{-2})$	$(+,\ 4.95\times10^{-2})$	$M_1^{\text{ts}}(+,\ 1.67\times10^{-2})$	$M_1^{\text{ts}}(-,\ 9.43\times10^{-3})$

Table 2: **Comparison of spatial biomarker methods.** Metrics include median p-value ($-\log_{10}$) and intra-cluster entropy; the phenotypic consistency is further validated by the high mean cosine similarity between the markers identified by our discovery algorithm and established baseline methods.

Metric	diffcyt [21]	MELD [20]	Squidpy [19]	**Ours**
Median p-value ($-\log_{10}$)	1.354	0.844	1.114	**2.107**
Intra-cluster Entropy	0.408	0.458	0.429	**0.316**
Phenotypic Similarity (Mean Cosine)	0.875	0.941	0.907	-

weighted entropies of 1.658–1.689, compared with 1.534 for adaptive construction. Adaptive units had median size 124 cells and purity 0.500, versus 128 cells and purity 0.423 for the 800-pixel grid. This homogeneity gain came with lower cell coverage, 0.590 versus approximately 1.000. An oracle-selected 800-pixel grid subset at matched coverage reached entropy 1.523 and purity 0.507, indicating that selective retention of homogeneous tissue contributes to the observed full-grid difference. Across all 3,232 frozen candidates, median patient-burden correlations between grid and adaptive results were 0.518–0.600, distinguishing changes in downstream marker measurements from the construction-level homogeneity comparison.

5.4 Comparative Experiments

To demonstrate the superiority of our framework, we conducted comparative experiments against several state-of-the-art spatial analysis methods, including diffcyt [21], MELD [20], and Squidpy [19]. As shown in Table 2, our approach achieves the highest statistical significance with a median p-value ($-\log_{10}$) of 2.107, outperforming the baseline methods in identifying prognostic markers. Furthermore, our method yields the lowest intra-cluster entropy (0.316), indicating that the generated microenvironments possess superior structural homogeneity and spatial consistency compared to those constructed by existing spatial biomarker discovery tools.

5.5 Ablation Studies

To evaluate the robustness of our microenvironment representation and discovery algorithm, we conducted systematic ablation and stability analyses. Across varying clustering resolutions (k), we observed a trade-off: lower resolutions ($k = 8$) maximize cross-slide consistency (Table 3), while higher resolutions ($k = 32$) capture finer individual heterogeneity but suffer from reduced assignment stability. Our selected resolution ($k = 16$) strikes an optimal balance, maintaining high inter-slide distribution similarity (Median Cosine = 0.602) and superior stability against bootstrap re-fitting (Median NMI = 0.915) and random initialization (Median NMI = 0.883) (Table 4). Furthermore, applying discovery methods diffcyt [21], MELD [20] and Squidpy [19] to our upstream clusters yields high phenotypic consistency (Mean Cosine > 0.87), validating that our framework reliably refines raw spatial signals into stable, reproducible, and clinically relevant biomarkers (Table 2).

Table 3: **Cross-slide consistency of microenvironments.** Metrics: Pair Count (number of slide pairs), Cosine similarity, and JSD (Jensen-Shannon Divergence). Higher Cosine or lower JSD denotes greater consistency.

Resolution (k)	**Pair Count**	**Mean Cosine**	**Median Cosine**	**Mean JSD**	**Median JSD**
$k = 8$	378	0.642	0.805	0.487	0.427
$k = 16$	378	0.516	0.602	0.575	0.543
$k = 32$	378	0.373	0.353	0.665	0.654

Table 4: **Ablation of meta-cluster stability**. NMI and ARI metrics measure clustering similarity; higher values denote greater stability.

Resolution (k)	**Experiment**	**Mean NMI**	**Median NMI**	**Mean ARI**	**Median ARI**
$k = 8$	Bootstrap Fit	0.992	0.991	0.997	0.997
	Random Seed	0.998	0.999	0.999	0.999
$k = 16$ **(Ours)**	Bootstrap Fit	0.915	0.915	0.910	0.909
	Random Seed	0.908	0.883	0.890	0.857
$k = 32$	Bootstrap Fit	0.885	0.883	0.773	0.730
	Random Seed	0.915	0.901	0.839	0.825

6 Conclusion

We present a unified framework for multi-scale spatial biomarker discovery from H&E whole-slide images, integrating adaptive microenvironment construction, spatial feature decomposition, and distribution-level differential analysis. Our approach captures cellular compositions, multi-cellular coordination patterns within microenvironments, and tissue-level structures, enabling the identification of interpretable biomarkers across multiple cancer types. Experiments demonstrate the robustness, generalizability, and superior discriminative power of our method compared with existing approaches. This framework provides a scalable and explainable tool for spatial phenotyping, with potential applications in prognostic stratification and integrative multi-omics studies.

Limitations and Future Work. Our framework's performance depends on the accuracy and diversity of cell segmentation and classification; misclassified or rare cell types may affect microenvironment construction and biomarker discovery. Computational efficiency could limit scalability for very large WSI cohorts, motivating future acceleration strategies such as approximate neighbor search or GPU-based aggregation. Integrating multi-omics data and uncertainty-aware modeling could enhance robustness and biological interpretability, paving the way for more comprehensive characterization of tumor microenvironments across diverse cancer types. Marker-level nominal associations remain exploratory and require multiplicity-controlled independent validation before clinical use.

Ethics Statement. This study utilized de-identified, retrospective data. All procedures were reviewed and approved by the relevant institutional ethics committees (Protocol ID: IIT20240720B-R1). Informed consent was waived in accordance with applicable regulations.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (62376248), the Zhejiang Provincial Natural Science Foundation of China (LQN26F020006), the Ningbo Natural Science Foundation (2025QL002), and the Fundamental Research Funds for the Central Universities (226-2025-00240).

The authors declare no competing interests.

References

[1] F. Mungenast *et al.*, "Next-generation digital histopathology of the tumor microenvironment," *Genes*, vol. 12, no. 4, p. 538, 2021.

[2] H. Zhang *et al.*, "Deep learning-driven recognition of panoramic tumor microenvironment features in h&e sections and its application," *Journal for Immunotherapy of Cancer*, vol. 14, p. e014429, Apr 2026.

[3] J. Kather, A. Pearson, N. Halama, *et al.*, "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer," *Nature Medicine*, vol. 25, pp. 1054–1056, 2019.

[4] D. Schapiro, H. Jackson, S. Raghuraman, *et al.*, "histocat: analysis of cell phenotypes and interactions in multiplex image cytometry data," *Nature Methods*, vol. 14, pp. 873–876, 2017.

[5] Y. Yuan *et al.*, "Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling," *Science Translational Medicine*, vol. 4, no. 157, pp. 157ra143–157ra143, 2012.

[6] J. Saltz *et al.*, "Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images," *Cell Reports*, vol. 23, no. 1, pp. 181–193, 2018.

[7] H. Ali, A. Dariush, E. Provenzano, *et al.*, "Computational pathology of pre-treatment biopsies identifies lymphocyte density as a predictor of response to neoadjuvant chemotherapy in breast cancer," *Breast Cancer Research*, vol. 18, p. 21, 2016.

[8] S. Kunhoth, S. Al-maadeed, Y. Akbari, and R. M. Al Saady, "Computational methods for breast cancer molecular profiling using routine histopathology: A review," *Archives of Computational Methods in Engineering*, Mar 2025.

[9] A. Levy-Jurgenson, X. Tekpli, V. Kristensen, *et al.*, "Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer," *Scientific Reports*, vol. 10, p. 18802, 2020.

[10] H. Xu, J. R. Clemenceau, S. Park, J. Choi, S. H. Lee, and T. H. Hwang, "Spatial heterogeneity and organization of tumor mutation burden with immune infiltrates within tumors based on whole slide images correlated with patient survival in bladder cancer," *Journal of Pathology Informatics*, vol. 13, p. 100105, 2022.

[11] S. Brussee, G. Buzzanca, A. M. Schrader, and J. Kers, "Graph neural networks in histopathology: Emerging trends and future directions," *Medical Image Analysis*, vol. 101, p. 103444, 2025.

[12] D. Anand, S. Gadiya, and A. Sethi, "Histographs: graphs in histopathology," in *Medical Imaging 2020: Digital Pathology* (J. E. Tomaszewski and A. D. Ward, eds.), vol. 11320, p. 113200O, International Society for Optics and Photonics, SPIE, 2020.

[13] X. Meng and T. Zou, "Clinical applications of graph neural networks in computational histopathology: A review," *Computers in Biology and Medicine*, vol. 164, p. 107201, 2023.

[14] M. C. Lau *et al.*, "Tumor-immune partitioning and clustering algorithm for identifying tumor-immune cell spatial interaction signatures within the tumor microenvironment," *PLoS Computational Biology*, vol. 21, p. e1012707, Feb 2025.

[15] S. Jing, H. Wang, P. Lin, *et al.*, "Quantifying and interpreting biologically meaningful spatial signatures within tumor microenvironments," *npj Precision Oncology*, vol. 9, p. 68, 2025.

[16] A. Heindl, S. Nawaz, and Y. Yuan, "Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology," *Laboratory Investigation*, vol. 95, no. 4, pp. 377–384, 2015.

[17] G. Kumar *et al.*, "Spatial modelling of the tumor microenvironment from multiplex immunofluorescence images: methods and applications," *Frontiers in Immunology*, vol. 14, p. 1288802, 2023.

[18] P. Chen, M. Aminu, S. El Hussein, J. D. Khoury, and J. Wu, "Cellspatialgraph: Integrate hierarchical phenotyping and graph modeling to characterize spatial architecture in tumor microenvironment on digital pathology," *Software Impacts*, vol. 10, p. 100156, 2021.

[19] G. Palla *et al.*, "Squidpy: a scalable framework for spatial omics analysis," *Nature Methods*, vol. 19, pp. 171–178, 2022.

[20] D. B. Burkhardt *et al.*, "Quantifying the effect of experimental perturbations at single-cell resolution," *Nature Biotechnology*, vol. 39, pp. 619–629, 2021.

[21] L. M. Weber, M. Nowicka, C. Soneson, and M. D. Robinson, "diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering," *Communications Biology*, vol. 2, p. 183, 2019.

[22] F. Giraud-Sauveur, Q. Blampey, H. Benkirane, *et al.*, "Sthelar, a multi-tissue dataset linking spatial transcriptomics and histology for cell type annotation," *Scientific Data*, vol. 13, p. 665, 2026.

[23] T. C. G. A. R. Network, J. N. Weinstein, E. A. Collisson, *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

A Datasets and Prognostic Grouping

All H&E whole-slide image (WSI) data used in this study were obtained from publicly available sources (TCGA [23]) and private institutional collections (reviewed and approved by the relevant institutional ethics committees Protocol ID: IIT20240720BR1), with clinical diagnoses and follow-up information. Table 5 retains the source-archive inventory, rather than the analysis denominators. The marker-mining runs used 28 LIHC, 22 BLCA, and 20 BRCA slides; evaluable external LIHC patient counts vary by marker and endpoint.

Patients were stratified into favorable- and adverse-prognosis groups based on clinical outcomes: individuals who experienced recurrence within 24 months or death within 60 months were classified as having adverse prognosis, whereas those who did not experience recurrence and survived beyond 60 months, or were censored with follow-up extending past 60 months, were considered to have favorable prognosis. Recurrence and death events were determined according to recorded follow-up times, and censored patients refer to those who remained alive without recurrence at the end of observation. All LIHC, BRCA, and BLCA patients were classified using this unified criterion, providing a consistent stratification framework for spatial biomarker discovery and prognostic analysis. This approach ensures comparability across cancer types and institutions, and establishes a clear foundation for computing log fold change (LogFC) and conducting Mann-Whitney U tests to evaluate the statistical significance of microenvironmental feature.

Table 5: Source-archive inventory and recorded overall survival summaries across cancer types. These are not the marker-analysis denominators. The 870 TCGA files are not a verified unique-patient count.

Cancer	**Inst.**	**Slides**	**Patients**	**OS status**		**Avg. OS**
				Alive	Dead	
LIHC	Private	319	204	137	67	1916
(Liver Hepatocellular Carcinoma)	TCGA	870	–	241	130	813
BLCA (Bladder Urothelial Carcinoma)	Private	51	51	30	21	890
BRCA (Breast Invasive Carcinoma)	Private	107	78	42	36	2190

B Parameter Selection Strategy

The parameters for formal marker mining are chosen to balance sensitivity, stability, and interpretability of microenvironmental patterns. The local neighborhood radius r_{local} is set to capture meaningful cellular composition within each microenvironment, while R_{contact} defines adjacency for cluster growth, and $R_{\text{hard-limit}}$ constrains the spatial expansion of a single cluster to avoid spanning excessively large tissue regions. Root nodes are selected using R_{nms} to control seed density, preventing overly fragmented clusters or excessive sparsity. Minimum cluster size min-cluster-cells filters out small, unstable clusters that may arise from noise, without losing rare microenvironments in smaller samples.

Subspace sizes subspace-sizes are chosen to capture simple pairwise interactions and higher-order multi-cellular configurations, while subspace-entry-threshold ensures that only subspaces with meaningful contributions from all included cell types are considered. Neighborhood statistics for marker discovery are computed adaptively using k_{nn} and adaptive radius steps to accommodate both dense and sparse local patterns. Minimum support requirements min-neighborhood-count and min-group-count prevent spurious markers caused by insufficient data, and balance-groups corrects for differences in sample sizes across prognostic groups. Marker significance is enforced by $\min |\text{LogFC}|$ and p-value thresholds to guarantee both biological relevance and statistical robustness. Coverage of markers is restricted by max-marker-fraction, and extreme values are excluded using interval quantiles. Finally, overlapping marker regions are merged according to merge-radius-factor to reduce redundancy while preserving local heterogeneity.

C Classifier Training and Results

On the STHELAR random10k128 dataset [22], a systematic comparison of multiple cell classifiers was conducted to determine the optimal front-end for downstream microenvironment modeling. The experiments included ConvNeXt Tiny, ConvNeXt Tiny with Hard-negative SupCon, EfficientNet-B0, and ResNet50. All models were pretrained on ImageNet and trained on 128×128 single-cell crops with strong data augmentation and balanced sampling, using square-root inverse class weighting and slight label smoothing to improve macro-level metrics. Evaluation metrics included Top-1 accuracy, macro F1, and Top-3 accuracy, supplemented by multi-granularity diagnostics to assess performance across fine- and coarse-grained cell types.

Results showed that baseline ConvNeXt Tiny achieved a Top-1 accuracy of 0.6003 and a macro F1 of 0.5503. EfficientNet-B0 underperformed on macro F1 and fine-grained category distinction, particularly for commonly confused classes such as NK_T versus B/Plasma and Myeloid versus Fibroblast/Myofibroblast. Incorporating Hard-negative SupCon into ConvNeXt Tiny yielded only a marginal macro F1 improvement to 0.5509, with Top-1 accuracy of 0.5971, indicating limited benefit at the current sample scale. ResNet50 also lagged behind ConvNeXt Tiny in macro F1. The frozen downstream model uses the ConvNeXt Tiny hard-negative variant to generate cell probability distributions. Details are shown in Tab. 6. These metrics use the official grouped patch-level split, which can include different patches from one WSI in different partitions; they are not a strict cross-WSI benchmark.

Nine categories were chosen to cover major immune, stromal, epithelial, and specialized functional components of the tumor microenvironment. They provide a balance between biological interpretability and practical classification complexity, capturing the most relevant cell types for spatial pattern analysis while avoiding overly fine-grained distinctions that could reduce model robustness.

- **NK_T cells**: Cytotoxic lymphocytes involved in immune surveillance and tumor killing; critical for capturing immune-tumor interactions.
- **Plasma cells**: Antibody-secreting cells derived from B lymphocytes; important for understanding local humoral immune response.
- **Myeloid cells**: Includes macrophages and dendritic cells; participate in inflammation, antigen presentation, and tissue remodeling.
- **Blood vessel**: Endothelial structures forming the vasculature; essential for modeling nutrient delivery, hypoxia, and tumor angiogenesis.
- **Fibroblast_Myofibroblast**: Stromal cells producing extracellular matrix and supporting tissue architecture; influence tumor invasion and mechanical properties.
- **Epithelial cells**: Parenchymal tumor or normal epithelial cells; core component for defining tumor regions and tissue organization.
- **Specialized cells**: Tissue-specific cells not captured by other categories; included to preserve rare or unique microenvironmental features.
- **Melanocyte**: Pigment-producing cells; relevant in certain cancer types (e.g., melanoma) and for spatial context in tissue.
- **Necrosis**: Regions of dead or dying tissue; indicate hypoxic or damaged areas, important for delineating tumor boundaries and stress zones.

Table 6: Baseline comparison of cell classifiers on the STHELAR random10k128 grouped patch-level split. Metrics are shown as rows for easier cross-model comparison and rounded from recorded results.

Metric	ConvNeXt Tiny	ConvNeXt Tiny HN-SupCon	EfficientNet-B0	ResNet50
Pretrained	ImageNet	ImageNet	ImageNet	ImageNet
Loss	CE	SupCon+HN	CE	CE
Top-1	0.6003	0.5971	0.5834	0.5855
Macro-F1[†]	0.5503	0.5509	0.5330	0.5429

[†]Macro-F1 calculated for 9-cell type classification; CE = cross-entropy; SupCon+HN = supervised contrastive with hard-negative mining; HN-SupCon indicates the hard-negative SupCon variant of ConvNeXt Tiny.

D Microenviroment Cluster Generation

Algorithm 1: Adaptive Multi-level Microenvironment Construction

Input: WSI with segmented cells $\{v_i\}$ at coordinates (x_i, y_i), cell features $F_i = f_{encoder}(v_i)$, pre-trained classifier $T_i = f_{classifier}(F_i)$, parameters $r, R_{nms}, R_{contact}, Threshold_{sim_base}, \sigma, R_{hard_limit}$

Output: Multi-level microenvironment clusters $\mathcal{C}$ with composition vectors V_C

foreach *cell* v_i *in WSI* **do**

  Compute K-dimensional neighborhood vector $P_i = [p_1, ..., p_K]$ using KD-Tree radius search r;

  Compute local purity $S^{(i)}_{seed} = 1/(-\sum_{k=1}^{K} p_k \log p_k)$;

Sort cells by S_{seed} descending to create list $\mathcal{L}$;

Initialize cluster list $\mathcal{C} = \emptyset$;

while $\mathcal{L}$ *not empty* **do**

  Select top unmasked cell v_{top} as new seed;

  Initialize cluster C with centroid $\bar{P}_C = P_{top}$;

  Mask all neighbors within radius R_{nms} in $\mathcal{L}$;

  Add C to $\mathcal{C}$;

foreach *cluster* $C \in \mathcal{C}$ **do**

  Initialize BFS queue with seed node;

  while *queue not empty and cluster radius* $< R_{hard_limit}$ **do**

    Pop node v_i from queue;

    Find neighbors v_j within $R_{contact}$ using KD-Tree;

    foreach *neighbor* v_j *not yet assigned* **do**

      Compute cosine similarity $\cos(\theta) = \frac{P_j \cdot \bar{P}^n_C}{\|P_j\| \|\bar{P}^n_C\|}$;

      Adjust threshold: $\mathcal{T}_C(n) = \mathcal{T}_{\text{base}} + \sigma \log(n)$;

      if $\cos(\theta) \geq \mathcal{T}_C(n)$ **then**

        Assign v_j to cluster C;

        Update centroid $\bar{P}^n_C \leftarrow \frac{|C| \cdot \bar{P}_C + P_j}{|C|+1}$;

        Add v_j to BFS queue;

    Resolve cluster collisions using score:

$$Score(v_j, C) = \omega D_M(v_j, C) + (1 - \omega) \frac{||\mathbf{xy}_j - \mathbf{xy}_{\text{top},C}||_2}{R_{\text{current},C}}$$

    Reassign node if competing cluster has lower score;

  Compute final cluster composition vector $V_C = \sum_{v_i \in C} P_i / |C|$;

return $\mathcal{C}$ *with* V_C

The figure 5 presents example results of cell-type predictions on H&E patches across multiple cancer types. For each tissue region, the first column shows the original H&E image, the second column

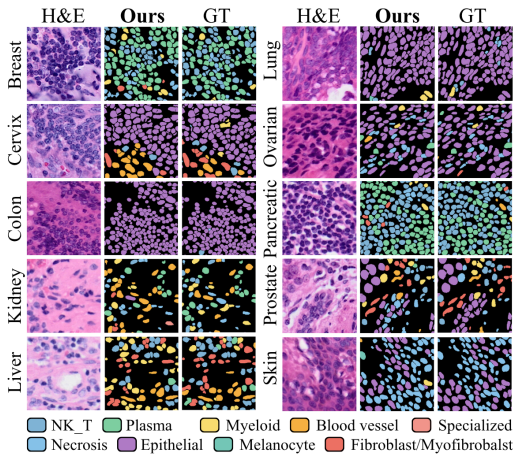

Figure 5: **Example visualizations of cell-type predictions on H&E patches across multiple cancer types.** For each tissue region, the first column shows the raw H&E image, the second column shows the predicted cell-type segmentation by our method, and the third column shows the corresponding ground-truth annotation (GT). Different cell types are color-coded as indicated in the legend. The results demonstrate accurate and spatially coherent cell-type predictions across diverse tissue morphologies, highlighting the model's ability to capture heterogeneous cellular compositions at single-cell resolution.

illustrates the predicted cell-type segmentation generated by the proposed method, and the third column depicts the corresponding ground-truth annotation (GT). Different cell types are color-coded according to the legend, including immune, stromal, epithelial, and specialized cell populations. These visualizations demonstrate that the method can accurately identify individual cells while preserving spatial organization and local neighborhood structures. The predicted segmentations closely match the GT, capturing heterogeneous cellular compositions and subtle morphological variations across diverse tissue types. This supports the reliability of the predicted cellular maps for downstream tasks, including microenvironment clustering, spatial analysis, and biomarker discovery, providing a robust foundation for high-resolution spatial phenotyping studies.

E Additional Experimental Settings

All experiments were run on a shared Linux workstation/server. The main LIHC pipeline used CPU-based preprocessing and graph construction with multi-threaded numerical libraries (OpenBLAS/OMP/MKL/NumExpr capped at 8 threads for graph/ablation steps). A 32-core CPU node with 64-128 GB RAM is sufficient to reproduce the experiments; peak memory for marker mining and ablation was typically below 16 GB, while raw WSI/cell-feature extraction can require higher I/O bandwidth and approximately 10-30 GB RAM depending on parallelism. GPU acceleration was optional and was not the primary bottleneck for the reported experiments.

Preprocessing: Extraction of cellular features and microenvironment graph construction for the LIHC dataset (28 slides, 97,366 microenvironment clusters) took approximately 30 hours on the above hardware. Most of this time was spent on patch/cell feature extraction from WSI and patchmask files; graph construction itself completed within approximately 1 hour.

Algorithm ablation: The stability and consistency tests, including 10 bootstrap refits and 4 random-seed initializations across the evaluated meta-cluster resolutions, required approximately 2 hours on the aforementioned hardware.

Biomarker discovery: KDTree-based differential marker mining and slide-level scoring for the LIHC cohort completed in approximately 1-2 hours. The grid-search baseline over 2.4M input rows completed in approximately 3 minutes, while additional external baseline methods were run separately.

Total pipeline: The end-to-end execution from raw WSI preprocessing to vectorized spatial partitioning and differential biomarker discovery was completed in approximately 36 hours. Including optional ablation and baseline comparison experiments, the full reproducibility run required approximately 38-40 hours.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please refer to Section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This work is empirical and methodological; no theoretical theorems or formal proofs are included.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Full implementation details, parameter settings, dataset sources, and evaluation protocols are provided in Section 4, 5, Appendix A, B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Private clinical datasets are under ethical approval restrictions and cannot be fully open-access; code will be released upon acceptance for academic reproducibility. Public datasets including TCGA [23] and STHELAR [22] are publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: All hyperparameters, model architectures, training pipelines, and clustering settings are clearly reported in Section 5 and Appendix B,C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Statistical significance is reported using p-values, log-fold-changes, and log-rank tests for survival analysis in Section 5 and Table 1 - 4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We strictly adhere to the ethical standards outlined in the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to Section 1.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The proposed method does not involve high-risk models or public data release requiring additional safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets (TCGA [23], STHELAR [22]) and baseline methods (diffcyt [21], MELD [20], Squidpy [19]) are properly cited and comply with their open-access licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new public datasets or code assets are released at submission time; full documentation will be provided upon acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human subject experiments beyond retrospective histopathology data analysis.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Ethical approval is explicitly stated in the Ethics Statement (Section 6); all data are de-identified retrospective clinical data with waived informed consent.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used in any core method, model design, or experimental pipeline; only standard academic writing assistance is applied if any.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.