SpaCRD: Multimodal Deep Fusion of Histology and Spatial Transcriptomics for Cancer Region Detection

Shuailin Xue[1,2], **Jun Wan**[3], **Lihua Zhang**[4], **Wenwen Min**[1,2*]

[1]School of Information Science and Engineering, Yunnan University, Kunming 650500, China
[2]Yunnan Key Laboratory of Intelligent Systems and Computing, Yunnan University, Kunming 650500, China
[3]School of Information Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China
[4]School of Artificial Intelligence, School of Computer Science, Wuhan University, Wuhan 430072, China
xueshuailin@stu.ynu.edu.cn, junwan2014@whu.edu.cn, zhanglh@whu.edu.cn, minwenwen@ynu.edu.cn

Abstract

Accurate detection of cancer tissue regions (CTR) enables deeper analysis of the tumor microenvironment and offers crucial insights into treatment response. Traditional CTR detection methods, which typically rely on the rich cellular morphology in histology images, are susceptible to a high rate of false positives due to morphological similarities across different tissue regions. The groundbreaking advances in spatial transcriptomics (ST) provide detailed cellular phenotypes and spatial localization information, offering new opportunities for more accurate cancer region detection. However, current methods are unable to effectively integrate histology images with ST data, especially in the context of cross-sample and cross-platform/batch settings for accomplishing the CTR detection. To address this challenge, we propose SpaCRD, a transfer learning-based method that deeply integrates histology images and ST data to enable reliable CTR detection across diverse samples, platforms, and batches. Once trained on source data, SpaCRD can be readily generalized to accurately detect cancerous regions across samples from different platforms and batches. The core of SpaCRD is a category-regularized variational reconstruction-guided bidirectional cross-attention fusion network, which enables the model to adaptively capture latent co-expression patterns between histological features and gene expression from multiple perspectives. Extensive benchmark analysis on 23 matched histology-ST datasets spanning various disease types, platforms, and batches demonstrates that SpaCRD consistently outperforms existing eight state-of-the-art methods in CTR detection.

Code — https://github.com/wenwenmin/SpaCRD

1 Introduction

In clinical diagnosis and medical research, cancer tissue regions (CTR) detection is a critical step in developing treatment strategies for oncology patients (Lawrence et al. 2023). It not only aids in delineating surgical margins and precisely delivering radiation doses, but also provides essential spatial references for tumor microenvironment analysis (Khalighi et al. 2024). Prior research has commonly relied on manual annotations from pathologists and traditional anomaly detection algorithms based on histology images analysis for CTR detection (Shmatko et al. 2022; Zingman et al. 2024). However, the former is limited by its high cost and time-consuming nature, while the latter suffers from poor accuracy due to the misleading morphological similarities across different tissue regions and inconsistent staining quality in histology images (Xue et al. 2025). As a result, neither is an ideal solution for CTR detection in clinical diagnosis and biomedical research.

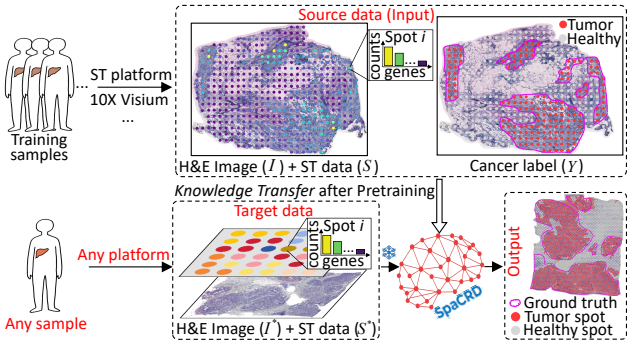

Figure 1: SpaCRD achieves accurate CTR detection across diverse samples and platforms&batches by deeply integrating histology images and ST data through transfer learning.

As a revolutionary technology, spatial transcriptomics (ST) enables comprehensive profiling of transcripts across entire tissue sections while preserving their spatial locations, offering unprecedented capabilities for characterizing spatial heterogeneity and abnormal tissue architecture (Rao et al. 2021; Tian, Chen et al. 2023). However, to precisely localize sequencing positions, ST requires additional processing steps during sequencing, which inevitably introduce background noise into the gene expression (Janssen et al. 2023). ST-based algorithms (Benjamin et al. 2024; Ferri-Borgogno et al. 2023; Shen et al. 2022) for various tasks are inevitably affected by the substantial noise inherent in ST data, which often prevents them from achieving optimal performance (Zahedi et al. 2024).

In addition, current tissue annotation methods (Zhang et al. 2024; Hu et al. 2023), which rely heavily on expert-defined prior knowledge, also detect CTR by aggregating the expression of marker genes. However, identifying re-

*Corresponding author.

liable markers often requires extensive domain knowledge and experimental validation, and many cancer types lack well-defined markers that have been systematically identified, which further limits the generalizability and applicability of such approaches.

Effectively integrating histology and ST data to overcome morphological ambiguities and ST noise remains a major challenge in CTR detection (Maan et al. 2025). While some multimodal methods have been proposed, they suffer from critical shortcomings. For instance, SpaCell (Tan et al. 2020) fuses histology and ST data through simple feature concatenation, neglecting cross-modal interactions and global spatial context. STANDS (Xu et al. 2024) and MEATRD (Xu et al. 2025) follow the paradigm of traditional visual anomaly detection (Liu et al. 2023), but their reliance on reconstruction errors is ill-suited for structured cancer regions, which differ fundamentally from sparse anomalies in natural images (Seferbekova et al. 2023). Furthermore, these methods often fail to generalize across datasets due to batch heterogeneity, highlighting the need for more robust solutions.

Transfer learning offers a promising solution by leveraging knowledge learned from well-annotated source datasets to improve performance on heterogeneous target domains (Huang et al. 2025). For example, several studies have used single-cell RNA-seq datasets from similar tissues as the source domain to guide cell type classification in ST datasets (target domain), effectively transferring cell-type signatures to spatial contexts (Yan et al. 2025; Hao et al. 2021). In CTR detection tasks, the target samples often originate from different platforms and experimental batches, resulting in substantial technical and biological variability. Motivated by the success of transfer learning, we apply transfer learning to align heterogeneous ST datasets and improve the generalizability of CTR detection across platforms and batches.

In this study, we propose **SpaCRD**, a multimodal deep fusion framework that integrates histology images and **Spa**tial transcriptomics data for **C**ancer **R**egion **D**etection. By leveraging transfer learning and a powerful multimodal deep fusion module, SpaCRD effectively mitigates technical and batch variations among affected individuals, achieving accurate and consistent CTR detection performance across different samples, platforms, and batches. Specifically, SpaCRD comprises a pretrained pathology foundation model, a modality-alignment representation learning, and a category-regularized **V**ariational **R**econstruction-guided **B**idirectional **C**ross-**A**ttention fusion network (VRBCA), which collectively enable end-to-end CTR detection from multimodal inputs. To the best of our knowledge, SpaCRD is the first framework that combines multimodal deep fusion with transfer learning for CTR detection. In summary, our contributions can be summarized as follows:

- We propose SpaCRD, a novel CTR detection framework leveraging multimodal deep fusion and transfer learning.
- SpaCRD mitigates technical and batch effects by aligning samples of the same disease type into a consistent representation space, thus enabling CTR detection across diverse samples and platforms&batches.
- We design a VRBCA network that reduces modality discrepancies, filters out noise, and stabilizes the fusion of ST data and histology images from complementary perspectives, while comprehensively modeling interactions among neighboring spots to facilitate the generation of compact and class-specific multimodal embeddings.
- Extensive benchmark on eleven breast cancer and twelve colorectal cancer datasets demonstrate that SpaCRD consistently outperforms eight state-of-the-art (SOTA) methods in CTR detection.

2 Methods [1]

As shown in Figure 2, our proposed SpaCRD framework consists of three main training stages: (1) **Modality-alignment representation learning.** We employ a pretrained pathology foundation model UNI (Chen et al. 2024) to extract informative histology image features. Meanwhile, we adopt a CLIP-based contrastive learning strategy (Radford et al. 2021) to align histology and ST modalities in a shared embedding space, effectively narrowing the modality gap and preparing for subsequent fusion. (2) **VRBCA Fusion Network.** VRBCA aims to learn compact and informative representations of the interactions between histology and ST modalities. (3) **Cancer likelihood estimation.** SpaCRD estimates a cancer likelihood score for each spot based on the compact representation yielded by VRBCA. The overall framework is summarized in the algorithm in Supplementary Material S2.

2.1 Modality-Alignment Representation Learning

First, we crop image patches from histology images based on spatial coordinates of each spot in ST data. Patch size is determined by spot diameter in ST data and the pixel resolution of the histology images (Jaume et al. 2024a). We adopt UNI, a pathology-specific foundation model pretrained on large-scale histology data, as our histological feature extractor. Due to its demonstrated ability to capture fine-grained histological structures, we skip the fine-tuning step to reduce computational overhead. Let $I = \left\{ I_i \mid I_i \in \mathbb{R}^{l \times l \times 3} \right\}_{i=1}^{n}$ denote the set of image patches corresponding to all spots, where n is the total number of spots, and $l = d/r$, with d and r representing the spot diameter and the pixel resolution of the histology image, respectively. Then passing the entire set of patches I through the UNI model yields the H&E embeddings $X^{\text{img}} = \left\{ \mathbf{x}_1^{\text{img}}, \ldots, \mathbf{x}_n^{\text{img}} \right\}$ where

$$\mathbf{x}_i^{\text{img}} = f_{\text{UNI}}\left(I_i\right), \quad i = 1, \ldots, n. \tag{1}$$

Let $X^{\text{gene}} = \left\{ \mathbf{x}_1^{\text{gene}}, \ldots, \mathbf{x}_n^{\text{gene}} \right\}$ represent the set of gene expression profiles in the ST data, where $\mathbf{x}_i^{\text{gene}}$ denotes the normalized expression vector of spot i. Given the substantial heterogeneity between image embeddings and gene expression data (Jaume et al. 2024b), we employ a contrastive learning strategy to align the two modalities. By optimizing the contrastive loss, the model encourages paired image features and gene expression vectors from the same spatial

[1] Related Work is in Supplementary Material S1 due to space limitation.

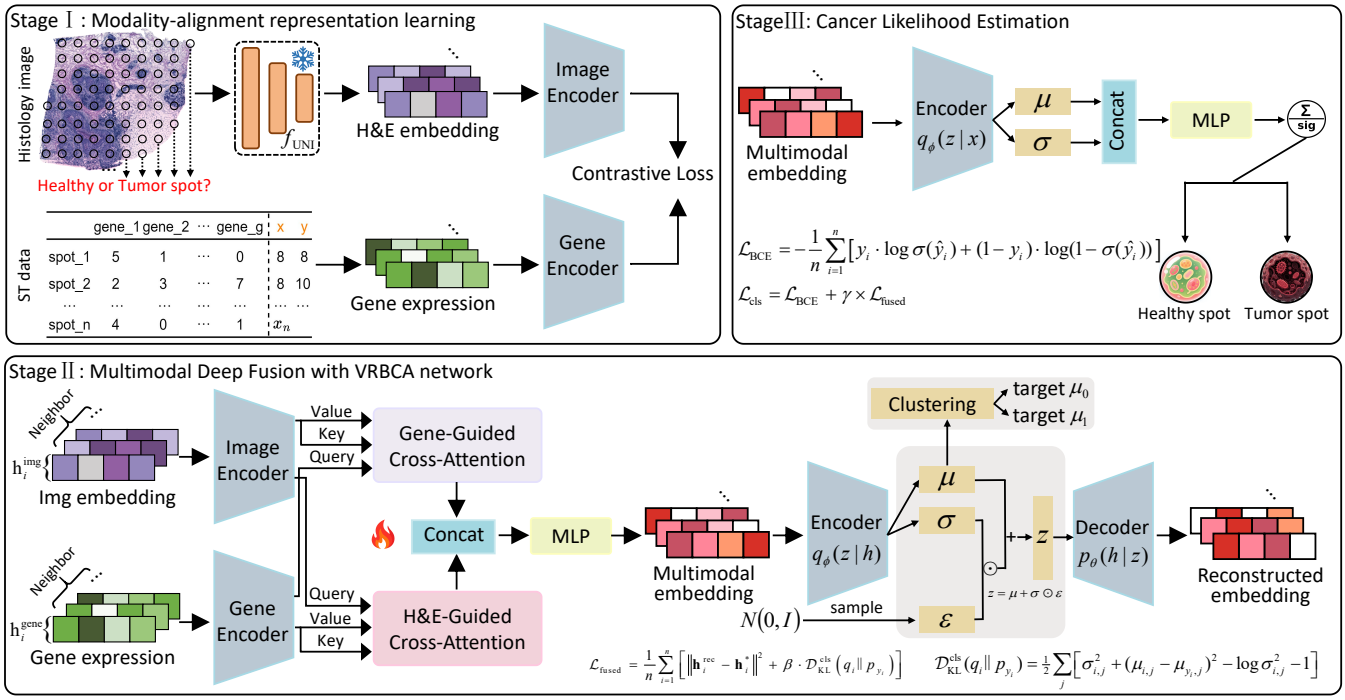

Figure 2: The framework of SpaCRD. Stage I: UNI is used to extract histology features, while modality-alignment representation learning aligns histology and ST modalities into a shared embedding space. Stage II: VRBCA integrates aligned histology and ST features into a compact and class-consistent embedding that captures biologically relevant cross-modal interactions. Stage III: The learned representation is used to estimate cancer likelihood scores for each spot.

location to be close in the latent space, while pushing apart those from different locations. This alignment reduces inconsistencies during the subsequent fusion process, ensuring a more stable and effective integration. We design two lightweight three-layer MLP-based encoders—an image encoder f_{c1} and a gene encoder f_{c2}—to perform the aforementioned alignment task:

$$\mathbf{h}_i^{\text{img}} = f_{c1}(\mathbf{x}_i^{\text{img}}), \quad \mathbf{h}_i^{\text{gene}} = f_{c2}(\mathbf{x}_i^{\text{gene}}). \tag{2}$$

To measure similarity between latent representations, we construct a similarity matrix using their scaled dot-product:

$$\mathbf{S}_{ij} = \frac{\mathbf{h}_i^{\text{img}} \cdot \mathbf{h}_j^{\text{gene}}}{\tau}, \quad \text{where } \|\mathbf{h}_i^{\text{img}}\| = \|\mathbf{h}_j^{\text{gene}}\| = 1, \tag{3}$$

with τ being a temperature parameter. Here, $\mathbf{S}_{ii}$ denotes the similarity between the i-th image and its paired gene expression, while $\mathbf{S}_{ij}$ for $i \neq j$ represents the similarity between unmatched pairs. Two directional InfoNCE loss components, $\mathcal{L}_{\text{img}\to\text{gene}}$ and $\mathcal{L}_{\text{gene}\to\text{img}}$, are independently computed using cross-entropy over similarity logits:

$$\mathcal{L}_{\text{img}\to\text{gene}} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp(\mathbf{S}_{ii})}{\sum_{j=1}^{n}\exp(\mathbf{S}_{ij})}, \tag{4}$$

$$\mathcal{L}_{\text{gene}\to\text{img}} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp(\mathbf{S}_{ii})}{\sum_{j=1}^{n}\exp(\mathbf{S}_{ji})}. \tag{5}$$

Finally, the total contrastive loss for this stage is defined as:

$$\mathcal{L}_{\text{contrast}} = \alpha \times \mathcal{L}_{\text{img}\to\text{gene}} + (1-\alpha) \times \mathcal{L}_{\text{gene}\to\text{img}}, \tag{6}$$

where $\alpha \in [0,1]$ controls the balance between the two directional losses. For all experiments, α is fixed at 0.5.

2.2 VRBCA Fusion Network

Following modality-alignment representation learning, the encoded features are fed into the VRBCA network. Overall, VRBCA utilizes a bidirectional cross-attention (BCA) mechanism to comprehensively model interactions between spots and to integrate the aligned features from multiple perspectives, followed by a category-regularized variational autoencoder (RVAE) for filtering out noise and promoting the generation of compact and class-specific embeddings. We define two independent Cross-Attention (CA) modules with identical architectures: gene-guided and H&E-guided CA blocks. Each module employs m parallel attention heads. For the i-th head, the query, key, and value matrices are derived through linear projections of the inputs:

$$Q_i = QW_i^Q, K_i = KW_i^K, V_i = VW_i^V. \tag{7}$$

Here, the shapes of W_i^Q, W_i^K, and W_i^V depend on the input modality and output embedding dimension d. The output of each head is computed as:

$$h_i = \text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_h}}\right) V_i, \tag{8}$$

$$\text{CA}(Q,K,V) = (h_1 \parallel h_2 \parallel \cdots \parallel h_m)\, W, \tag{9}$$

where $\parallel$ denotes the concatenation of outputs from all attention heads, and W is a shared projection matrix.

Metric	Method	Cross-samples											
		CRC_A1	CRC_A2	CRC_B1	CRC_B2	CRC_C1	CRC_C2	CRC_D1	CRC_D2	CRC_E1	CRC_E2	CRC_G1	CRC_G2
AUC	SimpleNet	0.491±0.163	0.518±0.098	0.488±0.036	0.481±0.055	0.440±0.047	0.488±0.070	0.518±0.050	0.435±0.036	0.429±0.129	0.580±0.103	0.506±0.082	0.463±0.047
	Spatial-ID	0.594±0.068	0.563±0.091	0.518±0.016	0.496±0.043	0.513±0.057	0.503±0.082	0.510±0.029	0.526±0.040	0.457±0.109	0.453±0.093	0.458±0.042	0.525±0.033
	STAGE	0.544±0.088	0.516±0.116	0.517±0.037	0.520±0.078	0.488±0.084	0.502±0.059	0.498±0.086	0.490±0.055	0.354±0.129	0.466±0.086	0.488±0.106	0.537±0.117
	STANDS	0.540±0.045	0.530±0.043	0.466±0.030	0.506±0.021	0.587±0.053	0.549±0.057	0.564±0.053	0.510±0.022	0.609±0.075	0.643±0.098	0.504±0.027	0.600±0.070
	TESLA	0.496±0.016	0.527±0.004	0.436±0.006	0.548±0.007	0.716±0.011	0.593±0.002	0.664±0.006	0.582±0.036	0.517±0.007	0.635±0.024	0.662±0.004	0.721±0.006
	iStar	0.444±0.025	0.436±0.027	0.517±0.002	0.514±0.005	0.733±0.004	0.638±0.004	0.654±0.008	**0.672**±0.061	0.505±0.014	0.593±0.051	0.781±0.002	0.799±0.005
	MEATRD	0.505±0.007	0.517±0.008	0.511±0.005	0.548±0.002	0.765±0.008	0.539±0.009	0.498±0.007	0.466±0.013	0.541±0.005	0.904±0.011	0.507±0.006	0.550±0.006
	SpaCell-Plus	0.821±0.012	0.762±0.010	0.678±0.006	0.713±0.004	0.761±0.014	0.713±0.012	0.625±0.034	0.472±0.018	0.799±0.024	0.816±0.015	0.638±0.028	0.733±0.016
	SpaCRD(ours)	**0.953**±0.002	**0.925**±0.004	**0.824**±0.005	**0.897**±0.003	**0.895**±0.007	**0.880**±0.010	**0.789**±0.020	0.603±0.022	**0.966**±0.006	**0.961**±0.007	**0.853**±0.005	**0.888**±0.003
AP	SimpleNet	0.312±0.079	0.356±0.070	0.688±0.027	0.703±0.039	0.239±0.019	0.425±0.042	0.869±0.016	0.888±0.010	0.342±0.067	0.598±0.081	0.165±0.034	0.188±0.015
	Spatial-ID	0.366±0.023	0.401±0.103	0.707±0.019	0.712±0.025	0.285±0.045	0.441±0.070	0.866±0.012	0.912±0.012	0.352±0.085	0.512±0.059	0.141±0.011	0.215±0.023
	STAGE	0.300±0.054	0.335±0.092	0.686±0.025	0.712±0.044	0.257±0.047	0.419±0.043	0.852±0.031	0.913±0.010	0.295±0.052	0.504±0.064	0.164±0.045	0.240±0.089
	STANDS	0.319±0.038	0.351±0.039	0.664±0.020	0.713±0.017	0.349±0.056	0.452±0.053	0.888±0.016	0.911±0.005	0.475±0.074	0.676±0.087	0.167±0.019	0.305±0.071
	TESLA	0.317±0.008	0.421±0.024	0.574±0.004	0.613±0.005	0.376±0.004	0.536±0.006	0.874±0.003	0.813±0.017	0.427±0.004	0.642±0.022	0.425±0.002	0.360±0.004
	iStar	0.252±0.007	0.290±0.008	0.695±0.002	0.717±0.003	0.410±0.002	0.503±0.003	0.900±0.002	0.911±0.010	0.372±0.006	0.597±0.033	0.349±0.007	0.433±0.007
	MEATRD	0.272±0.005	0.329±0.005	0.693±0.003	0.742±0.005	0.518±0.023	0.439±0.007	0.861±0.005	0.902±0.002	0.403±0.005	0.892±0.015	0.194±0.003	0.236±0.004
	SpaCell-Plus	0.714±0.035	0.695±0.016	0.731±0.008	0.738±0.018	0.577±0.020	0.670±0.005	0.773±0.026	0.836±0.022	0.748±0.009	0.797±0.023	0.520±0.036	0.624±0.011
	SpaCRD(ours)	**0.856**±0.009	**0.853**±0.010	**0.904**±0.014	**0.951**±0.001	**0.749**±0.018	**0.815**±0.026	**0.951**±0.010	**0.936**±0.006	**0.935**±0.008	**0.954**±0.014	**0.619**±0.030	**0.730**±0.005
F1	SimpleNet	0.290±0.138	0.351±0.061	0.678±0.012	0.701±0.022	0.210±0.031	0.412±0.057	0.865±0.007	0.910±0.002	0.306±0.112	0.606±0.059	0.151±0.061	0.166±0.028
	Spatial-ID	0.359±0.044	0.364±0.086	0.688±0.005	0.705±0.026	0.283±0.066	0.418±0.064	0.864±0.006	0.918±0.002	0.326±0.123	0.510±0.078	0.116±0.019	0.217±0.049
	STAGE	0.294±0.093	0.327±0.109	0.721±0.016	0.726±0.031	0.226±0.084	0.417±0.045	0.816±0.011	0.911±0.002	0.242±0.107	0.520±0.054	0.150±0.080	0.234±0.116
	STANDS	0.300±0.048	0.344±0.041	0.671±0.011	0.711±0.010	0.326±0.064	0.465±0.043	0.866±0.003	0.915±0.002	0.466±0.071	0.638±0.081	0.160±0.031	0.298±0.079
	TESLA	0.412±0.010	0.345±0.015	0.427±0.022	0.106±0.021	0.513±0.007	0.612±0.004	0.853±0.002	0.809±0.029	0.546±0.030	0.594±0.033	0.384±0.006	0.528±0.018
	iStar	0.332±0.037	0.318±0.011	0.103±0.027	0.058±0.019	0.576±0.003	0.667±0.003	0.872±0.005	0.744±0.150	0.434±0.026	0.572±0.092	0.531±0.009	0.608±0.006
	MEATRD	0.270±0.011	0.320±0.010	0.690±0.004	0.723±0.002	0.526±0.017	0.452±0.014	0.866±0.002	0.911±0.001	0.406±0.012	0.851±0.016	0.156±0.011	0.239±0.013
	SpaCell-Plus	0.592±0.021	0.633±0.008	0.714±0.020	0.696±0.006	0.598±0.016	0.573±0.012	0.726±0.036	0.823±0.008	0.710±0.025	0.743±0.016	0.516±0.016	0.576±0.021
	SpaCRD(ours)	**0.802**±0.006	**0.805**±0.009	**0.836**±0.013	**0.881**±0.005	**0.683**±0.025	**0.777**±0.010	**0.911**±0.002	**0.923**±0.001	**0.894**±0.009	**0.919**±0.006	**0.599**±0.015	**0.694**±0.007

Table 1: Quantitative evaluation of SpaCRD for CTR detection on twelve colorectal cancer datasets compared to baselines. Each reported values are means ± standard deviations over five independent runs. Best results in bold, second-best underlined.

Metric	Method	Cross-samples								Cross-platforms&batches		
		STHBC_A	STHBC_B	STHBC_C	STHBC_D	STHBC_E	STHBC_F	STHBC_G	STHBC_H	ViHBC	XeHBC	IDC
AUC	SimpleNet	0.536±0.126	0.468±0.157	0.476±0.132	0.528±0.143	0.429±0.122	0.536±0.063	0.571±0.101	0.454±0.132	0.468±0.076	0.531±0.106	0.612±0.098
	Spatial-ID	0.544±0.045	0.534±0.135	0.471±0.083	0.451±0.085	0.477±0.102	0.496±0.107	0.558±0.094	0.438±0.117	0.436±0.067	0.593±0.081	0.502±0.080
	STAGE	0.536±0.126	0.468±0.157	0.476±0.132	0.528±0.143	0.429±0.122	0.536±0.063	0.571±0.101	0.474±0.190	0.444±0.067	0.575±0.054	0.605±0.068
	STANDS	0.535±0.053	0.449±0.058	0.510±0.008	0.507±0.038	0.524±0.027	0.575±0.036	0.564±0.046	0.451±0.072	0.629±0.102	0.532±0.037	0.466±0.092
	TESLA	0.724±0.008	0.586±0.018	0.551±0.012	0.605±0.011	0.627±0.002	0.705±0.022	0.697±0.014	0.729±0.009	0.672±0.025	0.748±0.026	0.576±0.021
	iStar	0.788±0.013	0.674±0.020	0.591±0.017	0.668±0.016	0.741±0.004	0.672±0.024	0.714±0.017	0.681±0.011	0.736±0.032	0.842±0.038	0.531±0.009
	MEATRD	0.604±0.007	0.963±0.003	0.783±0.010	0.837±0.006	0.442±0.013	0.468±0.023	0.774±0.029	0.614±0.024	0.517±0.002	0.472±0.004	0.496±0.002
	SpaCell-Plus	0.929±0.010	0.950±0.012	0.773±0.019	0.844±0.009	0.668±0.056	0.774±0.056	0.826±0.021	0.812±0.017	0.784±0.028	0.818±0.034	0.803±0.009
	SpaCRD(ours)	**0.979**±0.002	**0.993**±0.002	**0.795**±0.025	**0.952**±0.003	**0.909**±0.011	**0.913**±0.008	**0.923**±0.004	**0.969**±0.004	**0.900**±0.029	**0.931**±0.008	**0.891**±0.006
AP	SimpleNet	0.907±0.038	0.226±0.062	0.724±0.070	0.511±0.097	0.536±0.083	0.870±0.024	0.412±0.082	0.287±0.062	0.603±0.047	0.570±0.054	0.662±0.073
	Spatial-ID	0.909±0.017	0.264±0.122	0.729±0.038	0.453±0.047	0.565±0.076	0.855±0.042	0.383±0.075	0.285±0.036	0.571±0.051	0.405±0.035	0.582±0.044
	STAGE	0.907±0.038	0.226±0.062	0.724±0.070	0.511±0.097	0.536±0.083	0.870±0.024	0.412±0.082	0.309±0.155	0.559±0.039	0.529±0.047	0.643±0.053
	STANDS	0.889±0.019	0.179±0.018	0.731±0.023	0.451±0.037	0.557±0.024	0.859±0.017	0.345±0.029	0.290±0.016	0.692±0.094	0.575±0.060	0.573±0.070
	TESLA	0.842±0.004	0.354±0.012	0.803±0.004	0.657±0.006	0.683±0.008	0.829±0.006	0.576±0.016	0.568±0.004	0.754±0.028	0.620±0.012	0.697±0.007
	iStar	0.956±0.003	0.332±0.020	0.783±0.009	0.632±0.011	0.761±0.004	0.903±0.007	0.543±0.029	0.460±0.005	0.788±0.021	0.723±0.014	0.621±0.007
	MEATRD	0.916±0.017	0.948±0.007	0.916±0.005	0.832±0.013	0.554±0.010	0.848±0.008	0.653±0.024	0.459±0.011	0.631±0.005	0.317±0.008	0.595±0.003
	SpaCell-Plus	0.991±0.002	0.853±0.036	0.904±0.007	0.845±0.011	0.757±0.043	0.949±0.017	0.745±0.049	0.745±0.013	0.846±0.015	0.726±0.018	0.847±0.006
	SpaCRD(ours)	**0.998**±0.000	**0.980**±0.003	**0.923**±0.011	**0.961**±0.001	**0.911**±0.013	**0.983**±0.003	**0.867**±0.007	**0.947**±0.005	**0.930**±0.037	**0.859**±0.021	**0.914**±0.006
F1	SimpleNet	0.903±0.006	0.197±0.127	0.738±0.046	0.499±0.103	0.535±0.067	0.863±0.008	0.385±0.099	0.260±0.081	0.604±0.044	0.592±0.087	0.672±0.059
	Spatial-ID	0.905±0.009	0.225±0.164	0.740±0.029	0.437±0.068	0.569±0.063	0.859±0.013	0.385±0.080	0.227±0.069	0.590±0.030	0.430±0.430	0.611±0.055
	STAGE	0.903±0.006	0.197±0.127	0.738±0.046	0.499±0.103	0.535±0.067	0.863±0.008	0.385±0.099	0.330±0.172	0.597±0.039	0.437±0.034	0.674±0.037
	STANDS	0.898±0.005	0.111±0.043	0.718±0.019	0.453±0.031	0.566±0.015	0.863±0.003	0.341±0.050	0.079±0.035	0.699±0.059	0.604±0.046	0.582±0.049
	TESLA	0.820±0.015	0.476±0.020	0.404±0.005	0.579±0.022	0.643±0.004	0.663±0.018	0.676±0.016	0.498±0.008	0.641±0.034	0.586±0.016	0.510±0.026
	iStar	0.803±0.028	0.483±0.026	0.389±0.026	0.536±0.044	0.701±0.006	0.700±0.027	0.618±0.023	0.582±0.009	0.668±0.065	0.689±0.020	0.135±0.033
	MEATRD	0.907±0.022	0.886±0.021	**0.843**±0.007	0.752±0.008	0.536±0.013	0.850±0.005	0.648±0.017	0.453±0.027	0.630±0.001	0.333±0.007	0.597±0.003
	SpaCell-Plus	0.956±0.003	0.778±0.044	0.827±0.016	0.730±0.013	0.672±0.042	0.902±0.010	0.668±0.043	0.645±0.022	0.766±0.025	0.667±0.036	0.758±0.009
	SpaCRD(ours)	**0.975**±0.002	**0.926**±0.012	0.818±0.004	**0.878**±0.009	**0.866**±0.011	**0.929**±0.005	**0.789**±0.007	**0.860**±0.009	**0.867**±0.014	**0.795**±0.013	**0.854**±0.008

Table 2: Quantitative evaluation of SpaCRD for CTR detection on eleven breast cancer datasets compared to baselines. Each reported values are means ± standard deviations over five independent runs. Best results in bold, second-best underlined.

Next, let $H_i^{\text{img}} = \left[\mathbf{h}_i^{\text{img}}, \mathbf{h}_{i_1}^{\text{img}}, \ldots, \mathbf{h}_{i_k}^{\text{img}}\right]^T$ and $H_i^{\text{gene}} = \left[\mathbf{h}_i^{\text{gene}}, \mathbf{h}_{i_1}^{\text{gene}}, \ldots, \mathbf{h}_{i_k}^{\text{gene}}\right]^T$ denote the input matrices for spot i, where $\mathbf{h}_i^{\text{img}}$ and $\mathbf{h}_i^{\text{gene}}$ denote the contrastively aligned embeddings of H&E image and gene expression for spot i, respectively, and $\{i_1, \ldots, i_k\} \subset \mathcal{N}(i)$ denotes its set of neighboring spots. The operation of the BCA module proceeds:

$$Z_i^{\text{img}} = \text{CA}_{\text{img}}(H_i^{\text{img}}, H_i^{\text{gene}}, H_i^{\text{gene}}), \tag{10}$$

$$Z_i^{\text{gene}} = \text{CA}_{\text{gene}}(H_i^{\text{gene}}, H_i^{\text{img}}, H_i^{\text{img}}), \tag{11}$$

$$\mathbf{h}_i^* = \text{MLP}(Z_i^{\text{img}}[0] \parallel Z_i^{\text{gene}}[0]), \tag{12}$$

where h_i^* denotes the multimodal representation of spot i, fused by the BCA module based on the features aligned through contrastive learning.

To further refine the fused multimodal representation h_i^*, the training of VRBCA network is guided by a category-regularized variational reconstruction objective, which imposes semantic structure on the latent space and facilitates the learning of class-discriminative multimodal features. Specifically, we extend the VAE framework by introducing learnable class-specific latent centers, forming a structure-

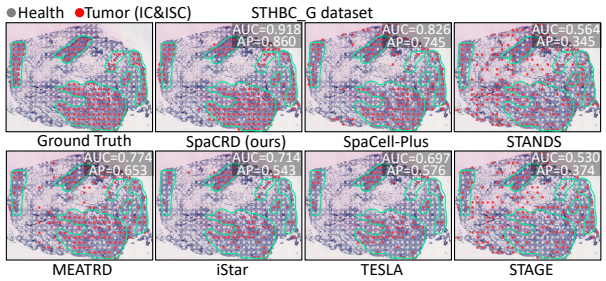

Figure 3: ATR detection results of SpaCRD and other baselines on the STHBC_G dataset. Green outlines indicate pathologist-annotated CTR. Gray and red dots represent normal and cancerous spots, respectively.

aware latent prior. VRBCA encodes the fused multimodal representation into latent variables via an encoder f_{enc}, producing mean $\boldsymbol{\mu}_i$ and log-variance $\log \boldsymbol{\sigma}_i^2$. A latent vector $\mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$ is sampled and decoded by f_{dec} to reconstruct the multimodal representation, yielding $\hat{h}_i^*$:

$$\boldsymbol{\mu}_i, \log \boldsymbol{\sigma}_i^2 = f_{\text{enc}}\left(\mathbf{h}_i^*\right), \tag{13}$$

$$z_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma} \odot \epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{14}$$

$$\hat{\mathbf{h}}_i^* = f_{\text{dec}}\left(z_i\right). \tag{15}$$

The training objective for the fused representation is defined as:

$$\mathcal{L}_{\text{fused}} = \frac{1}{n} \sum_{i=1}^{n} \left[\left\| \hat{\mathbf{h}}_i^* - \mathbf{h}_i^* \right\|^2 + \beta \cdot \mathcal{D}_{\text{KL}}^{\text{cls}}\left(q_i \,\|\, p_{y_i}\right) \right], \tag{16}$$

$$\mathcal{D}_{\text{KL}}^{\text{cls}}(q_i \| p_{y_i}) = \tfrac{1}{2} \textstyle\sum_j \left[\sigma_{i,j}^2 + (\mu_{i,j} - \mu_{y_i,j})^2 - \log \sigma_{i,j}^2 - 1 \right], \tag{17}$$

where $q_i = \mathcal{N}(\boldsymbol{\mu}_i, \text{diag}(\boldsymbol{\sigma}_i^2))$ is the approximate posterior, $p_{y_i} = \mathcal{N}(\boldsymbol{\mu}_{y_i}, \mathbf{I})$ is the class-specific Gaussian prior with learnable mean $\boldsymbol{\mu}_{y_i}$ based on the label $y_i \in \{0, 1\}$. In all experiments, we set $\beta = 0.5$.

2.3 Cancer Likelihood Discriminator

After the modality-aware fusion by the VRBCA network, the cancer likelihood discriminator estimates the probability of each spot being cancerous. Specifically, the mean $\boldsymbol{\mu}_i$ and log-variance $\log \boldsymbol{\sigma}^2$ produced by the trained encoder f_{enc} of VRBCA, are concatenated and fed into a two-layer MLP classifier to predict cancer likelihood. The loss function for the discriminator combines BCE loss with the fused loss:

$$\mathcal{L}_{\text{BCE}} = -\tfrac{1}{n} \textstyle\sum_{i=1}^{n} \left[y_i \cdot \log \sigma(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \sigma(\hat{y}_i)) \right], \tag{18}$$

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{BCE}} + \gamma \times \mathcal{L}_{\text{fused}}, \tag{19}$$

where $\hat{y}_i$ denotes the predicted score by the classifier indicating how likely the spot is cancerous. The hyperparameters γ is fixed to 0.1 in the experiments. At inference time, benefiting from the regularized latent representations, our model effectively separates cancerous and normal spots in the predicted score space. To automatically determine a classification threshold, we fit a two-component Gaussian Mixture Model (GMM) to the distribution of predicted scores and use the intersection point between the two Gaussian components as the decision threshold, as detailed in Supplementary Material S3.

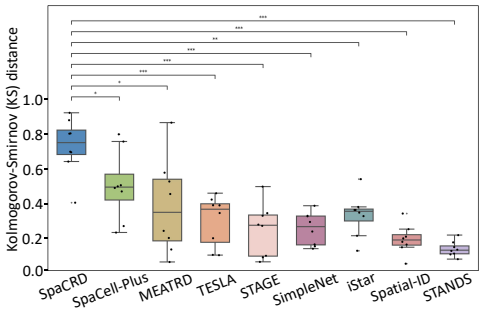

Figure 4: Comparison of the KS distances between predicted cancer likelihood distributions in healthy and tumor regions.

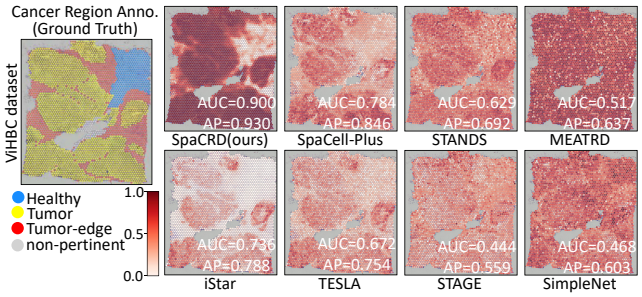

Figure 5: Visualization of cancer likelihood scores predicted by SpaCRD and other baselines on the ViHBC dataset.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate SpaCRD on five datasets comprising a total of 23 matched histology-ST datasets (termed tissue sections). Among them, two multi-section datasets-STHBC (Andersson et al. 2021) and CRC (Valdeolivas et al. 2024)-are used for cross-sample evaluation, while the remaining three—10XHBC, XeHBC (Janesick et al. 2023), and IDC-are used for cross-platforms&batches evaluation (see Supplementary Material S4 for detailed dataset description and preprocessing of the 23 tissue sections).

Implementation Details. We conducted all experiments using a single NVIDIA RTX 3090 GPU (24GB), with the development environment based on PyTorch 2.1.1 and Python 3.11.5. Detailed network architectures, training schedules, and implementation settings are provided in Supplementary Material S5.

Baselines and Evaluation Metrics. To evaluate the performance of SpaCRD, we compared it with eight SOTA methods, including five multimodal-based: MEATRD (Xu et al. 2025), STANDS (Xu et al. 2024), SpaCell (Tan et al. 2020), iStar (Zhang et al. 2024), TESLA (Hu et al. 2023); two ST-based: STAGE (Li et al. 2024), Spatial-ID (Shen et al. 2022); and one image-based: SimpleNet (Liu et al. 2023). For evaluation metrics, AUC, AP, F1-score, and KS distance are used to assess the performance of all methods. Supplementary Materials S6 and S7 provide the detailed descriptions for baselines and evaluation metrics.

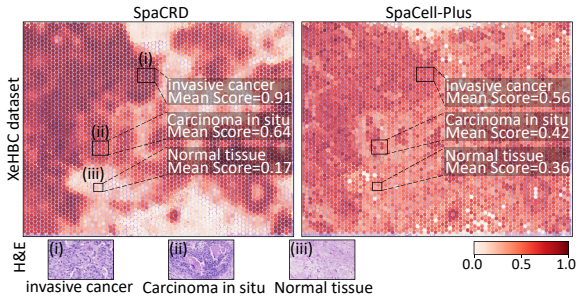

Figure 6: Cancer likelihood scores predicted by SpaCRD and the best-performing baseline method, SpaCell-Plus on the XeHBC dataset.

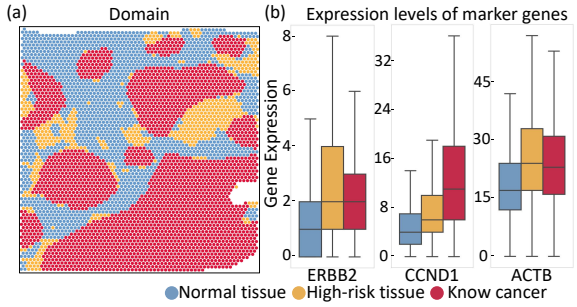

Figure 7: In the IDC dataset, SpaCRD-predicted high-score spots labeled non-cancerous exhibit elevated breast cancer marker expression, suggesting potential early lesions.

3.2 Cross-Samples Cancer Region Detection

To evaluate the cross-samples CTR detection capability of SpaCRD, we conducted leave-one-out cross-validation on the twelve colorectal cancer (CRC_[A1-G2]) and eight human breast cancer (STHBC_[A-H]) datasets. As shown in Table 1 and 2 (Cross-samples), SpaCRD achieves the best performance in AUC, AP, and F1-score, surpassing all baselines. Across 20 datasets covering both breast cancer and colorectal cancer, SpaCRD consistently outperforms the second-best method, achieving an average improvement of 13.5%, 14.1%, and 14.0% in AUC, AP, and F1-score, respectively. Methods that rely on prior knowledge, such as iStar and TESLA, perform suboptimally, possibly due to limitations in generalizing predefined features across datasets. Similarly, methods like MEATRD and STANDS, which adopt training strategies inspired by conventional anomaly detection in computer vision, struggle to capture the continuous and structured nature of cancer regions. SpaCell-Plus lacks effective integration of histology and ST data, resulting in inferior performance compared with our method. We further demonstrate the superiority of SpaCRD over the baselines by visualizing the CTR detection results on the STHBC datasets (Figure 3 and Supplementary Material S8.1). Finally, we evaluated the distributional divergence between predicted cancer likelihood scores in healthy and tumor regions across the eight STHBC datasets using the Kolmogorov–Smirnov (KS) distance. As shown in Figure 4, SpaCRD exhibited the strongest separation between healthy and tumor regions across most STHBC datasets (Median: SpaCRD=0.754, SpaCell-Plus=0.494, MEATRD=0.348).

3.3 Cross-Platforms&batches Cancer Region Detection

We further evaluated the performance of SpaCRD on the challenging task of cross-platforms&batches CTR detection. We trained SpaCRD on the STHBC dataset generated from the ST platform and tested it on datasets from other platforms, including ViHBC (Visium), IDC (Visium) and XeHBC (Xenium) datasets. As shown in Table 2 (Cross-platforms&batches), SpaCRD achieves superior performance over all baselines across all test datasets, with average improvements of 12.1%, 11.8%, and 13.8% in AUC, AP, and F1-score, respectively, compared to the second-best method. In addition, we comparative analyzed the predicted cancer likelihood scores of SpaCRD and other baseline methods on the ViHBC dataset. As shown in Figure 5, SpaCRD accurately distinguishes cancer regions from healthy tissues. Moreover, compared to baselines, it effectively separates the tumor-edge in certain regions, with predicted scores falling between those of the cancer and healthy spots. SpaCell-Plus and STANDS also identified ATR regions; however, the separation between cancer and healthy regions was less distinct compared to SpaCRD. iStar and TESLA were able to predict parts of the cancer regions; however, their predictions were incomplete and failed to capture the full extent of the lesions. Distribution analysis using histograms and violin plots further confirms SpaCRD's ability to distinguish between biologically distinct tissue regions (see Supplementary Material S8.2). Consistent with the findings on the ViHBC dataset, visualization analysis of the XeHBC and IDC datasets shows that the cancer likelihood scores predicted by SpaCRD are highly consistent with ground truth (Supplementary Material S8.1). These results suggest SpaCRD mitigates technical and batch effects by integrating image and ST data into a unified latent space, enabling robust, generalizable CTR detection across diverse platforms and batches.

3.4 Downstream Analysis: Detection of Potential Lesion Regions

We systematically evaluated SpaCRD's ability to stratify cancer severity by analyzing its predicted likelihood scores. On the XeHBC dataset generated from Xenium platform, SpaCRD achieved average scores of 0.91 (invasive cancer), 0.64 (carcinoma in situ), and 0.17 (normal tissue), effectively distinguishing malignancy levels (Figure 6). Heatmap visualization confirmed clear stratification, with invasive regions showing the highest intensity, followed by carcinoma in situ and normal tissue. In contrast, SpaCell-Plus produced less discriminative scores (0.56, 0.42, and 0.36, respectively), failing to separate carcinoma in situ from normal tissue. Other baselines also exhibited dispersed, inconsistent score distributions (Supplementary Material S8.1). Notably, these distinctions are often imperceptible in histology images, suggesting SpaCRD's scores may reflect tumor aggressiveness. Finally, we examined the spots in the IDC

Extractor	Cross-samples						Cross-platforms&batches								
	STHBC			CRC			10XHBC			IDC			XeHBC		
	AUC	AP	F1	AUC	AP	F1	AUC	AP	F1	AUC	AP	F1	AUC	AP	F1
w/ Swin-Tiny	0.880	0.923	0.838	0.820	0.795	0.793	0.689	0.751	0.728	0.501	0.601	0.601	0.718	0.465	0.501
w/ ResNet50	0.878	0.908	0.836	0.830	0.800	0.794	0.637	0.697	0.701	0.596	0.649	0.657	0.656	0.698	0.691
w/ HIPT	0.885	0.912	0.827	0.822	0.796	0.793	0.784	0.829	0.786	0.399	0.525	0.546	0.771	0.736	0.742
w/ UNI (ours)	**0.929**	**0.946**	**0.880**	**0.869**	**0.854**	**0.810**	**0.900**	**0.930**	**0.867**	**0.891**	**0.914**	**0.854**	**0.931**	**0.859**	**0.795**

Table 3: Ablation study of histology feature extractors conducted across all datasets used in this study.

Model	HBC(eleven datasets)			CRC(twelve datasets)		
	AUC	AP	F1	AUC	AP	F1
Image-based	0.789	0.752	0.733	0.606	0.557	0.553
ST-based	0.832	0.815	0.747	0.782	0.793	0.755
w/o BCA	0.849	0.833	0.788	0.797	0.774	0.759
w/o RVAE	0.887	0.898	0.817	0.831	0.816	0.795
w/o VRBCA	0.815	0.796	0.734	0.771	0.746	0.717
w/o CL	0.892	0.886	0.828	0.824	0.807	0.776
Ours	**0.923**	**0.934**	**0.869**	**0.869**	**0.854**	**0.810**

Table 4: Ablation study of modalities and fusion modules conducted across all datasets used in this study.

Metric	Parameter α			Parameter β		
	0.0	0.5	1.0	0.1	0.5	1.0
AUC	0.908	**0.923**	0.916	0.863	**0.923**	0.920
AP	0.912	**0.934**	0.929	0.895	**0.934**	0.931
F1	0.843	**0.869**	0.863	0.828	**0.869**	0.867
Metric	Parameter γ			neighboring spots		
	0.0	0.1	1.0	4	6	10
AUC	0.772	**0.923**	0.919	0.908	**0.923**	0.912
AP	0.768	**0.934**	0.931	0.896	**0.934**	0.925
F1	0.794	**0.869**	0.856	0.844	**0.869**	0.858

Table 5: Sensitivity analysis of key hyperparameter across eleven breast cancer datasets. Gray : default settings.

dataset that received high predicted scores from SpaCRD but were annotated as non-cancerous in the manual labels (Figure 7(a), orange spots). Remarkably, these spots exhibit significantly elevated expression of canonical breast cancer marker genes (e.g., ERBB2 (Harari et al. 2000), CCND1 (Valla et al. 2022), and ACTB (Majidzadeh-A et al. 2011), Figure 7(b)), compared to normal tissue. This suggests that SpaCRD captures spatial regions warranting clinical investigation and potentially of pathological relevance.

3.5 Ablation Studies

To assess the contributions of key components in SpaCRD, we conducted ablation studies on all datasets, focusing on the Histology feature extractor, unimodal inputs, and module ablations. All reported results are averaged over five independent runs. **Since some experiments involve aggregating scores across multiple datasets, we report the overall mean values without standard deviations for consistency.**

Impact of Extractors. We replaced UNI with ResNet50 (He et al. 2016) and Swin-Tiny (Liu et al. 2021) pretrained on natural images, and HIPT (Chen et al. 2022) pretrained on pathological images. As shown in Table 3, UNI yields the best performance. Notably, the performance of other extractors drops markedly in the cross-platforms&batches evaluation, likely due to large variations in histology images that hinder their ability to capture core fine-grained features.

Impact of Modalities and Fusion Modules. We assessed the contributions of different input modalities and the fusion module by (i) using only histology or ST data, and (ii) removing key fusion components, including the multimodal deep fusion module (BCA), the regularized reconstruction-guided denoising module (RVAE), and the modality-alignment representation learning (CL). As shown in Table 4, removing any modality input or module results in suboptimal performance.

3.6 Robustness and Efficiency Analysis

- **Sensitivity Analysis**: We conducted sensitivity analyses on eleven breast cancer datasets, covering key hyperparameter α, β, and γ, as well as the number of neighboring spots selected in the BCA model (Table 5). Detailed analyses are provided in Supplementary Material S8.3.
- **Empirical Efficiency Analysis**: We analyzed the cost of SpaCRD and baselines by evaluating the number of parameters, runtime, and memory usage. The detailed results in Supplementary Material S8.4 indicate that SpaCRD maintains an acceptable computational cost.
- **Small-Sample Training Analysis**: Additional experiments show that even when training set contains less than 10% of the spots in test set, SpaCRD maintains stable performance in cross-platform CTR detection, demonstrating strong generalization and data efficiency (see Supplementary Material S8.5 for detailed results).

4 Conclusion

In this study, we present SpaCRD, a multimodal deep fusion and transfer learning-based framework that integrates histology images and ST data for accurate and generalizable CTR detection across samples and platforms&batches. SpaCRD leverages the proposed VRBCA fusion module, synergistically optimized through contrastive learning objectives, to dynamically integrate histology images and gene expression features in both image-to-gene and gene-to-image directions. By jointly attending to both gene-to-image and image-to-gene signals and modeling interactions among neighboring spots, VRBCA ensures comprehensive feature integration, while its variational reconstruction mechanism filters out noise and promotes compact and class-consistent embeddings. Extensive evaluations across diverse datasets from multiple platforms demonstrate the strong generalization capability of SpaCRD. In addition, SpaCRD may serve as a promising tool for medical researchers by enabling the stratification of cancer severity and revealing spatial regions that may warrant further clinical investigation. Supplementary Materials can be found at the code link.

Acknowledgments

The work was supported in part by the Program of Yunnan Key Laboratory of Intelligent Systems and Computing (No. 202405AV340009) and the National Natural Science Foundation of China (Nos. 62262069, 62571555).

References

Andersson, A.; Larsson, L.; Stenbeck, L.; Salmén, F.; Ehinger, A.; Wu, S. Z.; Al-Eryani, G.; Roden, D.; Swarbrick, A.; Borg, Å.; et al. 2021. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature communications*, 12(1): 6012.

Benjamin, K.; Bhandari, A.; Kepple, J. D.; Qi, R.; Shang, Z.; Xing, Y.; An, Y.; Zhang, N.; Hou, Y.; Crockford, T. L.; et al. 2024. Multiscale topology classifies cells in subcellular spatial transcriptomics. *Nature*, 630(8018): 943–949.

Chen, R. J.; Chen, C.; Li, Y.; Chen, T. Y.; Trister, A. D.; Krishnan, R. G.; and Mahmood, F. 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16144–16155.

Chen, R. J.; Ding, T.; Lu, M. Y.; Williamson, D. F.; Jaume, G.; Song, A. H.; Chen, B.; Zhang, A.; Shao, D.; Shaban, M.; et al. 2024. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3): 850–862.

Ferri-Borgogno, S.; Zhu, Y.; Sheng, J.; Burks, J. K.; Gomez, J. A.; Wong, K. K.; Wong, S. T.; and Mok, S. C. 2023. Spatial transcriptomics depict ligand–receptor cross-talk heterogeneity at the tumor-stroma interface in long-term ovarian cancer survivors. *Cancer research*, 83(9): 1503–1516.

Hao, Y.; Hao, S.; Andersen-Nissen, E.; Mauck, W. M.; Zheng, S.; Butler, A.; Lee, M. J.; Wilk, A. J.; Darby, C.; Zager, M.; et al. 2021. Integrated analysis of multimodal single-cell data. *Cell*, 184(13): 3573–3587.

Harari; et al. 2000. Molecular mechanisms underlying ErbB2/HER2 action in breast cancer. *Oncogene*, 19(53): 6102–6114.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, J.; Coleman, K.; Zhang, D.; Lee, E. B.; Kadara, H.; Wang, L.; and Li, M. 2023. Deciphering tumor ecosystems at super resolution from spatial transcriptomics with TESLA. *Cell Systems*, 14(5): 404–417.

Huang, T.; Liu, T.; Babadi, M.; Jin, W.; and Ying, R. 2025. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. In *International Conference on Machine Learning*, 1–16. PmLR.

Janesick, A.; Shelansky, R.; Gottscho, A. D.; Wagner, F.; Williams, S. R.; Rouault, M.; Beliakoff, G.; Morrison, C. A.; Oliveira, M. F.; Sicherman, J. T.; et al. 2023. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature communications*, 14(1): 8353.

Janssen, P.; Kliesmete, Z.; Vieth, B.; Adiconis, X.; Simmons, S.; Marshall, J.; McCabe, C.; Heyn, H.; Levin, J. Z.; Enard, W.; et al. 2023. The effect of background noise and its removal on the analysis of single-cell expression data. *Genome biology*, 24(1): 140.

Jaume, G.; Oldenburg, L.; Vaidya, A.; Chen, R. J.; Williamson, D. F.; Peeters, T.; Song, A. H.; and Mahmood, F. 2024a. Transcriptomics-guided slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9632–9644.

Jaume, G.; Vaidya, A.; Chen, R. J.; Williamson, D. F.; Liang, P. P.; and Mahmood, F. 2024b. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11579–11590.

Khalighi, S.; Reddy, K.; Midya, A.; Pandav, K. B.; Madabhushi, A.; and Abedalthagafi, M. 2024. Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. *NPJ precision oncology*, 8(1): 80.

Lawrence, R.; Watters, M.; Davies, C. R.; Pantel, K.; and Lu, Y.-J. 2023. Circulating tumour cells for early detection of clinically relevant cancer. *Nature Reviews Clinical Oncology*, 20(7): 487–500.

Li, S.; Gai, K.; Dong, K.; Zhang, Y.; and Zhang, S. 2024. High-density generation of spatial transcriptomics with STAGE. *Nucleic Acids Research*, 52(9): 4843–4856.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z.; Zhou, Y.; Xu, Y.; and Wang, Z. 2023. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20402–20411.

Maan, H.; Ji, Z.; Sicheri, E.; Tan, T. J.; Selega, A.; Gonzalez, R.; Krishnan, R.; WANG, B.; and Campbell, K. R. 2025. Multi-modal disentanglement of spatial transcriptomics and histopathology imaging. In *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2025*, 1–35. PmLR.

Majidzadeh-A; et al. 2011. TFRC and ACTB as the best reference genes to quantify Urokinase Plasminogen Activator in breast cancer. *BMC research notes*, 4(1): 215.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rao, A.; Barkley, D.; França, G. S.; and Yanai, I. 2021. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871): 211–220.

Seferbekova, Z.; Lomakin, A.; Yates, L. R.; and Gerstung, M. 2023. Spatial biology of cancer evolution. *Nature Reviews Genetics*, 24(5): 295–313.

Shen, R.; Liu, L.; Wu, Z.; Zhang, Y.; Yuan, Z.; Guo, J.; Yang, F.; Zhang, C.; Chen, B.; Feng, W.; et al. 2022. Spatial-ID: a cell typing method for spatially resolved transcriptomics via transfer learning and spatial embedding. *Nature communications*, 13(1): 7640.

Shmatko, A.; Ghaffari Laleh, N.; Gerstung, M.; and Kather, J. N. 2022. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nature cancer*, 3(9): 1026–1038.

Tan, X.; Su, A.; Tran, M.; and Nguyen, Q. 2020. SpaCell: integrating tissue morphology and spatial gene expression to predict disease cells. *Bioinformatics*, 36(7): 2293–2294.

Tian, L.; Chen, F.; et al. 2023. The expanding vistas of spatial transcriptomics. *Nature Biotechnology*, 41(6): 773–782.

Valdeolivas, A.; Amberg, B.; Giroud, N.; Richardson, M.; Gálvez, E. J.; Badillo, S.; Julien-Laferrière, A.; Túrós, D.; Voith von Voithenberg, L.; Wells, I.; et al. 2024. Profiling the heterogeneity of colorectal cancer consensus molecular subtypes using spatial transcriptomics. *NPJ precision oncology*, 8(1): 10.

Valla, M.; Klæstad, E.; Ytterhus, B.; and Bofin, A. M. 2022. CCND1 amplification in breast cancer-associations with proliferation, histopathological grade, molecular subtype and prognosis. *Journal of mammary gland biology and neoplasia*, 27(1): 67–77.

Xu, K.; Lu, Y.; Hou, S.; Liu, K.; Du, Y.; Huang, M.; Feng, H.; Wu, H.; and Sun, X. 2024. Detecting anomalous anatomic regions in spatial transcriptomics with STANDS. *Nature Communications*, 15(1): 8223.

Xu, K.; Wu, Q.; Lu, Y.; Zheng, Y.; Li, W.; Tang, X.; Wang, J.; and Sun, X. 2025. Meatrd: Multimodal anomalous tissue region detection enhanced with spatial transcriptomics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 12918–12926.

Xue, S.; Zhu, F.; Chen, J.; and Min, W. 2025. Inferring single-cell resolution spatial gene expression via fusing spot-based spatial transcriptomics, location, and histology using GCN. *Briefings in Bioinformatics*, 26(1): bbae630.

Yan, C.; Zhang, Y.; Feng, J.; et al. 2025. Triple-effect correction for Cell Painting data with contrastive and domain-adversarial learning. *Nature Communications*, 16(1): 6886.

Zahedi, R.; Ghamsari, R.; Argha, A.; Macphillamy, C.; Beheshti, A.; Alizadehsani, R.; Lovell, N. H.; Lotfollahi, M.; and Alinejad-Rokny, H. 2024. Deep learning in spatially resolved transcriptomics: a comprehensive technical view. *Briefings in Bioinformatics*, 25(2): bbae082.

Zhang, D.; Schroeder, A.; Yan, H.; Yang, H.; Hu, J.; Lee, M. Y.; Cho, K. S.; Susztak, K.; Xu, G. X.; Feldman, M. D.; et al. 2024. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature Biotechnology*, 42(9): 1372–1377.

Zingman, I.; Stierstorfer, B.; Lempp, C.; and Heinemann, F. 2024. Learning image representations for anomaly detection: application to discovery of histological alterations in drug development. *Medical Image Analysis*, 92: 103067.