A Foundational Generative Model for Cross-platform Unified Enhancement of Spatial Transcriptomics

Xiaofei Wang[1*]; Hanyu Liu[2*]; Ningfeng Que[3*]; Chenyang Tao[1]; Yu Jiang[4]; Yixuan Jiang[5]; Pinan Zhu[5]; Junze Zhu[6]; Xiaoyang Li[5]; Stephen Price[1]; Jianguo Xu[4]; Jianzhong Xi[3†]; Xinjie Wang[7†]; Chao Li[1,2,8,9†]

1. Department of Clinical Neurosciences, University of Cambridge, UK

2. School of Science and Engineering, University of Dundee, UK

3. State Key Laboratory of Natural and Biomimetic Drugs, Department of Biomedical Engineering, College of Future Technology, Peking University, Beijing, China

4. Department of Neurosurgery, West China Hospital, Sichuan University, Chengdu, China

5. College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China

6. School of Computer Science and Technology, Tongji University, Shanghai, China

7. Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Science, Shenzhen, China

8. Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK

9. School of Medicine, University of Dundee, UK

*Equal contributions

†Correspondence: Chao Li (cl647@cam.ac.uk); Xinjie Wang (wangxinjie@caas.cn); Jiangzhong Xi (jzxi@pku.edu.cn)

Abstract

Spatial transcriptomics (ST) enables *in situ* mRNA profiling but remains limited by spatial resolution, sensitivity, histological alignment, and mis-profiling in complex tissues. Most enhancement methods target a single challenge using an auxiliary modality, e.g., super-resolution using hematoxylin and eosin (H&E) images and sensitivity enhancement with single-cell RNA-seq (scRNA-seq). However, most ignore integration across modalities and interdependence across challenges, yielding biologically inconsistent reconstructions. Here we introduce FOCUS, a foundational generative model for cross-platform unified ST enhancement, conditioned on H&E images, scRNA-seq references, and spatial co-expression priors. FOCUS uses a modular design for *multimodal* integration, and a cross-challenge coordination strategy to target co-occurring defects, enabling joint challenge optimization. FOCUS was trained and benchmarked on >1.7 million H&E-ST pairs and >5.8 million single-cell profiles, demonstrating state-of-the-art performance on both isolated and coupled challenges across ten platforms. We utilized FOCUS in elucidating the niche characterization in papillary craniopharyngioma and uncovering spatial heterogeneity in primary and metastatic head and neck squamous cell carcinoma.

Introduction

Spatial transcriptomics (ST) enables *in situ* profiling across tissues; however, recent benchmark[1–3] reveal persistent coupled limitations across platforms, including spatial resolution, profiling sensitivity, molecular-histology alignment, and the balance of profiling across tissue complexity, which hinder its biological utility and interpretability.

First, there is an inherent trade-off between spatial resolution and sequencing depth. Recent platforms, e.g., VisiumHD, offer subcellular resolution but at the cost of per-spot transcripts, limiting profiling accuracy. **Second**, due to limited tissue RNA, inefficient capture, and throughput bottlenecks[3,4], existing platforms remain low in sensitivity. Despite experimental strategies such as increasing barcoding density[5], profiling bias persists. **Third**, spatial fidelity is often compromised by artifacts due to lateral mRNA diffusion during tissue permeabilization in sequencing-based ST and repeated rounds of hybridization in imaging-based ST[2,3,6,7]. The resulting spatial distortions complicate morphology-guided tasks. **Finally**, in complex tissues such as cancer, cellular packing causes blended transcripts from neighboring cells, even when there is perfect spatial alignment[8–10]. This signal mixing reduces profiling precision for rare cells or subtle niches.

Emerging studies have sought to address these limitations by including other modalities to enhance ST. One strategy leverages paired hematoxylin and eosin (H&E)-stained images to enhance ST resolution[11–13] or spatially align histology with molecular signals[14–16]. Other strategies involve single-cell RNA sequencing (scRNA-seq) to enhance profiling sensitivity[4,17,18]. However, each modality carries intrinsic limitations: scRNA-seq is often profiled on different samples from ST and may not reflect the cell composition or tissue architecture of ST; whereas H&E images, though structurally detailed, lack the molecular resolution to distinguish transcriptionally distinct but morphologically similar cells. Integrating complementary modalities, therefore, represents a promising yet under-explored opportunity to enhance ST. This integration is particularly important since major ST challenges are often **coupled**. For example, low resolution may blur tissue boundaries and cause poor alignment, while high cellular complexity can restrict the capture of per-

cell molecules, leading to low profiling sensitivity. Without unified optimization across challenges, existing methods risk yielding biologically inconsistent reconstructions.

Generative models, especially recent diffusion models[19,20], promise to jointly tackle multiple ST challenges via incorporating complementary modalities as generation conditions. Hence, we propose FOCUS (FOundational generative model for Cross-platform Unified enhancement of Spatial transcriptomics), a diffusion-based framework that reconstructs high-quality ST by jointly optimizing multiple tasks. Built on the foundational pre-trained encoders[21,22], FOCUS integrates ST with paired H&E images, cell masks, scRNA-seq references, and spatial co-expression priors to simultaneously enhance resolution, sensitivity, ST-H&E alignment, and precision in complex tissues. Through challenge-tailored modular design and coordinated cross-challenge interaction, it transforms fragmented, challenge-specific enhancements into unified ST reconstruction.

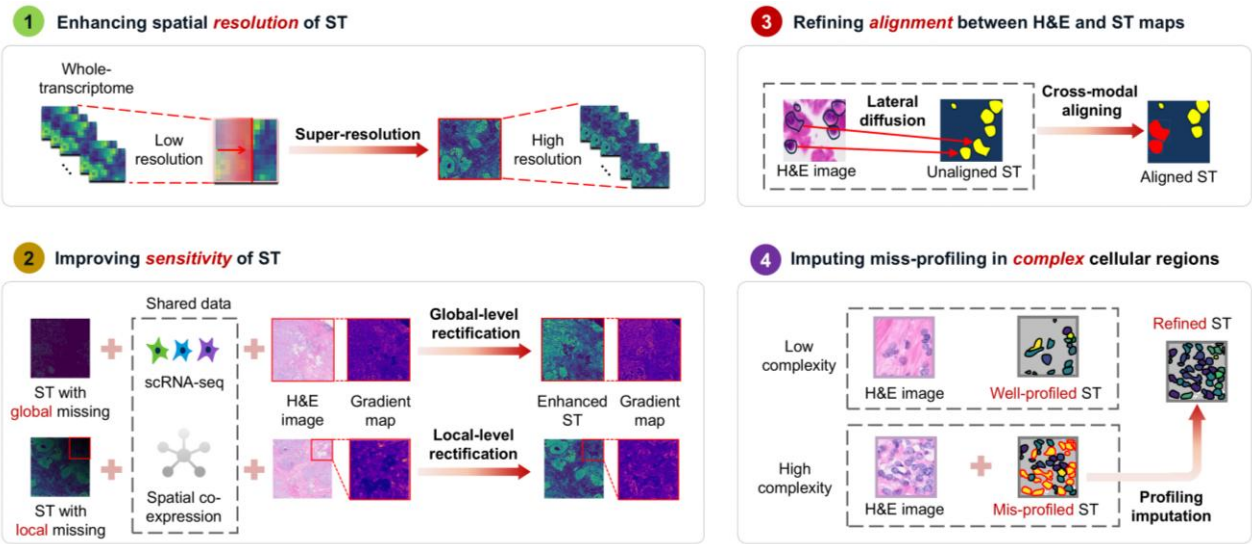
FOCUS is benchmarked on ten ST platforms, with over 1.7 million paired H&E-ST patches and 5.8 million scRNA-seq cell profiles. FOCUS outperforms both native platforms and state-of-the-art methods, consistently demonstrating improvement across various challenges and downstream tasks, such as spatial domain characterization and cell-cell communications. Finally, we utilized FOCUS on two real-world scenarios, thereby validating its strong generalization and translational potential. FOCUS is open source and publicly available at https://github.com/LHY1007/FOCUS/.

Results

2.1 Study overview

In this study, we proposed a foundational generative model, FOCUS, to jointly enhance ST spatial resolution, sensitivity, alignment with histological contexts, and profiling in complex regions in a single, unified framework (Fig. 1a, Extended Figs. 1-2). Built on a diffusion-based backbone, FOCUS adopts challenge-tailored modules and coordinated cross-challenge interactions to achieve unified optimization across all interdependent challenges.

FOCUS was trained and validated across eight sequencing-based and two imaging-based platforms, on both human (*Homo sapiens*) and mouse (*Mus musculus*) specimens (Fig. 1b). Comprehensive validation was performed for each challenge (Fig. 1c). Furthermore, to assess real-world utility, we fine-tuned it on unseen tissue context or platform: **(i)** a papillary craniopharyngioma (PCP) dataset, a rare heterogeneous suprasellar tumor, to localize disease-relevant transcriptional patterns, and **(ii)** primary and metastatic human head and neck squamous cell carcinoma (HNSCC) dataset generated from Open-ST[23], a recent platform profiling consecutive tissue sections as a novel ST technology, evaluating cross-platform generalization.



Extended Fig. 1 | Schematic illustration of challenges in current ST platforms that are amenable to computational mitigation: (1) coarse spatial resolution, (2) limited expression sensitivity, (3) misalignment with histological structure, and (4) mis-profiling in complex tissues

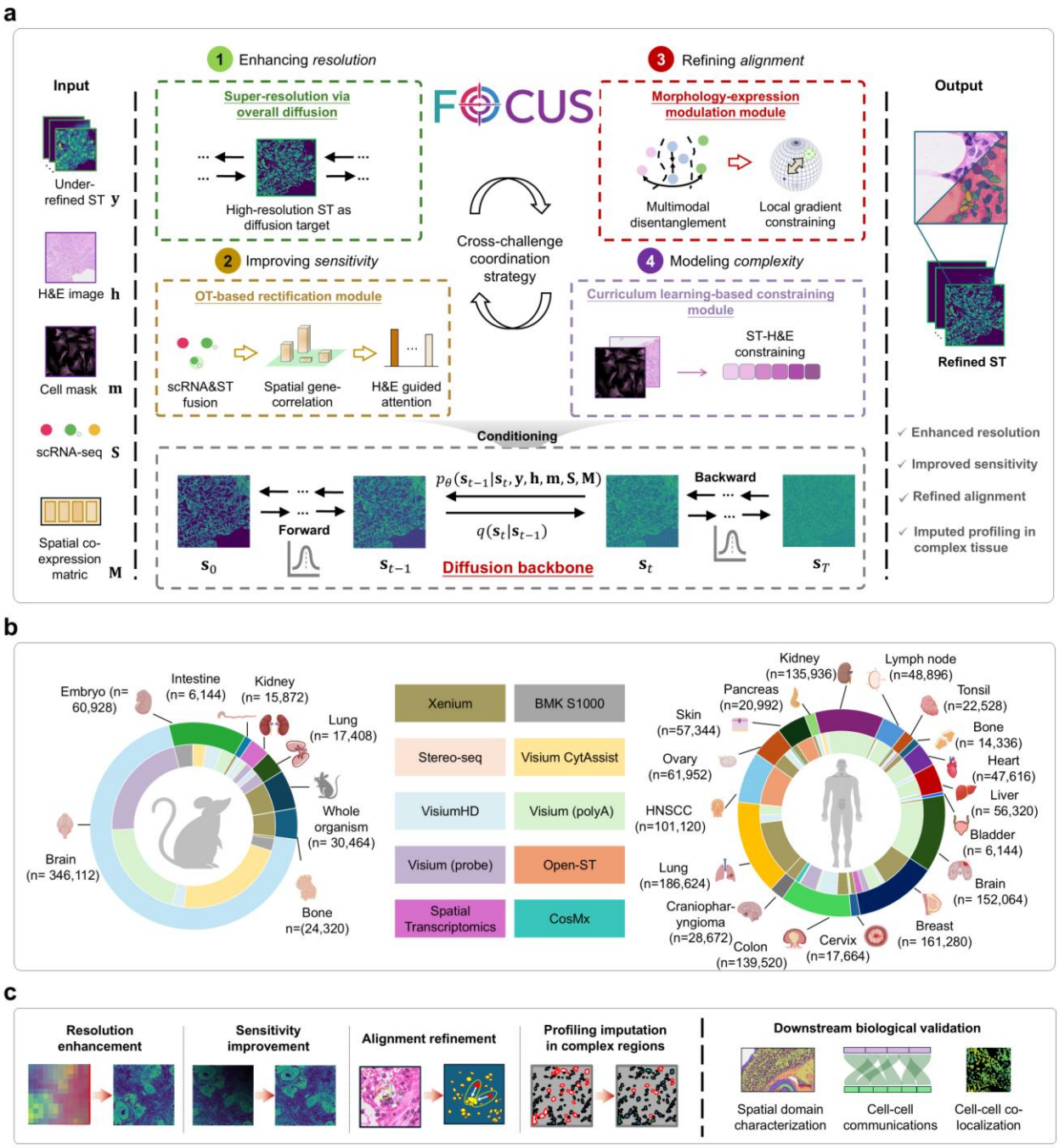
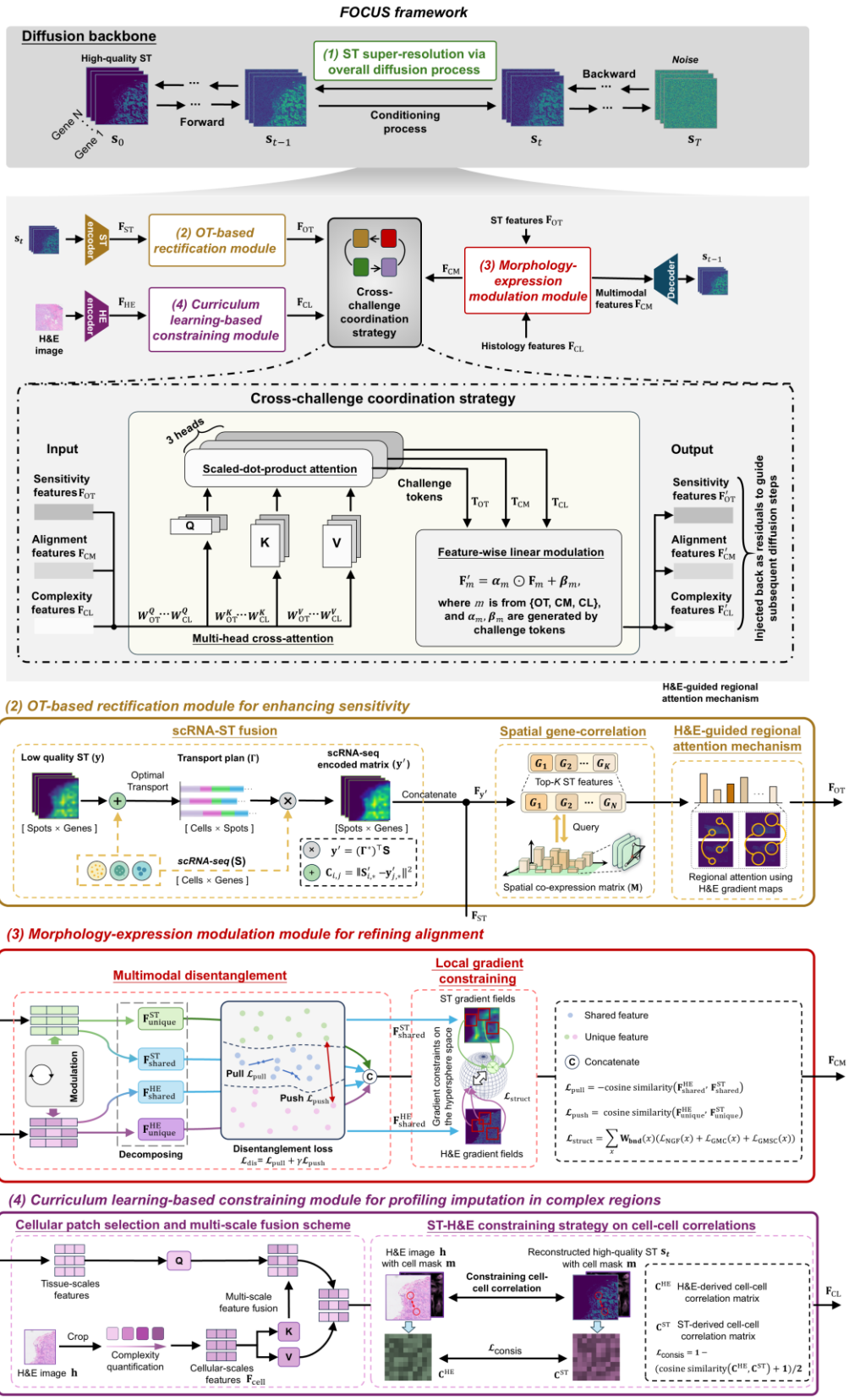


Fig. 1 | Model design, and validation for ST enhancement. a, Model overview. FOCUS is a diffusion-based generative model that leverages ST and H&E encoders[21,22] pretrained on large-scale, cross-tissue data to extract robust multimodal features. It integrates multimodal conditions as inputs, including under-refined ST maps, paired H&E images

with cell segmentation masks, scRNA-seq references, and spatial gene co-expression matrices. Each challenge is addressed through tailored modules, with a cross-challenge coordination strategy enabling module interaction for coherent improvement across challenges. **b**, Large-scale, cross-platform multimodal dataset. In total, we assemble 6,876 paired ST-H&E images (corresponding to over 1.7 million patches) with matched cell segmentation masks, referenced scRNA-seq from public resources (over 5.8 million scRNA-seq cell profiles; Table S1), and precomputed spatial gene co-expression matrices. The data collection spans ten ST platforms, including eight sequencing-based (Visium (probe-based and polyA-based), Visium Cytassist, VisiumHD, Spatial Transcriptomics, Stereo-seq, BMK S1000, and Open-ST) and two imaging-based (Xenium and CosMx) platforms, and two species (human and mouse), comprising 17 normal and 17 cancer tissues, with whole-transcriptome profiles available for both species. **c,** Benchmarking and validation across challenges and downstream tasks, including spatial domain characterization, cell-cell communications, and cell-cell co-localization.


FOCUS framework
Diffusion backbone
High-quality ST
Gene N
Gene 1
Forward
Backward
Noise
(1) ST super-resolution via overall diffusion process
Conditioning process
(2) OT-based rectification module
(3) Morphology-expression modulation module
(4) Curriculum learning-based constraining module
Cross-challenge coordination strategy
ST encoder
HE encoder
H&E image
ST features F_{OT}
Histology features F_{CL}
Multimodal features F_{CM}
Decoder
Cross-challenge coordination strategy
Input
Sensitivity features F_{OT}
Alignment features F_{CM}
Complexity features F_{CL}
3 heads
Scaled-dot-product attention
Challenge tokens
Multi-head cross-attention
Feature-wise linear modulation
$F'_m = \alpha_m \odot F_m + \beta_m$,
where m is from {OT, CM, CL}, and α_m, β_m are generated by challenge tokens
Output
Injected back as residuals to guide subsequent diffusion steps
H&E-guided regional attention mechanism
(2) OT-based rectification module for enhancing sensitivity
scRNA-ST fusion
Low quality ST (y)
Transport plan (Γ)
scRNA-seq encoded matrix (y')
Optimal Transport
[Spots × Genes]
[Cells × Genes]
[Spots × Genes]
scRNA-seq (S)
Concatenate
Spatial gene-correlation
Top-K ST features
Query
Spatial co-expression matrix (M)
H&E-guided regional attention mechanism
Regional attention using H&E gradient maps
(3) Morphology-expression modulation module for refining alignment
Multimodal disentanglement
Modulation
Decomposing
Pull $\mathcal{L}_{pull}$
Push $\mathcal{L}_{push}$
Disentanglement loss
$\mathcal{L}_{dis} = \mathcal{L}_{pull} + \gamma \mathcal{L}_{push}$
Local gradient constraining
ST gradient fields
H&E gradient fields
Gradient constraints on the hypersphere space
Shared feature
Unique feature
Concatenate
$\mathcal{L}_{pull} = -\text{cosine similarity}(F^{HE}_{shared}, F^{ST}_{shared})$
$\mathcal{L}_{push} = \text{cosine similarity}(F^{HE}_{unique}, F^{ST}_{unique})$
$\mathcal{L}_{struct} = \sum_x W_{bnd}(x)(\mathcal{L}_{NGF}(x) + \mathcal{L}_{GMC}(x) + \mathcal{L}_{GMSC}(x))$
(4) Curriculum learning-based constraining module for profiling imputation in complex regions
Cellular patch selection and multi-scale fusion scheme
Tissue-scales features
Multi-scale feature fusion
Crop
Complexity quantification
Cellular-scales features F_{cell}
ST-H&E constraining strategy on cell-cell correlations
H&E image h with cell mask m
Reconstructed high-quality ST S_t with cell mask m
Constraining cell-cell correlation
$\mathcal{L}_{consis}$
C^{HE} H&E-derived cell-cell correlation matrix
C^{ST} ST-derived cell-cell correlation matrix
$\mathcal{L}_{consis} = 1 - (\text{cosine similarity}(C^{HE}, C^{ST}) + 1)/2$

Extended Fig. 2 | Illustration of the FOCUS framework. FOCUS builds on a diffusion backbone. In the forward process, high-quality ST $\mathbf{s}_0$ is gradually noised to $\mathbf{s}_T$ in T steps, and the reverse process learns to denoise. At each reverse step t, FOCUS learns to predict $\mathbf{s}_{t-1}$ from $\mathbf{s}_t$ with a conditioning process. This process uses multimodal pretrained encoders[21,22] (encoding ST features $\mathbf{s}_t$ and H&E image $\mathbf{h}$ to $\mathbf{F}_{\mathrm{ST}}$ and $\mathbf{F}_{\mathrm{HE}}$, respectively), and comprises both a challenge-tailored design and a cross-challenge coordination strategy: **(1)** By taking high-resolution ST maps as the diffusion generation target, FOCUS is able to achieve ST resolution enhancement. **(2)** The optimal transport (OT)-based rectification module enhances expression sensitivity. First, the under-refined ST $\mathbf{y}$ is mapped with reference scRNA-seq $\mathbf{S}$ of the same tissue context by solving an OT-based cell-spot mapping problem. The resulting transport plan maps high-sensitivity single-cell expression onto the ST spots, producing the scRNA-informed ST feature $\mathbf{F}_{\mathbf{y}'}$. We then refine $\mathbf{F}_{\mathbf{y}'}$ using precomputed spatial co-expression matrices ($\mathbf{M}$) to reflect gene-gene spatial relationships. Finally, an H&E-guided regional attention mechanism, which upweights high-gradient tissue boundaries, yields the rectified ST feature $\mathbf{F}_{\mathrm{OT}}$. **(3)** Morphology-expression modulation module refines H&E-ST alignment. First, a multimodal disentanglement strategy splits the ST and H&E features $\mathbf{F}_{\mathrm{OT}}$ and $\mathbf{F}_{\mathrm{CL}}$ into shared and modality-specific components using a loss $\mathcal{L}_{\mathrm{dis}}$ that makes shared features similar and modality-specific features distinct. Local gradient constraining is then imposed on the shared structural features to correct residual geometric misalignment between H&E and ST. **(4)** Curriculum learning-based constraining module imputes profiling in complex regions. Complexity scores are calculated from H&E and used to select cellularly complex patches, whose tissue- and cell-scale features ($\mathbf{F}_{\mathrm{HE}}$ and $\mathbf{F}_{\mathrm{cell}}$) are fused to $\mathbf{F}_{\mathrm{CL}}$ in a multi-scale manner. Within selected patches, ST-derived cell-cell correlation matrices are constrained to match H&E-derived correlations via a cosine-similarity-based consistency loss $\mathcal{L}_{\mathrm{consis}}$, improving expression precision in cellular heterogeneous regions. **Cross-challenge coordination strategy**: To enable interaction and joint enhancement across challenges, features from each module ($\mathbf{F}_{\mathrm{OT}}, \mathbf{F}_{\mathrm{CL}}, \mathbf{F}_{\mathrm{CM}}$) are converted into compact tokens, interacting via multi-head cross-attention. The resulting tokens produce feature-wise scaling and shifting coefficients that recalibrate features across modules. The final coordinated features ($\mathbf{F}'_{\mathrm{OT}}, \mathbf{F}'_{\mathrm{CL}}, \mathbf{F}'_{\mathrm{CM}}$) are reintroduced into the conditioning process as residuals to guide subsequent diffusion steps.

2.2 Model, data, and benchmarking

2.2.1 Model design

Backbone and coordinated modular design

Generative models are designed to learn underlying distributions and can naturally incorporate multimodal conditions, enabling to tackle multiple ST challenges within a single framework. Compared to conventional models[24–26], diffusion models provide stable training on heterogeneous, multi-platform ST by gradually adding noise to the data and learning to reconstruct original signals in a reverse process[27]. With high-resolution ST as the generation target, diffusion models naturally support resolution enhancement. FOCUS further introduces challenge-tailored modules to integrate complementary conditions, including paired H&E images with cell segmentation masks, and scRNA-seq references with spatial gene co-expression priors (Extended Fig. 2). To ensure comprehensive representation across modalities, FOCUS is built on two large-scale pretrained encoders to capture morphological and transcriptional patterns: **1)** an H&E encoder[21] pretrained on 1.3 billion patches covering 31 tissues, and **2)** an ST encoder[22] pretrained on 30 million spatially profiled spots. A cross-challenge coordination strategy is embedded in the reverse diffusion process, enabling selective module combinations and providing flexibility.

Optimal transport (OT)-based rectification module

To boost sensitivity, we designed an OT-based cell-spot mapping and fused ST with scRNA-seq, thereby transferring high-sensitivity single-cell expression onto ST spots[28]. To further model gene relationships, we devise a spatial gene-correlation strategy based on precomputed co-expression matrices, motivated by our observation that spatial context stabilizes gene correlations across samples (Extended Fig. 3a-b; Supplementary section 1).

Finally, we introduced an H&E-guided regional attention mechanism that upweights areas with high H&E gradients, where significant local changes in H&E imaging indicate tissue structures, thereby further improving sensitivity in morphologically well-defined regions.

Morphology-expression modulation module

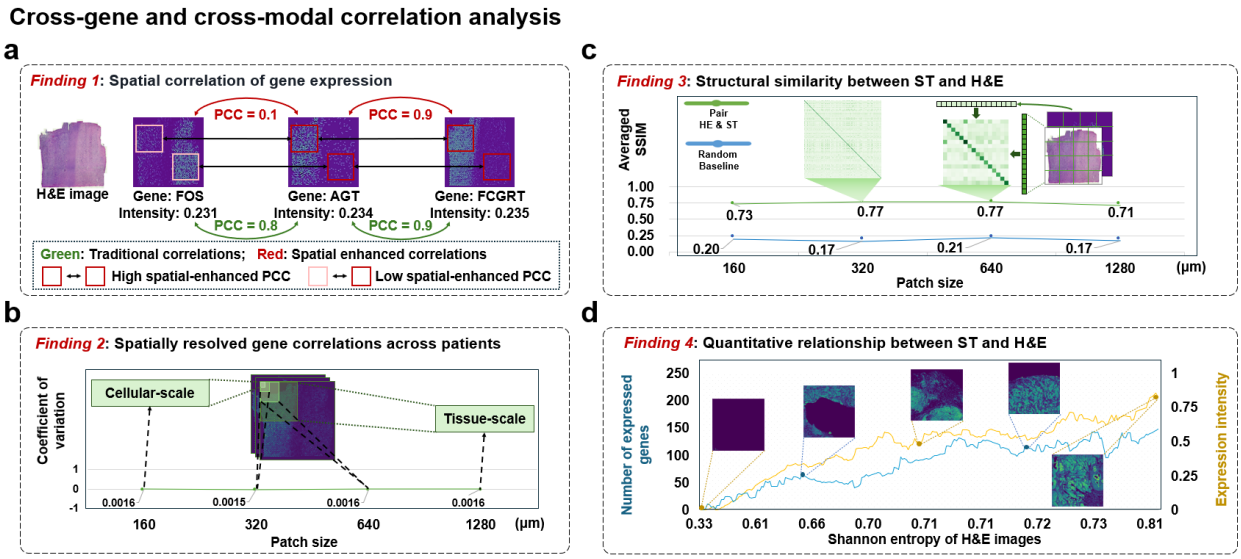
To reduce misalignment, we devised a multimodal disentangling strategy to separate modality-shared structural features from modality-specific patterns between H&E images and ST maps. We imposed local gradient constraining on the shared features to correct geometric misalignment, grounded in our finding that ST maps and H&E images exhibit high spatial consistency across cellular to tissue scales (Extended Fig. 3c-d).

Curriculum learning-based constraining module

To reduce transcript mixing in complex areas, FOCUS first quantifies complexity based on H&E features and gradually focuses on complex areas. Within these regions, H&E-derived cell-cell correlations constrains ST-derived correlations, as cellular morphology is relatively more stable and less affected by tissue complexity compared to ST profiling[14].

Cross-challenge coordination strategy

The cross-challenge coordination strategy further models co-occurring challenges (e.g., low resolution with poor alignment shown in Fig. S1) inside the diffusion conditioning process (Extended Fig. 2). For each module, the generated features are compressed into module-specific tokens, with multi-head cross-attention[29] applied over tokens, so that each module conditions its update on the others. The attended tokens are then converted to module-specific scaling and shifting coefficients that modulate the original feature maps, providing coordinated updates on modules. See Supplementary section 2 (Fig. S2) for an ablation analysis of the distinct modules and cross-challenge coordination strategy.



Extended Fig. 3 | Cross-gene and cross-modal correlation analysis supporting FOCUS modular designs. a, Spatial information refines traditional gene correlations. For example, in Xenium human breast cancer samples, the Pearson correlation coefficient (PCC) between *FOS* and *AGT* drops from 0.8 (global expression PCC) to 0.1 when spatial PCC is used, indicating that their similar global expression masks different spatial patterns. This implies that spatial gene co-expression priors should be explicitly incorporated into ST generation. **b,** Spatial gene co-expression correlation matrices are highly consistent across patients, as reflected by their low variation across patients. **c,** Average structure similarity index measure (SSIM) between paired H&E and ST patches (green) *versus* randomly matched patches (blue, baseline) across patch sizes. High SSIM values reflect strong structural consistency between ST and H&E images. **d,** Structural heterogeneity of spatial expression (indicated by Shannon entropy) increases when the number of expressed genes (blue) and average expression intensity (yellow) increase. This indicates that the spatial expression complexity can be reflected by the corresponding histological complexity.

2.2.2 Datasets for training and validation

We assembled a large-scale, cross-modal dataset of 6,876 paired tissue-level ST-H&E images (2,560 × 2,560 μm) and over 1.7 million patches (160 × 160 μm) with cell masks generated by a cell-segmentation algorithm[30]. Each tissue is linked to scRNA-seq profiles (over 5.8 million reference cells) and precomputed corresponding spatial gene co-expression matrices. Detailed dataset description is in Tables S1-3.

The dataset comprises 17 normal and 17 cancer tissue contexts from human and mouse samples. These include sequencing-based: Visium (probe-based and polyA-based)[31], Visium CytAssist[32], and VisiumHD[33] (all Visium platforms are products of 10X Genomics), Spatial Transcriptomics[34] (Spatial Transcriptomics AB), Stereo-seq[35] (BGI), BMK S1000[36] (BMKGENE), and Open-ST[23], and image-based platforms: Xenium[37] (10X Genomics)

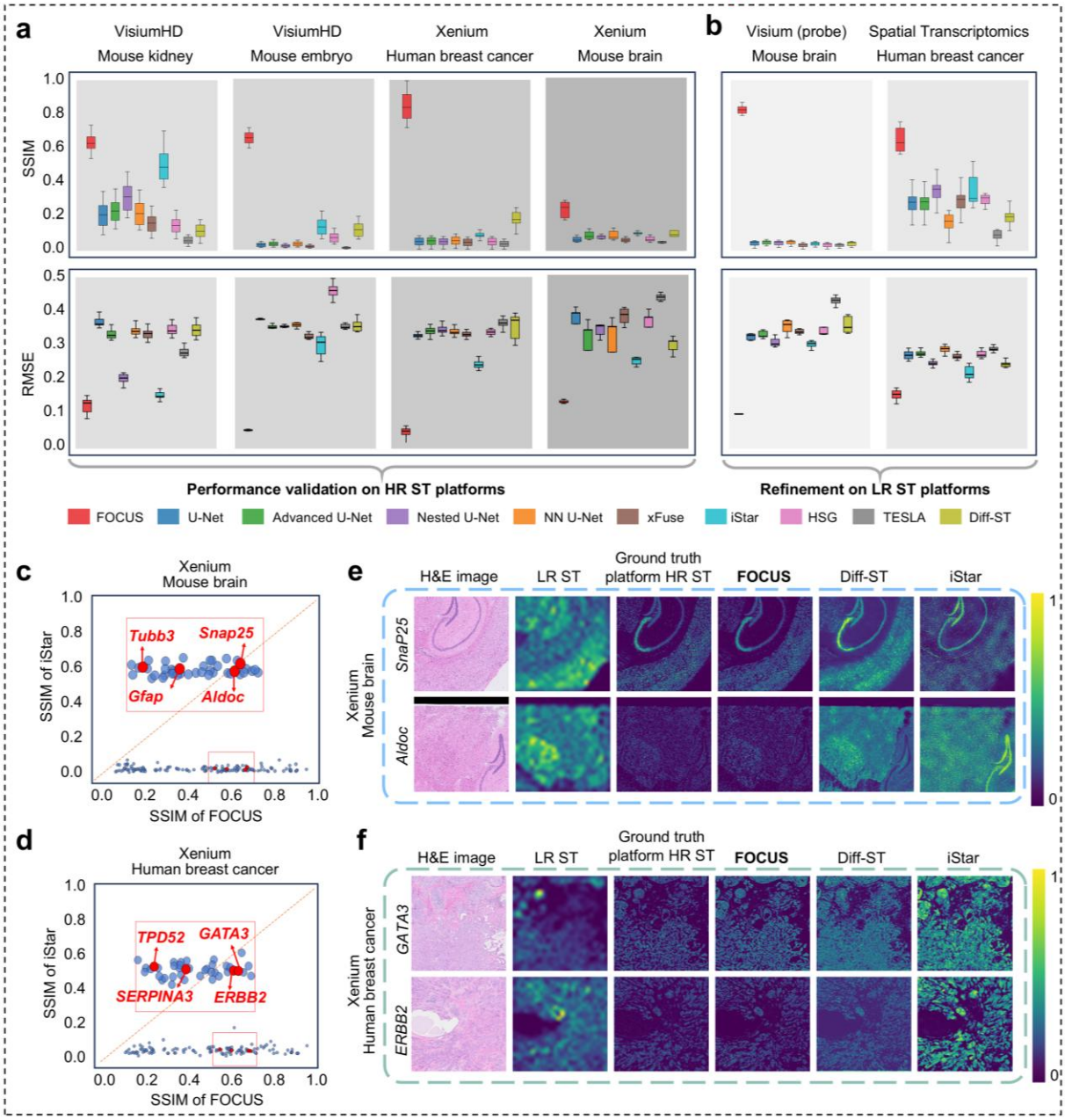
and CosMx[38] (Nanostring). Detailed dataset splitting for training and validation are in Methods.

2.2.3 Benchmarking settings

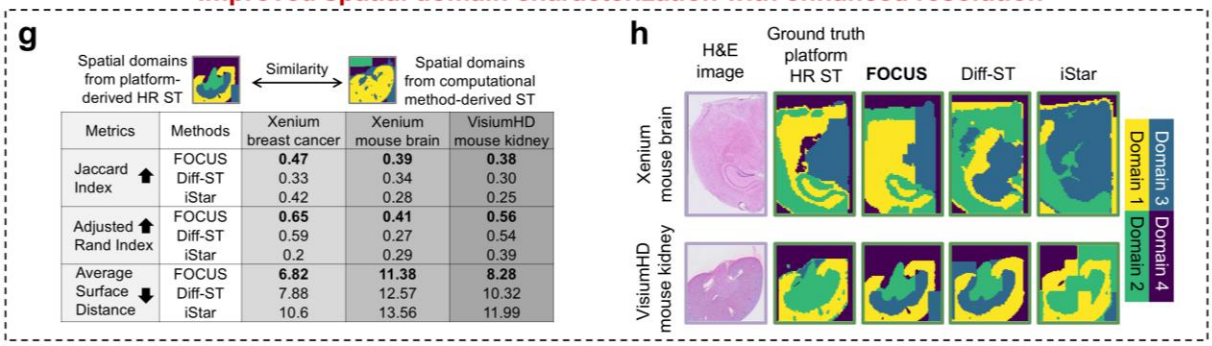
We benchmarked FOCUS on four challenges, comparing both native platforms and challenge-specific baselines[1] (Sections 2.3-2.6). Detailed benchmarking results across all platforms are presented in Tables S4-7. Further, we compared FOCUS with state-of-the-art foundational generative models adapted for ST, demonstrating the value of ST-specific design (Fig. S3; Supplementary sections 4).

[1] Definition of metrics used for evaluation for each challenge is in Supplementary section 3.

Resolution enhancement benchmarking



Improved spatial domain characterization with enhanced resolution



Metrics	Methods	Xenium breast cancer	Xenium mouse brain	VisiumHD mouse kidney
Jaccard Index ↑	FOCUS	**0.47**	**0.39**	**0.38**
	Diff-ST	0.33	0.34	0.30
	iStar	0.42	0.28	0.25
Adjusted Rand Index ↑	FOCUS	**0.65**	**0.41**	**0.56**
	Diff-ST	0.59	0.27	0.54
	iStar	0.2	0.29	0.39
Average Surface Distance ↓	FOCUS	**6.82**	**11.38**	**8.28**
	Diff-ST	7.88	12.57	10.32
	iStar	10.6	13.56	11.99

Fig. 2 | Evaluation of FOCUS in resolution enhancement. (**a-f**)**, Benchmarking resolution enhancement performance. a-b**, Boxplots of SSIM (top) and RMSE (bottom) for FOCUS and nine competing methods, in scenarios where high-resolution (HR) ST (**a**) or only low-resolution (LR) ST are available (**b**). FOCUS consistently achieves the best, indicating superior performance. **c-d,** Scatter plots comparing gene-wise SSIM by FOCUS (x-axis) and iStar[13] (best-performing; y-axis) for mouse brain (**c**) and human breast cancer (**f**) on Xenium. Canonical tissue marker genes are highlighted in red, indicating that FOCUS more effectively recapitulates their spatial patterns than iStar. **e-f,** Representative spatial patterns of marker genes for mouse brain (*Snap25, Aldoc*; **e**) and human breast cancer (*GATA3, ERBB2*; **f**). Columns show the H&E image, LR ST, ground-truth HR ST, and ST generated by FOCUS, and Diff-ST[12] and iStar[13] (top two best-performing), where FOCUS-generated ST most matches the HR ST maps. **(g-h), Improved spatial domain characterization with enhanced resolution. g**, Quantitative comparisons of FOCUS-derived *versus* competing methods-derived spatial domains, using the Jaccard index, adjusted Rand index, and average surface distance, show that FOCUS most accurately recovers the true spatial organization. **h**, Illustration of spatial domains inferred from super-resolved ST by FOCUS, Diff-ST, and iStar, with H&E images shown as reference.

2.3 FOCUS enhances ST resolution

By using high-resolution (HR) ST as the generation target and conditioning on H&E images with rich cellular morphology, FOCUS achieves resolution enhancement. We compared FOCUS with generic image enhancement (U-Net[39], Advanced U-Net[40], Nested U-Net[41] , nnU-Net[42]) and ST-specific methods[2] (xFuse[43], TESLA[11], HSG[44], iStar[13], Diff-ST[12]) (Fig. 2). Performance was evaluated at a 10× scale-up[3], from 100 μm/pixel to single-cell resolution of 10 μm/pixel, using root mean square error (RMSE) and structural similarity index measure (SSIM) to assess pixel-wise accuracy and structural preservation.

First, we benchmarked FOCUS on two HR platforms (VisiumHD [mouse kidney, embryo], and Xenium [human breast cancer, mouse brain]) using synthetic low-resolution (LR) ST (Fig. 2a). Performance is evaluated against native HR platforms as the ground truth. Across tissues and platforms, FOCUS consistently outperformed competing methods, reducing RMSE by at least 0.05 and increasing SSIM by at least 0.18 (both $P<0.001$).

Then, we utilized FOCUS to enhance platforms with native low-resolution (LR) (Visium (probe) [mouse brain], Spatial Transcriptomics [human breast cancer]) (Fig. 2b). As HR ST is unavailable, following prior work[12,13], we downsampled the enhanced ST to LR and

[2] To ensure fairness, all methods are provided with matched H&E images and low-resolution ST.

[3] FOCUS can achieve arbitrary enlargement times. This study only evaluated FOCUS at 10x scale.

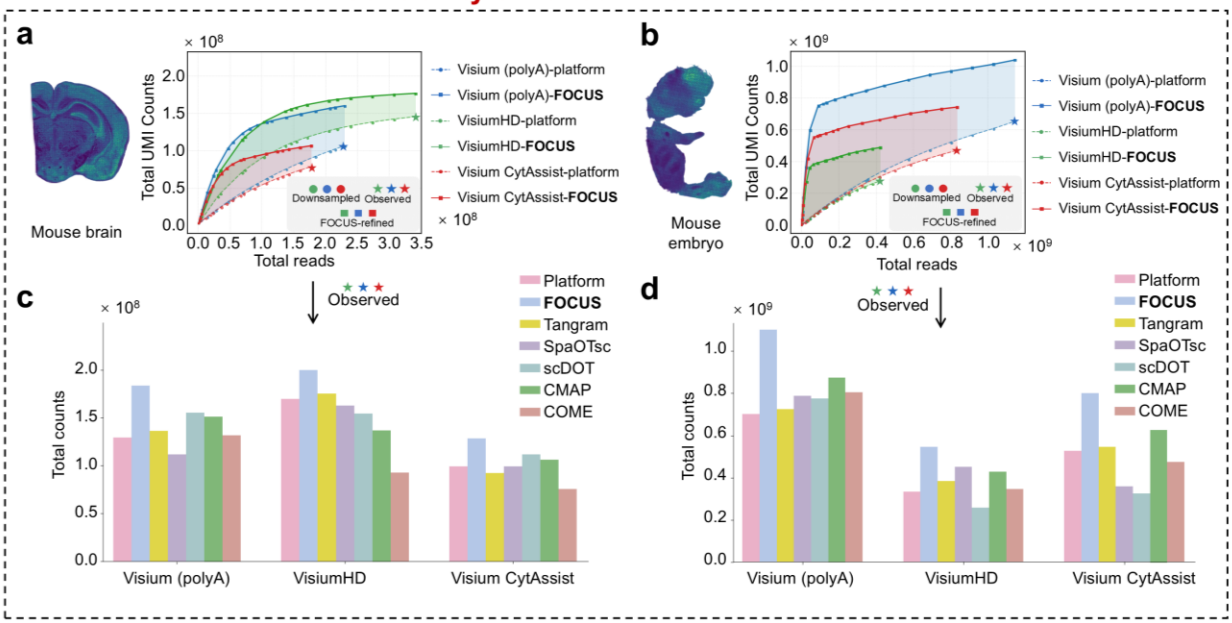
calculated SSIM and RMSE with native LR. FOCUS consistently performed the best in all comparisons, with SSIM gains of 0.80 and 0.28, and RMSE reductions of 0.20 and 0.18 (all $P < 0.001$) on Visium (probe) and Spatial Transcriptomics, respectively, highlighting real-world utility.

Further, we assessed gene-wise spatial fidelity, on Xenium for mouse brain and human breast cancer (Fig. 2c-f). Per-gene SSIM scatter plots show that most genes have higher SSIM values by FOCUS compared to competing methods, including key markers (e.g., *Tubb3, Snap25, Gfap,* and *Aldoc* in mouse brain, *TPD52, GATA3, SERPINA3,* and *ERBB2* in human breast cancer, and *Umod*, *Cdh16*, *Sdhc*, *Igfbp7* and *Cd63* in mouse kidney) (Fig. 2c-d, Fig. S4). Visualization further confirms that FOCUS provides sharper, anatomically coherent gene patterns than competing methods, reflecting its H&E-ST alignment with H&E structure. (Fig. 2e-f).

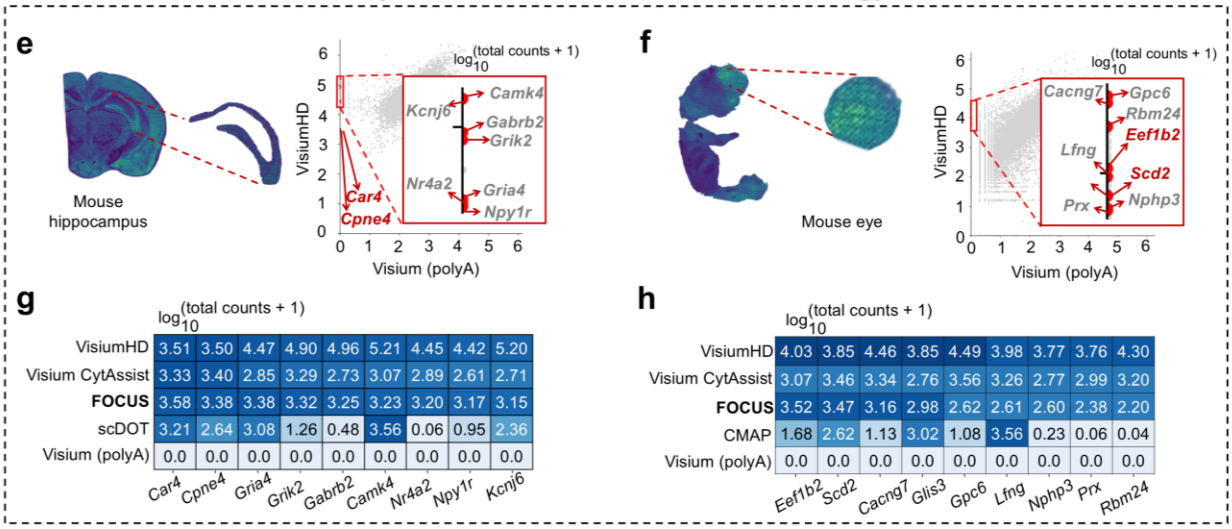
Finally, we evaluated whether FOCUS-enhanced resolution improves downstream analysis sensitive to resolution. We focused on spatial domain characterization, as low-resolution spots mix neighboring regions, blurring domain boundaries. Spatial domains inferred from FOCUS-reconstructed HR ST best matched domains from platform HR ST across tissues, with the highest Jaccard and adjusted Rand indices, and the lowest surface distance[4] (Fig. 2g-h), across both sequencing- (VisiumHD) and imaging-based (Xenium) platforms. This implies that FOCUS-enhanced resolution enables more accurate and robust delineation in the spatial domain.

[4] These three metrics are commonly used for evaluating spatial domain similarity.

Global Sensitivity enhancement for entire tissues



Sensitivity enhancement in known morphology



	Car4	Cpne4	Gria4	Grik2	Gabrb2	Camk4	Nr4a2	Npy1r	Kcnj6
VisiumHD	3.51	3.50	4.47	4.90	4.96	5.21	4.45	4.42	5.20
Visium CytAssist	3.33	3.40	2.85	3.29	2.73	3.07	2.89	2.61	2.71
FOCUS	3.58	3.38	3.38	3.32	3.25	3.23	3.20	3.17	3.15
scDOT	3.21	2.64	3.08	1.26	0.48	3.56	0.06	0.95	2.36
Visium (polyA)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

	Eef1b2	Scd2	Cacng7	Glis3	Gpc6	Lfng	Nphp3	Prx	Rbm24
VisiumHD	4.03	3.85	4.46	3.85	4.49	3.98	3.77	3.76	4.30
Visium CytAssist	3.07	3.46	3.34	2.76	3.56	3.26	2.77	2.99	3.20
FOCUS	3.52	3.47	3.16	2.98	2.62	2.61	2.60	2.38	2.20
CMAP	1.68	2.62	1.13	3.02	1.08	3.56	0.23	0.06	0.04
Visium (polyA)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Validation of the added value of H&E for sensitivity beyond scRNA

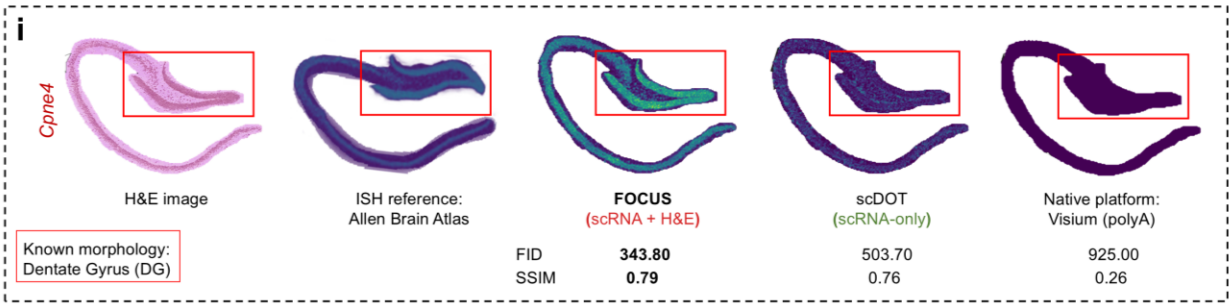


Fig. 3 | Validation of FOCUS in sensitivity enhancement. (**a-d**)**, Global sensitivity enhancement for entire tissues. a-b,** Comparison of FOCUS-enhanced with native platform-derived total counts, at both observed full set and downsampled sequencing depth (reads), on various platforms (Visium (polyA), VisiumHD, and Visium CytAssist) for mouse brain (**a**) and embryo (**b**). FOCUS increases profiling sensitivity with higher total counts across platforms. **c-d,** FOCUS outperforms scRNA-based competing methods in enhancing ST sensitivity, measured as higher total counts at the observed depth, for mouse brain (**c**) and embryo (**d**). (**e-h**)**, Sensitivity enhancement for regions of known morphology. e-f,** Total counts of detected genes are compared between Visium (polyA) (x-axis) and VisiumHD (y-axis) in mouse hippocampus (**e**) and embryonic eye (**f**). Each dot represents a gene, shown in grey. Genes with high expression in VisiumHD but undetectable in Visium (polyA) are highlighted in red and labeled. **g-h,** Heatmaps showing $\log_{10}$-expression for genes not captured by Visium (polyA) and recovered by FOCUS and consistently captured by reference platforms (VisiumHD and Visium CytAssist) for mouse hippocampus (**g**) and eye (**h**). **i**, **Contribution of H&E for sensitivity beyond scRNA.** Spatial expression of *Cpne4* along the dentate gyrus (DG) in mouse hippocampus, comparing native Visium (polyA), FOCUS and scDOT (using only scRNA-seq), with the paired H&E and *in situ* hybridization reference from the Allen Brain Atlas (ABA). Compared with scDOT, FOCUS further exploits H&E structural guidance and reconstructs DG architecture, achieving a lower FID and higher SSIM with the ABA reference.

2.4 FOCUS improves ST sensitivity

By integrating scRNA-seq, spatial gene co-expression priors, and H&E guidance in OT-based rectification module, FOCUS enhances ST sensitivity. Following prior works[3], we quantified sensitivity with total unique molecular identifier (UMI) counts, which represent the mRNA capture efficiency under fixed assay conditions and tissue areas[2,3]. Additionally, we also evaluated recovery of poorly captured genes with high-sensitivity *in situ* hybridization (ISH) as ground truth. We benchmarked FOCUS under two real-world scenarios as follows.

Global enhancement in whole tissue scale

First, we validated FOCUS in whole tissue scale, using widely studied mouse brain and embryo, across platforms of Visium (polyA), Visium CytAssist, and VisiumHD. FOCUS increased total UMI counts under both the observed full set and downsampled[3] sequencing depth (reads) across platforms (Fig. 3a-d). For example, in the observed data, total counts increased from 1.08×10^8 to 1.61×10^8 ($P < 0.001$) for Visium (polyA) mouse brain, and from 2.87×10^8 to 4.95×10^9 ($P < 0.001$) for VisiumHD mouse embryo. On the downsampled data, FOCUS yielded larger relative gains at lower depths (e.g., a 86% increase at the downsampled 8.68×10^7 reads *versus* 35% at the observed 1.77×10^8 reads

for Visium CytAssist mouse brain) (Fig. 3a-b), highlighting its effectiveness in low-throughput settings.

For extensive benchmarking, we further compared FOCUS with five competing methods (Tangram[4], SpaOTsc[45], scDOT[46], CMAP[47], and COME[48]) that integrate scRNA-seq for ST sensitivity enhancement. At observed sequencing depth, FOCUS achieved the highest total UMI counts among all competitors, consistently across platforms (Fig. 3c-d). Beyond increase of total counts, FOCUS also preserved global expression patterns across the transcriptome, with high Pearson correlation coefficients (PCC) between FOCUS and native platforms (e.g., 0.83, 0.92, and 0.83 on Visium (polyA), VisiumHD and Visium CytAssist for mouse brain), consistently surpassing all competing methods (Extended Fig. 4), likely driven by incorporating spatial co-expression priors.

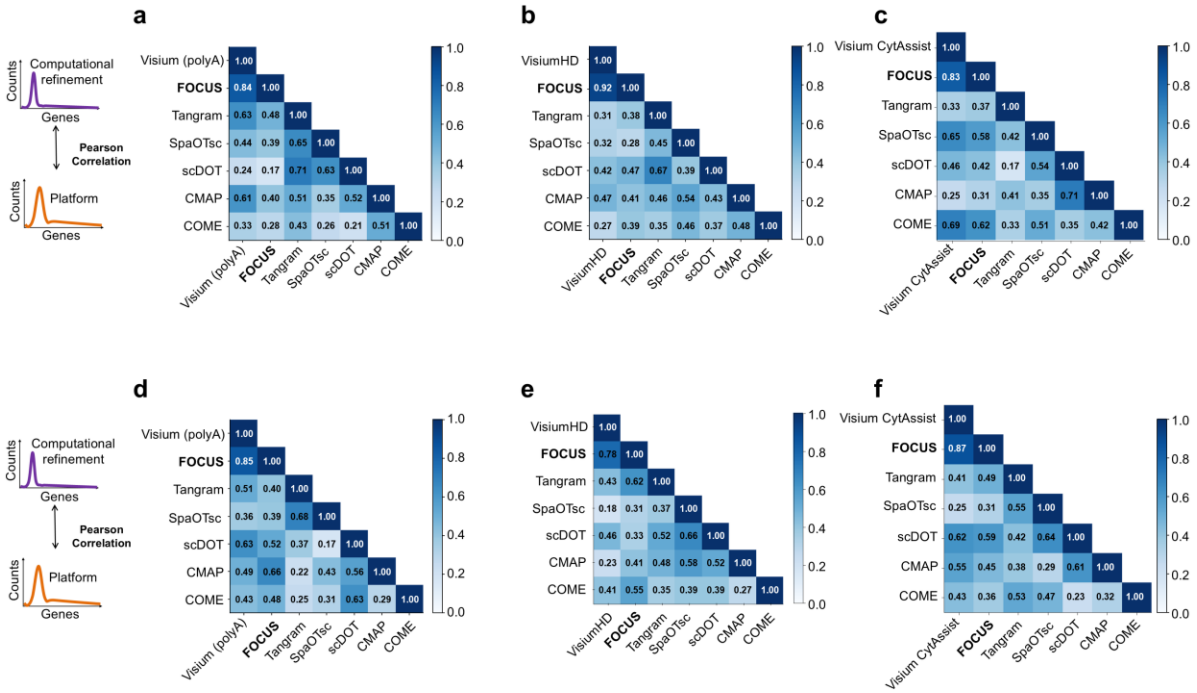
Local enhancement in regions of known morphology

Further, we validated FOCUS in regions with distinct morphology, including the mouse hippocampus and embryonic eye (Fig. 3e-h). We first segmented the hippocampus and eye regions from the whole tissues. Through pairwise comparisons across platforms, we identified marker genes undetectable in Visium (polyA) but consistently expressed across platforms of VisiumHD and Visium CytAssist, e.g., *Car4, Cpne4, Gria4, Grik2, Gabrb2, Camk4, Nr4a2, Npy1r and Kcnj6* in mouse hippocampus (Fig. 3e, Fig. S5a). FOCUS restored their expression to levels concordant with referenced platforms, e.g., achieving PCC of 0.73 (95%CI 0.70-0.75, $P<0.001$) with VisiumHD (Fig. 3g). Similar results were achieved in the embryonic eyes (Fig. 3f,h, Fig. S5b).

Validation of H&E morphology in sensitivity enhancement

Finally, to assess the contribution of H&E to sensitivity enhancement, we compared FOCUS with the best-performing competing method, scDOT, which uses only scRNA-seq. Analyses focused on the dentate gyrus (DG) region in the mouse hippocampus, a region with high-contrast tissue structure. FOCUS-enhanced ST better preserved DG structure, while also matching the Allen Brain Atlas *in situ* hybridization reference[49], assessed by Fréchet Inception Distance (FID) and SSIM (Fig. 3i). For *Cpne4*, a synaptic

marker gene with clear DG expression[50], FOCUS achieved the lowest FID and highest SSIM (FID 343.80, SSIM 0.79), outperforming scDOT (FID 503.70, SSIM 0.76; $P < 0.001$), demonstrating the added value of H&E guidance for sensitivity enhancement.



Extended Fig. 4 | FOCUS best preserves global expression patterns of platform-derived ST. PCC between native and FOCUS-enhanced ST are 0.83, 0.92, and 0.83 for mouse brain (**a**, Visium (polyA); **b**, VisiumHD; **c**, Visium CytAssist) and 0.85, 0.78, and 0.87 for mouse embryo (**d**, Visium (polyA); **e**, VisiumHD; **f**, Visium CytAssist), outperforming all competing methods and demonstrating robust cross-platform enhancement of sequencing sensitivity.

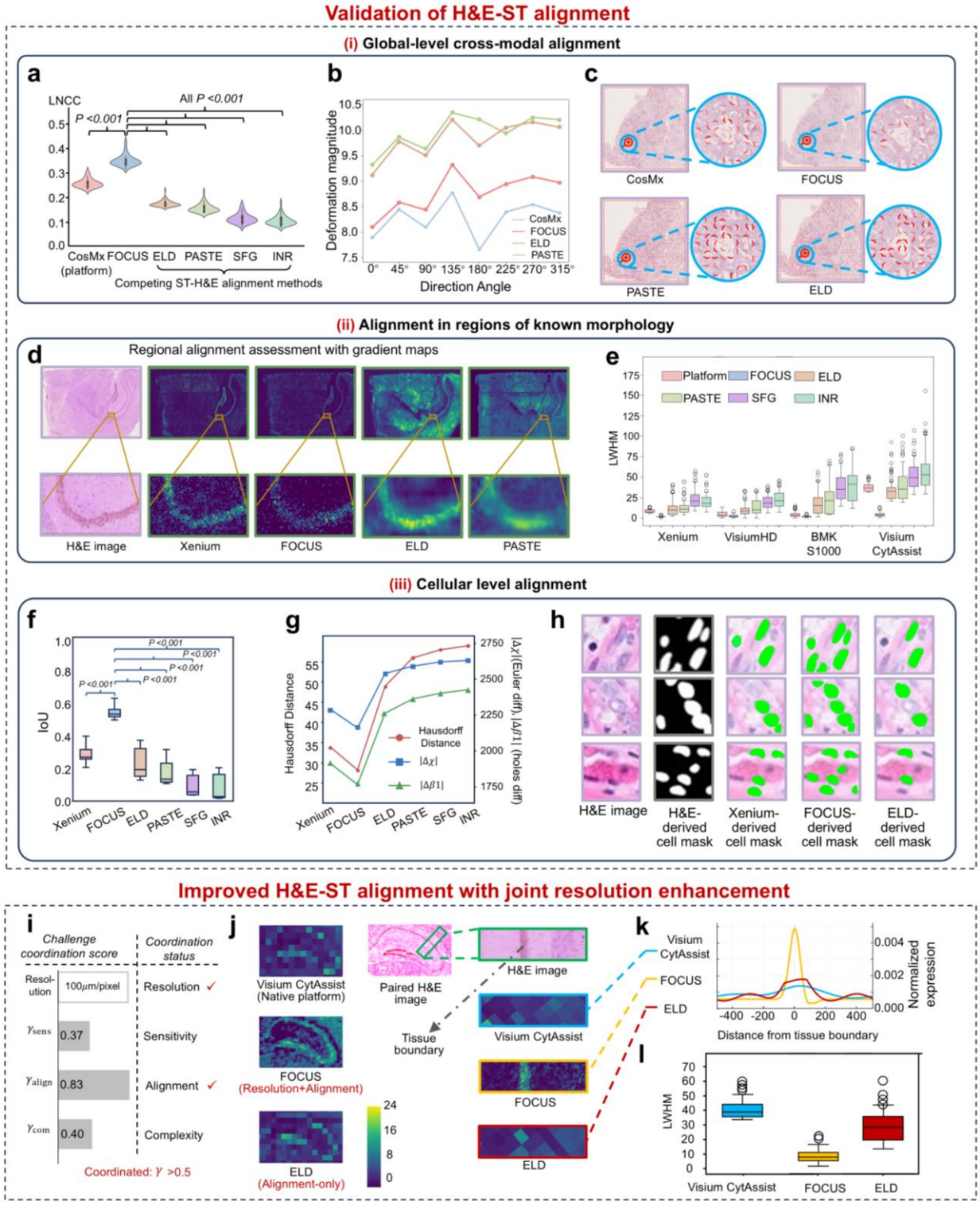
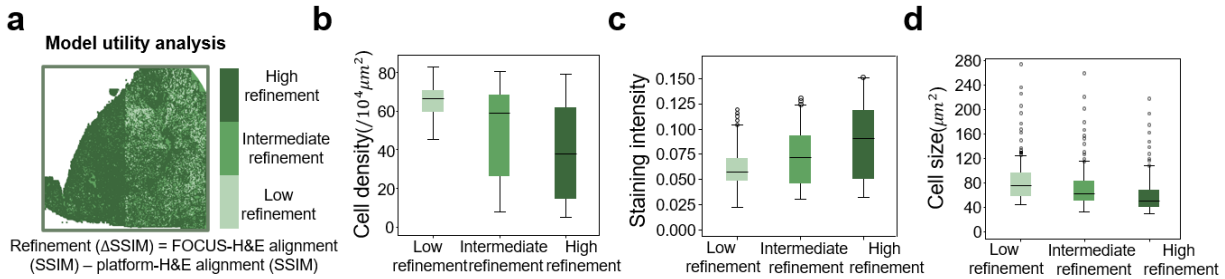


Fig. 4 | Evaluation of FOCUS in aligning ST with its histology context. a-c, Global alignment. Overall H&E-ST alignment (quantified by LNCC; **a**) and deformation magnitudes across deformation directions (**b**). FOCUS yields consistently highest LNCC and lowest deformations comparing with native CosMx platform and competing methods, indicating improved alignment with less warping. **c**, Illustration of deformation fields of CosMx, FOCUS, and competing

methods. **d-e,** Regional alignment. **d,** Gradient visualization of H&E and ST in mouse hippocampus, with FOCUS best preserving the known laminar morphology. **e**, Left half-width at half-maximum[3] (LWHM) distribution across four platforms, where FOCUS consistently performs the best, indicating the sharpest and most localized gene expression. **f-h,** Cellular level alignment. Pixel-wise (**f**) and topological (**g**) discrepancies between H&E- and ST-derived cell masks. FOCUS yields the smallest differences, indicating its closest agreement in cellular morphology and tissue topology. **h,** Representative cellular patches with H&E and corresponding cell masks derived from native platform, FOCUS and compared methods. **(i-l)**, **Improved alignment with joint resolution enhancement, validated on Visium CytAssist mouse brain. i,** The challenge coordination score shows co-occurrence of low resolution and poor alignment, indicating joint optimization in FOCUS. **j-l,** Spatial expression patterns in the mouse hippocampus for the native platform, FOCUS, and ELD (**j**), with alignment quality quantified by normalized expression (**k**) and LWHM (**l**). Compared with ELD, which optimizes alignment only, FOCUS-generated ST better aligns with tissue structures in H&E, demonstrating the benefit of joint resolution enhancement for alignment.



Extended Fig. 5 | Model utility analysis relating alignment enhancement (ΔSSIM) to histological features on CosMx. ΔSSIM denotes the increase in cross-modal SSIM of FOCUS-generated ST relative to the native platform. **a**, tissue map stratified into low, intermediate and high ΔSSIM regions. **b-d**, distributions of cell density (**b**), staining intensity (**c**), and cell size (**d**) across these groups show that alignment enhancement is greatest in regions with stronger staining, lower cell density and smaller cells, indicating that FOCUS most effectively corrects alignment in well-defined cellular structures.

2.5 FOCUS improves H&E-ST alignment

2.5.1 Standalone validation of H&E-ST alignment

FOCUS improves the spatial alignment of mRNA locations with H&E structures across platforms and spatial levels, enabled by its morphology-expression modulation module. We tested the accuracy at global (tissue deformation), regional (alignment in known morphology), and cellular (local, lateral mRNA diffusion) levels (Fig. 4). For comprehensive evaluation, we compared with state-of-the-art image registration methods adapted for H&E-ST alignment (PASTE[51], INR[52], SFG[53]), and ELD[54], specially designed for H&E-ST alignment.

Global validation in whole tissue scale

First, we benchmarked FOCUS on a whole tissue scale, based on mouse brain (CosMx) and human breast cancer (Xenium and BMK 1000). We followed prior work[52–55] and used local normalized cross-correlation (LNCC) and SSIM to quantify structural consistency between H&E and ST gradient maps. On CosMx, FOCUS outperformed both the native platform and competing methods, achieving best LNCC (0.352, 95%CI 0.332-0.376; P<0.001) and SSIM (0.386, 95%CI 0.347-0.420; P<0.001) (Fig. 4a, Fig. S6). Similar alignment gains by FOCUS were observed on BMK S1000 and Xenium (Fig. S7a, Fig. S8a-b), demonstrating its platform-agnostic performance.

For robustness validation, we then examined the deformation magnitude across different direction angles (from 0° to 315°) on both CosMx and Xenium. FOCUS produced the lowest H&E-ST deformation magnitudes in all angles, compared to both native platforms and competing methods (mean deformation magnitude: FOCUS 8.31 *versus* native platform 8.78, ELD 9.77, and PASTE 9.91 on CosMx; FOCUS 9.29 *versus* native platform 9.77, ELD 10.88, and PASTE 10.99 on Xenium) (Fig. 4b-c, Fig. S8c-d), indicating robust alignment with less warping of the underlying histology.

Regional validation in known morphology

We next evaluated lateral diffusion correction in selected anatomical structures, i.e., hippocampal region in mouse brain. Two metrics were used: SSIM between H&E and ST gradient maps (Fig. S9), and the left-width at half-maximum[3] (LWHM) of expression peaks (Fig. S10). For LWHM, we focused on histological regions where spatial expressions should exhibit a sharp contrast, with high expression in one subregion and minimal to no expression in the rest. Across platforms, FOCUS consistently narrowed lateral spread and sharpened boundary transitions compared to native platforms, as reflected by LWHM (Xenium: FOCUS 5.21, platform 9.78; VisiumHD: FOCUS 6.12, platform 3.73; BMKS 1000: FOCUS 4.57, platform 7.84; Visium CytAssist: FOCUS 6.68, platform 38.26), while also outperforming competing methods (Fig. 4d-e).

Cellular alignment validation

Further, we assessed cellular-level alignment on human breast cancer patches (Xenium), because misalignment often arises in complex tumor regions where lateral diffusion mixes signals between neighboring cells (Fig. 4f-h). FOCUS increased Intersection over Union between cell masks derived from H&E and ST maps to 0.57 (95%CI 0.53-0.59; $P<0.001$), outperforming Xenium (0.28, 95%CI 0.26-0.29; $P<0.001$), and competing methods of ELD (0.24, 95%CI 0.21-0.29; $P<0.001$), PASTE (0.19, 95%CI 0.17-0.23; $P<0.001$), SFG (0.11, 95%CI 0.07-0.15; $P<0.001$), and INR (0.07, 95%CI 0.03-0.08; $P<0.001$) (Fig. 4f). In addition, FOCUS generated smallest geometric and topological discrepancies between H&E- and ST-derived cell masks, evaluated by Hausdorff distance, Euler and Holes difference (Fig. 4g). Finally, visual inspection confirmed that FOCUS minimized directional offsets between predicted gene expression and H&E-derived cell boundaries (Fig. 4h).

Utility analysis across histological features

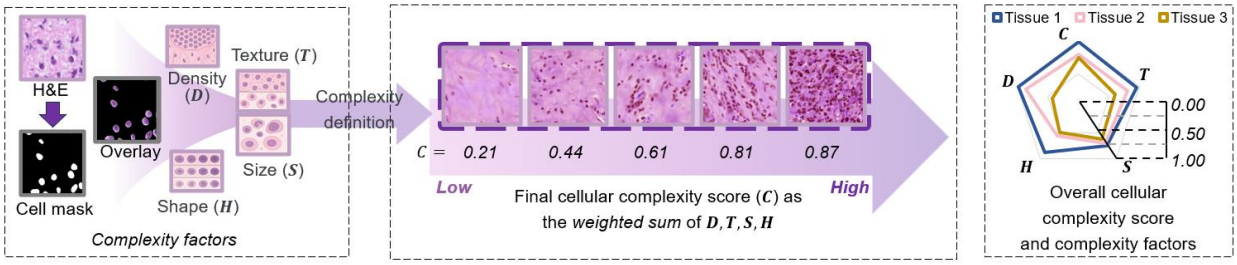
Moreover, we conducted a *model utility analysis* that relates FOCUS's alignment gains to histological contexts. We used ΔSSIM, defined as the SSIM increase of FOCUS over the native platform. We observe that regions with higher ΔSSIM tended to show lower cell density (mean 63, 51, 39 cells per $10^4 \mu m^2$ for low-, intermediate-, and high ΔSSIM regions, respectively; $P<0.001$), higher staining intensity (mean 0.061, 0.072, and 0.095, respectively; $P<0.001$), and smaller cell sizes[5] (mean 79, 68, 52 μm^2, respectively; $P<0.001$; Extended Fig. 5). Together, these patterns show that FOCUS's alignment gains occur in specific, interpretable histological contexts, directly linking improved H&E-ST alignment to underlying tissue features.

2.5.2 Joint enhancement of resolution and alignment

Finally, we evaluated joint enhancement of co-occurring challenges of resolution and alignment, because LR ST blurs boundaries in ST, possibly hindering accurate alignment with H&E structures (Fig. 4i-l). The native LR Visium CytAssist mouse brain was used. The activated challenge-coordination score (Activated: >0.5; Methods) confirmed that FOCUS jointly optimizes resolution and alignment, benefiting from hippocampal anatomy

[5] The metrics were calculated with the segmented cells from H&E images.

in H&E (Fig. 4i). Compared with the best-performing competing method, ELD, which optimizes alignment only, FOCUS-generated ST aligned better with H&E (Fig. 4j), showing a sharper expression distribution (P < 0.001; Fig. 4k), and achieving a much lower average LWHM (8.5 *versus* 39.2 for ELD; P < 0.001; Fig. 4l). These results demonstrate the benefit of joint resolution and alignment enhancement.



Extended Fig. 6 | Schematic definition of tissue complexity. For a given tissue patch, complexity is determined by cell density and heterogeneity (cell-type diversity), with the latter captured by variations in cell size, shape, and texture.

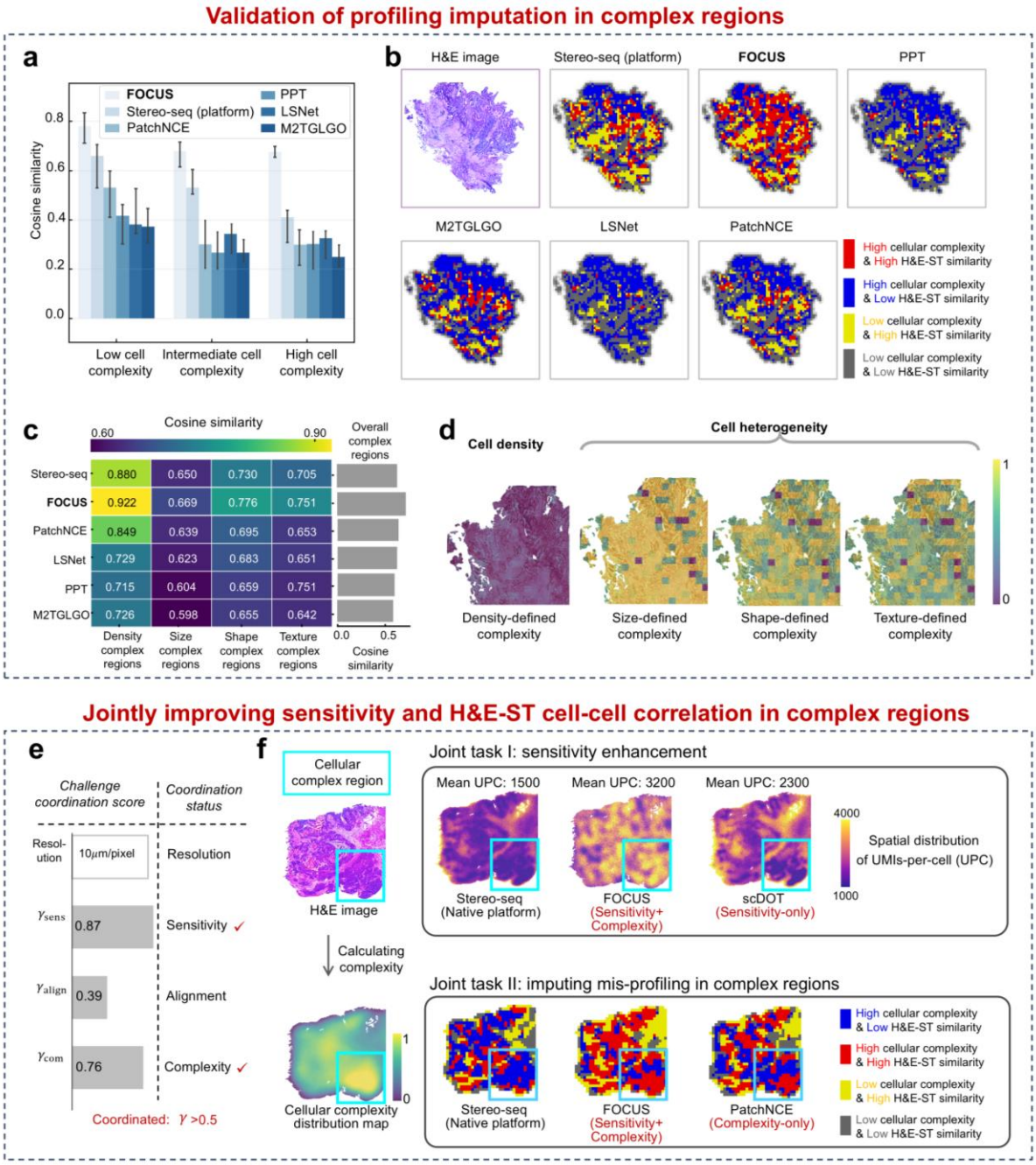


Fig. 5 | Enhancement of profiling precision in complex regions. a, Cosine similarity between cell-cell correlation matrices derived from H&E and ST (from the native Stereo-seq platform, FOCUS and competing methods) across low, intermediate, and high complexity regions. FOCUS achieves the greatest improvement in cosine similarity, especially in high-complexity regions. **b**, Spatial maps of complexity and H&E-ST feature consistency, with pixels classified by joint categories of complexity and feature consistency; FOCUS yields the largest extent of high-complexity regions with high H&E-ST feature consistency. **c-d**, Model robustness analysis. **c**, Cosine similarity (defined in **a**) in high-complexity regions *determined* by various factors of cell density, size, shape, or texture. FOCUS consistently maintains the highest similarity compared to the native platform and competing methods, demonstrating its robust profiling performance. **d**,

Illustration of spatial distribution of factor-specific complexity in a human colon cancer tissue section on Stereo -seq. **(e-f)**, **Joint enhancement of challenges in sensitivity and complexity, tested on Stereo-seq ovarian cancer. e,** The challenge coordination score reveals co-occurring challenges sensitivity and complexity, triggering joint optimization in FOCUS. **f,** Left: H&E image with the corresponding cellular complexity map, where low sensitivity (quantified by UMIs per cell) and poor H&E-ST feature consistency co-occur. Right: performance comparison with single-task baselines, i.e., scDOT for sensitivity and PatchNCE for complexity. FOCUS performs best on both tasks, demonstrating the benefit of joint optimization.

2.6 FOCUS imputes profiling in complex regions

2.6.1 Standalone validation of complexity challenge

Guided by H&E-derived cell-cell correlations, FOCUS imputes mis-profiling in complex regions. Following prior work[56,57], we quantified cellular complexity using representative histological features, including cell density and spatial variations in texture, size, and shape (Extended Fig. 6; Methods). We benchmarked FOCUS against M2TGLGO[58], designed for H&E-based ST generation and state-of-the-art image-to-image translation methods (ASP[59], PPT[60], and LSNet[61]) adapted for imputing mis-profiling in ST. We quantified the correction of mixing artifacts by the cosine similarity between H&E-derived and ST-derived cell-cell correlation matrices.

Benchmarking was conducted on Stereo-seq of human colon cancer, with pronounced heterogeneity in cellular complexity across tumor and infiltrative areas. On Stereo-seq, FOCUS achieved the highest cosine similarity across areas of low (0.78, 95%CI 0.70-0.84; P<0.001), intermediate (0.66, 95%CI 0.62-0.71; P<0.001), and high complexity[6] (0.66, 95%CI 0.64-0.69; P<0.001; Fig.5 a-b), over native platform and competing methods. Of note, FOCUS achieved the largest cosine similarity improvement in more complex areas prone to high mixing artifacts, with the increase of 0.25 in high-complexity *versus* 0.13 and 0.14 in low- and intermediate-complexity (Fig. 5a). We also observed similar improvement in mouse brain on Xenium (Fig. S11a), indicating broad utility with normal tissues. These results suggest the usefulness of a curriculum learning-based

[6] Defined by dividing the complexity score into three equal tertiles (low, intermediate, high).

constraining module, which quantifies tissue complexity and prioritizes constraining between H&E- and ST-based cell-cell correlations in complex areas.

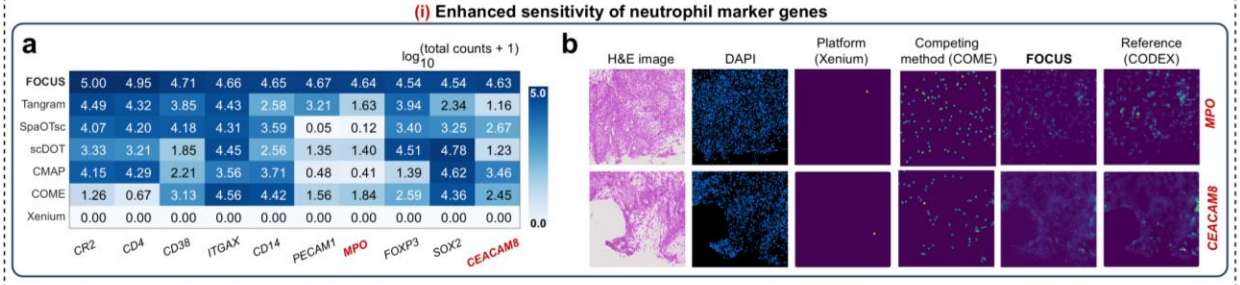
To further assess *robustness*, FOCUS was evaluated in variously defined complexity based on factors of cell density as well as cell size, shape, and texture. For high-complexity regions defined by each factor, FOCUS consistently achieved the highest cosine similarity on both Stereo-seq human colon cancer (density 0.922, size 0.669, shape 0.776, texture 0.751) and Xenium mouse brain (density 0.841, size 0.815, shape 0.823, texture 0.838) (Fig. 5c–d; Fig. S12), outperforming both native platforms and competing methods. Consistent gains across complexity factors indicate robust profiling performance of FOCUS in diverse complexity.

2.6.2 Joint validation of sensitivity and complexity challenges

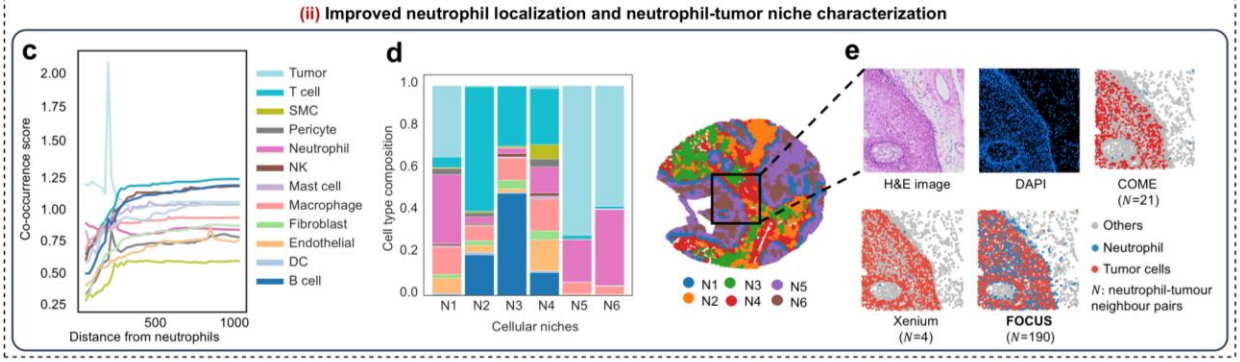
Finally, we validated FOCUS in the co-occurring challenges of sensitivity and complexity, as high cellular complexity limits per-cell molecule capture (e.g., UMIs), thereby reducing sensitivity. We used Stereo-seq ovarian cancer data, which contain both low-sensitivity[3] and complex tumor regions (Fig. 5e-f). As shown in Fig. 5f, high-complexity areas exhibit fewer UMIs per cell and low H&E-ST feature consistency. The challenge-coordination score (Methods) confirmed that FOCUS detects and jointly mitigates both challenges (Fig. 5e). Compared with the best single-task baselines, i.e., scDOT for sensitivity and PatchNCE for complexity, FOCUS achieved a higher per-cell UMI count (3,200 *versus* 2,300 for scDOT; $P < 0.001$) and a larger area with high H&E-ST similarity in complex regions (82% *versus* 53% for PatchNCE; $P < 0.001$) (Fig. 5f). These results demonstrate the benefit of joint optimization across sensitivity and complexity via the challenge-coordinated strategy.

Neutrophil-tumor niche characterization in papillary craniopharyngioma

(i) Enhanced sensitivity of neutrophil marker genes

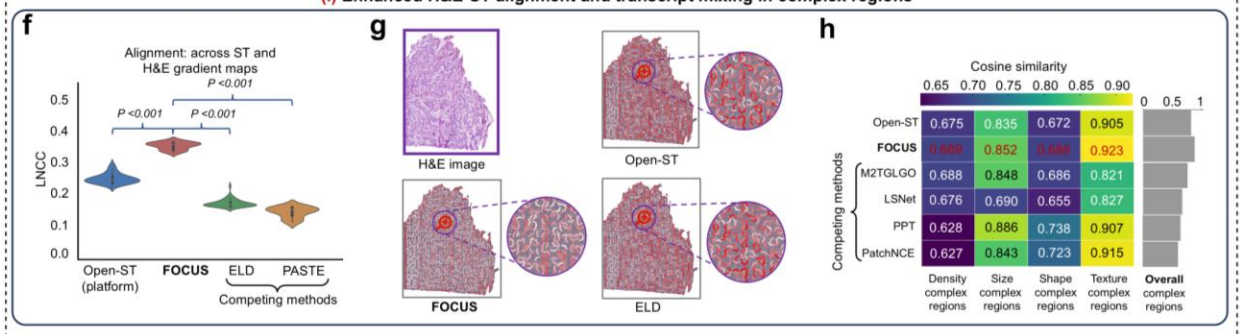


(ii) Improved neutrophil localization and neutrophil-tumor niche characterization

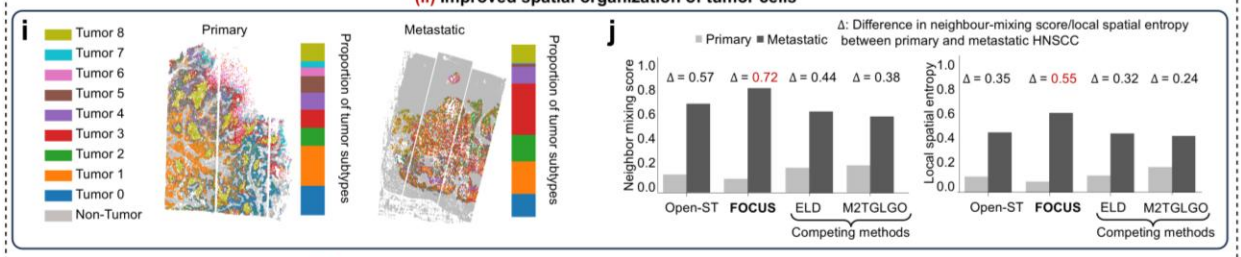


Spatially constrained heterogeneity characterization in primary and metastatic HNSCC

(i) Enhanced H&E-ST alignment and transcript mixing in complex regions



(ii) Improved spatial organization of tumor cells



(iii) Clarified transcriptomic heterogeneity based on cell-cell communications

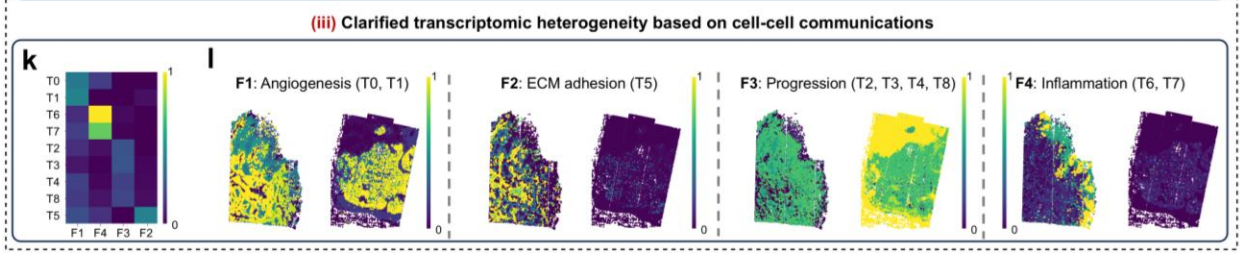


Fig. 6 | Real-world utility validation of FOCUS. (a-e) Neutrophil-tumor niche characterization for PCP on Xenium. a-b, Sensitivity enhancement. **a,** Heatmap of $\log_{10}$-expression for genes poorly detected by Xenium, including neutrophil markers *MPO* and *CEACAM8*. FOCUS recovers these genes and shows stronger expression than competing methods. **b,** Spatial distributions of *MPO* and *CEACAM8* for native platform-derived and method-enhanced ST, with FOCUS most faithfully recapitulating the coherent neutrophil patterns observed in CODEX. **c-e,** Improved neutrophil-tumor niche characterization. **c**, Co-occurrence scores between neutrophils and other cell types for FOCUS-enhanced ST. Tumor cells are the dominant neighbors of neutrophils, revealing pronounced neutrophil-tumor niches. **d**, Neutrophil-centered cellular niches (N1-N6) obtained by ST clustering. All niches enriched with tumor cells (N1, N4 and N6) consistently confirm the neutrophil-tumor co-localization patterns inferred in **c. e**, Example region containing all three tumor-enriched niches identified in **d**. **(f-l) Spatially constrained heterogeneity characterization in primary and metastatic HNSCC on Open-ST. f-h,** Enhanced H&E-ST alignment and feature consistency in complex regions. **f**, H&E-ST alignment quantified by LNCC, comparing FOCUS with Open-ST and competing methods. **g**, Illustration of deformation fields, indicating improved alignment with less warping in FOCUS. **h**, Cosine similarity between H&E- and ST-derived cell-cell correlation matrices in complex regions. FOCUS shows the highest similarity, indicating improved profiling precision in complex regions. **i-j,** Improved spatial organization of tumor cells. **i**, Spatial maps of tumor transcriptomic states (T0-T8) from FOCUS-enhanced ST in primary and metastatic HNSCC. The primary tumor shows large, coherent domains, whereas the metastasis displays intermixed patterns. **j**, Spatial intermixing degree of spatial clustered tumor states derived from Open-ST, FOCUS, and competing methods, evaluated by neighbor mixing score[62] and local spatial entropy[63]. **k-l,** Clarified transcriptomic heterogeneity based on cell-cell communications. **k**, Loadings of tumor states on four communication factors (F1-F4) using spatial ligand-receptor scores (Methids). **l**, Spatial distribution of F1-F4, corresponding to signaling programs of angiogenesis, ECM adhesion, tumor progression, and inflammation.

2.7 FOCUS elucidates neutrophils-tumor niche characterization in papillary craniopharyngioma

Beyond performance evaluation, we further utilized FOCUS in a real-world application of characterizing neutrophils-tumor niches in PCP. PCP is a rare suprasellar tumor with a characteristically neutrophil-rich inflammatory microenvironment[64]. However, in the tumor microenvironment, activated neutrophils show transient transcriptional activity and short lifespan, making mRNAs unstable and less sensitive to capture[65].

To characterize neutrophils-tumor niches, we first used FOCUS to enhance the detection sensitivity, particularly neutrophil markers such as *MPO* and *CEACAM8,* by fine-tuning FOCUS on the PCP dataset (Xenium) (Fig. 6a-b). Of note, the poorly expressed genes were confirmed using high-sensitivity spatial proteomics (CODEX[66], profiled using adjacent tissue sections of those for Xenium) as reference, shown in Fig. S13. FOCUS markedly increased the profiling sensitivity (total counts) over the native platform and

competing methods (Fig. 6a). FOCUS-enhanced ST also showed much higher structural concordance with CODEX as a protein-level validation: for *MPO*, SSIM of 0.58 between FOCUS and CODEX *versus* 0.05 between the native platform and CODEX (P<0.001); for *CEACAM8*, 0.51 *versus* 0.03 (P<0.001) (Fig. 6b, Fig. S14).

Based on the enhanced ST, FOCUS improved neutrophil localization and further helped reveal co-localized neutrophils and tumor cells. The distance-based cell co-occurrence analysis showed that tumor cells were the dominant neighbors of neutrophils (Fig. 6c). This pattern is particularly observed in three niches enriched with tumor cells (Niche 1, 4, and 6; Fig. 6d), derived from ST clustering analysis. In an example region containing all three niches, FOCUS revealed dense neutrophil clusters at tumor borders and within tumor nests (neutrophil-tumor neighbor pairs N=190), largely missed by the native platform (N=4) and the best-performing competing method, COME (N=21) (Fig. 6e). These results indicate the usefulness of FOCUS in identifying local inflammatory niches and potentially informing disease courses and treatment responses[64]

2.8 FOCUS uncovers spatially constrained heterogeneity in primary and metastatic HNSCC on Open-ST

Open-ST[67] is a recent high-resolution platform that profiles consecutive sections. Despite strengths, Open-ST has reported anisotropic lateral diffusion and inaccurate cell localization[67], leading to H&E-ST misalignment and transcript mixing in complex regions. We thereby applied FOCUS to Open-ST data, enabling clearer characterization of the spatially constrained transcriptomic heterogeneity in primary HNSCC and its metastatic lymph node (Fig. 6f-l), where subtle cell-cell localizations and interactions are critical for interpreting metastasis.

First, FOCUS jointly improved H&E-ST alignment and transcript mixing in complex regions compared with native Open-ST and competing methods (Fig. 6f-h, Figs. S15-16). Using the enhanced ST, we clustered ~45,000 tumor cells from primary and metastatic lesions into eight transcriptomic states (T0-T8). In the primary tumor, these states formed spatially coherent domains, whereas in the lymph node, these states were highly intermixed (Fig. 6i), consistent with the disrupted spatial architecture of metastasis[68,69].

We then quantified this difference with the neighbor-mixing score and local spatial entropy (Supplementary section 3.6). FOCUS showed the largest primary-to-metastatic increase in spatial intermixing (neighbor-mixing improvement = 0.72 *versus* 0.57 for Open-ST, 0.44 for ELD, and 0.38 for M2TGLGO; entropy improvement = 0.55 *versus* 0.35, 0.32, and 0.24, respectively; Fig. 6j, Fig. S17), indicating that FOCUS best captures preserved compartmentalization in the primary tumor and increased mixing in the metastasis, likely through FOCUS's improved alignment and profiling in complex regions.

Finally, based on the FOCUS-derived cell states, we performed cell-cell communication analysis to dissect spatially constrained transcriptomic heterogeneity (Fig. 6k-l). We decomposed ligand-receptor interactions into latent factors and recovered four stable non-negative programs corresponding to angiogenesis, ECM adhesion, progression, and inflammation (Fig. 6l, Fig. S18), linking the distinct spatial architecture to functional pathways supporting growth, invasion, and immune modulation.

Discussion

FOCUS is a diffusion-based generative foundation model that jointly mitigates four interdependent ST challenges, including resolution, sensitivity, H&E-ST alignment and profiling precision in complex regions. Its performance rests mainly on three key aspects: (i) integration of multiple ST-complement modalities, (ii) coordinated interaction across challenges, and (iii) cross-platform foundational training.

First, FOCUS exploits multimodal complementary information within a modular architecture to jointly tackle multiple ST challenges. It uses scRNA-seq references, spatial co-expression, and H&E-gradient guidance to boost sensitivity; H&E-derived cellular morphology and H&E-ST spatial concordance to sharpen resolution and alignment; and H&E-guided cell-cell correlation constraints to reduce signal mixing in complex regions. **Second**, a cross-challenge coordination strategy targets defects that tend to co-occur. For instance, low-resolution ST blurs tissue boundaries and hampers alignment with H&E structures; the benefit of joint resolution and alignment enhancement is shown in Section 2.5.2. Additionally, the coordination design also enables flexible module combinations and for different real-world needs. **Finally**, FOCUS builds on large-scale pretrained H&E and ST encoders[21,22] that capture rich cellular morphology and spatial expression patterns. Further, FOCUS was jointly trained across ten imaging- and sequencing-based platforms, possibly leveraging strengths while compensating limitations of each platform.

We benchmarked FOCUS on four challenges across all platforms, comparing both native platforms and challenge-specific baselines. Beyond challenge validation, we further fine-tuned it on unseen tissue context or platform to assess real-world utility: **(i)** for PCP on Xenium, FOCUS recovered spatial patterns of neutrophil markers such as *MPO* and *CEACAM8*, and characterized tumor-neutrophil niches that was largely invisible from the native platform; **(ii)** for primary and metastatic HNSCC on Open-ST, FOCUS rectified misalignment and transcript mixing in complex tumor regions, leading to clearer tumor subtypes, neighborhood patterns and communication programs across primary and metastatic sites. Together, these examples show that FOCUS is not only an ST enhancer but also a useful tool for revealing biologically interpretable findings in real-world datasets.

Computational methods have been proposed for each challenge, yet remain intrinsically difficult. For **resolution** enhancement, H&E-guided approaches[11,12,70] improve spatial detail but cannot capture transcriptomic heterogeneity, as morphologically similar cells can differ markedly in gene expression. **Sensitivity** can be increased by integrating scRNA-seq with ST[4,17,18]. Despite effective, most methods require paired and annotated single-cell references and are not capable of considering tissue heterogeneity[71]. H&E-ST **alignment** methods[15,16,72,73] mainly align features but rarely correct lateral mRNA diffusion, leading to geometric misalignment between molecular and histological structures. Decomposition-based approaches for signal mixing in **complex** regions[10,74] are largely designed for low-resolution ST and do not readily extend to newer high-resolution platforms. **Collectively**, most challenge-tailored methods ignore to integrate multiple complementary modalities or jointly tackle these coupled challenges within a single framework.

In addition to models targeted for specific challenges, recent explored generative foundation models[19,20] show potential for joint ST enhancement across challenges. We conducted a comparison study with representative generative foundation models (Supplementary section 4), and found FOCUS performs the best across all challenges. This could be because they primarily optimize visual realism and do not explicitly resolve ST-specific issues, such as transcriptional mixing

In summary, FOCUS is the first computational method for jointly mitigating multiple challenges of resolution, sensitivity, alignment and profiling precision in complex regions. We believe FOCUS will be an invaluable analysis tool for the enhancement of ST from present or future platforms. There are also multiple avenues of possible extensions for FOCUS. One is to extend ST to spatial multi-omics. Most recently, technologies have been proposed for extending ST to other spatial omics, including spatial proteomics, and spatial multi-omics, such as spatial epigenome-transcriptome. Systematic evaluation of current experiment-based spatial multi-omics technologies should be conducted, while the tailored computational rectification model should be designed accordingly. Thus, we plan to extend FOCUS to generate enhanced spatial multi-omics with incorporation of single-cell multi-omics and cross-omics correlation graph as prior knowledge. We also

aim to extend FOCUS's functionality with data from serial tissue sections. Such input multi-section integration can involve sections with the same data modalities or even mosaic integration of different modalities captured by different sections, with outputting 3D multi-omics atlas.

Reference

1. Systematic benchmarking of high-throughput subcellular spatial transcriptomics platforms across human tumors | Nature Communications. https://www.nature.com/articles/s41467-025-64292-3.
2. Wang, H. *et al.* Systematic benchmarking of imaging spatial transcriptomics platforms in FFPE tissues. *bioRxiv* 2023.12.07.570603 (2023) doi:10.1101/2023.12.07.570603.
3. You, Y. *et al.* Systematic comparison of sequencing-based spatial transcriptomic methods. *Nat. Methods* **21**, 1743–1754 (2024).
4. Biancalani, T. *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat. Methods* **18**, 1352–1362 (2021).
5. Cao, J. *et al.* Decoder-FFPE-seq enables sensitive, genome-wide spatial transcriptomics of archival tissues at single-cell resolution. 2025.09.12.675967 Preprint at https://doi.org/10.1101/2025.09.12.675967 (2025).
6. Wirth, J. *et al.* Spatial transcriptomics using multiplexed deterministic barcoding in tissue. *Nat. Commun.* **14**, 1523 (2023).
7. Ren, J. *et al.* Spatiotemporally resolved transcriptomics reveals the subcellular RNA kinetic landscape. *Nat. Methods* **20**, 695–705 (2023).
8. Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R. & Haque, A. An introduction to spatial transcriptomics for biomedical research. *Genome Med.* **14**, 68 (2022).
9. Kennedy-Darling, J. *et al.* Chapter 4 - Highly multiplexed spatial protein data using CODEX technology. in *Revealing Uncharted Biology With Single Cell Multiplex Proteomic Technologies* (ed. Fantl, W. J.) 93–118 (Academic Press, 2024). doi:10.1016/B978-0-12-822209-6.00001-1.
10. Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* **40**, 517–526 (2022).
11. Hu, J. *et al.* Deciphering tumor ecosystems at super resolution from spatial transcriptomics with TESLA. *Cell Syst.* **14**, 404-417.e4 (2023).
12. Wang, X., Huang, X., Price, S. & Li, C. Cross-Modal Diffusion Modelling for Super-Resolved Spatial Transcriptomics. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024* (eds Linguraru, M. G. et al.) 98–108 (Springer Nature Switzerland, Cham, 2024). doi:10.1007/978-3-031-72384-1_10.
13. Zhang, D. *et al.* Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nat. Biotechnol.* **42**, 1372–1377 (2024).
14. Zhang, P. *et al.* Thor: a platform for cell-level investigation of spatial transcriptomics and histology. *Nat. Commun.* **16**, 7178 (2025).
15. Huang, J. *et al.* Bridging cell morphological behaviors and molecular dynamics in multi-modal spatial omics with MorphLink. *Nat. Commun.* **16**, 5878 (2025).
16. Lin, Y. *et al.* ST-Align: A Multimodal Foundation Model for Image-Gene Alignment in Spatial Transcriptomics. Preprint at https://doi.org/10.48550/arXiv.2411.16793 (2024).
17. Sun, E. D., Ma, R., Navarro Negredo, P., Brunet, A. & Zou, J. TISSUE: uncertainty-calibrated prediction of single-cell spatial transcriptomics improves downstream analyses. *Nat. Methods* **21**, 444–454 (2024).
18. Vahid, M. R. *et al.* High-resolution alignment of single-cell and spatial transcriptomes with CytoSPACE. *Nat. Biotechnol.* **41**, 1543–1548 (2023).
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-Resolution Image Synthesis With Latent Diffusion Models. in 10684–10695 (2022).
20. Podell, D. *et al.* SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. in (2023).
21. Xu, H. *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
22. Wang, C. *et al.* scGPT-spatial: Continual Pretraining of Single-Cell Foundation Model for Spatial Transcriptomics. 2025.02.05.636714 Preprint at https://doi.org/10.1101/2025.02.05.636714 (2025).
23. Schott, M. *et al.* Open-ST: High-resolution spatial transcriptomics in 3D. *Cell* **187**, 3953-3972.e26 (2024).
24. Yang, L. *et al.* Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Comput Surv* **56**, 105:1-105:39 (2023).

25. Goodfellow, I. J. *et al.* Generative Adversarial Nets. in *Advances in Neural Information Processing Systems* vol. 27 (Curran Associates, Inc., 2014).
26. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. Preprint at https://doi.org/10.48550/arXiv.1312.6114 (2022).
27. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) 234–241 (Springer International Publishing, Cham, 2015). doi:10.1007/978-3-319-24574-4_28.
28. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
29. Vaswani, A. *et al.* Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2023).
30. Cellpose: a generalist algorithm for cellular segmentation | Nature Methods. https://www.nature.com/articles/s41592-020-01018-x.
31. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
32. Visium CytAssist Spatial Gene Expression Reagent Kits User Guide | Official 10x Genomics Support. *10x Genomics* https://www.10xgenomics.com/support/cytassist-spatial-gene-expression/documentation/steps/library-construction/visium-cyt-assist-spatial-gene-expression-reagent-kits-for-ffpe.
33. Oliveira, M. F. de *et al.* High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer. *Nat. Genet.* **57**, 1512–1523 (2025).
34. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
35. Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777-1792.e21 (2022).
36. gd-admin. BMKMANU S1000 Spatial Transcriptome. *https://www.bmkgene.com/* https://bmkgene.com:443/bmkmanu-s1000-spatial-transcriptome-product/.
37. Xenium In Situ Platform. *10x Genomics* https://www.10xgenomics.com/platforms/xenium (2024).
38. He, S. *et al.* High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat. Biotechnol.* **40**, 1794–1806 (2022).
39. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) 234–241 (Springer International Publishing, Cham, 2015). doi:10.1007/978-3-319-24574-4_28.
40. Oktay, O. *et al.* Attention U-Net: Learning Where to Look for the Pancreas. Preprint at https://doi.org/10.48550/arXiv.1804.03999 (2018).
41. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (eds Stoyanov, D. et al.) 3–11 (Springer International Publishing, Cham, 2018). doi:10.1007/978-3-030-00889-5_1.
42. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
43. Bergenstråhle, L. *et al.* Super-resolved spatial transcriptomics by deep data fusion. *Nat. Biotechnol.* **40**, 476–479 (2022).
44. Pang, M., Su, K. & Li, M. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. 2021.11.28.470212 Preprint at https://doi.org/10.1101/2021.11.28.470212 (2021).
45. Cang, Z. & Nie, Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat. Commun.* **11**, 2084 (2020).
46. Nguyen, N. D. *et al.* scDOT: optimal transport for mapping senescent cells in spatial transcriptomics. *Genome Biol.* **25**, 288 (2024).
47. Ke, J. *et al.* High-resolution mapping of single cells in spatial context. *Nat. Commun.* **16**, 6533 (2025).

48. Wei, X. *et al.* COME: contrastive mapping learning for spatial reconstruction of single-cell RNA sequencing data. *Bioinformatics* **41**, btaf083 (2025).
49. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
50. von Ziegler, L. M., Selevsek, N., Tweedie-Cullen, R. Y., Kremer, E. & Mansuy, I. M. Subregion-Specific Proteomic Signature in the Hippocampus for Recognition Processes in Adult Mice. *Cell Rep.* **22**, 3362–3374 (2018).
51. Zeira, R., Land, M., Strzalkowski, A. & Raphael, B. J. Alignment and integration of spatial transcriptomics data. *Nat. Methods* **19**, 567–575 (2022).
52. Wolterink, J. M., Zwienenberg, J. C. & Brune, C. Implicit Neural Representations for Deformable Image Registration. in *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning* 1349–1359 (PMLR, 2022).
53. Ge, L., Wei, X., Hao, Y., Luo, J. & Xu, Y. Unsupervised Histological Image Registration Using Structural Feature Guided Convolutional Neural Network. *IEEE Trans. Med. Imaging* **41**, 2414–2431 (2022).
54. Ekvall, M. *et al.* Spatial landmark detection and tissue registration with deep learning. *Nat. Methods* **21**, 673–679 (2024).
55. Wang, X., Price, S. & Li, C. Multi-task Learning of Histology and Molecular Markers for Classifying Diffuse Glioma. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023* (eds Greenspan, H. et al.) 551–561 (Springer Nature Switzerland, Cham, 2023). doi:10.1007/978-3-031-43990-2_52.
56. Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
57. Feng, L. *et al.* Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a multicentre observational study. *Lancet Digit. Health* **4**, e8–e17 (2022).
58. Shi, H., Chi, C., Wan, P., Zhang, D. & Shao, W. Multi-modal Topology-embedded Graph Learning for Spatially Resolved Genes Prediction from Pathology Images with Prior Gene Similarity Information. in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 20810–20819 (2025). doi:10.1109/CVPR52734.2025.01938.
59. Li, F., Hu, Z., Chen, W. & Kak, A. Adaptive Supervised PatchNCE Loss for Learning H&E-to-IHC Stain Translation with Inconsistent Groundtruth Image Pairs. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023* (eds Greenspan, H. et al.) 632–641 (Springer Nature Switzerland, Cham, 2023). doi:10.1007/978-3-031-43987-2_61.
60. Zhang, W. *et al.* High-Resolution Medical Image Translation via Patch Alignment-Based Bidirectional Contrastive Learning. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024* (eds Linguraru, M. G. et al.) 178–188 (Springer Nature Switzerland, Cham, 2024). doi:10.1007/978-3-031-72083-3_17.
61. Wang, A., Chen, H., Lin, Z., Han, J. & Ding, G. LSNet: See Large, Focus Small. in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 9718–9729 (2025). doi:10.1109/CVPR52734.2025.00908.
62. Yang, T. *et al.* SPIAT: An R package for the Spatial Image Analysis of Cells in Tissues. 2020.05.28.122614 Preprint at https://doi.org/10.1101/2020.05.28.122614 (2020).
63. Altieri, L., Cocchi, D. & Roli, G. The use of spatial information in entropy measures. Preprint at https://doi.org/10.48550/arXiv.1703.06001 (2017).
64. Jia, Y. *et al.* Immune infiltration in aggressive papillary craniopharyngioma: High infiltration but low action. *Front. Immunol.* **13**, (2022).
65. Zhang, F. *et al.* Neutrophil diversity and function in health and disease. *Signal Transduct. Target. Ther.* **9**, 343 (2024).
66. Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968-981.e15 (2018).
67. Schott, M. *et al.* Open-ST: High-resolution spatial transcriptomics in 3D. *Cell* **187**, 3953-3972.e26 (2024).
68. Karimi, E. *et al.* Single-cell spatial immune landscapes of primary and metastatic brain tumours. *Nature* **614**, 555–563 (2023).

69. Xu, K. *et al.* Microenvironment components and spatially resolved single-cell transcriptome atlas of breast cancer metastatic axillary lymph nodes. *Acta Biochim. Biophys. Sin.* **54**, 1336–1348 (2022).
70. Zhang, D. *et al.* Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nat. Biotechnol.* **42**, 1372–1377 (2024).
71. He, S. *et al.* Starfysh integrates spatial transcriptomic and histologic data to reveal heterogeneous tumor–immune hubs. *Nat. Biotechnol.* **43**, 223–235 (2025).
72. Zhang, P. *et al.* Thor: a platform for cell-level investigation of spatial transcriptomics and histology. *Nat. Commun.* **16**, 7178 (2025).
73. Coleman, K. *et al.* Resolving tissue complexity by multimodal spatial omics modeling with MISO. *Nat. Methods* **22**, 530–538 (2025).
74. Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661–671 (2022).
75. Long, Y. *et al.* Deciphering spatial domains from spatial multi-omics with SpatialGlue. *Nat. Methods* **21**, 1658–1667 (2024).
76. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
77. Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. https://ieeexplore.ieee.org/document/9880275.
78. Weigert, M. & Schmidt, U. Nuclei Instance Segmentation and Classification in Histopathology Images with Stardist. in *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)* 1–4 (2022). doi:10.1109/ISBIC56247.2022.9854534.
79. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. Preprint at https://doi.org/10.48550/arXiv.1506.02640 (2016).
80. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
81. Sinkhorn, R. A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. *Ann. Math. Stat.* **35**, 876–879 (1964).
82. Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777-1792.e21 (2022).
83. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. in (2018).
84. Dhariwal, P. & Nichol, A. Diffusion Models Beat GANs on Image Synthesis. in *Advances in Neural Information Processing Systems* vol. 34 8780–8794 (Curran Associates, Inc., 2021).
85. Comprehensive Integration of Single-Cell Data: Cell. https://www.cell.com/cell/fulltext/S0092-8674%2819%2930559-8.
86. Palla, G. *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).
87. Dimitrov, D. *et al.* Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nat. Commun.* **13**, 3224 (2022).
88. Nirmal, A. J. & Sorger, P. K. SCIMAP: A Python Toolkit for Integrated Spatial Analysis of Multiplexed Imaging Data. *J. Open Source Softw.* **9**, 6604 (2024).

Methods

1. Data processing

ST pre-processing

To preprocess the ST data from native platforms, we followed prior works[75] and filtered pixels expressing fewer than 2% of the total gene numbers and genes expressing fewer than 2% of total pixels. Next, the gene expression counts were log-transformed and normalized by library size via the SCANPY package[76].

For single-cell–resolution platforms, we generated paired high- and low-resolution ST maps for training and validation: 10 μm/pixel (single-cell) and 100 μm/pixel, both covering a 2,560 × 2,560 μm tissue field. For intrinsically low-resolution platforms, only 100 μm/pixel maps over the same field of view were used for model validation.

H&E image preprocessing and cell segmentation

Following previous works[12,13], whole-slide H&E images were first converted to RGB and cropped to the tissue-covered area using a simple tissue mask (intensity thresholding and morphological closing). For each whole-slide image, we cropped a tissue-level field of view of 2,560 × 2,560 μm and tiled it into non-overlapping cellular-scale patches of 160 × 160 μm, consistent with typical local cellular organisation in histology[77]. H&E patches with low tissue coverage or strong blur (estimated from gradient magnitude) were discarded. Patches were stored at 0.5 μm/pixel for cell segmentation, and downsampled to 10 μm/pixel (approximately single-cell resolution) for FOCUS training and validation.

Nuclei were segmented on the 0.5 μm/pixel patches using a pretrained StarDist2D[78] model, which produces instance-level probability maps and polygonal outlines. Predictions were converted to nuclei masks with default thresholds, small artifacts were removed, and masks were isotopically dilated in a non-overlapping manner to approximate full-cell boundaries. The resulting nuclei and cell masks were stitched back to the tissue field and visually inspected on representative regions.

Pair-wise alignment of H&E and ST maps

To obtain ground-truth H&E-ST alignment for FOCUS training and validation, we used a two-step landmark-based registration, following previous works[23]. First, H&E images and ST maps were rescaled to a common resolution. Candidate landmark pairs were generated automatically: fiducial beads or printed markers visible in both modalities were detected with a YOLO[79]-based object detector.

Second, these landmarks were manually enhanced in an interactive viewer, removing mismatches and adding points in morphologically complex regions (e.g. large vessels, ducts, region boundaries). The final set of 100-500 landmarks per sample (depending on image size) was used to estimate a global affine transform followed by a thin-plate-spline warp, yielding high-quality registered H&E-ST pairs for FOCUS training and evaluation.

Spatial gene co-expression matrix

We constructed a tissue-type–specific spatial gene co-expression matrix as prior knowledge. For each sample of a given tissue context, we flattened the 2D ST map into 1D expression profiles for each gene and computed the PCC between all gene pairs to obtain a sample-level co-expression matrix. We then averaged these matrices across samples to yield the final spatial gene co-expression matrix for that tissue context.

scRNA-seq processing

For each tissue context, we collected scRNA-seq datasets from public repositories (Table S1). Raw count matrices and metadata were converted to AnnData objects and processed with SCANPY package. Following standard quality control (QC), we removed low-quality cells (few detected genes, low total unique molecular identifier (UMI) counts or high mitochondrial fraction) and low-quality genes (expressed in only a small number of cells), with filtering thresholds determined by QC-metric distributions. Counts were then library-size-normalized so that each cell had the same total UMI count. Finally, to enable integration with ST data, we intersected the gene sets of each scRNA-seq dataset and its corresponding ST dataset and kept only shared genes in the scRNA-seq data.

2. Foundational generative model of FOCUS

2.1 Diffusion framework and multimodal conditioning

FOCUS is a conditional diffusion model that generates high-quality (HQ) ST maps $\mathbf{s_0}$ from multimodal inputs and biological prompts (Extended Fig. 2). HQ ST maps $\mathbf{s_0}$ serves as the generation targets. Under-refined low-quality (LQ) ST maps $\mathbf{y}$ and paired H&E images $\mathbf{h}$ provide sample-specific conditions. Cell masks $\mathbf{m}$, obtained by segmenting $\mathbf{h}$, impose local structural constraints. Reference scRNA-seq matrices $\mathbf{S}$ of the corresponding tissue context are used to rectify and project LQ ST expression. A precomputed spatial co-expression matrix (SCM) $\mathbf{M}$ acts as a memory bank encoding gene-gene spatial relationship. Finally, to incorporate biological prior knowledge, we used BioBERT[80] with fully frozen parameters to encode natural language prompts $\mathbf{P}$ describing the platform, species, tissue context and gene set.

Given a multi-channel HQ ST map $\mathbf{s_0}$ (channels[7] correspond to genes), the forward diffusion process gradually adds Gaussian noise over T diffusion steps following a noise variance schedule $\beta_1, \dots, \beta_T$. As step increases in the forward process, the input image gets noisier, with the noised image at step t defined as $\mathbf{s}_t = \sqrt{1-\beta_t}\mathbf{s}_{t-1} + \sqrt{\beta_t}\epsilon_t$ where $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a standard Gaussian noise vector. A key characteristic of the forward process is its ability to sample $\mathbf{s}_t$ at any arbitrary timestep t in closed form. Using the notation $\alpha_t \coloneqq 1 - \beta_t$ and, $\bar{\alpha} := \prod_{j=1}^{t} \alpha_j$, we have

$$\mathrm{q}(\mathbf{s}_t|\mathbf{s}_0) = \mathcal{N}\left(\mathbf{s}_t; \sqrt{\bar{\alpha}_t}\mathbf{s}_0, \right)(1-\bar{\alpha}_t)\mathbf{I}$$

Thus, for sufficiently large T, $\mathbf{s}_T$ approximates an isotropic Gaussian distribution.

The reverse process learns to iteratively remove noise from the noised image $\mathbf{s}_T$. We parameterized the reverse transition

$$p_\theta(\mathbf{s}_{t-1}|\mathbf{s}_t, \mathbf{y}, \mathbf{h}, \mathbf{m}, \mathbf{S}, \mathbf{M}, \mathbf{P}) = \mathcal{N}(\mathbf{s}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{s}_t, \mathbf{y}, \mathbf{h}, \mathbf{m}, \mathbf{S}, \mathbf{M}, \mathbf{P}), \sigma_t^2\mathbf{I})$$

where θ denotes trainable parameters and σ_t^2 represents the fixed variance. Further, with parameterization, learning μ_θ is replaced by learning the noise $\boldsymbol{\epsilon}_\theta(\mathbf{s}_t, \mathbf{y}, \mathbf{h}, \mathbf{m}, \mathbf{S}, \mathbf{M}, \mathbf{P}, t)$, enabling more stable training. Once trained, we sampled HQ ST maps by starting from

[7] Considering the training efficiency, we used the default setting of 5 genes per iteration, consistent with prior works[12].

Gaussian noise $\mathbf{s}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and running the learned reverse process conditioned on the multimodal inputs and prompts.

On top of this diffusion backbone and multimodal features from large-scale pretrained encoders, FOCUS introduces three challenge-targeted modules: (**i**) an optimal-transport (OT)-based rectification module that fuses high-sensitivity scRNA-seq with large-scale spatial co-expression knowledge to boost sensitivity; (**ii**) a morphology-expression modulation module that aligns and disentangles ST and H&E features by enforcing local gradient consistency and feature-level correspondence; and (**iii**) a curriculum learning-based constraining module that progressively focuses on structurally complex regions and aligns expression-based and morphology-based cell-cell correlations.

2.2 Multimodal pretraining

FOCUS builds on two large-scale pretrained encoders, including (1) a cross-tissue H&E encoder capturing universal cellular morphology and (2) a cross-platform ST encoder encoding spatial expression patterns, which were used in model training during the reverse diffusion process. Specifically, at timestep t in the reverse diffusion process, the H&E image $\mathbf{h}$ and reconstructed HQ ST map $\mathbf{s}_t$ are encoded into latent representations $\mathbf{F}_{\mathrm{HE}}$ and $\mathbf{F}_{\mathrm{ST}}$, respectively.

- **Histology encoder.** To obtain universal histological features, we initialized a transformer-based H&E encoder with weights from the pathology foundation model Prov-GigaPath[21]. A lightweight matching layer is appended to map Prov-GigaPath feature dimensions and scales into FOCUS's intermediate latent space.
- **ST encoder.** To capture generic spatial transcriptomic patterns, the ST encoder is initialized from the spatial transcriptomics foundation model scGPT-spatial[22]. As with the histology encoder, a small matching layer performs dimensionality mapping and stabilization so that the pre-trained weights can be directly inherited while producing features in the shared latent space.

2.3 OT-based rectification module for sensitivity enhancement

To enhance sensitivity, we designed an OT-based rectification module with three coupled components: (1) an OT-based scRNA-ST fusion stage that fuses high-sensitivity scRNA-seq into ST, (2) a spatial gene-correlation strategy that utilizes a gene memory bank incorporating spatial co-expression matrices, and (3) an H&E-guided regional attention mechanism that emphasizes expression sensitivity in high gradient tissue boundaries of morphologically well-defined regions .

OT-based scRNA-ST fusion

Let $\mathbf{y} \in \mathbb{R}^{\{n_{\{spots\}} \times n_{\{genes\}}\}}$ denote the observed ST expression tensor (reshaped from the spatial map) and $\mathbf{S} \in \mathbb{R}^{\{n_{\{cells\}} \times n_{\{genes\}}\}}$ the matched scRNA-seq matrix. We built a cost matrix $\mathbf{C} \in \mathbb{R}^{\{n_{\{cells\}} \times n_{\{spots\}}\}}$ with entries capturing the squared Euclidean distance between spot and cell expression profiles. Using uniform marginals for spots and cells, we solved an entropically regularized optimal transport problem:

$$\mathbf{\Gamma}^* = argmin_{\mathbf{\Gamma}} < \mathbf{\Gamma}, \mathbf{C} > -\lambda H(\mathbf{\Gamma})$$

where $H(\mathbf{\Gamma})$ is the entropy of the transport plan and λ controls the degree of regularization. We then computed $\mathbf{\Gamma}^*$ with the Sinkhorn algorithm[81] and project the scRNA-seq reference onto the spatial grid:

$$\mathbf{y}' = (\mathbf{\Gamma}^*)^\top \mathbf{S}$$

The rectified expression $\mathbf{y}'$ serves as a globally enhanced ST condition, providing high-sensitivity and denoised gene expression, which is used as input to the following spatial gene-correlation.

The spatial gene-correlation strategy and H&E-guided regional attention mechanism are detailed in Fig. S19 and Supplementary section 6.

2.4 Morphology-expression modulation module for aligning ST with H&E

To align ST with H&E images, we introduced a morphology-expression modulation module (Fig. S20) that **(1)** disentangles shared and unique representations across modalities, and **(2)** enforces local structural consistency across modality-shared representations.

Multimodal disentanglement strategy

Given multimodal feature maps $\mathbf{F}_{\mathrm{OT}}$ and $\mathbf{F}_{\mathrm{CL}}$ encoded from the OT-based rectification module and curriculum learning-based constraining module, we first performed Euclidean normalization,

$$\mathbf{F}_{\mathrm{align}}^{\mathrm{ST}} = \frac{\mathbf{F}_{\mathrm{OT}}}{\parallel \mathbf{F}_{\mathrm{OT}} \parallel_2}, \qquad \mathbf{F}_{\mathrm{align}}^{\mathrm{HE}} = \frac{\mathbf{F}_{\mathrm{CL}}}{\parallel \mathbf{F}_{\mathrm{CL}} \parallel_2}.$$

Then, we modulated $\mathbf{F}_{\mathrm{align}}^{\mathrm{ST}}$ and $\mathbf{F}_{\mathrm{align}}^{\mathrm{HE}}$ to disentangle shared and unique information between them. For H&E-to-ST modulation, we considered a local $k \times k$ neighborhood around each pixel x in the ST feature map, extracting patches $\mathbf{P}_x^{\mathrm{ST}}$ and $\mathbf{P}_x^{\mathrm{HE}}$ from $\mathbf{F}_{\mathrm{align}}^{\mathrm{ST}}$ and $\mathbf{F}_{\mathrm{align}}^{\mathrm{HE}}$, respectively. A small fully connected filter $\mathbf{f}_\theta$ takes the concatenated patch as input and outputs a local modulation kernel:

$$\mathbf{w}_x = \mathbf{f}_\theta(\mathbf{P}_x^{\mathrm{ST}}, \mathbf{P}_x^{\mathrm{HE}}),$$

which is applied to update the ST feature at x,

$$\tilde{\mathbf{F}}^{\mathrm{ST}}(x) = \mathbf{w}_x * \mathbf{P}_x^{\mathrm{ST}}.$$

A symmetric modulation is applied for ST-to-H&E, yielding $\tilde{\mathbf{F}}^{\mathrm{HE}}$.

We then decomposed the modulated features ($\tilde{\mathbf{F}}^{\mathrm{ST}}$ and $\tilde{\mathbf{F}}^{\mathrm{HE}}$) into shared and unique components ($\mathbf{F}_{\mathrm{shared}}^{\mathrm{HE}}$, $\mathbf{F}_{\mathrm{shared}}^{\mathrm{ST}}$, $\mathbf{F}_{\mathrm{unique}}^{\mathrm{HE}}$, $\mathbf{F}_{\mathrm{unique}}^{\mathrm{ST}}$). Based on these features, a disentangling loss

$$\mathcal{L}_{\mathrm{dis}} = -\text{cosine similarity}(\mathbf{F}_{\mathrm{shared}}^{\mathrm{HE}}, \mathbf{F}_{\mathrm{shared}}^{\mathrm{ST}}) + \gamma \text{cosine similarity}(\mathbf{F}_{\mathrm{unique}}^{\mathrm{HE}}, \mathbf{F}_{\mathrm{unique}}^{\mathrm{ST}})$$

maximizes cosine similarity between shared features while encouraging uniqueness of modality-specific components ($\gamma > 0$). The final multimodal fused representation

$$\mathbf{F}_{\mathrm{CM}} = [\, \mathbf{F}_{\mathrm{shared}}^{\mathrm{HE}}, \mathbf{F}_{\mathrm{shared}}^{\mathrm{ST}}, \mathbf{F}_{\mathrm{unique}}^{\mathrm{HE}}, \mathbf{F}_{\mathrm{unique}}^{\mathrm{ST}} \,]$$

serves as a rich conditioning input to the diffusion reverse process, providing structurally aligned and biologically disentangled guidance for generating HQ ST maps.

Local gradient constraining

Then, we conducted local gradient constraining on the modality-shared features $\mathbf{F}_{\text{shared}}^{\text{HE}}$ and $\mathbf{F}_{\text{shared}}^{\text{ST}}$ for structural alignment between H&E and ST maps. Specifically, gaussian noise is first added to $\mathbf{F}_{\text{shared}}^{\text{ST}}$ and the result is re-normalized to form a local ring distribution on the hypersphere, improving robustness to perturbations.

We then mapped the features to modality-independent structural maps $\mathbf{S}^{\text{ST}}$ and $\mathbf{S}^{\text{HE}}$ via shared convolutional layers. Non-cellular regions in the H&E map are masked out with the cellular mask to yield a boundary-emphasized map $\mathbf{S}_{\text{cell}}^{\text{HE}}$. Horizontal and vertical Sobel operators are applied to compute gradient fields $\nabla\mathbf{S}^{\text{ST}}$ and $\nabla\mathbf{S}_{\text{cell}}^{\text{HE}}$.

A spatial transformer network predicts a deformation field $\boldsymbol{\phi}$ that warps ST space to H&E space,

$$\mathbf{S}^{\text{ST}\to\text{HE}} = \mathbf{S}^{\text{ST}} \circ \boldsymbol{\phi}.$$

Based on $\nabla\mathbf{S}^{\text{ST}\to\text{HE}}$ and $\nabla\mathbf{S}_{\text{cell}}^{\text{HE}}$, we optimized three gradient-based consistency terms. First, the normalized gradient field (NGF) loss constrains edge orientations:

$$\mathcal{L}_{\text{NGF}} = \int_{\Omega} 1 - \frac{\langle \nabla\mathbf{S}^{\text{ST}\to\text{HE}}, \nabla\mathbf{S}_{\text{cell}}^{\text{HE}} \rangle + \delta}{\parallel \nabla\mathbf{S}^{\text{ST}\to\text{HE}} \parallel \ \parallel \nabla\mathbf{S}_{\text{cell}}^{\text{HE}} \parallel + \delta} \, dx,$$

where δ is a stability constant.

Second, the gradient magnitude consistency (GMC) loss aligns edge strengths:

$$\mathcal{L}_{\text{GMC}} = \int_{\Omega} \left\| \parallel \nabla\mathbf{S}^{\text{ST}\to\text{HE}} \parallel - \parallel \nabla\mathbf{S}_{\text{cell}}^{\text{HE}} \parallel \right\|^2 dx.$$

Third, we defined gradient magnitude similarity

$$GMS(x) = \frac{2a(x)b(x) + c}{a(x)^2 + b(x)^2 + c},$$

where $a(x) = \parallel \nabla\mathbf{S}^{\text{ST}\to\text{HE}}(x) \parallel$, $b(x) = \parallel \nabla\mathbf{S}_{\text{cell}}^{\text{HE}}(x) \parallel$ and c prevents division by zero. The gradient magnitude stability constraint (GMSC) is the standard deviation of GMS over Ω,

$$\mathcal{L}_{\text{GMSC}} = \text{Std}_{x\in\Omega}[GMS(x)],$$

penalizing abrupt local variations in gradient magnitude.

A boundary weighting map $\mathbf{W_{bnd}}$ was derived from the normalized gradient of the distance transform of the cellular mask, emphasizing structure around cell contours. The final structural loss is

$$\mathcal{L}_{\text{struct}} = \sum_{x} \mathbf{W_{bnd}}(x)(\mathcal{L}_{\text{NGF}}(x) + \mathcal{L}_{\text{GMC}}(x) + \mathcal{L}_{\text{GMSC}}(x)).$$

2.5 Curriculum learning-based constraining module for profiling imputation in complex regions

We balanced profiling performance across tissues with varying complexity by **(1)** a cellular patch selection and multi-scale fusion scheme, and **(2)** an ST-H&E constraining strategy on cell-cell correlations.

Cellular patch selection and multi-scale fusion scheme

The tissue-level H&E image $\mathbf{h}$ [2,560×2,560 μm] is cropped into non-overlapping cellular-level patches $\{\mathbf{p}_i\}$ [160×160 μm], which are separately encoded by the histology encoder into feature maps $\mathbf{F}_{\text{HE}}$ and $\{\mathbf{F}_i^{\text{cell}}\}$. For each patch, we computed an entropy-based image-structural complexity score

$$\mathrm{c}_i = \mathrm{Comp}(\mathbf{p}_i),$$

based on Shannon entropy of intensity distributions. Curriculum learning is implemented by increasing a complexity threshold over training so that early iterations focus on simple patches and later iterations progressively incorporate more complex regions. At each stage, only patches with c_i above the current threshold are retained, yielding selected cellular-scale features $\mathbf{F}_{\text{cell}}$.

To obtain overall H&E features $\mathbf{F}_{\text{CL}}$, we fused $\mathbf{F}_{\text{cell}}$ with tissue-scale features $\mathbf{F}_{\text{HE}}$ via cross-attention. After pixel flattening, queries $\mathbf{Q}$ is derived from tissue-scale features and keys/values $\mathbf{K}, \mathbf{V}$ from cellular-scale features, with learnable projections W_Q, W_K, W_V, producing

$$\mathbf{F}_{\mathrm{CL}} = \mathrm{Attention}(\mathbf{Q}W_Q, \mathbf{K}W_K, \mathbf{V}W_V).$$

ST and H&E constraining

To mitigate signal mixing in highly complex regions, we identified the cellular-level patch with the highest complexity score

$$\mathbf{p}^* = arg\ \max_{\mathbf{p}_i}\ \mathrm{c}_i,$$

and treated it as the patch most prone to expression mixing.

For each cell j identified within the cell segmentation mask of patch $\mathbf{p}^*$, we computed its spatially averaged embeddings using the modality-specific encoders, i.e., $\mathbf{e}_j^{\mathrm{HE}}$ from the H&E patch and $\mathbf{e}_j^{\mathrm{ST}}$ from the corresponding ST.

We then computed cell-cell correlation matrices for each modality ($\mathbf{C}^{\mathrm{HE}}$ for H&E image and $\mathbf{C}^{\mathrm{ST}}$ for ST) for cells in patch $\mathbf{p}^*$. The (j, k) entry is the Pearson correlation coefficient between the embeddings of cells j and k, i.e., $\mathbf{C}_{jk}^{\mathrm{HE}} = \mathrm{PCC}\left(\mathbf{e}_j^{\mathrm{HE}}, \mathbf{e}_k^{\mathrm{HE}}\right)$ for H&E and $\mathbf{C}_{jk}^{\mathrm{ST}} = \mathrm{PCC}(\mathbf{e}_j^{\mathrm{ST}}, \mathbf{e}_k^{\mathrm{ST}})$ for ST.

Finally, we encouraged the cell-cell correlation matrices across modalities to be similar, allowing histology-derived context to correct transcriptomic mis-profiling in complex regions. We imposed a cosine-similarity-based consistency loss

$$\mathcal{L}_{\mathrm{consis}} = 1 - \frac{cosine\ similarity(\mathbf{C}^{\mathrm{HE}}, \mathbf{C}^{\mathrm{ST}}) + 1}{2},$$

where the cosine similarity is computed after vectorizing the correlation matrices.

2.6 Cross-challenge coordination

To explicitly model cross-challenge coupling, we placed a lightweight cross-challenge coordinator inside the diffusion U-Net to mediate interaction among the expert modules targeted for individual challenges.

For each module m, we first compressed its spatial feature map $\mathbf{F}_m \in \mathbb{R}^{C_m \times H \times W}$ into a fixed-dimensional expert token $\mathbf{z}^{(m)} \in \mathbb{R}^d$ using global pooling followed by a linear

projection. The set of tokens $\{\mathbf{z}^{(m)}\}$ is then processed by a multi-head attention block, allowing each module to query the others and aggregate complementary information. This produces updated tokens $\hat{\mathbf{z}}^{(m)}$ that encode cross-challenge context, such as how misalignment interacts with local resolution or tissue heterogeneity.

Each updated token is mapped, via a multilayer perceptron, to channel-wise scaling and offset parameters $(\boldsymbol{\alpha}_m, \boldsymbol{\beta}_m)$ with the same dimensionality as $\mathbf{F}_m$. The original feature map is recalibrated in a feature-wise linear modulation (FiLM)-like manner,

$$\mathbf{F}'_m = \boldsymbol{\alpha}_m \odot \mathbf{F}_m + \boldsymbol{\beta}_m,$$

and a learnable residual gate mixes $\mathbf{F}'_m$ with $\mathbf{F}^{(m)}$ to preserve module specialization and maintain training stability. The coordinated features $\mathbf{F}'_m$ are finally injected back into the U-Net as residuals to guide subsequent diffusion steps.

3. Challenge coordination score

To indicate the detection and activation status of each challenge in FOCUS, we define a challenge coordination score. Specifically, for module m (sensitivity, alignment, or complexity), the score the score γ_m is computed from the channel-wise scaling parameters $\boldsymbol{\alpha}_m$ associated with the module output features $\mathbf{F}_m$. For a module m with C_m channels, we define

$$\gamma_m = sigmoid\left(\frac{1}{C_m}\sum_{c=1}^{C_m} \alpha_{m,c}\right).$$

A challenge is considered activated when $\gamma_m > 0.5$, indicating that the corresponding module is actively detecting and jointly mitigating that challenge.

For the resolution challenge, the score is derived directly from the spatial resolution of the input ST. We mark the resolution challenge as activated when the map resolution is finer than the single-cell resolution threshold of $10\ \mu$m/pixel.

4. Dataset splitting

Training and *internal validation*: All platforms with high-resolution[8] (HR) ST-H&E pairs were used for model training, with over 0.53 million (80% of total HR ST-H&E pairs) patches, including VisiumHD (144,384 patches), Stereo-seq (8,396 patches), BMK S1000 (24,576 patches), CosMx (18,227 patches), and Xenium (339,148 patches). For internal validation of ST challenges, we included around 1.09 million ST-H&E patches across nine platforms, except for Open-ST for external validation.

External validation: we fine-tuned and tested FOCUS on the previously unseen platform of Open-ST and an unseen clinical dataset of PCP (Xenium), consisting of 28,672 and 101,120 ST-H&E patches, respectively.

5. Foundational training and fine-tuning

Overall, the learning process of FOCUS comprises two stages: (1) cross-platform training with paired H&E-ST data to jointly address diverse challenges across ST platforms, and (2) lightweight fine-tuning on previously unseen platforms or tissue context to enhance real-world generalization.

5.1 Cross-platform training

FOCUS follows a conditional diffusion training paradigm, using HQ ST as the generation target and LQ ST, H&E, cell mask, scRNA-seq and spatial co-expression priors as the primary conditions. During training, Gaussian noise is added to the HQ ST map $\mathbf{s}_0$ according to the diffusion schedule to obtain a noisy sample $\mathbf{s}_t$. The model then predicts the injected noise $\boldsymbol{\epsilon}_t$ and progressively denoises $\mathbf{s}_t$ under multimodal constraints and biological prompts.

The overall training objective for the conditional diffusion model is a weighted multi-term loss. The primary term is the standard diffusion-based noise-prediction loss on HQ ST. Additional terms encourage (i) structural alignment between ST and H&E, and (ii) consistency between morphology-based and expression-based cell-cell relationships

[8] Due to the resolution requirement of the generation target in training diffusion models, we only included platforms providing single-cell resolution (i.e., ≤ 10 μm/pixel[82]) ST maps.

(see Sections 2.4-2.5 in Methods). The overall loss for training the conditional diffusion model ST super-quality is formulated as

$$\mathcal{L}_{\text{all}} = \lambda_1 \mathcal{L}_{\text{charb}} + \lambda_2 \mathcal{L}_{\text{struct}} + \lambda_3 \mathcal{L}_{\text{dis}} + \lambda_4 \mathcal{L}_{\text{consis}},$$

$$\text{where } \mathcal{L}_{\text{charb}} = \sqrt{(\boldsymbol{\epsilon}_\theta(\mathbf{s}_t, \mathbf{y}, \mathbf{h}, \mathbf{m}, \mathbf{S}, \mathbf{M}, \mathbf{P}, t) - \epsilon_t)^2 + \gamma^2}$$

In the above equation, Charbonnier loss $\mathcal{L}_{\text{charb}}$ is a smoother transition between $\mathcal{L}_1$ and $\mathcal{L}_2$ loss, facilitating more steady and accurate convergence during training. γ is a known constant and λ_1 , λ_2 , λ_3 , and λ_4 represent weights for four losses.

Implementation details. We trained FOCUS for 2000 epochs on seven NVIDIA RTX V100 32 GB GPUs, with batch size of 4 and learning rate 0.0001 with AdamW optimizer[83] with the weight decay. Following the sampling strategy in previous works[84], our method uses the sample steps of 1,000 in both forward and reverse processes. Key hyper-parameters are in Table S8, which are tuned to achieve the best performance in all challenges.

5.2 Fine-tuning for unseen platforms and tissue context

To adapt FOCUS to previously unseen ST platforms or tissue context, we employed a freeze-update fine-tuning strategy designed to preserve foundational cross-platform representations while enabling flexible, data-efficient specialization on new domains.

Specifically, to maintain the generic histological and ST representations acquired during large-scale cross-platform training, we kept U-Net encoder frozen. Only higher-level components are updated, including the decoder, the four expert modules, and the conditional fusion/match layers.

In addition, to achieve stable fine-tuning that preserved pre-trained features during transfer to unseen platforms or tissues, we applied Low-Rank Adaptation (LoRA) inside the expert modules. For each weight matrix $\boldsymbol{W} \in \mathbb{R}^{d \times h}$ in a module, we froze $\boldsymbol{W}$ and learn a low-rank update

$$\widehat{\boldsymbol{W}} = \boldsymbol{W} + \mathbf{AB},$$

where $\mathbf{A} \in \mathbb{R}^{d \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times h}$ with $r \ll \min(d, h)$. Only $\mathbf{A}$ and $\mathbf{B}$ are trainable, providing lightweight adaptation while preserving the geometry of the pre-trained features.

Finally, to further flexibly control the modules and achieve customized fine-tuning, we employed a routing mechanism based on expert tokens to combine modules on demand. In detail, we weighed up the residual updates of each module based on the expert tokens generated by the adapter and allowed the corresponding expert modules to be blocked when there was no corresponding demand. Each expert module $\mathbf{M}_i$ produces a residual update $R_i(\mathbf{F})$ to the shared latent features $\mathbf{F}$. A routing network reads the prompt and outputs coefficients α_i for all modules. The combined residual is

$$\mathbf{F}' = \mathbf{F} + \sum_i \alpha_i R_i(\mathbf{F})$$

Modules corresponding to challenges absent in each platform are effectively disabled by setting $\alpha_i = 0$.

6. Cellular complexity definition

To evaluate how well FOCUS corrects mis-profiling in regions of high complexity, we followed prior works[56,57] and defined complexity as a composite index summarizing two H&E-derived properties: (1) cell abundance (how crowded the patch is) and (2) morphological diversity (how heterogeneous the cells are). Formally, the overall complexity $C \in [0,1]$ is a convex combination of a normalized density score $D_{\text{norm}} \in [0,1]$ and a normalized heterogeneity score $H_{\text{norm}} \in [0,1]$:

$$C = \alpha\, D_{\text{norm}} + (1 - \alpha)\, H_{\text{norm}},$$

where α (=0.5) is a balancing weight between cell density and heterogeneity score. High C values therefore indicate patches that are both cell-dense and morphologically diverse, commonly associated with biologically active or pathologic regions (e.g., highly cellular, pleomorphic tumor areas).

Definitions of the cell density and heterogeneity score are as follows.

6.1 Cell density score

Cell density is quantified by normalizing the number of segmented cells N in a patch to an upper bound $N_{\max}$ on expected cell counts, yielding a score that increases with cellular crowding:

$$D_{\text{norm}} = \min\left(\frac{N}{N_{\max}}, 1\right).$$

Note that the cell number upper bound $N_{\max}$ is defined as:

$$N_{\max} = \frac{A_{\mu m^2}}{\pi r_{\min}^2},$$

where $A_{\mu m^2}$ is the patch area in μm^2, and $r_{\min}$ is a lower-bound on cell radius (set to 3 μm as a small lymphocyte-size cell).

6.2 Cell heterogeneity score

Cell heterogeneity is jointly determined by cell shape, size and texture variety, defined as follows.

Shape variety score: Shape heterogeneity captured variability in cellular morphology by summarizing the dispersion of three descriptors: eccentricity z_e, solidity z_s and circularity z_c , into a single score. The combined dispersion is:

$$\Delta_{\text{shape}} = \frac{|\, z_e \,| + |\, z_s \,| + |\, z_c \,|}{3}$$

and the corresponding heterogeneity score is:

$$S_{\text{shape}} = \frac{\Delta_{\text{shape}}}{1 + \Delta_{\text{shape}}},$$

where larger S_{shape} reflects more irregular and diverse nuclear shapes.

Size variety score: Size heterogeneity measured variation in nuclear size using the variability of log-transformed nuclear area $z_{\log A}$ and equivalent diameter z_d . The variability is given by:

$$\Delta_{\text{size}} = \frac{|\, z_{\log A} \,| + |\, z_d \,|}{2}$$

and the normalized score is:

$$S_{\text{size}} = \frac{\Delta_{\text{size}}}{1+\Delta_{\text{size}}},$$

Where patches with a wide range of nuclear sizes attain higher scores.

Texture variety score: The texture heterogeneity score evaluates diversity in chromatin staining patterns. It uses the deviations in hematoxylin entropy z_h and staining intensity variance z_v:

$$\Delta_{\text{texture}} = \frac{|\, z_h \,| + |\, z_v \,|}{2}$$

with the corresponding score:

$$S_{\text{texture}} = \frac{\Delta_{\text{texture}}}{1 + \Delta_{\text{texture}}},$$

where a higher S_{texture} indicates heterogeneous chromatin organization.

Combined heterogeneity: The overall heterogeneity score was obtained as a weighted combination of the shape, size and texture heterogeneity components and was constrained to lie in $[0, 1]$:

$$H_{\text{norm}} = \frac{w_{\text{shape}} S_{\text{shape}} + w_{\text{size}} S_{\text{size}} + w_{\text{texture}} S_{\text{texture}}}{w_{\text{shape}} + w_{\text{size}} + w_{\text{texture}}},$$

where w_{shape}, w_{size}, and w_{texture} are all set to 1.

7. Simulated complexity

Based on the complexity metric C, we first selected patches in the top 5% of C as high-complexity templates. For each selected patch, cell masks were obtained from H&E using our segmentation model. Starting from these masks, we synthetically varied complexity by applying different degrees of filtering of cells, thereby generating a series of patches that span from low to high complexity while preserving the same tissue context. The corresponding modified H&E image with the corresponding under-refined ST were then used to predict high-quality ST, allowing us to evaluate model performance under controlled changes in complexity.

8. Cell clustering with ST

Enhanced ST count matrices were processed with SCANPY package[76]. Library sizes were normalized to 10,000 counts per cell and log-transformed. Highly variable genes were selected using the Seurat v3 procedure[85], and the top 2,000 genes were retained. We performed principal component analysis with 50 components and constructed a K-nearest neighbor graph using the first 30 principal components (K = 15, Euclidean distance). Cluster annotation was performed by comparing top-ranked differentially expressed genes with canonical cell-type markers from literature and public databases.

9. Neighborhood enrichment analysis with ST

We used Squidpy package[86] to quantify spatial neighborhood enrichment between enhanced cell populations in primary HNSCC tumors and matched metastatic lymph nodes. For each sample, a spatial connectivity graph was built from spot coordinates. Neighborhood enrichment scores were computed by comparing the observed number of neighboring spot pairs for each cell-type pair with a null distribution generated by random permutation of cell labels. The resulting enrichment matrices were visualized as heatmaps, with the significant number denoting the number of cell-type pairs with statistically significant enrichment.

10. Spatial cell-cell communication analysis

Ligand-receptor-mediated cell-cell communication was inferred using LIANA package[87] with the integrated consensus resource. Consensus interaction scores across methods were combined via robust rank aggregation, and interaction pairs with a magnitude rank < 0.05 were retained. From these, we constructed a cell-by-ligand-receptor score matrix, applied L1 normalization and decomposed the matrix using non-negative matrix factorization. The number of factors was chosen by the elbow criterion. Spatial maps of factor scores were generated, and the dominant ligand-receptor pairs contributing to each factor were summarized.

11. Cell co-occurrence analysis with ST

In PCP enhanced ST maps, we assessed distance-dependent spatial relationships between neutrophils and other cell populations using Squidpy package[86]. For increasing distance around each neutrophil, we calculated the ratio of the conditional probability of observing a given cell type near neutrophils to its marginal probability in the tissue. Co-occurrence curves thus report enrichment (score > 1) or depletion (score < 1) of each cell population around neutrophils across spatial scales.

12. Spatial niches identification with ST

Spatial niches were defined using the radius-based neighborhood workflow implemented in the SCIMAP package[88]. For each cell, we computed the normalized composition of neighboring cell types within a fixed radius to generate a neighborhood-by–cell-type matrix, which was subsequently subjected to unsupervised clustering (k = 6) to identify recurrent spatial niches in the PCP microenvironment. The relative abundance of annotated cell populations within each niche was summarized and visualized using stacked barplots.

Data availability:

The collected cross-platform multimodal data (including paired ST and H&E images with scRNA-seq references, as well as the segmented cell masks and precomputed spatial gene co-expression matrix) from public sources are summarized in Tables S1-3, and the processed data are now available upon request.

Restrictions apply to the availability of the raw in-house data of PCP (Xenium), which was used with institutional permission through ethics approval for the current study, and are thus not publicly available. Please email all requests for academic use of raw and processed data to the corresponding author. All requests will be evaluated based on institutional and departmental policies to determine whether the data requested is subject to intellectual property or patient privacy obligations. Data can only be shared for non-commercial academic purposes and will require a formal material transfer agreement.

Code availability:

Codes are available at https://github.com/LHY1007/FOCUS/.

Acknowledgements:

This work was supported by the NIHR Brain Injury MedTech Co-operative, the NIHR Cambridge Biomedical Research Centre (NIHR203312), Addenbrooke's Charitable Trust. This work presents independent research funded by the National Institute for Health and Care Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. Xiaofei Wang reports financial support by China Scholarship Council. Chao Li reports financial support by the Guarantors of Brain.

Competing interests:

None.