VIO-ALIGN: A DECOUPLED VISION-OMICS ALIGNMENT METHOD FOR IN SITU CELL PHENOTYPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Routine hematoxylin and eosin (H&E) slides preserve both cellular morphology and surrounding tissue architecture, providing a practical basis for phenotyping cells in their native microenvironment. However, conventional vision models are mainly trained with morphology-defined labels, while molecularly supervised approaches often operate on cropped or prelocalized cells and therefore do not directly connect molecular supervision with raw-patch in situ phenotyping. ViO-Align is introduced as a decoupled vision–omics framework that transfers single-cell molecular supervision to H&E-only in situ cell phenotyping. In Stage I, matched H&E cell crops and transcriptomic profiles are aligned through morphology-conditioned gene gating to learn molecularly guided single-cell representations. In Stage II, cells are localized from uncropped H&E patches using count-guided flow, while center-masked microenvironment aggregation captures complementary local tissue context. Cross-stage alignment transfers Stage-I molecular supervision to raw-patch representations, and downstream phenotyping explicitly combines target-cell morphology with the corresponding microenvironment representation. Only H&E images are required at inference. Across four established cell phenotyping benchmarks, ViO-Align improves macro-F1 by 4.67 absolute points on average over the strongest strictly test-disjoint baseline on each dataset, while maintaining competitive cell localization performance. To more directly evaluate molecularly guided representation learning, five molecularly defined cell-state recognition tasks are further considered. ViO-Align achieves the highest AUROC on all five tasks and improves over the strongest vision baseline by up to 3.86 absolute AUROC points. The learned H&E representations also transfer to patient-disjoint breast cancer diagnosis. Overall, ViO-Align achieves strong and consistent performance across cell phenotyping, molecular-state recognition, and diagnostic transfer, demonstrating robust generalization across cellular and tissue-level tasks. Code is provided in the supplementary material to facilitate reproducibility.

1 INTRODUCTION

Accurate cell phenotyping in H&E-stained whole-slide images is important for studying tumor heterogeneity, immune infiltration, and therapy response. Yet visual representations are mainly optimized by morphology-defined task supervision and may underrepresent features associated with molecular or functional variation Bera et al. (2019); Kather et al. (2020); Saini et al.. Spatial multi-omics offers complementary biological supervision by co-registering histology with molecular measurements such as gene expression Ståhl et al. (2016); Williams et al. (2022).

Existing H&E cell-analysis methods span visual localization and classification, contextual modeling, and cross-modal supervision. Contour-, flow-, and transformer-based models rely mainly on visual supervision Graham et al. (2019); Hörst et al. (2024); Stringer et al. (2021); Cutler et al. (2022), while context-aware methods often use discrete cell graphs. Cross-modal approaches introduce spatial transcriptomic or molecular supervision He et al. (2020); Pang et al. (2021); Tejada-Lapuerta et al. (2025); Hou et al. (2025); Chadoutaud et al. (2026), but commonly assume regional, cropped-cell, or prelocalized-cell inputs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

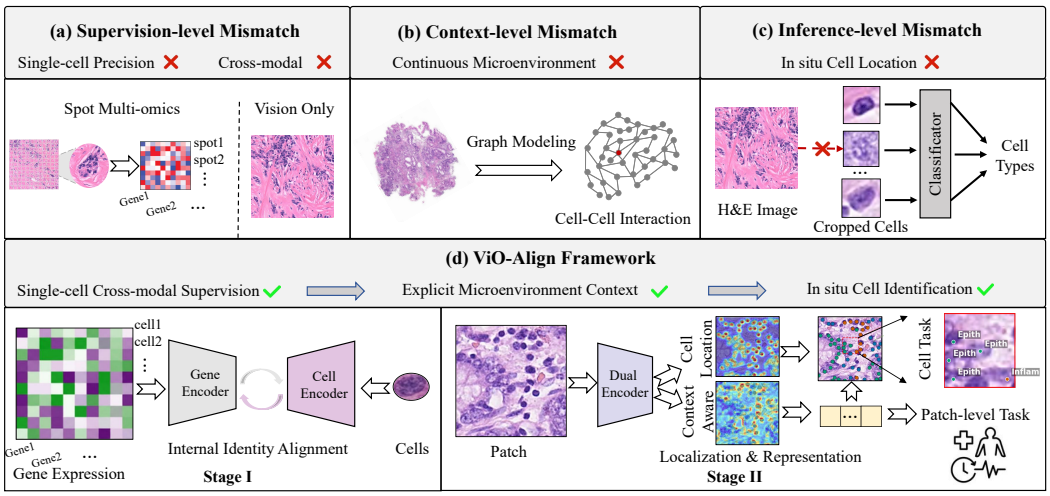

Figure 1: **Major methodological gaps for in situ cell phenotyping.** Existing paradigms only partially connect molecular supervision, microenvironment modeling, and raw-patch in situ inference. ViO-Align integrates these aspects through a decoupled two-stage framework that first learns molecularly guided cell representations and then performs context-aware cell identification from raw H&E patches.

As summarized in Figure 1, three gaps remain: visual representations receive limited molecular guidance, graph-based context may omit continuous histological cues, and molecularly supervised methods often assume cropped or prelocalized cells rather than raw H&E patches. Single-cell molecular guidance, continuous microenvironment modeling, and raw-patch in situ phenotyping therefore remain insufficiently connected.

To bridge these gaps, ViO-Align is introduced as a decoupled vision–omics alignment framework that transfers single-cell molecular guidance to H&E-only in situ cell phenotyping from raw tissue patches. Cell-level H&E–transcriptomics alignment is first performed on matched single cells, where Morphology-conditioned Gene Gating (MGG) uses visual morphology to modulate gene-derived latent representations before cross-modal alignment. Raw H&E patches are subsequently processed with Count-guided Flow Localization (CFL) to identify cell anchors and Microenvironment-aware Aggregation (MEA) to capture continuous local tissue context. Cross-stage alignment transfers the resulting molecularly guided single-cell representations to the raw-patch representation space, while downstream phenotype prediction explicitly combines target-cell morphology with its surrounding microenvironment. Only H&E images are required at inference.

The main contributions are summarized as follows:

- A cell-to-patch molecular supervision transfer framework is established for H&E histology. ViO-Align decouples cell-level vision–transcriptomics alignment from raw-patch representation learning, enabling molecular guidance learned from matched single cells to support H&E-only in situ cell phenotyping.
- Morphology-conditioned Gene Gating is introduced for cell-level H&E–transcriptomics alignment. MGG uses cell morphology to reweight gene-derived latent factors, reducing the influence of transcriptomic variation that is weakly reflected in H&E appearance.
- A raw-patch representation strategy combines count-guided flow localization with center-masked microenvironment aggregation. Target-cell morphology and continuous surrounding tissue context are represented separately and explicitly combined for context-aware phenotype prediction.
- Across four public cell-phenotyping benchmarks, ViO-Align improves macro-F1 by 4.67 absolute points on average over the strongest strictly test-disjoint baseline on each dataset, while maintaining a more compact model size than several recent strong baselines. Additional evaluation on five molecularly defined cell-state tasks yields the highest AUROC

among the compared H&E-only representations, and the learned representations further transfer to patient-disjoint breast cancer diagnosis.

2 Related Work

Vision-based Cell Localization and Phenotyping. Cell localization and phenotyping in H&E images have been addressed using contour-, flow-, and transformer-based models Graham et al. (2019); Stringer et al. (2021); Cutler et al. (2022); Hörst et al. (2026). These methods perform strongly but rely mainly on visual supervision. ViO-Align retains raw-patch localization while introducing molecular guidance and contextual phenotype prediction.

Microenvironment Context Modeling. Neighboring cells and local tissue organization provide complementary phenotypic evidence. Existing approaches often use cell graphs Jaume et al. (2021); Zhou et al. (2019); Lou et al. (2024a); Melton & Krishnan (2023), which discretize tissue into predefined nodes and edges and may omit continuous histological cues. ViO-Align instead aggregates local image-token context around each cell.

Cross-modal Biological Supervision. Spatial transcriptomics and multiplex imaging provide supervision beyond morphology-defined labels, including regional gene-expression prediction Pang et al. (2021); Zeng et al. (2022) and cellular-scale alignment Biancalani et al. (2021); Wei et al. (2022); Hou et al. (2025). Such methods commonly assume cropped or pre-localized cells. ViO-Align instead transfers molecularly guided single-cell representations to cells localized directly from raw H&E patches.

3 Methodology

ViO-Align separates single-cell molecular alignment, raw-patch supervision transfer, and downstream phenotype adaptation. As shown in Figure 2, Stage I learns molecularly guided cell references with MGG, while Stage II uses CFL for contour-independent localization and MEA for continuous local context. Environment representations are aligned with Stage-I references during pretraining; downstream adaptation combines them with direct target-cell morphology, preserving H&E-only deployment.

3.1 Cell-level Vision–Transcriptomics Alignment

Stage I learns molecularly guided single-cell H&E representations from registered cell pairs. For each cell ID i, an H&E crop $\mathbf{x}_i$ and its matched gene-expression profile $\mathbf{t}_i$ correspond to the same physical cell across modalities. Matched cells form positive image–transcriptomics pairs, whereas other cells in the mini-batch serve as negatives.

The image and transcriptomic encoders produce $\mathbf{z}_{\mathrm{img},i}$ and $\mathbf{z}_{\mathrm{tx},i}$, respectively. Directly aligning the complete transcriptomic representation with H&E morphology may introduce molecular variation that is weakly reflected in routine histology. We therefore introduce Morphology-conditioned Gene Gating (MGG), which uses the visual representation of each cell to modulate its gene-derived latent representation.

Specifically, MGG predicts a morphology-conditioned gate from the image representation:

$$\mathbf{g}_i = \sigma(\mathbf{W}_g \mathbf{z}_{\mathrm{img},i} + \mathbf{b}_g)\,, \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The gate is then applied element-wise to the transcriptomic representation:

$$\tilde{\mathbf{z}}_{\mathrm{tx},i} = \mathbf{z}_{\mathrm{tx},i} \odot \mathbf{g}_i, \tag{2}$$

where $\odot$ denotes element-wise multiplication.

Importantly, MGG operates on the latent transcriptomic representation rather than directly masking individual raw genes. The resulting $\tilde{\mathbf{z}}_{\mathrm{tx},i}$ therefore retains gene-derived components that are more compatible with the observed H&E morphology while attenuating components that are less visually reflected.

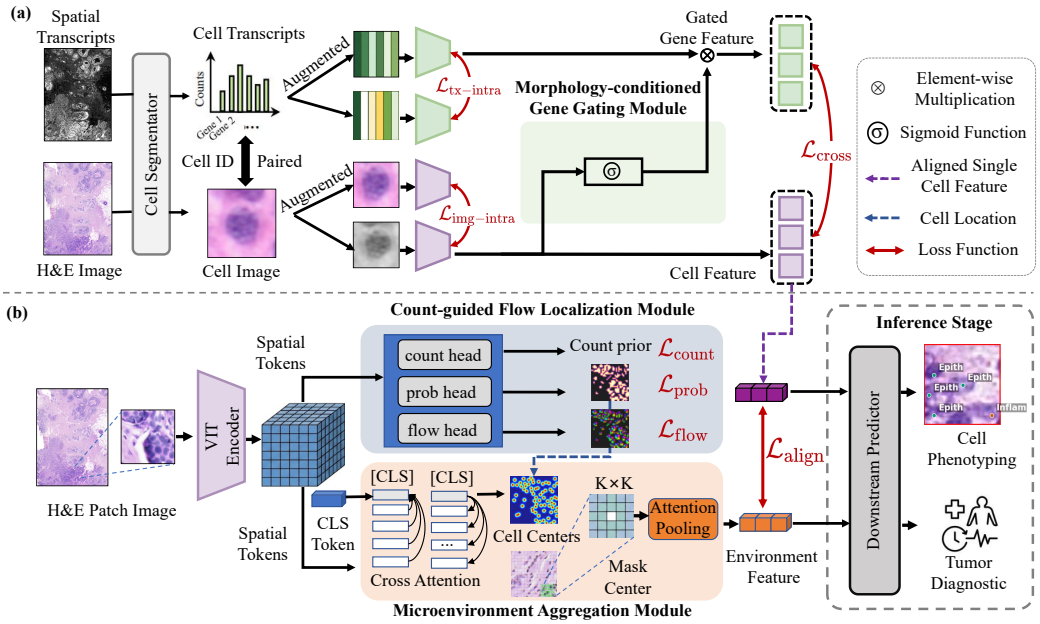

Figure 2: **Framework of the proposed ViO-Align. (a) Stage I: Cell-level Vision–Transcriptomics Alignment.** Paired H&E cell crops and transcriptomic profiles are aligned through morphology-conditioned gene gating to learn molecularly grounded single-cell references. **(b) Stage II: Raw-patch Localization and Molecular Supervision Transfer.** Uncropped H&E patches are processed for cell localization and microenvironment-aware representation learning. During Stage-II pretraining, environment representations are aligned with Stage-I references; during downstream adaptation, target-cell and environment representations are fused for classification. Only H&E images are required at inference.

For two representation sets $\mathbf{Z}_1$ and $\mathbf{Z}_2$, the InfoNCE objective is defined as

$$\mathcal{L}_{\mathrm{NCE}}(\mathbf{Z}_1, \mathbf{Z}_2, \tau) = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(\mathrm{sim}(\mathbf{z}_{1,i}, \mathbf{z}_{2,i})/\tau)}{\sum_{j \in \mathcal{B}} \exp(\mathrm{sim}(\mathbf{z}_{1,i}, \mathbf{z}_{2,j})/\tau)}. \tag{3}$$

The Stage-I objective combines image intra-modal consistency, transcriptomic intra-modal consistency, and cross-modal alignment:

$$\begin{aligned} \mathcal{L}_{\mathrm{I}} = & \mathcal{L}_{\mathrm{NCE}}(\mathbf{Z}_{\mathrm{img}}^{(1)}, \mathbf{Z}_{\mathrm{img}}^{(2)}, \tau_{\mathrm{intra}}) \\ & + \mathcal{L}_{\mathrm{NCE}}(\tilde{\mathbf{Z}}_{\mathrm{tx}}^{(1)}, \tilde{\mathbf{Z}}_{\mathrm{tx}}^{(2)}, \tau_{\mathrm{intra}}) \\ & + \frac{1}{2} \sum_{v=1}^{2} \mathcal{L}_{\mathrm{NCE}}(\mathbf{Z}_{\mathrm{img}}^{(v)}, \tilde{\mathbf{Z}}_{\mathrm{tx}}^{(v)}, \tau_{\mathrm{cross}}). \end{aligned} \tag{4}$$

After Stage-I training, the H&E-side representation

$$\mathbf{z}_{\mathrm{ref},i} = f_{\mathrm{img}}(\mathbf{x}_i)$$

is retained as the molecularly grounded reference for Stage-II supervision. Importantly, $\mathbf{z}_{\mathrm{ref}}$ is a Stage-I reference feature and is distinct from the target-cell representation $\mathbf{z}_{\mathrm{cell}}$ extracted later from predicted downstream instances.

3.2 Count-guided Flow Localization

Given an uncropped H&E patch $\mathbf{X}$, the Stage-II encoder produces spatial features $\mathbf{S} \in \mathbb{R}^{h \times w \times d}$ and a global `[CLS]` representation $\mathbf{u}$. CFL contains lightweight prediction heads for foreground

probability, flow, center response, and patch-level cell count:

$$\mathbf{P} = g_{\text{prob}}(\mathbf{S}), \quad \mathbf{V} = g_{\text{flow}}(\mathbf{S}), \quad \mathbf{H}_c = g_{\text{center}}(\mathbf{S}), \quad \hat{N} = g_{\text{count}}(\mathbf{u}). \tag{5}$$

The foreground map identifies likely cellular regions, while the flow field moves foreground pixels toward their associated instance centers. The center response provides direct center supervision, and the count prediction supplies a patch-level density prior for filtering candidate detections.

During inference, foreground pixels are iteratively propagated as

$$\mathbf{p}^{(t+1)} = \mathbf{p}^{(t)} + \mathbf{V}(\mathbf{p}^{(t)}). \tag{6}$$

The converged points are clustered into candidate instances and subsequently filtered using foreground, center, and count predictions. The retained instances provide spatial anchors for downstream target-cell extraction and microenvironment modeling.

3.3 Microenvironment-aware Aggregation

For each localized center $\mathbf{c}_k$, MEA samples a $K \times K$ local token window

$$\mathbf{H}_k = \text{Sample}(\mathbf{S}, \mathbf{c}_k).$$

Unlike graph-based context modeling based only on predefined cell nodes, the sampled token window preserves continuous histological cues around the cell, including neighboring nuclei, stromal texture, extracellular matrix, and local tissue organization.

To reduce redundant encoding of the target cell in the environment branch, the center token is masked to obtain $\widetilde{\mathbf{H}}_k$. The global patch representation acts as the query, whereas the center-masked environment tokens provide keys and values:

$$\mathbf{q}_k = \mathbf{W}_q \mathbf{u}, \qquad \mathbf{K}_k = \widetilde{\mathbf{H}}_k \mathbf{W}_k, \qquad \mathbf{V}_k = \widetilde{\mathbf{H}}_k \mathbf{W}_v. \tag{7}$$

The environment representation is then computed by

$$\mathbf{z}_{\text{env},k} = \text{softmax}\left(\frac{\mathbf{q}_k \mathbf{K}_k^\top}{\sqrt{d}}\right)\mathbf{V}_k. \tag{8}$$

The resulting $\mathbf{z}_{\text{env},k}$ summarizes local tissue context. Masking the center token reduces direct target-specific shortcuts in the environment branch, while target-cell appearance is retained independently through $\mathbf{z}_{\text{cell},k}$ during downstream adaptation.

3.4 Cross-stage Molecular Supervision Transfer

Stage-II pretraining transfers molecularly guided single-cell supervision from Stage I to cells embedded in raw H&E patches. For each registered cell k, the Stage-II environment representation $\mathbf{z}_{\text{env},k}$ is paired with the Stage-I reference $\mathbf{z}_{\text{ref},k}$ of the same physical cell. Other cell references serve as negatives, while a memory bank $\mathcal{M}$ increases the number of negative references:

$$\mathcal{L}_{\text{align}} = -\frac{1}{K}\sum_{k=1}^{K} \log \frac{e^{\text{sim}(\mathbf{z}_{\text{env},k}, \mathbf{z}_{\text{ref},k})/\tau}}{e^{\text{sim}(\mathbf{z}_{\text{env},k}, \mathbf{z}_{\text{ref},k})/\tau} + \sum_{m \in \mathcal{M}^-} e^{\text{sim}(\mathbf{z}_{\text{env},k}, m)/\tau}}. \tag{9}$$

Here, $\mathbf{z}_{\text{ref}}$ comes from the Stage-I H&E encoder and $\mathbf{z}_{\text{env}}$ from the contextualized raw patch. Registered cell identity defines positive cross-stage pairs without phenotype labels, so Stage-I features serve as supervisory anchors rather than downstream inputs. Transcriptomics is therefore unnecessary at inference.

3.5 Stage-II Pretraining Objective

Stage-II pretraining jointly optimizes localization and cross-stage alignment:

$$\mathcal{L}_{\text{II}} = \lambda_{\text{prob}}\mathcal{L}_{\text{prob}} + \lambda_{\text{flow}}\mathcal{L}_{\text{flow}} + \lambda_{\text{center}}\mathcal{L}_{\text{center}} + \lambda_{\text{count}}\mathcal{L}_{\text{count}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}. \tag{10}$$

The first four terms supervise foreground probability, flow, center response, and cell count, respectively, whereas $\mathcal{L}_{\text{align}}$ transfers molecularly guided Stage-I supervision to the Stage-II representation. Detailed target construction and optimization settings are provided in Appendix C.

Table 1: **Comparison of cell-level representation capacity.** All values denote classification accuracy (%) under the same linear probing protocol. `Avg.` denotes the average accuracy across four datasets.

Input modals	Implementation	CoNSeP	PanNuke	MoNuSAC	GLySAC	Avg.
H&E only	Volta	53.46	56.49	82.21	34.30	56.62
H&E + IF/protein	AutoCell	65.92	**77.87**	87.96	46.54	69.57
H&E + Transcriptomics	ViO-Align	**81.30**	77.08	**91.43**	**60.80**	**77.65**

3.6 Downstream Cell Phenotyping and Inference

Downstream phenotype adaptation is separated from Stage-II cross-stage pretraining, and the contrastive alignment against $\mathbf{z}_{\mathrm{ref}}$ is not recomputed.

For each predicted instance k, the corresponding target-cell H&E crop $\hat{\mathbf{x}}_k$ is encoded using the Stage-I image encoder:

$$\mathbf{z}_{\mathrm{cell},k} = f_{\mathrm{img}}(\hat{\mathbf{x}}_k). \tag{11}$$

Unlike $\mathbf{z}_{\mathrm{ref},k}$, which provides supervision for a registered cell during Stage-II pretraining, $\mathbf{z}_{\mathrm{cell},k}$ is extracted from the currently predicted cell instance during downstream adaptation and inference.

The target-cell representation is combined with its environment representation through element-wise addition followed by normalization:

$$\mathbf{z}_{\mathrm{fuse},k} = \mathrm{normalize}(\mathbf{z}_{\mathrm{cell},k} + \mathbf{z}_{\mathrm{env},k}), \qquad \hat{y}_{\mathrm{cell},k} = \mathrm{MLP}_{\mathrm{cell}}(\mathbf{z}_{\mathrm{fuse},k}). \tag{12}$$

The fused representation combines target-cell morphology with surrounding tissue context. During joint fine-tuning, the visual encoders, localization modules, and phenotype predictor are optimized with task-specific supervision; the transcriptomic encoder and MGG are not used, and inference is H&E-only.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. Upstream pretraining uses 15 registered Xenium H&E–transcriptomics datasets Janesick et al. (2023): Stage I uses 128×128 cell crops and Stage II uncropped 256×256 patches. Downstream H&E-only evaluation uses CoNSeP Graham et al. (2019), PanNuke Gamper et al. (2019), MoNuSAC Verma et al. (2021), and GLySAC Doan et al. (2022) for cell phenotyping, additional Xenium cohorts for molecularly defined cell-state recognition, and BreaKHis Spanhol et al. (2015) for diagnostic transfer. F_d, macro phenotype F1 (F_{avg}), class-wise F1, and cell-state AUROC are reported. Stage I/II use 90/10 train–validation splits; downstream train/test partitions are 27/14 (CoNSeP), 34/25 (GLySAC), 168/41 (MoNuSAC), and Fold 1/Fold 3 (PanNuke). Validation is drawn only from training data, with test data excluded from model selection.

Baselines. For cell localization and phenotyping, ViO-Align is compared with DIST Naylor et al. (2018), DSF-CNN Graham et al. (2020), SONNET Doan et al. (2022), CGT Lou et al. (2024a), SENC Lou et al. (2024b), CP-Mamba Zhang et al. (2025), HoVer-Net Graham et al. (2019), and CellViT++ Hörst et al. (2026). Volta Nakhli et al. (2024) and AutoCell Li et al. (2025) are used for Stage-I linear probing.

Implementation. Stage I uses a ResNet-50-style H&E encoder and 4-layer GeneTransformer; Stage II uses a 6-layer ViT with stride 8, a 9×9 context window, and a memory bank. For Table 2, both visual encoders are unfrozen and jointly fine-tuned with CFL and the phenotype predictor; the transcriptomic encoder and MGG are omitted downstream. Training uses at most 100 epochs (minimum 10) with early-stopping patience 8. Further details appear in Appendices A and C.

4.2 Effectiveness of Representations on Downstream Tasks

Cell-level representation analysis. Table 1 shows that ViO-Align achieves the highest average linear-probing accuracy and performs best on CoNSeP, MoNuSAC, and GLySAC. AutoCell

Table 2: **Performance comparison with baselines on four cell phenotyping benchmarks.** F_d denotes detection F1-score and F_{avg} denotes macro-average phenotype F1-score. Results for DIST, DSF-CNN, SONNET, CGT, SENC, and CP-Mamba follow the previously reported CP-Mamba comparison. HoVer-Net and CellViT++ are evaluated using the same benchmark splits and evaluation pipeline as ViO-Align. CellViT++ on PanNuke is omitted because its released pretrained backbone is not strictly test-disjoint from the Fold-3 test set. Best and second-best results are shown in **bold** and underlined, respectively.

Methods	Parameters	GLySAC		CoNSeP		MoNuSAC		PanNuke	
		F_d	F_{avg}	F_d	F_{avg}	F_d	F_{avg}	F_d	F_{avg}
DIST	9.2M	80.11	39.47	71.39	48.10	60.48	44.29	71.80	32.24
DSF-CNN	3.7M	82.79	42.56	66.31	33.06	82.11	61.40	78.28	40.88
SONNET	41.8M	83.29	45.67	73.96	54.69	83.63	66.24	80.02	50.05
CGT	37.4M	85.76	47.59	74.22	53.69	81.59	63.46	78.68	50.89
SENC	37.4M	84.59	47.30	73.17	53.49	75.82	62.08	76.78	49.21
CP-Mamba	33.8M	86.94	48.52	75.33	56.80	82.14	67.55	81.70	52.87
HoVer-Net	54.7M	86.79	57.97	75.76	63.12	82.58	63.95	80.61	50.89
CellViT++	46.8M	86.37	56.13	76.03	62.11	46.15	29.50	–	–
ViO-Align	19.6M	**87.07**	**68.68**	**76.18**	**64.22**	**83.96**	**73.68**	**82.22**	**53.61**

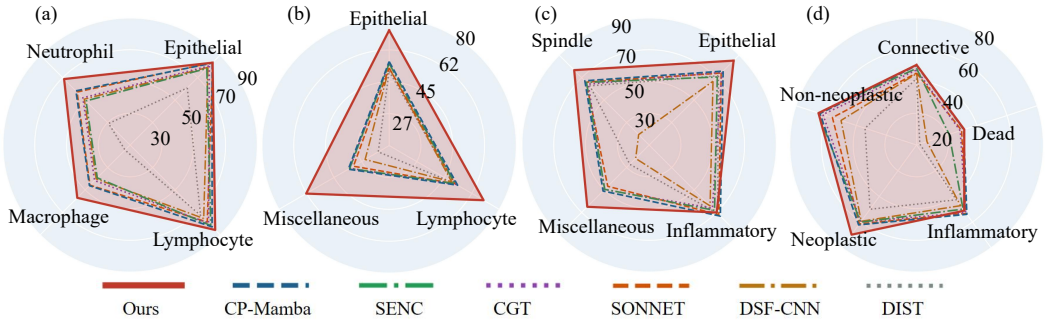

Figure 3: **Class-wise F1-score comparison.** **(a)** MoNuSAC, **(b)** GLySAC, **(c)** CoNSeP, and **(d)** PanNuke.

is slightly stronger on PanNuke, potentially benefiting from additional large-scale H&E self-supervision and domain adaptation.

In-situ cell localization and phenotyping. Table 2 shows the highest strictly test-disjoint F_{avg} on all four datasets. Against the strongest eligible baseline, gains are 10.71, 1.10, 6.13, and 0.74 points on GLySAC, CoNSeP, MoNuSAC, and PanNuke, averaging 4.67 points, while localization remains competitive with 19.6M parameters. Figure 3 shows improvements across multiple phenotype families, including epithelial, lymphoid, myeloid, and neoplastic cells.

Label efficiency. With only 1–25% of downstream labels, ViO-Align remains competitive and at 25% exceeds the reported full-label comparator on three benchmarks (Appendix D.3), indicating useful transfer under limited phenotype supervision.

Molecularly defined cell-state recognition. Five additional binary tasks evaluate whether H&E representations retain information associated with molecularly defined cellular states. They cover proliferative versus activated B cells, breast-tumor hypoxia/glycolysis, IFN/inflammatory activity, proliferative versus generic tumor cells, and luminal versus basal expression programs. As shown in Table 3, ViO-Align achieves the highest AUROC on all five tasks among the evaluated H&E-only baselines.The B-cell task uses three reactive lymphoid specimens with leave-one-specimen-out evaluation, whereas hypoxia, IFN/inflammation, and proliferative-versus-generic tasks use non-overlapping spatial blocks within S1_Mid. Labels come from predefined transcriptomic programs or Xenium cell-group annotations; inference uses H&E representations only.

Table 3: **H&E-only recognition of molecularly defined cell states.** AUROC is reported separately for each task. B-cell and luminal–basal tasks use specimen-held-out evaluation; the three S1_Mid breast tasks use same-section spatial-block holdout and are not averaged across protocols.

Cell-state task	ViO-Align	CTransPath	CellViT	HoVer-Net
B-cell proliferation vs. activation	**0.6862**	0.6636	0.6779	0.6533
Breast hypoxia/glycolysis	**0.7411**	0.6737	0.7000	0.7025
Breast IFN/inflammation	**0.6970**	0.6281	0.6701	0.6526
Breast proliferative vs. generic tumor	**0.6585**	0.6252	0.6432	0.5937
Breast luminal vs. basal program	**0.6003**	0.5793	0.5883	0.5675

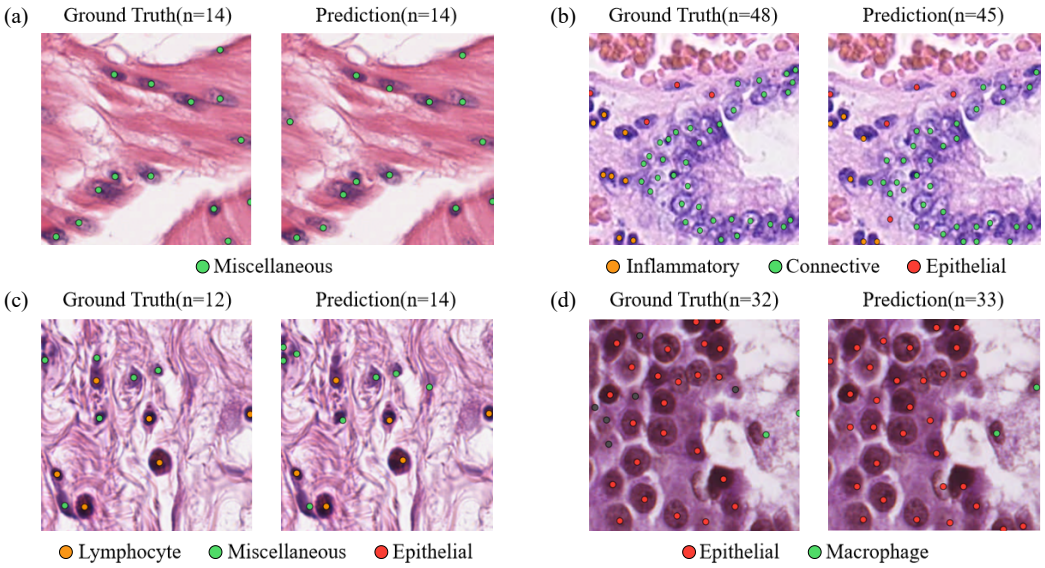

Figure 4: **Visualization of in situ cell localization and phenotyping.** Representative H&E patches from four benchmarks are shown: (a) CoNSeP, (b) PanNuke, (c) MoNuSAC, and (d) GLySAC. Colored dots denote localized cell centers and phenotype categories. The cell count n is reported for each patch.

Qualitative results. Figure 4 presents representative localization and phenotyping results across the four benchmarks. ViO-Align shows close spatial agreement with the ground-truth cell centers across regions with varying cell densities and phenotype compositions, while largely preserving the annotated distribution of different cell populations. Remaining errors mainly arise from missed or additional detections in crowded regions and occasional local phenotype mismatches. These results qualitatively complement the quantitative evaluation and demonstrate consistent in situ localization and phenotyping across heterogeneous H&E tissue regions.

Patch-level diagnostic transfer. With frozen Stage-II features and five-fold patient-disjoint evaluation on BreaKHis, ViO-Align obtains $90.29 \pm 3.87\%$ accuracy, $91.32 \pm 3.16\%$ balanced accuracy, $89.31 \pm 3.90\%$ macro-F1, $95.09 \pm 4.94\%$ ROC-AUC, $88.42 \pm 6.27\%$ sensitivity, and $94.22 \pm 5.35\%$ specificity, supporting transfer beyond cell phenotyping.

4.3 Ablation Studies

Effect of MGG. Removing MGG reduces Stage-I Top-1 retrieval from 92.88% to 33.14%. Table 4 further shows that MGG improves average matched-cell macro-F1 from 65.60 to 68.00, supporting both cross-modal alignment and downstream representation transfer.

Table 4: **Ablation of MGG.** Values denote matched-cell macro-F1 (%). Impr. denotes the absolute improvement brought by MGG.

Setting	CoNSeP	GLySAC	MoNuSAC	PanNuke	Avg.
w/o MGG	60.65	63.16	69.34	69.23	65.60
w/ MGG	**64.90**	**65.46**	**71.86**	**69.79**	**68.00**
Impr.	+4.25	+2.30	+2.52	+0.56	+2.40

Table 5: **Ablation of CFL.** F_d denotes detection F1 (%) and F_{avg} denotes macro-F1 (%). Impr. denotes the absolute improvement brought by CFL.

Setting	CoNSeP		GLySAC		MoNuSAC		PanNuke		Avg.	
	F_d	F_{avg}	F_d	F_{avg}	F_d	F_{avg}	F_d	F_{avg}	F_d	F_{avg}
w/o CFL	59.91	57.28	72.84	55.46	50.71	38.62	68.37	44.92	62.96	49.07
w/ CFL	**76.18**	**64.22**	**87.07**	**68.68**	**83.96**	**73.68**	**82.22**	**53.61**	**82.36**	**65.05**
Impr.	+16.27	+6.94	+14.23	+13.22	+33.25	+35.06	+13.85	+8.69	+19.40	+15.98

Table 6: **Ablation of MEA.** Values denote classification accuracy (%) under fixed ground-truth cell centers. Impr. denotes the absolute improvement brought by MEA.

Setting	CoNSeP	GLySAC	MoNuSAC	PanNuke	Avg.
w/o MEA	81.30	60.80	91.43	77.08	77.65
w/ MEA	**83.63**	**73.94**	**97.33**	**78.86**	**83.44**
Impr.	+2.33	+13.14	+5.90	+1.78	+5.79

Effect of CFL. Compared with center-offset regression, CFL improves average detection F1 and macro-F1 by 19.40 and 15.98 points, respectively (Table 5). Joint localization–alignment training further improves detection F1 by 1.54 points on Stage-II validation and 2.46 points on unseen data before fine-tuning (Appendix D.4).

Effect of MEA. With ground-truth centers fixed, MEA improves classification accuracy by 5.79 points on average across four datasets (Table 6). Because target-cell evidence is retained in both settings, this comparison isolates the complementary contribution of the surrounding tissue context rather than replacing direct cell recognition. Performance is stable for $K = 4, 9, 16$, with only 0.62 points of average macro-F1 variation (Appendix D.6).

5 Conclusion

ViO-Align establishes a cell-to-patch molecular supervision transfer framework for H&E-only in situ cell phenotyping. Molecularly guided single-cell representations learned from paired H&E–transcriptomics data are transferred to raw tissue representations, where target-cell morphology and continuous microenvironment context are jointly exploited for phenotype prediction. Across four public cell-phenotyping benchmarks, ViO-Align consistently improves phenotype recognition while maintaining competitive localization performance and a relatively compact model size. Evaluation on five molecularly defined cell-state tasks further shows that the learned H&E representations retain information associated with molecular variation, while additional experiments demonstrate transfer to patient-disjoint breast cancer diagnosis and effectiveness under limited downstream supervision. Importantly, molecular measurements are required only during upstream pretraining, whereas downstream deployment relies exclusively on routine H&E images. These results support ViO-Align as a practical framework for transferring molecular guidance into scalable H&E-based cellular analysis across both cell-level and tissue-level tasks.

AI USE STATEMENT

Generative AI tools were used to assist with feedback on experimental presentation and research methodology, as well as with language refinement and restructuring of parts of the manuscript. They were also used to improve readability, summarize and reorganize author-provided material, and assist with LaTeX editing. All experimental designs, implementations, numerical results, figures, tables, scientific claims, and final methodological decisions were reviewed and verified by the authors against the underlying experiments and source materials. Generative AI tools were not used to generate experimental data or replace the authors' evaluation of the reported results. The authors take responsibility for the final content of this work, including all text and claims produced with the assistance of generative AI.

REFERENCES

Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nature reviews Clinical oncology, 16(11):703–715, 2019.

Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R Vanderburg, Åsa Segerstolpe, Meng Zhang, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. Nature methods, 18(11): 1352–1362, 2021.

Loïc Chadoutaud, Marvin Lerousseau, Daniel Herrero-Saboya, Julian Ostermaier, Jacqueline Fontugne, Emmanuel Barillot, and Thomas Walter. scellst predicts single-cell gene expression from h& e images. Nature Communications, 17(1):1194, 2026.

Kevin J Cutler, Carsen Stringer, Teresa W Lo, Luca Rappez, Nicholas Stroustrup, S Brook Peterson, Paul A Wiggins, and Joseph D Mougous. Omnipose: a high-precision morphology-independent solution for bacterial cell segmentation. Nature methods, 19(11):1438–1448, 2022.

Tan NN Doan, Boram Song, Trinh TL Vuong, Kyungeun Kim, and Jin T Kwak. Sonnet: A self-guided ordinal regression neural network for segmentation and classification of nuclei in large-scale multi-tissue histology images. IEEE Journal of Biomedical and Health Informatics, 26(7): 3218–3228, 2022.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In European congress on digital pathology, pp. 11–19. Springer, 2019.

Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical image analysis, 58:101563, 2019.

Simon Graham, David Epstein, and Nasir Rajpoot. Dense steerable filter cnns for exploiting rotational symmetry in histology images. IEEE Transactions on Medical Imaging, 39(12):4124–4136, 2020.

Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. Nature biomedical engineering, 4(8):827–834, 2020.

Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, et al. Cellvit: Vision transformers for precise cell segmentation and classification. Medical image analysis, 94:103143, 2024.

Fabian Hörst, Moritz Rempe, Helmut Becker, Lukas Heine, Julius Keyl, and Jens Kleesiek. Cellvit++: Energy-efficient and adaptive cell segmentation and classification using foundation models. Computer Methods and Programs in Biomedicine, pp. 109206, 2026.

Naiqiao Hou, Yue Yu, Zhaorun Wu, Yanqi Zhang, Jiayun Wu, Zhixing Zhong, Fangfang Xu, Zeyu Wang, Chaoyong Yang, Weihong Tan, et al. Spatiocell: Deep integration of histology and spatial transcriptomics for profiling the cellular microenvironment at single-cell level. bioRxiv, pp. 2025–04, 2025.

Amanda Janesick, Robert Shelansky, Andrew D Gottscho, Florian Wagner, Stephen R Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A Morrison, Michelli F Oliveira, Jordan T Sicherman, et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. Nature communications, 14(1):8353, 2023.

Guillaume Jaume, Pushpak Pati, Valentin Anklin, Antonio Foncubierta, and Maria Gabrani. Histocartography: A toolkit for graph analytics in digital pathology. In MICCAI Workshop on Computational Pathology, pp. 117–128. PMLR, 2021.

Jakob Nikolas Kather, Lara R Heij, Heike I Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M Niehues, Kai AJ Sommer, Peter Bankhead, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nature cancer, 1(8):789–799, 2020.

Zhe Li, Seyed Hossein Mirjahanmardi, Rasoul Sali, Feyisope Eweje, Matthew Gopaulchan, Leon Kloker, Xiaoming Zhang, Guoxin Li, Yuming Jiang, and Ruijiang Li. Automated cell annotation and classification on histopathology for spatial biomarker discovery. Nature Communications, 16 (1):6240, 2025.

Wei Lou, Guanbin Li, Xiang Wan, and Haofeng Li. Cell graph transformer for nuclei classification. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 3873–3881, 2024a.

Wei Lou, Xiang Wan, Guanbin Li, Xiaoying Lou, Chenghang Li, Feng Gao, and Haofeng Li. Structure embedded nucleus classification for histopathology images. IEEE Transactions on Medical Imaging, 43(9):3149–3160, 2024b.

Joshua Melton and Siddharth Krishnan. muxgnn: Multiplex graph neural network for heterogeneous graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9):11067–11078, 2023.

Ramin Nakhli, Katherine Rich, Allen Zhang, Amirali Darbandsari, Elahe Shenasa, Amir Hadjifaradji, Sidney Thiessen, Katy Milne, Steven JM Jones, Jessica N Mcalpine, et al. Volta: an environment-aware contrastive cell representation learning for histopathology. Nature Communications, 15(1):3942, 2024.

Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. IEEE transactions on medical imaging, 38 (2):448–459, 2018.

Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. BioRxiv, pp. 2021–11, 2021.

Agampreet Saini, Supragya Gandotra, Abhijit Kumar, Pratik Dutta, and Tirthankar Ghosal. Artificial intelligence for spatial transcriptomics: A scoping review of architectures and models. In NeurIPS 2025 Workshop for Imageomics: Discovering Biological Knowledge from Images Using AI.

Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. Ieee transactions on biomedical engineering, 63 (7):1455–1462, 2015.

Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science, 353 (6294):78–82, 2016.

Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. Nature methods, 18(1):100–106, 2021.

Alejandro Tejada-Lapuerta, Anna C Schaar, Robert Gutgesell, Giovanni Palla, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Till Richter, et al. Nicheformer: a foundation model for single-cell and spatial omics. Nature methods, pp. 1–14, 2025.

Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, Simon Graham, Quoc Dang Vu, Mieke Zwager, Shan E Ahmed Raza, Nasir Rajpoot, et al. Monusac2020: A multi-organ nuclei segmentation and classification challenge. IEEE Transactions on Medical Imaging, 40(12):3413–3423, 2021.

Runmin Wei, Siyuan He, Shanshan Bai, Emi Sei, Min Hu, Alastair Thompson, Ken Chen, Savitri Krishnamurthy, and Nicholas E Navin. Spatial charting of single-cell transcriptomes in tissues. Nature biotechnology, 40(8):1190–1199, 2022.

Cameron G Williams, Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo, and Ashraful Haque. An introduction to spatial transcriptomics for biomedical research. Genome medicine, 14(1):68, 2022.

Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. Briefings in Bioinformatics, 23(5):bbac297, 2022.

Ye Zhang, Zijie Fang, Yifeng Wang, Lingbo Zhang, Xianchao Guan, and Yongbing Zhang. Category prompt mamba network for nuclei segmentation and classification. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pp. 10284–10292, 2025.

Yanning Zhou, Simon Graham, Navid Alemi Koohbanani, Muhammad Shaban, Pheng-Ann Heng, and Nasir Rajpoot. Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images. In Proceedings of the IEEE/CVF international conference on computer vision workshops, pp. 0–0, 2019.

A Dataset Details and Preprocessing Protocols

A.1 Upstream Spatial Transcriptomics Cohort

A large-scale pan-tissue spatial transcriptomics cohort is compiled from the 10x Genomics Xenium platform to establish intrinsic morphology–omics alignment. The cohort spans 15 whole-slide datasets, including 10 cancerous and 5 non-cancerous or reactive tissues, as summarized in Table 7.

Table 7: Composition of the 10x Genomics Xenium pre-training cohort.

Dataset	Organ	Condition	Num Cells	Num Genes
Colon Normal (FFPE)	Colon	Non-diseased	270,912	541
Heart Normal (FFPE)	Heart	Non-diseased	26,356	541
Kidney Normal	Kidney	Non-diseased	97,550	541
Liver Normal (FFPE)	Liver	Non-diseased	239,247	541
Tonsil Reactive Hyperplasia	Tonsil	Reactive	1,349,530	541
Breast Cancer Rep1	Breast	Cancer	166,541	541
Brain GBM (FFPE)	Brain	Cancer	816,625	541
Colorectal Cancer (FFPE)	Colon	Cancer	388,132	541
Breast Ductal Adenocarcinoma	Breast	Cancer	234,959	541
Kidney Cancer + Protein	Kidney	Cancer	465,312	516
Lung Cancer (FFPE)	Lung	Cancer	278,456	541
Ovary Cancer (FF)	Ovary	Cancer	205,049	541
Bone Marrow ALL	Bone Marrow	Cancer	225,675	541
Pancreatic Cancer (FFPE)	Pancreas	Cancer	190,965	538
Skin Melanoma (FFPE)	Skin	Cancer	106,897	541

Table 8: Details of downstream cell phenotyping and macroscopic diagnostic benchmarks.

Dataset	Task	Organ/Tissue	Images	Annotated Nuclei	Classes
GLySAC	Cell Phenotyping	Gastric	59 (1000^2)	30,975	3 (Lymphocyte, Epithelial, Misc)
CoNSeP	Cell Phenotyping	Colorectal	41 (1000^2)	24,319	4 (Misc, Inflammatory, Epithelial, Spindle)
MoNuSAC	Cell Phenotyping	Multi-organ	209 (Variable)	31,411	4 (Epithelial, Lymphocyte, Macrophage, Neutrophil)
PanNuke	Cell Phenotyping	Pan-cancer (19)	7,899 (256^2)	189,744	5 (Neoplastic, Non-neoplastic, Inflammatory, Connective, Dead)
BreaKHis	Macro Diagnostics	Breast	7,909 (256^2)	N/A (82 Patients)	2 (Benign vs. Malignant)

Gene Matrix and Registration. Transcript profiles from `transcripts.parquet` are filtered for valid cells and converted into a compressed sparse row (CSR) matrix using raw counts, without global normalization, to preserve feature magnitude. Transcriptomic coordinates p_{xenium} are mapped to H&E pixel coordinates p_{he} using the provided affine transformation matrix M:

$$p_{\text{he}} = p_{\text{xenium}} \times M_{\text{inv}}^{T}.$$

Crop Extraction. For Stage I, single-cell crops are generated by expanding precise cell bounding boxes by a factor of 1.3, with a minimum crop size of 32×32, followed by padding and resizing to 128×128. For Stage II, uncropped 256×256 H&E patches are extracted, and polygon-area centroids computed from `nucleus_boundaries` via the shoelace formula are used as geometric anchors.

A.2 Downstream Benchmarks

Table 8 summarizes the four public cell phenotyping datasets and the BreaKHis macroscopic diagnostic benchmark. The fixed train/test partitions are 27/14 images for CoNSeP, 34/25 for GLySAC, 168/41 for MoNuSAC, and Fold 1/Fold 3 for PanNuke; validation subsets are drawn only from the corresponding training partitions. Test samples are never used for tuning, early stopping, or checkpoint selection. All downstream cell-phenotyping datasets are evaluated at nominal $40\times$ scale (approximately 0.25–0.275 μm/pixel), consistent with the level-0 Xenium H&E images used upstream (approximately 0.2738 μm/pixel).

B Algorithmic Pipeline of ViO-Align

To clarify the overall training and inference procedure, the complete ViO-Align pipeline is summarized in Algorithm 1.

Algorithm 1 Overall Pipeline of ViO-Align

Require: Paired single-cell H&E crops $\{\mathbf{x}_i\}$, matched transcriptomic profiles $\{\mathbf{t}_i\}$, raw H&E patches $\{\mathbf{X}\}$, registered cell centers $\{\mathbf{c}_i\}$, and cell IDs.

Ensure: Localized cell centers and phenotype predictions from H&E patches.

▷ **(I) Cell-level Vision–Transcriptomics Alignment**

for each mini-batch of paired cells **do**

Encode H&E crops:

$$\mathbf{z}_{\mathrm{img},i} = f_{\mathrm{img}}(\mathbf{x}_i).$$

Encode transcriptomic profiles:

$$\mathbf{z}_{\mathrm{tx},i} = f_{\mathrm{tx}}(\mathbf{t}_i).$$

Generate morphology-conditioned gates:

$$\mathbf{g}_i = \sigma(\mathbf{W}_g \mathbf{z}_{\mathrm{img},i} + \mathbf{b}_g).$$

Obtain gated transcriptomic representations:

$$\tilde{\mathbf{z}}_{\mathrm{tx},i} = \mathbf{z}_{\mathrm{tx},i} \odot \mathbf{g}_i.$$

Optimize image intra-modal, transcriptomic intra-modal, and cross-modal contrastive objectives.

end for

Store the molecularly guided H&E reference:

$$\mathbf{z}_{\mathrm{ref},i} = f_{\mathrm{img}}(\mathbf{x}_i).$$

▷ **(II) Raw-patch Localization and Molecular Supervision Transfer**

for each mini-batch of registered raw H&E patches **do**

Extract spatial and global patch representations:

$$(\mathbf{S}, \mathbf{u}) = F_{\mathrm{patch}}(\mathbf{X}).$$

Predict foreground probability, flow field, center response, and patch-level cell count:

$$\mathbf{P}, \mathbf{V}, \mathbf{H}_c, \hat{N} = g_{\mathrm{CFL}}(\mathbf{S}, \mathbf{u}).$$

Compute foreground, flow, center, and count localization losses.

for each registered cell center $\mathbf{c}_k$ **do**

Sample a local token window around $\mathbf{c}_k$ from $\mathbf{S}$.

Mask the center token to reduce direct target-cell information in the environment branch.

Aggregate the center-masked local context to obtain

$$\mathbf{z}_{\mathrm{env},k}.$$

Retrieve the corresponding Stage-I reference $\mathbf{z}_{\mathrm{ref},k}$ using the same cell ID.

end for

Align $\mathbf{z}_{\mathrm{env},k}$ with $\mathbf{z}_{\mathrm{ref},k}$ using contrastive learning and a memory bank.

Jointly optimize localization losses and $\mathcal{L}_{\mathrm{align}}$.

end for

▷ **(III) Downstream Phenotype Adaptation**

for each downstream H&E patch **do**

Localize cell instances using CFL.

for each localized cell instance k **do**

Extract the corresponding target-cell crop $\hat{\mathbf{x}}_k$ and encode its morphology:

$$\mathbf{z}_{\mathrm{cell},k} = f_{\mathrm{img}}(\hat{\mathbf{x}}_k).$$

Obtain the associated environment representation $\mathbf{z}_{\mathrm{env},k}$ using MEA.

Fuse target-cell morphology and local context:

$$\mathbf{z}_{\mathrm{fuse},k} = \mathrm{normalize}\left(\mathbf{z}_{\mathrm{cell},k} + \mathbf{z}_{\mathrm{env},k}\right).$$

Predict the phenotype:

$$\hat{y}_{\mathrm{cell},k} = \mathrm{MLP}_{\mathrm{cell}}(\mathbf{z}_{\mathrm{fuse},k}).$$

end for

Fine-tune the visual encoders, localization modules, and phenotype predictor with downstream supervision.

end for
▷ **(IV) H&E-only Inference**
Given an unseen H&E patch $\mathbf{X}$, extract patch representations with F_{patch}.
Use CFL to obtain candidate cell centers from $\mathbf{P}$, $\mathbf{V}$, $\mathbf{H}_c$, and $\hat{N}$.
for each localized cell instance k **do**
    Encode the target-cell crop as $\mathbf{z}_{\text{cell},k}$.
    Use MEA to obtain $\mathbf{z}_{\text{env},k}$.
    Compute

$$\mathbf{z}_{\text{fuse},k} = \text{normalize}\left(\mathbf{z}_{\text{cell},k} + \mathbf{z}_{\text{env},k}\right).$$

    Predict

$$\hat{y}_{\text{cell},k} = \text{MLP}_{\text{cell}}(\mathbf{z}_{\text{fuse},k}).$$

end for
return Localized cell centers and phenotype labels.

C Extended Implementation and Architectural Details

ViO-Align is implemented in PyTorch. The architectural configurations for Stage I and Stage II are summarized in Table 9, and the main training hyperparameters are listed in Table 10.

Downstream Fine-tuning Protocol. For the four cell-phenotyping benchmarks, the Stage-I H&E encoder and Stage-II patch encoder are initialized from ViO-Align pretraining and remain unfrozen during downstream optimization together with CFL and the phenotype predictor. The transcriptomic encoder and MGG are not used downstream. Training uses at most 100 epochs, a minimum of 10 epochs, and early-stopping patience of 8. Model selection is based only on the validation subset drawn from the training partition. HoVer-Net and CellViT++ use the same benchmark splits, class mappings, center-matching rule, model-selection procedure, and evaluator as ViO-Align; the remaining Table 2 baseline values follow the previously reported CP-Mamba comparison.

Flow Field Generation. The continuous 2D target flow field V^* is generated by treating ground-truth centers as absorbing sinks. For foreground pixels, V^* is defined as the normalized spatial gradient toward the corresponding centroid. During training, a masked mean squared error loss enforces the target flow within cell boundaries, while an L_2 regularization term ($\gamma = 0.1$) suppresses spurious background flow. At inference time, Euler integration over 100 iterations is used to recover sub-pixel center estimates.

D Additional Quantitative and Qualitative Analyses

D.1 Cross-Modal Alignment and Tissue-Level Organization

Figure 5 provides a qualitative view of the Stage I representation space. The embeddings organize primarily by tissue domain rather than by modality, with image and transcriptomic embeddings from the same biological source occupying nearby regions. This pattern is consistent with effective cross-modal alignment while preserving meaningful inter-tissue structure. Figure 6 further shows that the aligned representation supports accurate tissue- or cancer-type discrimination, with residual confusion largely restricted to semantically or histologically related categories.

D.2 Biological Plausibility of Image-Conditioned Gene Gating

To assess whether the image-conditioned gating module captures biologically meaningful transcriptomic cues, gradient-based feature importance scores are analyzed for the gated gene representation. As summarized in Table 11 and Figure 7, the module consistently prioritizes genes with known relevance to the corresponding tissue context, including lineage-associated markers, immune-related genes, and tissue-specific functional programs. While this analysis is not a causal validation, it provides supportive evidence that the gating mechanism is not merely exploiting superficial morphological correlations, but is instead responsive to biologically plausible molecular signals associated with local tissue environments. Importantly, the 64 latent dimensions are not interpreted as individual genes. Raw-gene attribution is computed through the differentiable path from each input gene

Table 9: Architectural configuration of ViO-Align.

Stage	Module	Architectural Details
Stage I	H&E Cell Encoder	ModifiedResNet2, blocks=(3,4,6,3), base width=64. The output feature is pooled and projected to $\mathbb{R}^{64}$.
	Transcriptomics Encoder	GeneTransformer with 4 layers, 4 heads, $d_{\text{model}} = 256$, and dropout=0.1. The transcriptomic `[CLS]` token is projected to $\mathbb{R}^{64}$.
	Morphology-conditioned Gene Gating Module (MGG)	Two-layer MLP bottleneck ($64 \to 512 \to 64$) with BatchNorm and sigmoid activation. The morphology feature gates the transcriptomic feature to suppress morphology-irrelevant molecular variation.
Stage II	Shared ViT Patch Encoder	6 transformer blocks, 6 heads, $d_{\text{model}} = 384$, patch size 8×8, and MLP ratio 4.0. The encoder outputs a spatial token grid and a global `[CLS]` token.
	Count-guided Flow Localization Module (CFL)	Lightweight dense heads predict the foreground probability map, 2D flow field, and center heatmap from the spatial token grid. Each dense head uses a 3×3 projection, 256 hidden channels, 2 residual blocks with GELU, and upsampling.
	Count Predictor	MLP over the global `[CLS]` token: $\text{Linear}_{384\to256} \to \text{GELU} \to \text{Linear}_{256\to1}$, producing the patch-level cell-count prior $\hat{N}$.
	Microenvironment-aware Aggregation Module (MEA)	Extracts a 9×9 local token window around each cell center, applies center masking to prevent direct copying of target-cell appearance, and performs attention pooling with hidden dimension 192.
	Downstream Predictors (CPP/DP)	CPP maps each environment-aware cell representation to phenotype logits. DP maps the global patch representation to diagnostic logits for patch-level transfer.

count through the transcriptomics encoder and MGG to the image–transcriptomics similarity score. Gradients therefore aggregate all paths from a raw gene through the gated latent factors; absolute attribution magnitudes are averaged across cells to rank the original genes.

D.3 Downstream Label Efficiency

To evaluate dependence on downstream phenotype annotations, the same ViO-Align initialization is adapted using only 1%, 5%, 10%, and 25% of the available training labels. Table 12 reports macro-average phenotype F1. The "Full-label comparator" column denotes the previously reported full-label reference used in the rebuttal comparison.

With at most 25% of the downstream annotations, ViO-Align exceeds the reported full-label comparator on GLySAC, CoNSeP, and PanNuke and approaches it on MoNuSAC. These results indicate that the pretrained representation can reduce, although not eliminate, the amount of task-specific downstream supervision required for adaptation.

D.4 Joint Localization–Alignment Training

The interaction between localization and cross-stage alignment is evaluated while keeping all localization losses unchanged. Table 13 compares localization-only Stage-II training with joint localization–alignment training. For the post-fine-tuning evaluations, both checkpoints undergo the same localization-only fine-tuning while alignment is frozen.

Table 10: Training hyperparameters and augmentation strategies.

Parameter	**Stage I** (Intrinsic Pre-training)	**Stage II** (In-Situ Fine-tuning)
Optimizer and LR	Adam, 2×10^{-4}	AdamW, 5×10^{-5} with weight decay 10^{-2} and StepLR decay
Batch Size and Epochs	Batch size 128, 50 epochs	Batch size 64, 300 epochs
Loss Objectives	InfoNCE ($\tau_{\text{hard}} = 0.5$, $\tau_{\text{soft}} = 1.0$)	$\mathcal{L}_{\text{prob}}$ (5.0), $\mathcal{L}_{\text{flow}}$ (20.0), InfoNCE ($\tau = 0.07$, $\lambda = 0.1$)
Image Augmentations	ResizedCrop, Flip, ColorJitter, Grayscale	Standard Macenko normalization
Gene Augmentations	Gaussian noise ($\sigma = 0.1$), feature masking ($p = 0.2$)	N/A
Misc. Configs	N/A	Memory bank (size 8192, EMA $\alpha = 0.5$), background flow regularization $\gamma = 0.1$

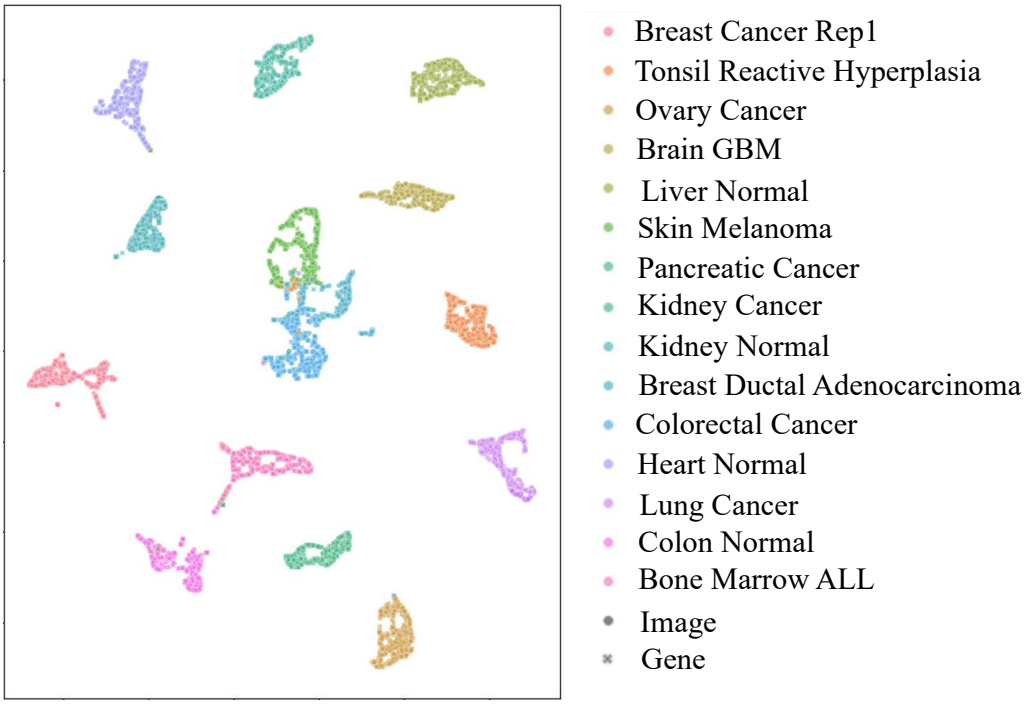

Figure 5: UMAP visualization of cell embeddings in Stage I. Image and gene representations align within tissue-specific manifolds.

The consistent gains indicate that cross-stage alignment provides auxiliary supervision to the shared Stage-II representation rather than acting only as an independent contextual objective.

D.5 Computational Efficiency

Figure 8 compares predictive performance against inference cost. Methods that rely on explicit contour modeling or graph construction typically incur substantially higher computational overhead at test time. In contrast, the Stage II design avoids an explicit segmentation-and-graph pipeline and

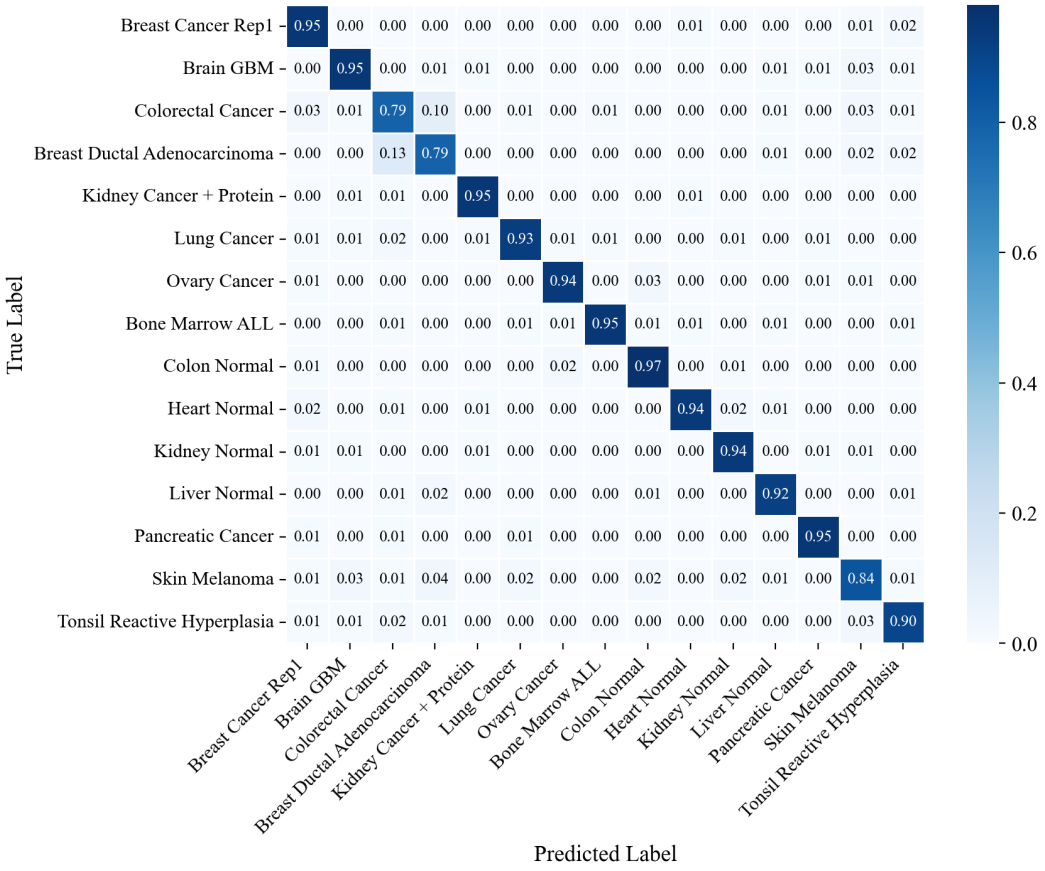

Figure 6: Confusion matrix for tissue/cancer-type prediction from Stage I embeddings. Most confusion is concentrated among biologically related domains.

completes inference within a practical runtime regime. Although some lightweight baselines may be faster, ViO-Align provides a favorable efficiency-accuracy balance, particularly in terms of detection quality relative to model complexity.

D.6 Sensitivity to Microenvironment Window Size

The contextual window size K is varied over $\{4, 9, 16\}$ to evaluate the sensitivity of MEA to microenvironment scale. As shown in Table 14, the preferred scale varies across datasets, but the average macro-F1 changes by only 0.62 points across the three settings. The default $K = 9$ is therefore used as an untuned intermediate scale.

D.7 Patch Encoder Design Comparison

Before Stage II training, different patch encoder designs are compared to select a suitable backbone for raw-patch cell localization. The comparison is conducted on the held-out test split of the original Xenium patch-level data, rather than on downstream benchmarks, because the goal is to evaluate localization capability and training practicality under the same upstream setting. As shown in Table 15, CellViT-256 achieves the highest F1-score, but its larger parameter size requires a smaller batch size and reduces training throughput. CTransPath is the most lightweight encoder, but its lower precision, recall, and F1-score indicate insufficient localization ability in this setting. Therefore, the evaluated pretrained encoders are not directly optimal for the Stage II framework, as they are either computationally heavy or less effective for fine-grained cell-center localization.

Table 11: Illustrative genes receiving high importance under image-conditioned gating across representative tissue contexts.

Tissue / Condition	Top Highlighted Genes	Biological Relevance and Function
Brain GBM	*S100B, PTN, VEGFA, EGFR*	Glial lineage markers, neoangiogenesis, and malignant proliferation drivers.
Colorectal Cancer	*EPCAM, GPX2, IGKC, JCHAIN*	Carcinoma-associated markers, anti-apoptotic programs, and plasma-cell-linked immune infiltration.
Normal Colon	*MUC12, CLCA1, CES2*	Mucus secretion, mucosal barrier maintenance, and xenobiotic metabolism.
Normal Liver	*TAT, CYP2A7, CYP3A4, CFB/HPX*	Liver-specific enzymes, drug metabolism, and plasma protein synthesis.
Normal Heart	*DES, S100A1, LPL*	Cardiac muscle structure, myocardial contraction, and fatty acid oxidation.
Reactive Tonsil	*MS4A1, BANK1, PTPRC, CCR7*	B-cell maturation, pan-leukocyte identity, and lymphocyte homing.

Table 12: **Downstream label efficiency.** Macro-average phenotype F1 (%) when ViO-Align is adapted with a fraction of the available downstream annotations.

Dataset	Full-label comparator	1%	5%	10%	25%
GLySAC	48.52	59.09	54.93	58.93	60.99
CoNSeP	56.80	41.92	44.44	48.28	64.96
MoNuSAC	67.55	43.02	51.94	50.89	64.29
PanNuke	52.87	34.63	43.53	50.27	60.86

Considering localization performance, model size, and training efficiency, the custom ViTGridEncoderCLS is adopted as the default Stage II patch encoder. It achieves competitive localization performance with fewer parameters than CellViT-256 and provides a better balance for subsequent count-guided flow localization and cross-stage vision-omics alignment.

D.8 Extended Qualitative Analysis

Additional qualitative results on all four benchmarks are provided in Figure 9. The visualizations compare ground-truth and predicted cell centers with phenotype labels, and report the number of cells above each patch. Overall, ViO-Align preserves the dominant cell composition and coarse spatial layout across different tissue appearances, including tumor regions, epithelial-stromal interfaces, immune-infiltrated areas, and sparse multi-cellular structures.

On PanNuke, the predicted cell counts are close to the ground truth, indicating stable center localization in heterogeneous pan-cancer patches. On CoNSeP, the model captures the main epithelial and inflammatory regions, although under-counting can occur in areas with weak nuclear boundaries or strong staining variation. On GLySAC, ViO-Align follows the major epithelial, lymphocyte, and miscellaneous-cell distributions, but the heterogeneous miscellaneous category remains challenging. On MoNuSAC, the model produces plausible predictions for epithelial and immune-related cells, while rare or ambiguous cell types are more prone to missed detections or unstable classification.

These examples are consistent with the quantitative results. ViO-Align generally provides reliable in situ localization and phenotype assignment, while the main failure cases arise from crowded nuclei, weakly separated boundaries, rare cell types, and visually ambiguous categories.

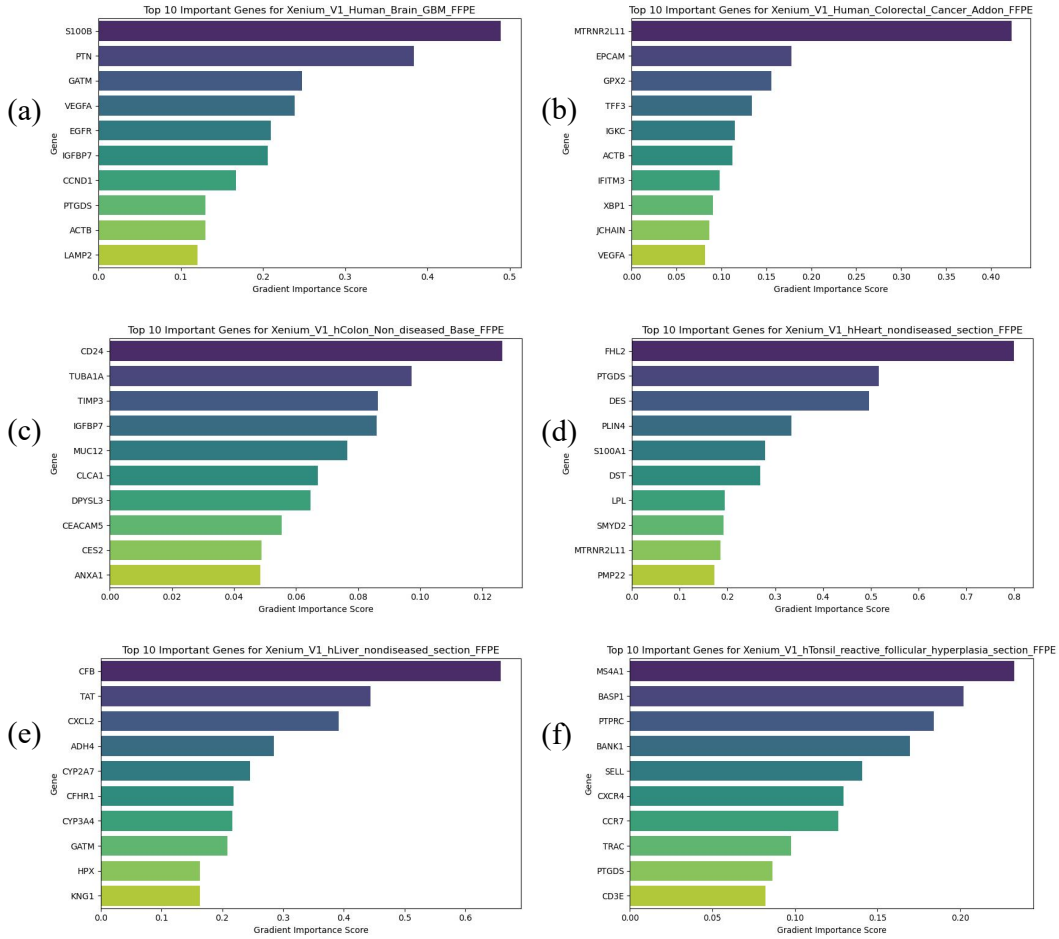

Figure 7: Representative examples of highly ranked genes under image-conditioned gating across six tissue contexts: (a) Brain GBM, (b) Colorectal Cancer, (c) Normal Colon, (d) Normal Heart, (e) Normal Liver, and (f) Reactive Tonsil. The highlighted genes are consistent with tissue identity, malignant state, metabolic specialization, contractile function, or immune activity, supporting the biological plausibility of the gating mechanism.

Table 13: **Effect of joint localization–alignment training.** Values are detection F1; gains are absolute percentage points.

Evaluation	Localization only	Joint	Gain
Stage-II validation	0.8584	0.8738	+1.54
Unseen data, before fine-tuning	0.7814	0.8060	+2.46
Unseen validation, after fine-tuning	0.8136	0.8204	+0.68
Unseen independent test	0.8447	0.8518	+0.72

E LIMITATIONS

Although ViO-Align achieves strong performance on multiple cell phenotyping benchmarks, several limitations remain. First, the framework relies on spatial transcriptomics supervision. Xenium acquisition is substantially more expensive and less widely available than routine H&E, although it is required only for upstream pretraining and not for downstream inference. Transcript assignment, cell

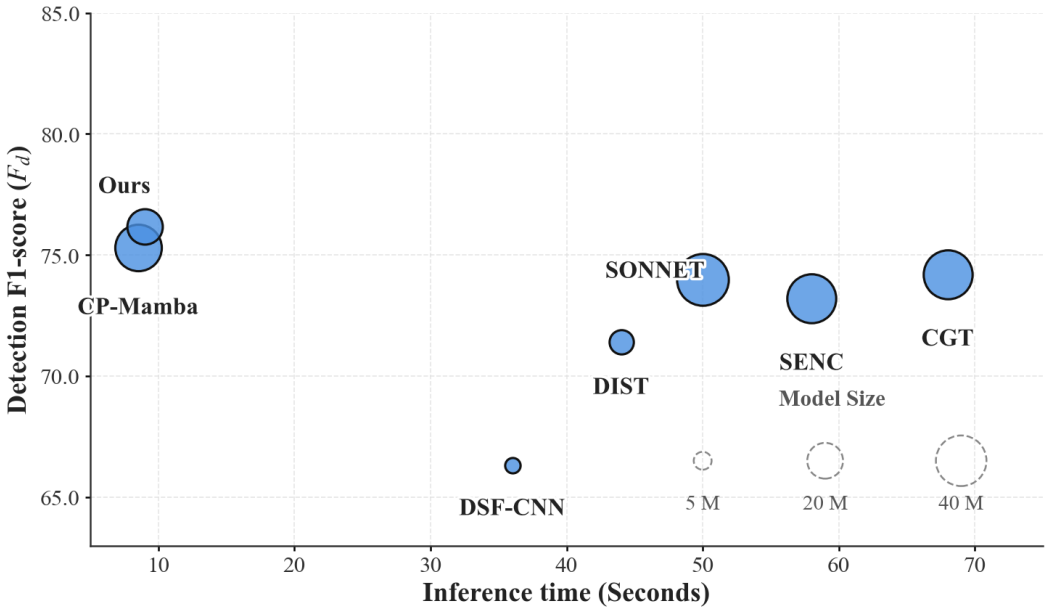

Figure 8: Inference efficiency versus detection performance (F_d). ViO-Align achieves a favorable balance between predictive quality and computational cost.

Table 14: **Sensitivity to context-window size.** Values are macro-average phenotype F1 (%).

K	CoNSeP	MoNuSAC	GLySAC	Avg.
4	**64.54**	71.05	65.58	67.06
9	63.47	69.38	**68.19**	67.01
16	63.14	**74.44**	65.32	**67.63**

boundary estimation, and H&E-omics registration may also introduce noise. These factors can affect image-omics alignment, especially in crowded regions or tissues with local deformation. Second, ViO-Align focuses on cell-center localization and phenotype prediction rather than full nuclear or cellular segmentation, so it does not provide boundary-level measurements such as nuclear shape, area, or contour-based morphology. Third, the microenvironment-aware aggregation module uses a fixed context window, whereas the optimal context scale may vary across tissue types, regions, and cell subtypes. Finally, current evaluation is based on public benchmarks and curated splits; larger multi-institutional validation is needed before applying the model to real clinical workflows.

Table 15: **Patch encoder design comparison on the original Xenium test split.** The comparison is conducted on held-out Xenium raw patches.

Method	Backbone / Configuration	Params	Precision	Recall	F1
`vitbase`	ViTGridEncoderCLS	19.59M	0.902	0.857	0.8790
`cellvit256`	CellViT-256	30.95M	**0.908**	**0.865**	**0.8862**
`ctranspath`	CTransPath	**7.70M**	0.868	0.840	0.8535

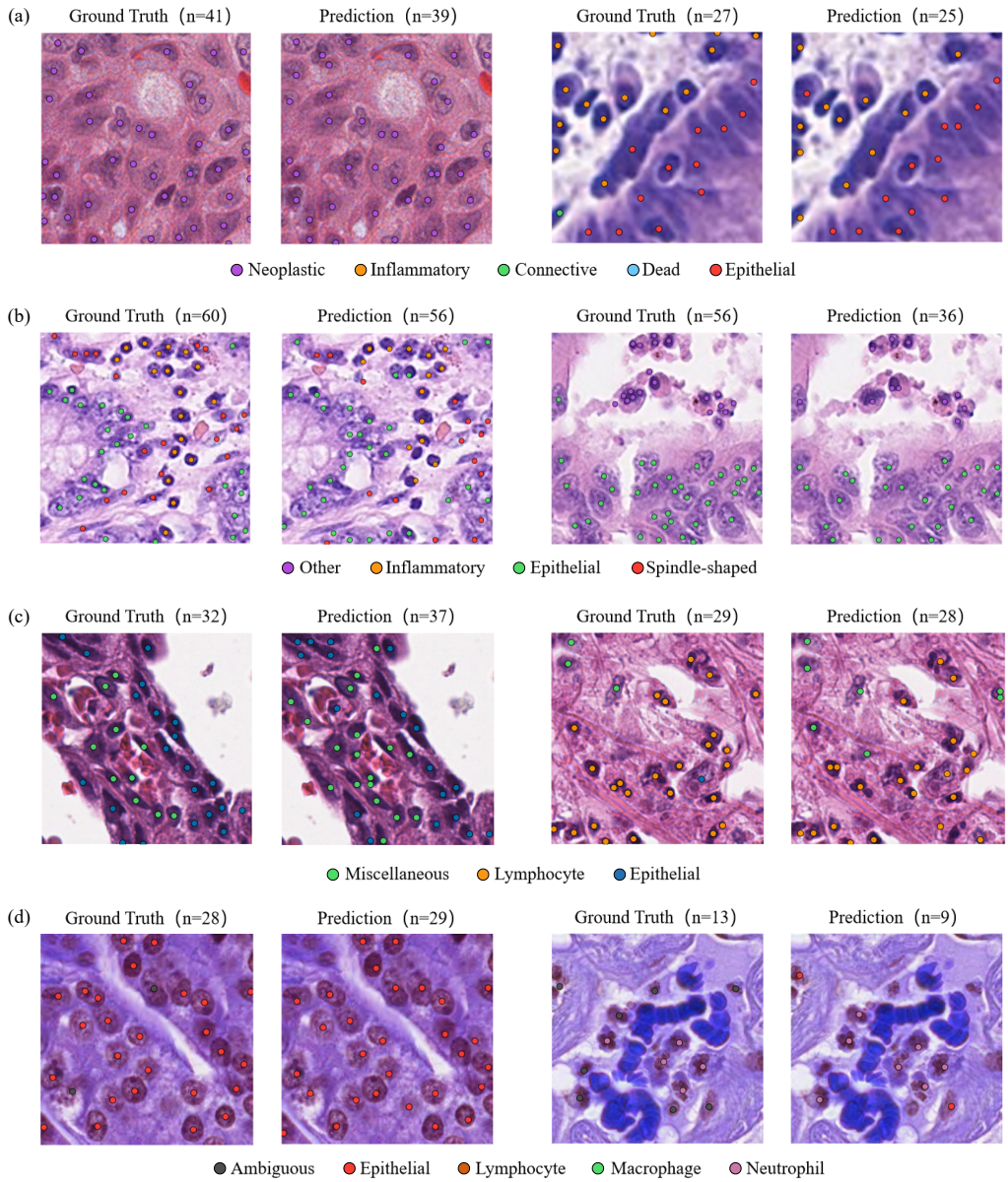

Figure 9: Extended qualitative results on four datasets. Rows (a)-(d) show representative examples from PanNuke, CoNSeP, GLySAC, and MoNuSAC, respectively. Each example compares ground-truth and predicted cell locations with class-specific color coding. ViO-Align generally preserves the dominant class composition and spatial distribution across datasets, while residual discrepancies mainly occur in crowded regions, weakly defined subtypes, and sparse minority classes.