PER-SPOT SEMANTIC CONDITIONING FOR HISTOLOGY-TO-SPATIAL-TRANSCRIPTOMICS GENERATION: CONTROLLABLE GENE EXPRESSION WITH A REGRESSION-ANCHORED GENERATOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting spatial transcriptomics (ST) from hematoxylin and eosin (H&E) images is typically performed using image encoders coupled with regression models trained under mean-squared error (MSE). Under MSE, the optimal predictor is the conditional mean, whose variance is necessarily lower than that of the true conditional distribution. We show that this structurally limits the dynamic range and spatial structure that regression can recover, not because of insufficient model capacity but because of the learning objective itself. Across HEST-bench tasks, regression compresses the predicted dynamic range to 0.74× that of the ground truth on average, with the strongest compression on the hardest tasks, while Pearson correlation largely hides this failure by remaining insensitive to spatial oversmoothing. This limitation persists even in high-capacity models and can be characterized through an aleatoric uncertainty decomposition, suggesting that H&E-to-ST models should be evaluated on dynamic range and spatial structure alongside correlation.

Motivated by this limitation, we reformulate H&E-to-ST prediction as conditional generation over the Visium spot grid. Our framework introduces user-specified per-spot semantic conditioning, enabling expression to be steered along biological axes that are not recoverable from morphology alone. On a morphology-invisible interferon-response axis (image predictability r=0.06), we achieve accurate directional control while evaluating only on genes disjoint from those defining the conditioning signal. A continuous per-spot scalar consistently outperforms a tertile one-hot representation. When the control axis is instead predicted from H&E, its correlation with the target remains low (r=0.13) and controllability collapses, demonstrating that the conditioning signal must be explicitly specified.

Finally, we couple conditional generation with a distribution-aware regression anchor,trained with a distributional (MMD) term so it does not collapse to the oversmoothed conditional mean,and let the flow restore residual spatial structure. On held-out tissues the anchored model stays competitive with strong predictors on localization while improving spatial-structure fidelity beyond what Pearson correlation captures. All experiments are conducted on held-out splits,cross-organ generalization is left for future work.

1 INTRODUCTION

Spatial transcriptomics (ST) measures gene expression at spatially resolved spots on a tissue section, but is costly,predicting it from the co-registered, inexpensive H&E image is therefore an active goal (He et al., 2020; Jaume et al., 2024), with value for biomarker discovery and for studying spatially structured biology such as tumor microenvironments and tissue zonation.

The dominant approach regresses each spot's patch embedding to its expression vector under an MSE objective. An MSE-optimal predictor returns the conditional mean $\mathbb{E}[\mathbf{y} \mid \mathbf{x}]$, whose variance is at most that of $\mathbf{y}$,predictions are therefore systematically over-smoothed, with bright high-expression peaks and structural zeros both pulled toward the mean. Because the field's headline metric,per-gene

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.

Pearson r,is invariant to a global rescaling of the predicted range, this amplitude collapse is invisible in reported numbers. We confirm the shrinkage persists even in a high-capacity accuracy model, establishing it as a property of the objective rather than of any single regressor.

If a single conditional mean cannot represent a high-variance, spatially structured distribution, a natural remedy is to model the distribution rather than its mean. Generative formulations take this route: diffusion-based Stem (Zhu et al., 2025) and flow-matching STFlow (Huang et al., 2025) cast expression prediction as conditional generation over the tissue, retaining the dynamic range a regressor collapses. This motivates our flow-matching generator over the spot grid. We are explicit about what generation does and does not provide: it restores amplitude and distributional structure, but on held-out tissue its position accuracy does not exceed a well-designed regressor. We therefore pair the generator with a regression prediction as an anchor, intending a division of labor in which the anchor supplies the low-frequency conditional-mean component that fixes position and the generator supplies residual spatial structure. We train the anchor with a distributional objective so it does not collapse to the over-smoothed conditional mean,the resulting anchored model stays competitive with strong predictors on position (§4.2) while additionally enabling per-spot semantic control (§4.3).

Controllability is where image-only generation falls short, and it is our focus. Every current generative H&E→ST model conditions only on the image, which inherits a hard limit from morphology itself: an expression axis that leaves little imprint on H&E cannot be recovered or steered from the image alone, however expressive the generator. Many biologically important axes are of this kind,we find, for example, an interferon-response axis nearly unpredictable from morphology yet a real, spatially organized signal. To reach such axes, we condition the generator on a per-spot semantic label placed on the spot grid, so conditioning varies spatially and can specify what a morphological encoder cannot see. The label is a control input, not a predicted one: at inference the user specifies it, placing this setting alongside controllable generation rather than blind prediction, so we do not claim it as free accuracy but ask only whether a specified axis is realized. Indeed, the label must be specified, not inferred: predicting the control axis from morphology yields near-zero axis correlation and collapses steering (§4.3), so a real semantic input is required rather than a self-derived proxy. To our knowledge no existing H&E→ST generator has such a per-spot semantic control channel.

Contributions.

- **Diagnosis.** We quantify over-smoothing as a structural, model-agnostic property of MSE regression across HEST-bench (mean std-ratio 0.74,a nonlinear multi-encoder MLP shrinks as much or more), show Pearson r cannot detect it, and bound recoverable signal with an aleatoric decomposition (§4.1).
- **Per-spot semantic control.** We introduce per-spot semantic conditioning for a spot-grid generator and show it steers expression along morphology-invisible axes that image-only generators cannot reach by construction, with a continuous scalar label beating a one-hot and verified on label-disjoint genes,the label is a specified input, since self-derived proxies fail (§4.3). A distribution-aware regression anchor keeps position competitive while relieving the amplitude collapse of plain MSE, and the flow residual improves spatial structure; position remains bounded by a morphology-predictability ceiling the anchor reaches rather than exceeds (§4.2, 4.4).

We also state scope separately from the contributions: control and structure recovery are shown on in-distribution held-out splits and on representative axes (metabolic zonation, IFN, and Wnt).

2 Related Work

Regression models for H&E-to-ST prediction and structural over-smoothing. Early work framed spatial gene expression prediction as supervised regression from histology image encoder to gene expression prediction. ST-Net (He et al., 2020) mapped CNN patch features to expression independently per spot, and HisToGene (Pang et al., 2021) applied a vision transformer to capture long-range context among patches. Subsequent methods enriched this with multi-scale and graph context by adding dynamic convolution in Hist2ST (Zeng et al., 2022) and propagating information over histology neighborhood graphs SEPAL (**?**). All predict expression across various modality, including image, text or graph, and none takes a per-spot semantic condition. Recent methods tried

to build pathology foundation-model encoders such as UNI (Chen et al., 2024), UNI-2h(Mahmood Lab, 2025) and GigaPath(Xu et al., 2024) and applied the encoder for gene expression as a downstream task. Despite architectural differences, these models share a training objective that minimize mean-squared error and therefore approximate the conditional mean of expression given morphology. Because the conditional mean has lower variance than the true conditional distribution, MSE regression compress dynamic range and flatten spatial heterogeneity, an over-smoothing that scale-invariant Pearson correlation does not reveal. We show this persists in high-capacity regressors and suggest to report the amplitude and spatial structure with the conventional correlation.

Generative H&E-to-ST prediction. Latest works casts gene expression prediction as conditional generation, which can retain the intrinsic variance that pointwise objectives discard. Stem (Zhu et al., 2025) samples spot-wise expression with conditional diffusion and STFlow(Huang et al., 2025) employs whole-slide flow matching with a ZINB source prior to synthesize expression on slide image condition. FLAG adds a spatial graph encoder and Gene Foundation Model alignment, introducing the Gene Structural Correlation (GSC) and Spatial Structural Correlation (SSC) metrics. Here we claim STFlow as the natural image-only baselines for the semantic-conditioning claim.

Controllable and spatially conditioned generation. Two lines are relevant to our per-spot conditioning. In non-spatial expression modeling, controllable generators steer transcriptomic state along semantic factors. scGen (Lotfollahi et al., 2019) and CPA (Lotfollahi et al., 2023) model perturbation responses, and SPATIA (Spatia authors, 2025) extends controllable generation to spatial cell phenotypes, though it operates at the single-cell and biological-transition level rather than on per-spot Visium expression. In parallel, the image-synthesis community conditions generators on spatially varying signals. SPADE (Park et al., 2019) introduced spatially-adaptive normalization, whose per-location modulation keeps a semantic layout from being washed out by uniform normalization, and ControlNet (**?**) conditions diffusion on dense layout maps. Our per-spot conditioning belongs to this spatially-adaptive family, and mechanically it is a per-spot, zero-initialized adaptive normalization described in §3. What is new is not spatially varying conditioning per se, but its use on ST expression with a continuous, ordered biological axis as the control rather than a categorical mask, together with its evaluation as controllability.

Informed source priors in flow matching. Anchoring a generator to a prior prediction is established in imaging (ResShift (Yue et al., 2023), SDEdit (Meng et al., 2022)). STFlow's data-shaped ZINB prior is the closest ST analog. Our anchor is instead an input-dependent regression prediction. We adopt the refinement mechanism and our novelty is the spatially-varying per-spot semantic control field steering a spot-grid H&E-to-ST generator.

3 METHOD

3.1 SETUP AND NOTATION

A slide is mapped to an $H \times W$ hexagonal-to-rectangular grid. Spot i occupies grid cell (y_i, x_i) with a valid-spot mask $\mathbf{M} \in \{0,1\}^{H\times W}$. Each spot carries an image feature $\mathbf{f}_i \in \mathbb{R}^{C_f}$ from one or more pathology foundation models (UNI (Chen et al., 2024), GigaPath, UNI2h, STVT) and, optionally, a per-spot semantic label $\ell_i \in \mathbb{R}^L$. Conditioning is assembled as a tensor $\mathbf{C} \in \mathbb{R}^{(C_f+L)\times H\times W}$,the target is $\mathbf{y} \in \mathbb{R}^{K\times H\times W}$ over K highly-variable genes (Fig. 1).

3.2 BACKBONE AND FLOW-MATCHING OBJECTIVE

The backbone is a 2D spot-token transformer over the grid with per-spot (spatially-adaptive), zero-initialized adaptive layer-norm conditioning. Unlike the global adaLN-Zero of DiT (Peebles & Xie, 2023), whose modulation parameters are shared across all tokens from a single conditioning vector, our modulation (scale/shift/gate) is produced independently for each spot from that spot's own conditioning token,it is therefore functionally a zero-initialized SPADE (Park et al., 2019), and the label channel modulates every spot separately. Spatial context is injected by k-nearest-neighbour-masked local attention on the grid (k=8), optionally weighting each neighbour by the cosine similarity of its features. We train a rectified-flow objective (Lipman et al., 2023; Liu et al.,

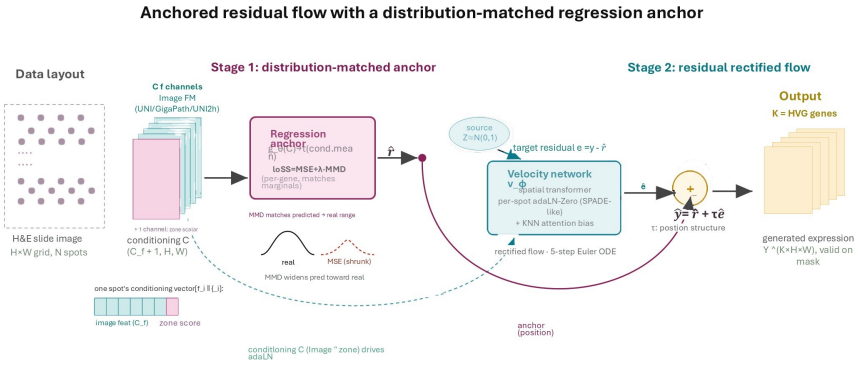

Figure 1: Data layout: a Visium slide becomes a masked (H, W) grid,each valid spot carries image-FM channels and optional per-spot label channels as conditioning, with an HVG expression target.

2023). With a source sample $\mathbf{z} \sim p_0$ and data $\mathbf{x}_0{=}\mathbf{y}$, the linear interpolant and constant target velocity are

$$\mathbf{x}_t = (1-t)\,\mathbf{z} + t\,\mathbf{x}_0, \qquad \mathbf{v}^\star = \mathbf{x}_0 - \mathbf{z}, \qquad t \sim \mathcal{U}(0,1), \tag{1}$$

and the network $v_\theta(\mathbf{x}_t, t, \mathbf{C})$ is trained with a masked velocity loss

$$\mathcal{L}_{\text{flow}} = \frac{\sum_i \mathbf{M}_i \, \|v_\theta(\mathbf{x}_t, t, \mathbf{C})_i - \mathbf{v}_i^\star\|^2}{K \sum_i \mathbf{M}_i}. \tag{2}$$

Sampling integrates the ODE $\dot{\mathbf{x}} = v_\theta$ from $t{=}0$ to 1 with a 5-step Euler scheme.

3.3 Per-spot semantic conditioning

The per-spot label ℓ_i is placed on the grid alongside $\mathbf{f}_i$ and drives the per-spot adaLN modulation, so conditioning varies spatially across the grid. Labels are defined from a set of LABEL genes only and evaluated on a disjoint set of EVAL genes (non-circularity). We represent the label as a single continuous per-spot scalar rather than a tertile one-hot: a monotone axis (e.g. periportal to pericentral zonation) loses information under one-hot, and the scalar is empirically both more accurate and more controllable (§4.3). The label is a user-specified control input,at inference it can be forced high or low across the grid to steer expression.

3.4 Distribution-aware regression anchor

We anchor the generator to a frozen regression prediction $\mathbf{r}$, produced by the same backbone in regression mode (a single forward at $t{=}0$ giving the conditional mean) on the image features augmented with a mean pool over each spot's $k{=}8$ spatial neighbours. A regression trained under MSE alone collapses to the conditional mean and over-shrinks amplitude (§4.1),we therefore train the anchor with an MSE term plus a per-gene maximum-mean-discrepancy (MMD) term that matches the predicted and true marginal expression distributions,

$$\mathcal{L}_{\text{anchor}} = \frac{\sum_i \mathbf{M}_i \|\mathbf{r}_i - \mathbf{y}_i\|^2}{K \sum_i \mathbf{M}_i} + \lambda \sum_{g=1}^{K} \mathrm{MMD}^2\big(\{r_{ig}\}_{i\in\mathcal{V}},\, \{y_{ig}\}_{i\in\mathcal{V}}\big), \tag{3}$$

where $\mathcal{V}$ is the set of valid spots and MMD uses a Gaussian kernel with a per-gene median-heuristic bandwidth. The distribution term relieves the shrinkage in the anchor itself, so amplitude and position improve together rather than trading off. The anchor is then frozen and the flow generates the standardized residual $(\mathbf{y} - \mathbf{r})$ from a Gaussian source (the residual is near zero-mean),the readout mixes the two,

$$\hat{\mathbf{y}} = \mathbf{r} + \tau\, \mathbf{v}_{\text{res}}, \tag{4}$$

where τ trades position (the anchor) against structure and dynamic range (the residual) and $\tau{=}0$ recovers the anchor alone. This keeps position at the regression level while the flow restores the spatial structure and amplitude that the conditional-mean anchor suppresses. The anchored benchmark

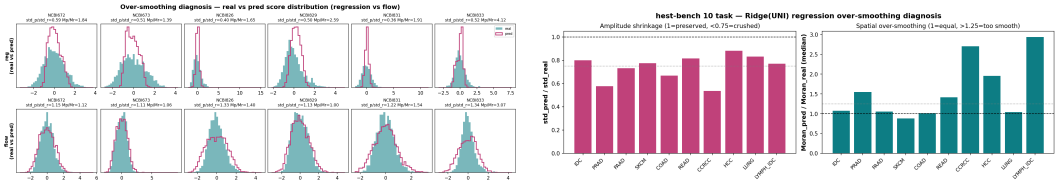

Figure 2: Over-smoothing is structural. Left: std-ratio and Moran-ratio diagnostics. Right: per-task shrinkage across HEST-bench,severity tracks difficulty.

(§4.2) is run without the semantic label, for a fair comparison to regression,the label enters only in §4.3.

4 EXPERIMENTS

Setup. Our experiments draw on two distinct data sources. Two gene protocols are used and kept strictly separate. (a) The head-to-head comparison against benchmark baselines (Table 1) follows the standard HEST-bench hvg50 protocol (top-50 HVG) over the full 10 tasks, since matching the 10-task set is the entire reason for adopting the HEST-bench protocol,any task subset would not be comparable to the published baselines. This uses patient-level held-out evaluation on the official HEST-bench splits,we verified that no patient appears in both train and test in any fold across all ten tasks. The interferon (IFN) and Wnt controllability axes are evaluated within HEST-bench tasks (CCRCC and READ, respectively). (b) Liver zonation is not a HEST-bench task: it uses 12 human liver Visium slides drawn from the full HEST-1k corpus (cross-slide train 8 / test 4), read at full transcriptome so that the axis-defining and evaluation gene sets are both available. We use HEST-1k liver rather than the HEST-bench HCC task because the latter is unusable for a zonation study on three counts: it contains only two slides (precluding a cross-slide split), its 50-gene benchmark panel retains just 1 of the 15 zonation genes (leaving the axis neither definable nor evaluable), and it is hepatocellular carcinoma, in which the normal periportal-pericentral lobular gradient is disrupted. We report metrics on separate axes: PCC (position), $\text{std}_{\text{pred}}/\text{std}_{\text{real}}$ (amplitude), and GSC/SSC and the Moran's-I ratio (structure).

4.1 OVER-SMOOTHING IS STRUCTURAL

Across 10 HEST-bench tasks, a Ridge/UNI regressor has mean std-ratio 0.74 (all tasks < 1), a Moran ratio of 1.56, and 25% of genes over-smoothed,severity co-varies with task difficulty (descriptive,CCRCC, PRAD, HCC, LYMPH_IDC worst) (Fig. 2). This is not an artifact of a weak model: a nonlinear multi-encoder MLP shrinks as much or more (std-ratio 0.28-0.44). Nor is it detected by correlation: quantile-matching restores the std-ratio toward 1 while leaving r unchanged. Decomposing what regression discards, the systematic part is partly recoverable (residual $r_{\text{sys}}{=}{+}0.24$ to $+0.35$ across three tasks) but the stochastic part is aleatoric ($r_{\text{sto}} \approx 0$), which bounds how much any generator can recover and grounds the ceiling of §4.4 (Fig. 3).

4.2 COMPARISON TO BASELINES

On the HEST-bench protocol (hvg50, patient-level split) we compare methods under an identical split and gene set (Table 1). Memory forces each generative model to drop the one slide-set it cannot fit in 24 GB,our anchor drops LYMPH_IDC, while STFlow and STPath drop IDC,so PCC is reported over each method's runnable tasks. On the nine tasks our anchor and the Ridge probes share (LYMPH_IDC excluded), the distribution-aware anchor reaches PCC 0.446, well above single-encoder probes (UNI/GigaPath+Ridge $0.366/0.364$),adding the flow residual ($\tau{=}0.5$) raises structure (SSC $0.362 \to 0.505$, GSC $0.413 \to 0.543$) at essentially unchanged position ($0.446 \to 0.437$), so the generative gain is structural rather than positional. The generative baselines land lower on position,STFlow 0.413 (IDC excluded), STPath 0.369, and per-spot Stem 0.254,but on their own task sets, so we treat them as references rather than a strict head-to-head. For reference, on the full ten-task protocol STFlow reports macro PCC 0.415 (we reproduce 0.413) and single-encoder probes

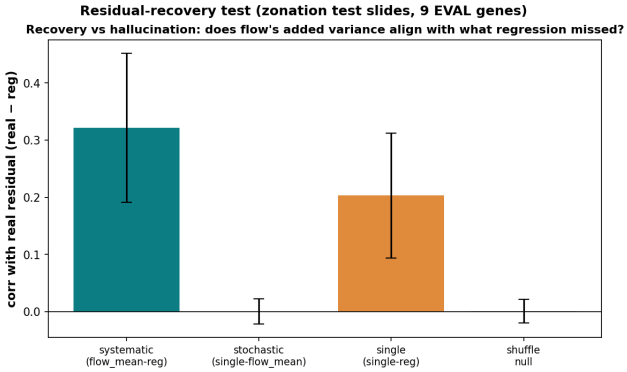

Figure 3: Aleatoric ceiling: the systematic part of what regression over-smooths is partly recoverable (r_{sys}), the stochastic part is not ($r_{\text{sto}} \approx 0$).

Table 1: HEST-bench (hvg50), our reproductions under an identical patient-level split and gene set; PCC (position), GSC/SSC (structure), single seed. Each generative model drops the one slide-set it cannot fit in 24 GB, so PCC is that method's macro over its runnable tasks (count listed),our anchor and the Ridge probes share the *same* nine tasks (LYMPH_IDC excluded) and are therefore directly comparable, where the distribution-aware anchor clearly leads and adding the flow residual (τ=0.5) raises structure at essentially unchanged position. STFlow/Stem structure metrics are being computed.

method	PCC	GSC	SSC	#tasks
Ours: distribution-aware anchor (τ=0)	**0.446**	0.413	0.362	9^a
Ours: anchor+flow (τ=0.5)	0.437	**0.543**	**0.505**	9^a
STFlow (our reproduction)	0.413	-	-	9^b
Stem (our reproduction, cond=UNI)	0.254	-	-	10
UNI+Ridge	0.366	0.534	0.354	9^a
GigaPath+Ridge	0.364	0.531	0.374	9^a

[a]LYMPH_IDC excluded (grid exceeds 24 GB under dense attention). [b]IDC excluded (STFlow OOM). [c]9 reproduced + IDC from HistoPrism.

report 0.344 (we reproduce 0.35), confirming that our pipeline matches the published protocol. We turn next to per-spot semantic control (§4.3).

4.3 Per-spot controllability

We condition on a per-spot label defined from LABEL genes and measure the effect on disjoint EVAL genes,steering is tested by forcing the label high vs. low across the grid, independent of any regression, isolating the label's causal effect. On liver zonation, forcing the zone label moves EVAL genes in the correct direction (Fig. 4, 5). Crucially, this holds for morphology-invisible axes, an interferon-response axis with image-predictability $r \approx 0.06$ is steered with directional accuracy 1.00 in both image-on and zone-only settings, while a Wnt axis is weaker,an axis an image-only model cannot steer because it cannot see it (Fig. 6).

Label representation. A single continuous scalar (from LABEL genes) gives the best accuracy and controllability, beating a tertile one-hot and other one-dimensional encodings (Table 2). The leakage-controlled number is 0.118: a score computed from EVAL genes inflates accuracy to 0.308 and is excluded, so the reported score uses LABEL genes disjoint from EVAL,the quantitative basis for non-circularity.

Self-derived labels fail. If instead the control axis is predicted from H&E and used as the label, the predicted axis correlates only 0.13 with the true axis and controllability collapses to 0.19. The label must therefore be specified, not inferred,direct evidence for a semantic control input.

Table 2: Per-spot label representation (liver zonation, held-out test). Accuracy = EVAL-gene PCC; direction = fraction of forced axes moved correctly. A continuous scalar wins on both. The leaky EVAL-derived score (grey) is excluded from claims.

representation	dim	accuracy (PCC)	direction
tertile one-hot	3	0.098	0.94
ordinal $\{1, 2, 3\}$	1	0.096	0.86
normalized $\{0, .5, 1\}$	1	0.086	0.75
continuous scalar (LABEL)	1	**0.118**	**0.97**
(leaky) EVAL-derived score	1	0.308	1.00

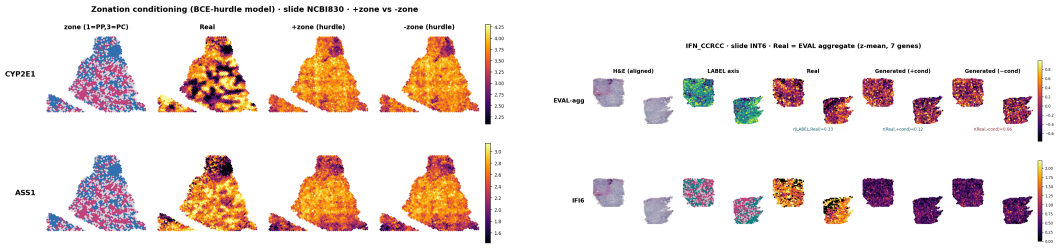

Figure 4: Directional control. Left: liver zonation,forcing the zone label moves pericentral/periportal EVAL genes the correct way. Right: interferon axis (morphology-invisible) is nonetheless steered.

4.4 Generalization analysis

On liver zonation (train 8 / test 4 slides) the flow alone overfits at the set level (train mean r=0.50, test 0.20, gap 0.30,Fig. 7),this is not a sampling artifact (single-sample r std $\pm$0.014, and the gap persists at the conditional mean). This motivates the anchor (§4.2) and does not contradict it,the two use different data and metrics (here, zonation flow-alone generalization,there, HEST-bench held-out with an anchor and the label off). Among interventions, expanding training data helps most (test $+0.036$ from 8$\rightarrow$11 slides), neighbor conditioning a little ($+0.024$), while batch-correction and condition-dropout give no test gain,condition-dropout lowered the train score without raising the test score, so a condition-memorization explanation is not supported and the residual gap reads as genuine domain shift and data scarcity. The held-out ceiling ($r \approx$ 0.2-0.3) matches image-predictability on unseen tissue.

5 Discussion and Limitations

Position accuracy is bounded by image predictability, not amplitude,the generative gain is calibrated realism, directional control, and structure recovery, not exceeding a well-designed regression on position. The anchor reaches this ceiling,it does not break it, and the aleatoric result (§4.1) explains why. Relative to FLAG, FLAG's gene list, split, and language-model embeddings are unpublished, so no valid quantitative head-to-head is possible,we do not compare to it numerically. Against STFlow we run an identical-split head-to-head (§4.2),reproducing its reported ten-task score (0.413 vs. 0.415) confirms the protocol match, and we compare position and structure on the shared task set (Table 1). We likewise reproduce the diffusion generator Stem under the same protocol. Data scarcity,our human liver zonation study uses 12 slides,limits generalization, and more slides is the primary lever.

Deployment note. Controllability is established by forcing an independent, user-specified label. Estimating the label at deployment (e.g. a zone score from the regression prediction) is viable only for morphology-predictable axes,for morphology-invisible axes it fails (axis correlation 0.13, steering collapses), so those require an externally provided label. A jointly optimized model that carries both the anchor and the control label is left to future work.

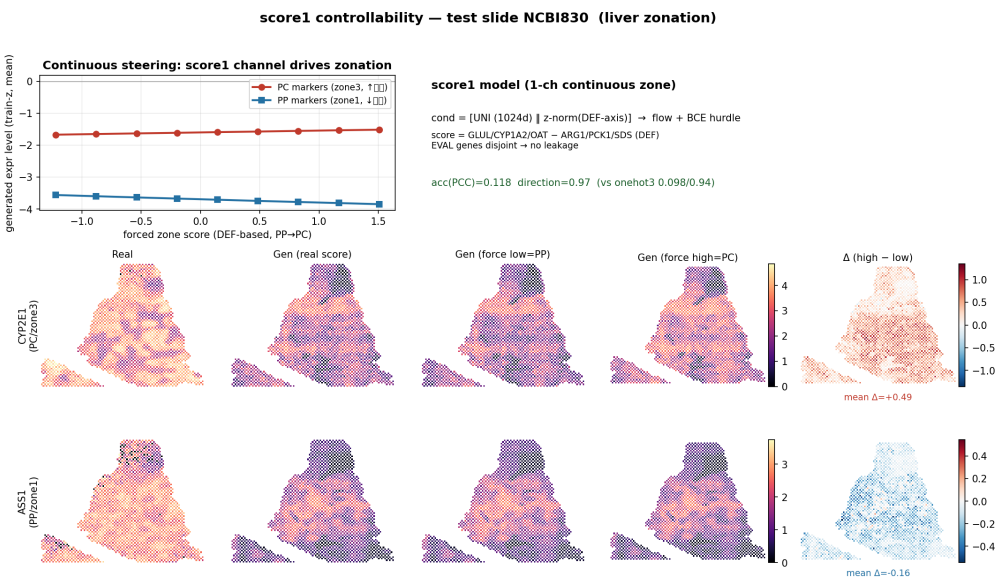

Figure 5: Continuous-scalar (score1) controllability on a held-out slide. Top: forcing the zone score monotonically raises pericentral markers and lowers periportal markers. Bottom: spatial maps of real, generated, and forced-low/high expression, with the high−low difference.

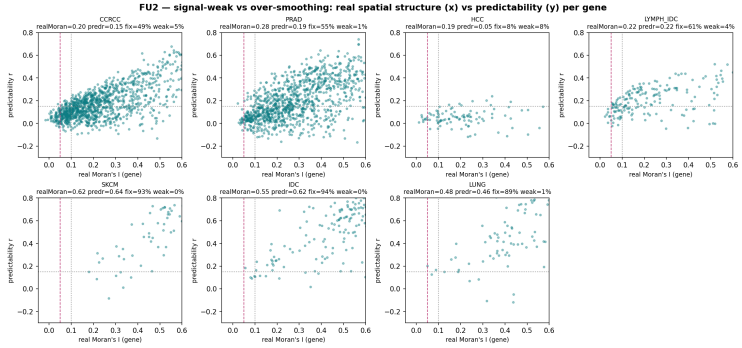

Figure 6: Which axes are controllable-fixable: real Moran's I and image-predictability separate over-smoothing-fixable axes from image-invisible ones.

6 Conclusion

Casting H&E→ST prediction as conditional generation with per-spot semantic conditioning yields a controllable model that steers expression along axes invisible to morphology, which image-only models cannot reach,separately, a distribution-aware regression anchor keeps position competitive on held-out splits while the flow improves spatial-structure fidelity. Position accuracy on unseen tissue remains bounded by image predictability,the anchor reaches this ceiling rather than exceeding it, which we report openly.

AI Use Statement

In accordance with the ICLR 2027 policy on the use of AI, we disclose our use of AI-based tools in the preparation of this work. We used a large language model-based coding assistant (Anthropic Claude, via the Claude Code interface) for the following tasks: implementing and refactoring experiment code (encoders, the flow-matching generator, the regression anchor and coupling variants, and evaluation harnesses),orchestrating and monitoring benchmark runs,implementing the figure-

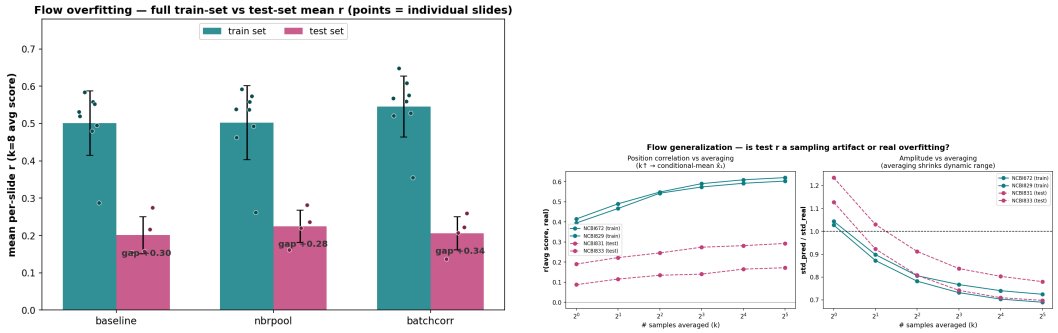

Figure 7: Generalization. Full train-set vs. test-set mean r across conditioning variants (left) and the train/test gap for the flow (right),the ~ 0.3 gap is not closed by conditioning changes.

generation code,aggregating and tabulating results,and editing the text of this paper. The assistant also helped brainstorm and refine analyses, such as the over-smoothing diagnosis and the moment/MMD auxiliary-loss experiments) and summarize prior work. AI tools were not used to generate synthetic data, and all reported numbers were produced by the described experiments on real HEST-1k / HEST-bench data.

The authors verified all AI-assisted outputs: we reviewed and tested generated code, re-ran and checked the reported metrics against the saved experiment artifacts, and confirmed that figures reflect the underlying results. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Richard J Chen et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

Bryan He et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, 2020.

Tinglin Huang et al. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. In *ICML*, 2025.

Guillaume Jaume et al. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *NeurIPS*, 2024.

Yaron Lipman et al. Flow matching for generative modeling. In *ICLR*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature Methods*, 2019.

Mohammad Lotfollahi et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 2023.

Mahmood Lab. UNI2-h: A vision transformer foundation model for histopathology. https://huggingface.co/MahmoodLab/UNI2-h, 2025. Hugging Face model card. ViT-H/14 successor checkpoint to UNI, trained on over 200M histopathology tiles.

Chenlin Meng et al. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022.

Minxing Pang, Kai Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*, 2021.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Spatia authors. Spatia: Controllable generation of spatial single-cell phenotypes. *arXiv:2507.04704*, 2025.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *NeurIPS*, 2023.

Yuansong Zeng et al. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 2022.

Sichen Zhu et al. Stem: Diffusion generative modeling for spatially resolved gene expression inference from histology images. In *ICLR*, 2025.