When Does Flow Matching Help Deterministic Multimodal Prediction: A Spatial Transcriptomics Study

Haoran Xu **Wei Yuan** **Xiao Han***
College of Biomedical Engineering
Sichuan University
Chengdu, China
xhr2023@stu.scu.edu.cn yuanw@stu.scu.edu.cn xiao_han@scu.edu.cn

Abstract

Flow matching is increasingly used beyond open-ended generation, including prediction tasks whose outputs are effectively deterministic, yet it remains unclear why flow matching helps when diversity is not the objective. We study this question in histology-conditioned spatial transcriptomics (ST) prediction, a challenging, sparse, high-dimensional task where thousands of spatially correlated spot-level observations arise from a limited number of heterogeneous tissue slides, and direct regression with pathology foundation features is already competitive. Repeated sampling with different initial Gaussian noises under the same histology input produces nearly identical predictions, indicating that output diversity does not explain the benefit of flow matching in this setting. We instead identify intermediate flow states as time-indexed target-side signals: they retain target-related information while remaining mixed with source noise. Their benefit emerges when this information is selectively routed into histology representations while preserving slide-level condition alignment. Guided by this mechanism, we instantiate conditional flow matching with a simple gated self-attention architecture to model interactions between histology tokens and noisy ST states, improving spatial-correlation metrics over regression and flow-based baselines. We further probe this mechanism through component ablations, context analyses, inference-time routing interventions, and additional evaluations in spatial proteomics and with a Qwen-gated-attention backbone. Together, these experiments provide a mechanistic account of when intermediate flow states can serve as structured target-side supervision in deterministic histology-conditioned spatial molecular prediction. Code is available **here**.

1 Introduction

Virtual spatial transcriptomics (ST) aims to infer spatially resolved gene expression directly from routine H&E-stained histology images, offering a computational alternative to costly spatial profiling assays. Recent methods span regression-based, retrieval-based, and generative approaches, with increasingly sophisticated architectures designed to capture complex morphology–expression relationships. However, the benefit of architectural complexity remains unclear. A recent unified benchmark, STP-BENCH, evaluated 21 methods across six cancer types and two ST platforms, revealing that most methods failed to outperform a simple linear-probing baseline when a shared pathology foundation encoder was used [4].

*Corresponding author: xiao_han@scu.edu.cn

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

Beyond architectural choice, virtual ST poses a distinct data-modeling challenge: its spot-level observations are spatially dependent and organized within heterogeneous tissue slides. A typical tissue slide contains thousands of spatially indexed spots, each paired with a corresponding histology patch and gene-expression profile. These spot-level observations are not independent: nearby locations exhibit spatially correlated expression, while spots across the same tissue section share broader histological and molecular context [19, 9]. Moreover, different tissue slides can exhibit distinct gene-expression distributions, reflecting their underlying tissue composition and microenvironment. Consequently, thousands of patch–spot pairs may originate from only a limited number of independent paired tissue contexts. The challenge is therefore not simply to predict gene expression independently for each patch, but to capture the spatial organization and slide-level expression distributions while generalizing across heterogeneous tissue contexts.

These characteristics motivate modeling the joint distribution of spatial gene-expression profiles across an entire tissue slide. Recent generative approaches [23, 9, 22] have explored spatially structured architectures, including spatial Transformer-based interactions, to model slide-level molecular distributions and dependencies among spots. However, we find that output diversity largely collapses under a fixed histology condition: repeated sampling with different initial Gaussian noises produces nearly identical predictions. This raises a fundamental question: if the desired conditional prediction is effectively deterministic, what does a generative formulation contribute beyond direct regression?

We hypothesize that the contribution of flow matching lies in the target-side information introduced during training. Specifically, conditional flow matching constructs time-indexed intermediate states by interpolating Gaussian noise with measured gene-expression profiles. These states carry target-related information while remaining mixed with source noise, potentially providing a form of target-side supervision beyond direct endpoint regression. Their usefulness may therefore depend on how selectively this information is extracted and integrated into histology representations.

To study this mechanism, we use a lightweight gated self-attention architecture that jointly processes histology tokens and noisy ST tokens, providing a controlled setting to investigate cross-modal and cross-location information flow without relying on a highly specialized ST architecture. Through context-matched-regression comparisons and mechanistic analyses, we examine how intermediate target states contribute to prediction. Our results indicate that their benefits depend on selective integration as shared tissue context, with slide-level context alignment being more important than exact spot-wise correspondence. We further examine the scope of these findings across additional spatial-molecular settings and model configurations.

Contributions. Our work provides a mechanistic investigation of flow matching in deterministic histology-conditioned spatial molecular prediction:

- We characterize the behavior of conditional flow matching in an effectively deterministic prediction setting, showing that repeated sampling under fixed histology inputs produces nearly identical predictions and that output diversity does not explain its observed benefit.
- We identify intermediate flow states as time-indexed target-side training signals whose usefulness depends on selective integration into histology representations.
- We conduct a comprehensive set of controlled experiments to validate this mechanistic account and disentangle the roles of intermediate target information, its integration, and slide-level context alignment across different settings and model configurations.

2 Generative Formulation for Spatial Transcriptomics Prediction

2.1 Flow Matching for Deterministic Conditional Prediction

We consider paired conditional prediction data $\mathcal{D} = \{(c_i, x_i)\}_{i=1}^n$, with $x_i \in \mathbb{R}^d$, drawn from a joint distribution $p_{\mathrm{data}}(c, x)$. Here c denotes the input condition and x denotes the target. For the flow-matching formulation, we denote the target by x_1 to emphasize its role as the terminal endpoint of a probability path.

A standard regression model learns a conditional predictor $f_\theta : \mathcal{C} \to \mathbb{R}^d$ by minimizing

$$\mathcal{L}_{\mathrm{reg}}(\theta) = \mathbb{E}_{(c,x_1)\sim p_{\mathrm{data}}} \left[\| f_\theta(c) - x_1 \|^2 \right]. \tag{1}$$

This objective directly learns the mapping $c \mapsto x_1$, using only the observed target as supervision.

We now formulate the same prediction problem using conditional flow matching [14]. Let $p_0 = \mathcal{N}(0, I_d)$ be a source distribution and let $x_1 \sim p_{\mathrm{data}}(\cdot \mid c)$ denote a target endpoint paired with condition c. The goal is to learn a conditional velocity field $u_\theta : \mathbb{R}^d \times [0,1] \times \mathcal{C} \to \mathbb{R}^d$ such that solving

$$\dot{x}(t) = u_\theta(x(t), t, c), \qquad x(0) \sim p_0, \tag{2}$$

defines a condition-dependent transport from p_0 to the target distribution $p_{\mathrm{data}}(\cdot \mid c)$. Given a training pair (c, x_1), we sample $x_0 \sim p_0$ and construct a linear interpolant to generate time-indexed noisy target states [2]

$$x_t = (1-t)x_0 + tx_1, \qquad x_0 \sim p_0, \quad x_1 \sim p_{\mathrm{data}}(\cdot \mid c), \quad t \sim U[0,1], \tag{3}$$

This path induces the conditional velocity

$$u^{\mathrm{cond}}(x_t, t, x_1) = x_1 - x_0 = \frac{x_1 - x_t}{1-t}. \tag{4}$$

Marginalizing over possible endpoints under the same condition gives the optimal conditional velocity

$$u^\star(x, t, c) = \mathbb{E}_{x_1 \mid x_t = x, t, c}\left[u^{\mathrm{cond}}(x, t, x_1)\right]. \tag{5}$$

The corresponding flow-matching objective is

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{c,t,x_t}\left[\left\|u_\theta(x_t, t, c) - u^\star(x_t, t, c)\right\|^2\right]. \tag{6}$$

Since $u^\star$ is generally intractable, conditional flow matching uses sampled conditional velocities:

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{\substack{(c,x_1)\sim p_{\mathrm{data}} \\ x_0 \sim p_0,\, t \sim U[0,1]}}\left[\left\|u_\theta(x_t, t, c) - (x_1 - x_0)\right\|^2\right]. \tag{7}$$

This formulation clarifies the distinction from direct regression. During training, x_1 is used not only as an endpoint loss target, but also to construct intermediate states x_t and the corresponding velocity targets. These noisy states expose the model to time-indexed target-side information while mixing it with source noise. At inference, the measured endpoint x_1 is unavailable; the model starts from $x(0) \sim p_0$ and integrates $u_\theta(x(t), t, c)$, where the current state $x(t)$ is updated by the solver.

2.2 Histology-Conditioned ST Prediction with Multi-Sample Token Coupling

We instantiate conditional flow matching in histology-conditioned spatial transcriptomics. For a slide with m spots, let $H = \{h_i\}_{i=1}^m$ and $X = \{x_i\}_{i=1}^m$, where h_i is the histology representation of the i-th tissue patch and $x_i \in \mathbb{R}^G$ is its paired expression vector over G genes. The slide-level goal is to predict the spatial transcriptomic profile conditioned on histology $X \sim p_{\mathrm{data}}(\cdot \mid H)$.

A direct spot-wise formulation would treat each histology–ST pair (h, x) as an independent conditional sample. For each pair, we set $x_1 = x$, $c = h$, and construct the intermediate state $x_t = (1-t)x_0 + tx_1$, with $x_0 \sim p_0$. However, this independent view ignores the fact that ST profiles are spatially organized: nearby and distant tissue locations may share tissue-level structure, gene co-expression patterns, and global histological context. We therefore introduce multi-sample token coupling. At each training step, we sample M histology–ST pairs and process them jointly as a multimodal token sequence $\mathcal{S}_t = [h_{i_1}, \ldots, h_{i_M}; x_{t,i_1}, \ldots, x_{t,i_M}]$. This joint processing allows each prediction to use a multi-location tissue context. Although training examples are initially formed from measured histology–ST pairs, the model operates on groups of tokens rather than isolated spot-wise pairs. Across epochs, reshuffling exposes each spot to different context groups, providing a simple way to learn cross-location and cross-modal dependencies. As we show in Section 4, the benefit of target-side states does not require exact spot-wise correspondence within a group.

We adopt an x-prediction parameterization with a velocity loss. Recent generative modeling studies suggest that predicting clean data can be more stable than predicting noise or velocity in high-dimensional spaces [13]; as shown by the additional comparison in Appendix C, this also holds for

[2] We use the linear interpolant for simplicity. Flow matching also supports more general probability paths and time parameterizations, such as diffusion-style paths, and cosine schedules [14, 21].

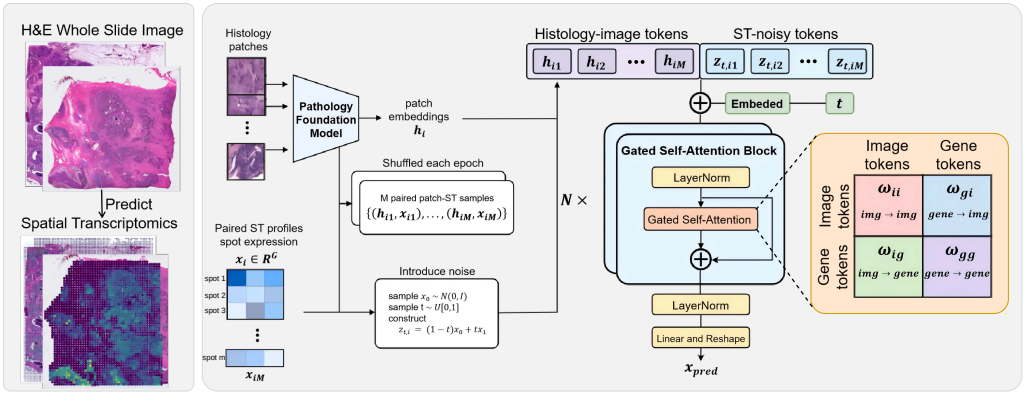

Figure 1: **Overview of the minimal gated self-attention architecture used to diagnose flow-matching gains.** Given an H&E whole-slide image, patch embeddings are extracted by a pathology foundation model and paired with measured ST profiles. At each epoch, patch-ST pairs are shuffled and grouped into M-sample contexts. For each group, noisy ST states are constructed by interpolating Gaussian noise with the measured expression profiles, then concatenated with histology tokens and processed by gated self-attention. The attention map is decomposed into image-image, image-ST, ST-image, and ST-ST blocks to inspect cross-modal and cross-location interactions.

sparse ST prediction. Specifically, the network predicts the clean endpoint $x_\theta = \mathrm{net}_\theta(x_t, t, c)$, and the velocity used for integration is recovered as $u_\theta(x_t, t, c) = (x_\theta - x_t)/(1-t)$. Since the target velocity is $u = (x_1 - x_t)/(1-t) = x_1 - x_0$, the velocity loss under this parameterization becomes

$$\mathcal{L}_{x\text{-pred},v\text{-loss}}(\theta) = \mathbb{E}_{\substack{(c,x_1)\sim p_{\mathrm{data}} \\ x_0\sim p_0,\ t\sim U[0,1]}} \left[\frac{1}{(1-t)^2} \left\| x_\theta - x_1 \right\|^2 \right]. \tag{8}$$

In practice, we compute the denominator as $\max(1-t, \epsilon)$ with $\epsilon = 10^{-4}$. This parameterization makes the connection to regression explicit. Both regression and x-prediction flow supervise the same endpoint x_1, but conditional flow matching exposes the model during training to the intermediate state x_t and time index t. This leads to the central question of our paper: *if the final task is deterministic, why does this generative formulation improve over regression?*

Appendix A provides complementary evidence on the spatial dependence and between-slide heterogeneity of the paired ST data. Section 3 then uses controlled experiments to disentangle when this target-side information is useful and when it is not.

3 Diagnosing the Source of Flow-Matching Gains

3.1 Minimal Self-Attention Architecture

To diagnose where the gain of flow matching comes from, we use a simple token architecture, as shown in Figure 1. The goal is not to introduce a specialized ST model, but to minimize architectural confounders and make the behavior of the flow formulation easier to inspect. In this controlled setting, regression and flow matching share nearly the same backbone, while the interactions among histology tokens, noisy ST states, and sampled tissue locations remain observable.

Following Section 2.2, each training step groups M shuffled histology–ST pairs, constructs noisy ST states from Gaussian sources and measured endpoints, and processes histology and ST-state tokens in one sequence. The output head predicts the clean ST endpoint. The initial training examples are measured histology–ST pairs, but each prediction is made in a multi-location token context rather than from an isolated spot-wise pair.

To inspect these interactions, we decompose attention into image-to-image, image-to-ST, ST-to-image, and ST-to-ST blocks. The full layer-wise attention visualization is provided in Appendix D; the main text focuses on controlled interventions that test the functional role of these routes.

3.2 Matched Comparisons and Controlled Ablations

We first ask whether the proposed formulation improves over strong regression when the histology features, data splits, and multi-token backbone are controlled. Table 1 compares the original pathology-feature regression, STFlow, Context-Matched Regression (CM-Regression), and our flow model across six cohorts. CM-Regression shares the feature representation, gated multi-token backbone, and splits with our model, but removes noisy ST states and time conditioning.

Table 1: **Matched cross-dataset comparison in PCC-All.** Results are averaged over predefined splits. CM-Regression is a stronger baseline we design, which shares the multi-token backbone and histology features with Ours, but removes noisy ST states. Complete metrics are reported in Appendix C, Tables 7 and 8.

Method	COAD	LYMPH	LUNG	PAAD	PRAD	CCRCC	Mean
Original Regression (PCA-Ridge)	0.4038	0.2950	0.5832	0.4862	0.3021	0.2481	0.3864
STFlow	0.4010	0.2615	0.6273	0.4852	0.2983	0.2588	0.3887
CM-Regression	0.4357	0.3083	0.6225	**0.4962**	0.3576	0.2827	0.4172
Ours	**0.4591**	**0.3233**	**0.6350**	0.4921	**0.3777**	**0.3140**	**0.4335**

Our simple gated self-attention architecture consistently outperforms the original regression baseline across all six cohorts, whereas STFlow improves over regression in only two. The comparison with CM-Regression further highlights the contribution of shared slide-level context through multi-spot coupling, while the additional gains of Ours point to the importance of gated routing of intermediate target states; we investigate these contributions in the following ablations.

To further disentangle these contributions, we construct a larger IDC-80K dataset to systematically examine the roles of intermediate target states, slide-level context, time conditioning, and gated interaction. For the ablations in Table 2, we use global random sampling rather than spatially local sampling; the effect of this sampling strategy is examined in Table 14. We perform controlled ablations by removing or altering each component while keeping the evaluation setting fixed.

Table 2: **Controlled ablations on IDC-80K.** The table compares the full model with controlled variants that alter intermediate target states, time conditioning, or the attention mechanism.

Method	Variant	PCC-200	PCC-100	PCC-50	MSE	MAE
Regression (PCA-Ridge)	–	0.635	0.804	0.874	6.546	2.124
STFlow	–	0.640	0.809	0.882	1.513	0.870
Gated self-attention	Full model	**0.660**	**0.824**	**0.888**	**1.208**	**0.803**
	With clean x	0.642	0.809	0.874	1.546	0.912
	With permuted x_t	0.657	0.819	0.884	1.229	0.813
	Without t embedding	0.651	0.808	0.883	1.366	0.859
Pure self-attention	Baseline	0.643	0.813	0.877	1.570	0.928
	Without x_t	0.644	0.812	0.877	1.503	0.896

The full model achieves the best performance across all reported metrics. Simply adding intermediate target states is insufficient: under pure self-attention, including x_t yields virtually no improvement over excluding it (0.643 vs. 0.644 PCC-200), whereas the full gated model reaches 0.660. Replacing intermediate states with clean targets or permuting x_t also reduces performance (0.642 and 0.657 PCC-200, respectively), while removing time conditioning leads to a further decrease to 0.651. These results highlight the complementary roles of gated interaction, intermediate target states, and their correspondence with the histology context, which we further investigate through context-alignment analyses.

4 When and How Do Intermediate States Help?

Before examining when intermediate states are useful, we first characterize the conditional output behavior of our model. Under fixed histology inputs, repeated sampling with different initial Gaussian noises produces nearly identical predictions. Across 20 initial-noise draws, the median maximum gene-wise discrepancy is 5.0×10^{-6} (95th percentile 1.0×10^{-5}; maximum 1.24×10^{-5}; Appendix D). This indicates that the learned conditional flow behaves as an effectively deterministic predictor, rather than exploiting sampling diversity to generate alternative outputs. We therefore investigate the conditions under which intermediate target states improve prediction and how their information should be integrated.

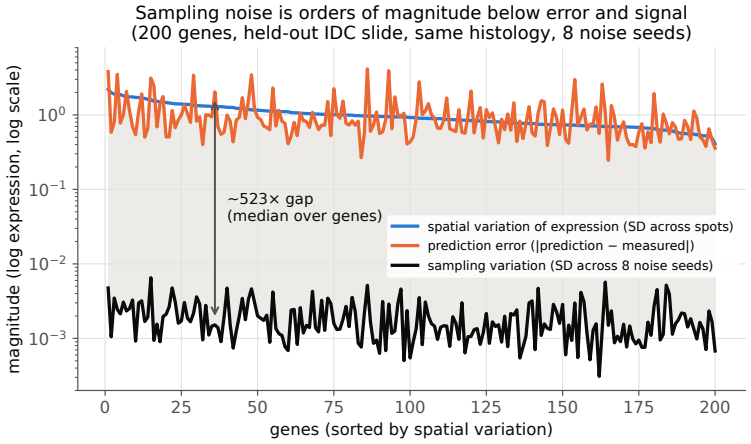

Figure 2: **Sampling noise is orders of magnitude below prediction error and spatial signal.** On a held-out IDC slide with fixed histology, we evaluate 200 genes across eight initial-noise seeds. The median sampling variation (standard deviation across seeds) is approximately 523 times smaller than the prediction error and is also much smaller than the spatial variation of measured expression.

4.1 When Does Flow Matching Help Across Data Regimes?

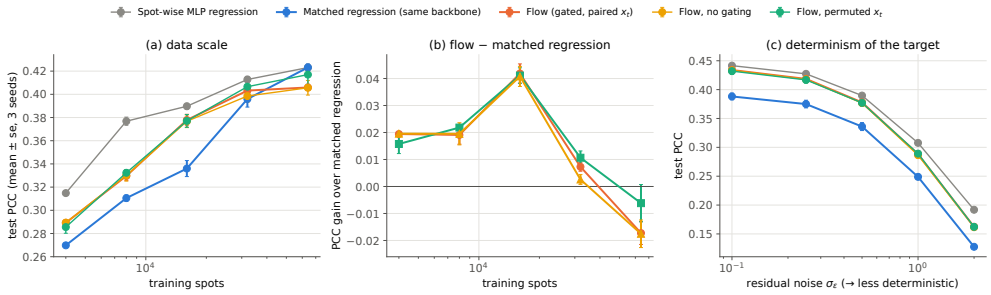

Figure 3: **The benefit of flow matching depends on paired-data scale and target determinism.** (a) Test PCC as the number of training spots increases. (b) PCC gain over context-matched regression. Gains are positive in several smaller paired-data regimes but diminish and can become negative with more training spots. (c) Test PCC as residual target noise increases (i.e., as the target becomes less deterministic). Curves show mean and standard error over three seeds where applicable.

We first examine how the relative performance of flow matching and context-matched regression changes with the amount of paired training data. As shown in Figure 3, flow matching shows advantages in several smaller-data regimes, whereas CM-Regression becomes increasingly competitive as more paired spots are available. The relative benefit of flow matching therefore depends on the training-data regime, rather than increasing monotonically with model or dataset scale.

This comparison suggests that intermediate target states can provide useful additional supervision when paired histology–ST observations are limited. However, as the amount of paired data increases, direct learning through the shared multi-spot architecture can become increasingly effective, narrowing the difference between the two formulations. These results motivate a closer examination of how intermediate states are exposed to the model and how their information is integrated with histology representations.

4.2 How Should Intermediate States Be Integrated?

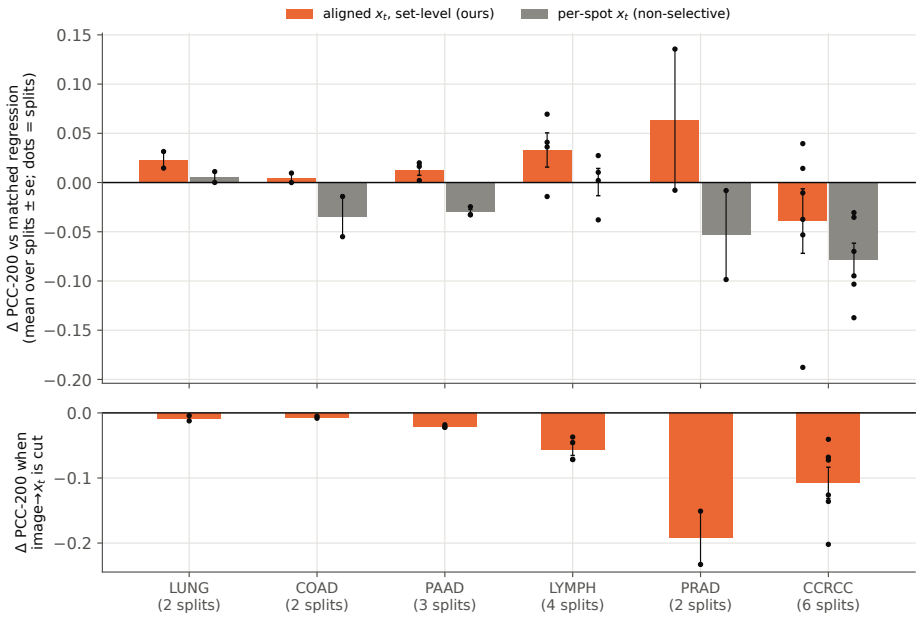

Figure 4: **Set-level aligned target states are more effective than non-selective per-spot exposure.** Top: change in PCC-200 relative to context-matched regression across six cohorts, aggregated over dataset splits (dots) and shown as mean ± s.e. Our aligned set-level design improves over matched regression on most cohorts, whereas per-spot x_t is less reliable and often degrades performance. Bottom: change in PCC-200 when the image-to-x_t route is cut in the aligned set-level model. The resulting drop shows that the gain depends on selectively routing histology information into target-side states, rather than simply making x_t available.

We next compare two ways of exposing the model to intermediate target states: our aligned set-level design, which provides multiple x_t tokens as shared context, and a non-selective per-spot alternative that injects x_t directly for each spot. As shown in Figure 4, aligned set-level x_t is generally more effective than per-spot exposure. Relative to context-matched regression, it improves PCC on most cohorts, whereas per-spot x_t is less reliable and often degrades performance. The gains are not uniform, however, and CCRCC remains a clear exception.

We further test whether these gains come simply from adding target-side states by cutting the image-to-x_t route. This substantially weakens the aligned set-level formulation, indicating that the benefit depends on selective routing from histology into target-side states rather than on the mere presence of x_t. These results show that intermediate states are most useful when treated as shared set-level context and selectively integrated with histology representations.

4.3 Tissue-Context Alignment Matters

The results in Figure 5 suggest that exact spot-wise correspondence is not required for effective set-level integration. We further examine the role of tissue context by comparing original target states, within-slide shuffled states, and states from different slides. Shuffling target states within the same slide has little effect on PCC, whereas using states from different slides leads to clear degradation.

These results indicate that preserving slide-level context is important, even when exact spot-wise correspondence is relaxed. Intermediate states can therefore provide useful shared molecular context across spots, provided that they remain consistent with the histology condition.

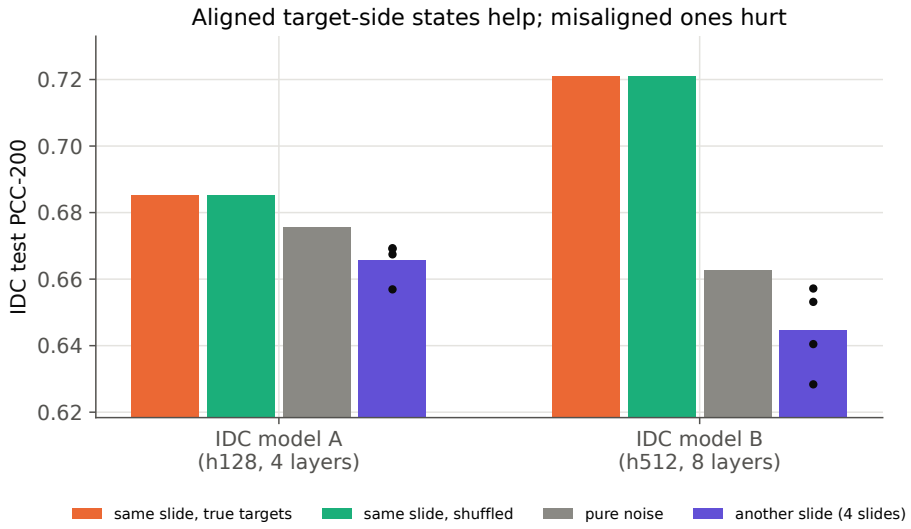

Figure 5: **Tissue-context alignment is more important than exact spot-wise pairing.** One-pass oracle probes on two IDC models at $t = 0.95$, comparing true target states, within-slide shuffled states, pure noise, and states from another slide. Dots indicate the four source slides used for misaligned states. This is a diagnostic probe using measured test expression.

4.4 Late-Time Per-Spot States Can Create a Shortcut

Although intermediate states can provide useful target-side information, exposing the model to states too close to the target endpoint can introduce a shortcut. We investigate this by varying the maximum training time $t_{\max}$ for per-spot state injection (Figure 6). Under velocity loss, increasing $t_{\max}$ from 0.3 to 0.95 reduces test PCC from 0.379 to 0.274, while the endpoint-loss variant is substantially less sensitive to this change.

Further analysis shows that later training states increase the model's reliance on oracle x_t, suggesting that near-endpoint states provide an overly direct route to target information rather than encouraging robust histology-conditioned prediction. This highlights that the usefulness of intermediate states depends not only on preserving tissue context, but also on avoiding overly direct access to target-side information.

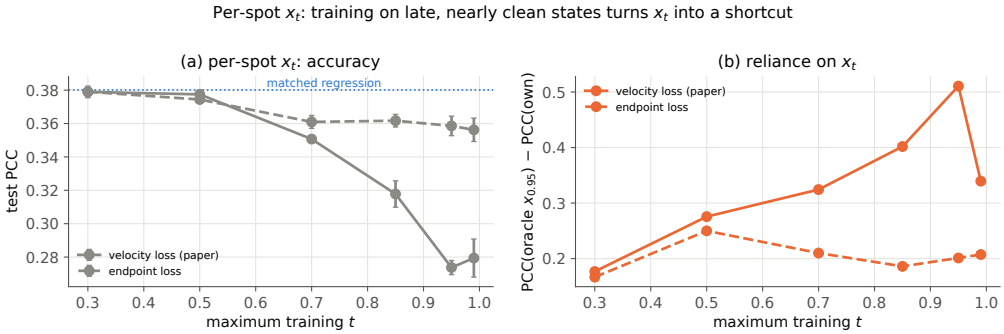

Figure 6: **Late-time per-spot states can become a shortcut.** Controlled simulation showing test PCC and reliance on oracle x_t as the maximum training time increases. Endpoint loss reduces the degradation observed with the paper's velocity-loss setup.

4.5 Which Information Pathways Does the Model Use at Inference Time?

Finally, we examine which information pathways the trained model relies on through inference-time attention interventions (Table 3). We selectively mask individual attention routes without retraining, evaluating both the minimal gated backbone and the Qwen-gated-attention [18] backbone.

Blocking histology-to-state attention (no_ig) consistently reduces PCC, by 0.041 on the minimal backbone and 0.033 on the Qwen-style backbone. In contrast, blocking the reverse state-to-histology route (no_gi) has negligible effects on PCC in both models. Interventions on target-state self-attention also produce smaller changes. This asymmetric response indicates that histology-to-state information routing plays a more important role in the learned predictions, with a consistent pattern across the two backbones. Notably, the effects on MSE and MAE are not always consistent with those on PCC, suggesting that this dependence is particularly evident in spatial-correlation performance.

Table 3: **Inference-time routing interventions on IDC.** `no_ig`: histology queries cannot attend to ST-state tokens; `no_gi`: the reverse direction; `no_gg`: no attention among ST-state tokens.

Backbone	Intervention	PCC	MSE	MAE
Ours	None	0.686	1.846	1.027
	`no_ig`	0.645	1.615	0.921
	`no_gi`	0.685	1.364	0.898
	`no_gg`	0.683	1.811	1.000
Qwen-style	None	0.681	1.590	0.878
	`no_ig`	0.648	1.383	0.854
	`no_gi`	0.681	1.578	0.876

5 Conclusion, limitation and broader impact

Conclusion. We investigate when and why flow matching helps effectively deterministic histology-conditioned spatial molecular prediction. Our findings show that its benefit lies not in sampling diversity, but in the use of intermediate flow states as structured target-side supervision. This benefit depends on the paired-data regime, shared tissue context, and selective integration of target-side information, while naive per-spot injection can introduce shortcuts. Together, these results provide a mechanistic understanding of how flow matching can improve deterministic ST prediction beyond direct regression.

Limitations. Our analysis focuses primarily on histology-conditioned ST prediction, and the observed gains are not uniform across datasets or spatial-molecular modalities. Although we additionally evaluate spatial proteomics and a Qwen-style backbone, broader generalization to other prediction tasks and architectures remains to be established.

Broader impact. Our findings may inform the design of more effective generative training strategies for structured biological prediction. However, computationally predicted molecular profiles require experimental validation and should not replace measured data in clinical decision-making.

Acknowledgments and Disclosure of Funding

This work was supported by the Fundamental Research Funds for the Central Universities (1082204112M91) and the Sichuan Provincial Natural Science Foundation (Grant No. 2026NSFSC0661). The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Quentin Bertrand, Anne Gagneux, Mathurin Massias, and Rémi Emonet. On the closed-form of flow matching: Generalization does not arise from target stochasticity. In *NeurIPS*, 2025.

[2] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

[3] Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11591–11600, 2024.

[4] Youngmin Chung, Ji Hun Ha, Andrew H Song, Cristina Almagro-Pérez, Chaeyoung Seo, Won Jun Suh, Jeong Won Beom, Kyoung Bin Oh, Eytan Ruppin, Faisal Mahmood, et al. Stp-bench: a unified systematic benchmark for virtual spatial transcriptomics from histopathology images. *arXiv preprint arXiv:2609.05956*, 2026.

[5] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568(7751):235–239, 2019.

[6] Aniruddha Ganguly, Debolina Chatterjee, Wentao Huang, Jie Zhang, Alisa Yurovsky, Travis Steele Johnson, and Chao Chen. Merge: Multi-faceted hierarchical graph-based gnn for gene expression prediction from whole slide histopathology images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15611–15620, 2025.

[7] 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.

[8] Shanshan He, Ruchir Bhatt, Carl Brown, Emily A Brown, Derek L Buhr, Kan Chantranuvatana, Patrick Danaher, Dwayne Dunaway, Ryan G Garrison, Gary Geiss, et al. High-plex imaging of rna and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nature biotechnology*, 40(12):1794–1806, 2022.

[9] Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Wengong Jin, and Rex Ying. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. *ICML*, 2025.

[10] Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Rex Ying, and Wengong Jin. Stpath: a generative foundation model for integrating spatial transcriptomics and whole-slide images. *NPJ Digital Medicine*, 8(1):659, 2025.

[11] Zahra Kadkhodaie, Florentin Guth, Eero Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations*, 2024.

[12] Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.

[13] Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

[14] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ICLR*, 2022.

[15] Artem Lukoianov, Chenyang Yuan, Justin Solomon, and Vincent Sitzmann. Locality in image diffusion models emerges from data statistics. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2025.

[16] Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature methods*, 19(5):534–546, 2022.

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

[18] Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *Advances in Neural Information Processing Systems*, 38:100092–100118, 2026.

[19] 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.

[20] Robert R Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L Marshall, Daniela J Di Bella, Paola Arlotta, Evan Z Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature biotechnology*, 39(3):313–319, 2021.

[21] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *TMLR*, 2024.

[22] Haoran Xu, Yang Liu, Wei Yuan, and Xiao Han. BioFlow: A Biologically Valid Support-Preserving Flow for Histology-Conditioned Spatial Transcriptomics Prediction . In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, volume LNCS 16891. Springer Nature Switzerland, September 2026.

[23] Sichen Zhu, Yuchen Zhu, Molei Tao, and Peng Qiu. Diffusion generative modeling for spatially resolved gene expression inference from histology images. *ICLR*, 2025.

A Spatial Dependence and Slide Heterogeneity in ST Data

The virtual ST setting has an important statistical structure: spots within the same tissue section are spatially related, while different slides may exhibit different molecular distributions. These observations motivate the multi-location token contexts introduced in Section 2.2. They characterize the paired data; by themselves, they do not establish an advantage of flow matching over a context-matched regression model.

A.1 Spatial Dependence Across Spots

Figure 7 summarizes the similarity of measured gene expression and UNI histology features as a function of spot separation. Both correlations decrease with spatial distance, with the strongest associations between nearby spots. Consequently, treating all spot-level measurements as independent overlooks the spatial dependence present in the prediction target and its histology condition.

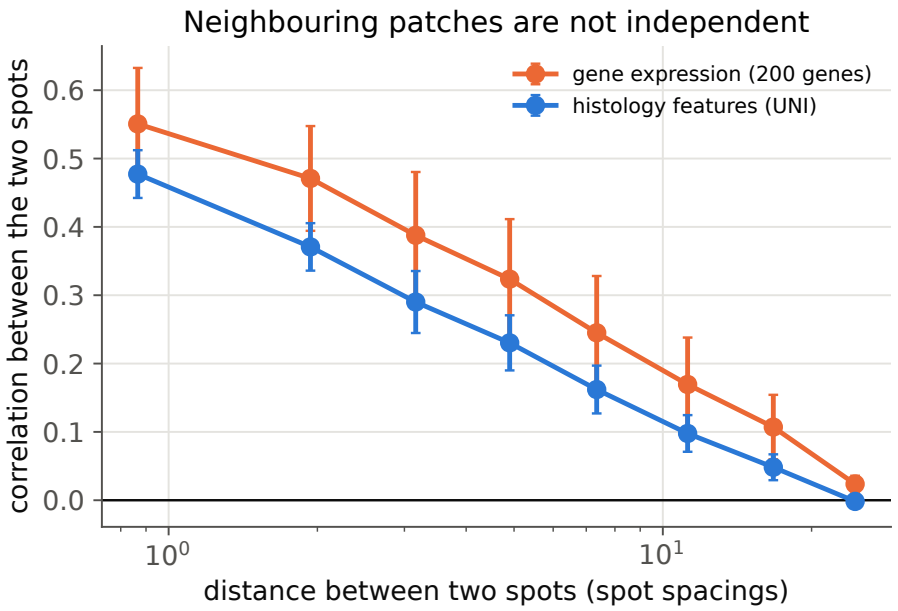

Figure 7: **Spatial dependence in the paired histology–ST data.** Correlation between pairs of spots as a function of their spatial separation (in spot spacings), shown for the measured 200-gene expression profiles and UNI histology features. Both exhibit decreasing similarity with distance, motivating multi-spot context modeling.

A.2 Heterogeneity Across Tissue Slides

Figure 8 illustrates differences among slides in the same ST collection. Spots from different slides show partially distinct regions in the PCA projection of measured gene expression, and their distributions of total log-expression differ. Thus, a large number of paired spots need not represent an equally large number of distinct tissue contexts. This slide-level heterogeneity motivates examining whether target states supply useful shared context and how important their alignment with the histology condition is (Section 4.3).

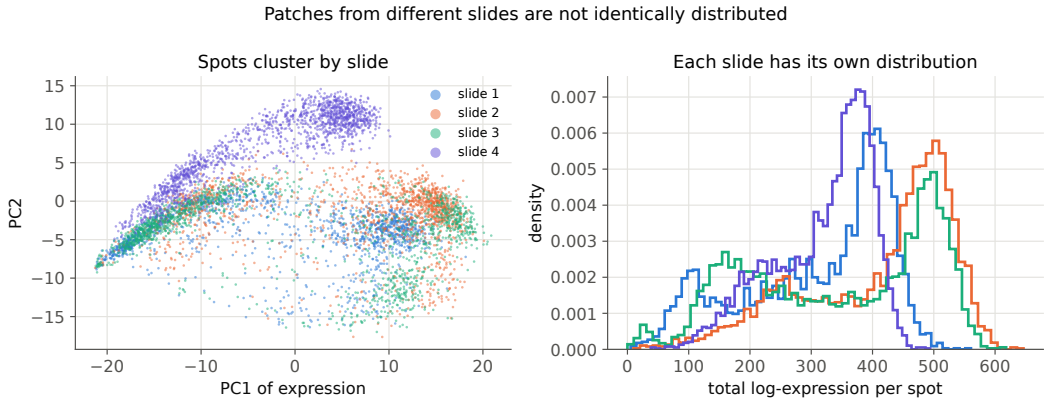

Figure 8: **Slide-level heterogeneity in spatial gene expression.** Left: PCA projection of spot-level expression colored by slide. Right: distributions of total log-expression per spot for the same slides. The displayed differences illustrate that measured spots are not identically distributed across tissue sections.

B Experimental Setup and Implementation Details

B.1 Implementation Details

To facilitate reproducibility, we provide the hyperparameter configuration used in our experiments in Table 4; exact configurations of each run are in the code repository. The cross-dataset comparison and IDC-80K ablations use the same expression representation and modeling setup, including the gene-rank embedding and non-negativity constraint. All runs use image-condition dropout ($p = 0.5$). Training sources are Gaussian; sampling in gene space starts from a zero-inflated negative binomial prior.

Table 4: Hyperparameter settings used in our experiments.

Hyperparameter	Value
Hidden dimension	128
Number of layers	4
Number of attention heads	4
Dropout	0.2
Attention dropout	0.2
Time modulation	concat
Flow parameterization	x-prediction, linear path
Image-condition dropout	0.5
Learning rate	5×10^{-4}
Random sample batch size	128
Integration steps	50

B.2 Dataset Statistics

We evaluate our method on seven spatial transcriptomics cancer cohorts spanning both Xenium and Visium platforms: IDC, COAD, LYMPH, LUNG, PAAD, PRAD, and CCRCC. As summarized in Table 5, the evaluation covers 48 patients, 80 spatial transcriptomics samples, and 356,368 measured spatial spots in total. This scale is important because histology-conditioned ST prediction must generalize not only across slides, but also across cancer types, profiling technologies, and spot-scale regimes. This diversity allows us to assess model robustness across different tissue contexts, platform-specific resolutions, and sample-size regimes.

Table 5: Dataset statistics for the spatial transcriptomics cohorts used in our experiments. We report the cancer type, number of patients, number of samples, spatial transcriptomics technology, average number of spots per sample, sample identifiers, and total number of spots.

Dataset	Patients	# Samples	Technology	Sample ID	Total Spots
COAD	3	4	Xenium	TENX111, TENX147, TENX148, TENX149	20316
LYMPH	4	4	Visium	NCBI681, NCBI682, NCBI683, NCBI684	19960
LUNG	2	2	Xenium	TENX118, TENX141	3888
PAAD	3	3	Xenium	TENX116, TENX126, TENX140	8340
PRAD	2	23	Visium	MEND139–MEND162	56442
CCRCC	24	24	Visium	INT1–INT24	67008
IDC-small	4	11	Xenium	TENX191–TENX202	80090
IDC-large	10	20	Xenium	TENX94–TENX99 TENX191–TENX202	180414

B.3 Computational Cost

All experiments were conducted on a single NVIDIA RTX PRO 6000 GPU. The total GPU memory is 97,887 MiB, while the observed memory utilization during training was typically around 10–20%. We report the wall-clock time for each dataset in Table 6. The reported time includes both training and evaluation, since the inference stage is comparatively short and difficult to measure reliably in isolation.

Table 6: Runtime on different datasets.

Dataset	Time (min)
COAD	169
PAAD	63
CCRCC	275
LUNG	89
LYMPH	78
PRAD	123
IDC-small	26
IDC-large	29

B.4 Clipping and Time Sampling

The factor $1/(1-t)$ appears when converting the predicted endpoint into a velocity. We compute this term with denominator clipping, $\max(1-t, 10^{-4})$, implemented as `(1-t).clamp_min(1e-4)`.

C Extended Benchmark Results

This section reports complete metrics, additional evaluation settings, and supplementary benchmark comparisons for the main experiments. Unless noted, results follow the data splits and evaluation conventions described in Appendix B.

C.1 Complete Matched Comparison and Seed Variability

Table 7 gives all metrics for the matched comparison in Table 1 (seed 1). Table 8 reports three-seed results and additional configuration results.

Table 7: Complete metrics, seed 1, mean over predefined splits; Macro weights datasets equally.

Dataset	Method	PCC	MSE	MAE	Spearman	LISA
COAD	Original Regression	.4038	6.1484	1.9321	.3896	.2121
	STFlow	.4010	1.8582	.9816	.3975	.2266
	CM-Regression	.4357	1.8213	.9843	.4270	.2348
	Ours	.4591	1.8812	.9958	.4476	.2475
LYMPH	Original Regression	.2950	.8935	.6735	.2817	.1499
	STFlow	.2615	.7133	.5779	.2772	.1237
	CM-Regression	.3083	.6711	.6073	.2965	.1573
	Ours	.3233	.5708	.5658	.3092	.2008
LUNG	Original Regression	.5832	5.2472	1.9273	.5379	.3649
	STFlow	.6273	1.1123	.7955	.5815	.4079
	CM-Regression	.6225	1.1005	.7967	.5733	.3977
	Ours	.6350	1.0741	.7898	.5929	.4187
PAAD	Original Regression	.4862	3.0708	1.3694	.4507	.2450
	STFlow	.4852	.9461	.7389	.4525	.2493
	CM-Regression	.4962	.8509	.7069	.4587	.2492
	Ours	.4921	.8554	.6980	.4635	.2490
PRAD	Original Regression	.3021	1.4182	.8019	.2953	.0789
	STFlow	.2983	.6735	.5779	.2694	.0933
	CM-Regression	.3576	.5986	.5755	.3440	.1333
	Ours	.3777	.6297	.5712	.3597	.1417
CCRCC	Original Regression	.2481	1.0870	.7667	.2419	.0777
	STFlow	.2588	.6266	.5880	.2743	.0803
	CM-Regression	.2827	.6038	.6040	.2780	.0833
	Ours	.3140	.5888	.5895	.3033	.0865
Macro	Original Regression	.3864	2.9775	1.2451	.3662	.1881
	STFlow	.3887	.9883	.7100	.3754	.1969
	CM-Regression	.4172	.9410	.7125	.3963	.2093
	Ours	.4335	.9333	.7017	.4127	.2240

Table 8: PCC-200, mean $\pm$ sd across seeds 1, 40, 2026. "Ours (gene space)" omits the gene-rank embedding and non-negativity constraint. Further seeds of the main configuration: LUNG 0.629, 0.631; PAAD 0.499.

Method	COAD	LYMPH	LUNG	PAAD	PRAD	CCRCC	Macro
STFlow	0.429±0.024	0.306±0.039	0.632±0.005	0.501±0.014	0.372±0.064	0.311±0.046	0.425
CM-Regression	0.438±0.003	0.308±0.000	0.624±0.001	0.496±0.001	0.357±0.015	0.290±0.008	0.419
Ours (gene space)	0.431±0.003	0.294±0.005	0.625±0.003	0.488±0.003	0.312±0.002	0.267±0.003	0.403

C.2 Controlled Ablation Variability

Table 2 reports values averaged over different random seeds. For our gated self-attention, standard errors PCC-200/PCC-100/PCC-50/MSE/MAE are 0.0014/0.0017/0.0019/0.081/0.029.

C.3 Statistical Significance Across Gene Panels

To assess whether the observed improvements are statistically reliable across datasets and random splits, we conducted paired significance tests at multiple gene-panel scales, including PCC-200, PCC-100, and PCC-50. Each paired sample corresponds to the same dataset–split unit, ensuring that our method and the compared baseline are evaluated under identical data partitions. This paired design removes confounding effects introduced by dataset difficulty and split variability.

For each metric scale, we compare our method against the original PCA-Ridge regression baseline and STFlow (not the later Context-Matched Regression baseline in the main text). Since higher Pearson correlation coefficient (PCC) indicates better predictive performance, we report one-sided paired

t-test p-values testing whether our method achieves higher PCC than the baseline. We additionally report Wilcoxon signed-rank test p-values as a non-parametric robustness check. To account for multiple comparisons across scales and baselines, we apply Benjamini–Hochberg false discovery rate correction to the paired t-test p-values.

As shown in Table 9, our method yields statistically significant improvements over the original PCA-Ridge regression baseline and STFlow across all three PCC scales. The improvements remain significant after multiple-comparison correction, suggesting that the gains are consistent across datasets and splits rather than being driven by a small number of favorable cases. Notably, the effect sizes are large when compared with the regression baseline and remain substantial against STFlow, indicating that the improvements are not only statistically significant but also practically meaningful.

Table 9: Statistical significance analysis across gene-panel scales. Each paired sample corresponds to the same dataset-split unit, resulting in $n = 19$ paired observations. Δ denotes the mean PCC improvement of our method over the compared baseline. We report one-sided paired t-test p-values, Wilcoxon signed-rank test p-values as a non-parametric robustness check, Benjamini-Hochberg corrected q-values, and paired effect size Cohen's d_z.

Metric	Comparison	Ours	Baseline	Δ	p_t^{one}	p_{wilcoxon}	q_{BH}	Cohen's d_z
PCC-200	Ours vs. Regression	0.3998	0.3529	+0.0469	2.22×10^{-4}	2.67×10^{-5}	5.69×10^{-4}	0.984
PCC-200	Ours vs. STFlow	0.3998	0.3530	+0.0468	1.34×10^{-3}	5.86×10^{-4}	2.19×10^{-3}	0.798
PCC-100	Ours vs. Regression	0.5107	0.4701	+0.0406	2.58×10^{-4}	3.62×10^{-5}	5.69×10^{-4}	0.968
PCC-100	Ours vs. STFlow	0.5107	0.4488	+0.0619	1.76×10^{-3}	8.47×10^{-4}	2.20×10^{-3}	0.770
PCC-50	Ours vs. Regression	0.5979	0.5593	+0.0386	2.33×10^{-4}	3.62×10^{-5}	5.69×10^{-4}	0.978
PCC-50	Ours vs. STFlow	0.5979	0.5347	+0.0632	1.53×10^{-3}	1.42×10^{-3}	2.19×10^{-3}	0.784

C.4 Gene-Panel Scaling and Prediction Parameterization

The full gene-panel scaling results and the comparison between endpoint (x-) prediction and vanilla velocity prediction are included here for completeness. These experiments support scalability and optimization choices but are not required for the main mechanism argument.

Table 10: Scaling to larger high-resolution IDC datasets and expanded gene panels.

Dataset	Genes	PCC-all	PCC-200	PCC-100	PCC-50	MSE	MAE
IDC-80k	200	0.660	0.660	0.820	0.885	1.333	0.849
	541 (Full)	0.475	0.752	0.846	0.885	0.618	0.515
IDC-180k	1025 (Full)	0.451	0.731	0.831	0.872	0.572	0.490

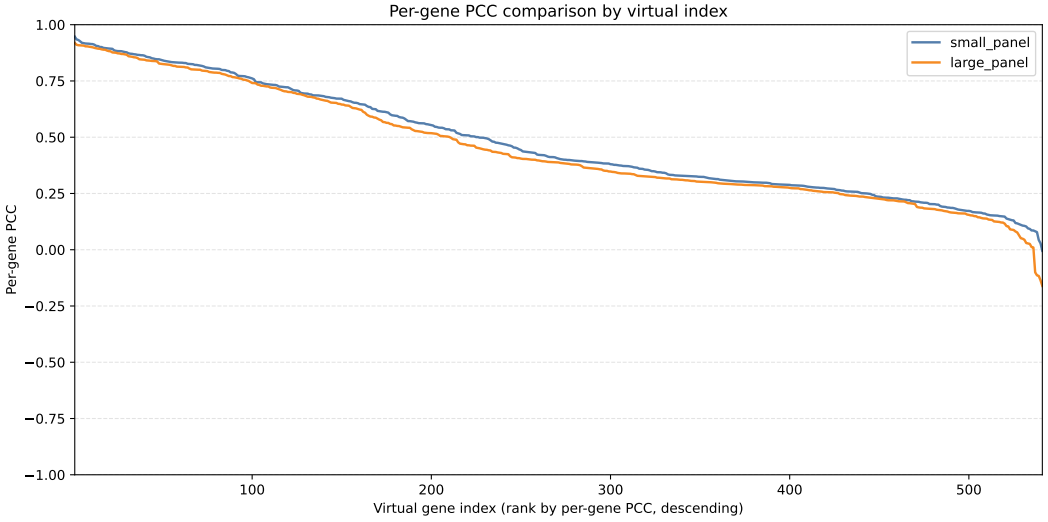

Figure 9: Expanding the gene panel preserves the per-gene predictive profile on shared genes.

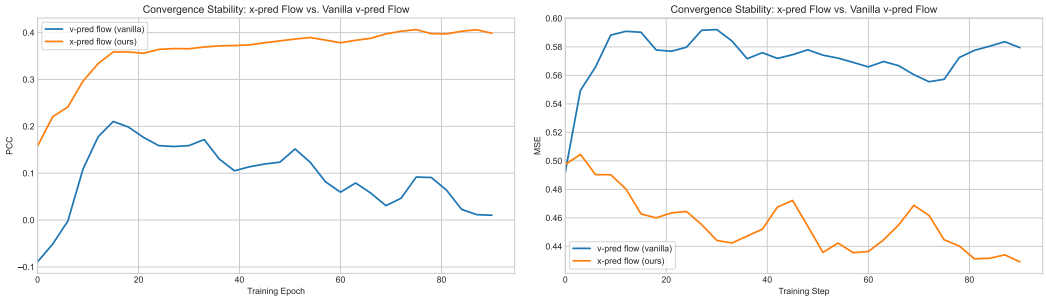

Figure 10: x-prediction yields more stable convergence than vanilla velocity prediction.

C.5 Capacity Scaling and a Second Gated Backbone

Table 11 varies width and depth on IDC, and replaces the backbone with a Qwen-style gated transformer [18].

Table 11: Capacity scaling on IDC with local $M = 128$.

Backbone	Width	Layers	LR	PCC-200	PCC-100	MSE	MAE
Ours	128	4	5×10^{-4}	0.686	0.836	1.819	0.989
Ours	256	4	5×10^{-4}	0.685	0.838	1.635	0.918
Ours	256	8	5×10^{-4}	0.674	0.832	1.616	0.917
Ours	256	8	1×10^{-4}	0.684	0.836	1.584	0.948
Ours	512	8	5×10^{-4}	0.666	0.821	1.404	0.875
Ours	512	8	5×10^{-5}	0.687	0.838	1.468	0.877
Qwen-style	256	4	1×10^{-4}	0.681	0.835	1.590	0.877
Qwen-style	384	8	5×10^{-5}	0.682	0.835	1.825	0.944

C.6 Additional Spatial-Proteomics Evaluation

We also evaluate histology-conditioned protein prediction (Table 12). The flow model outperforms CM-Regression on Visium PEX Tonsil, is comparable on CRC-CODEX and is below it on Visium PEX Pathology. The gain observed in ST therefore does not generally transfer to this modality in our evaluation.

Table 12: Spatial proteomics on held-out samples.

Dataset	Method	Markers	PCC	Spearman	MSE	MAE
Visium PEX Pathology	PCA-Ridge	35	0.693	0.664	0.223	0.312
	CM-Regression	35	**0.749**	**0.709**	**0.162**	**0.271**
	Ours	35	0.689	0.651	0.237	0.326
Visium PEX Tonsil	PCA-Ridge	35	0.563	0.605	0.236	0.296
	CM-Regression	35	0.606	0.638	**0.200**	**0.277**
	Ours	35	**0.619**	**0.651**	0.201	0.283
CRC-CODEX	CM-Regression	56	**0.274**	**0.280**	**3.101**	1.274
	Ours	56	0.269	0.275	3.112	**1.258**

C.7 Earlier 50-Gene Diagnostic Comparison

The following compact comparison was used during the initial diagnosis. The main text instead uses the later six-cohort matched comparison (Table 1); the values below are not substituted for that comparison.

Table 13: Cross-dataset PCC comparison on HEST-Bench (50-gene panels; earlier diagnostic run).

Model	PRAD	PAAD	LYMPH	IDC
Regression (PCA-Ridge)	0.294	0.481	0.256	0.635
STEM	0.261	0.434	0.203	0.628
STFlow	0.293	0.466	0.229	0.640
Ours (gated self-attention)	**0.366**	**0.493**	**0.310**	**0.660**

D Additional Mechanistic Analyses

The following visualizations and controlled probes complement the mechanistic findings in Section 4. Attention maps are descriptive, whereas route masking and oracle-state manipulation are diagnostic interventions; neither should be interpreted as a causal account of training on its own.

D.1 Set-Level and Per-Spot Integration Schematic

Figure 11 contrasts the two ways of exposing intermediate target states discussed in Section 4.2. In set-level routing, histology and target-state tokens are jointly processed as a shared context. In per-spot routing, each histology token is paired directly with its corresponding intermediate state, creating a more direct route for target copying.

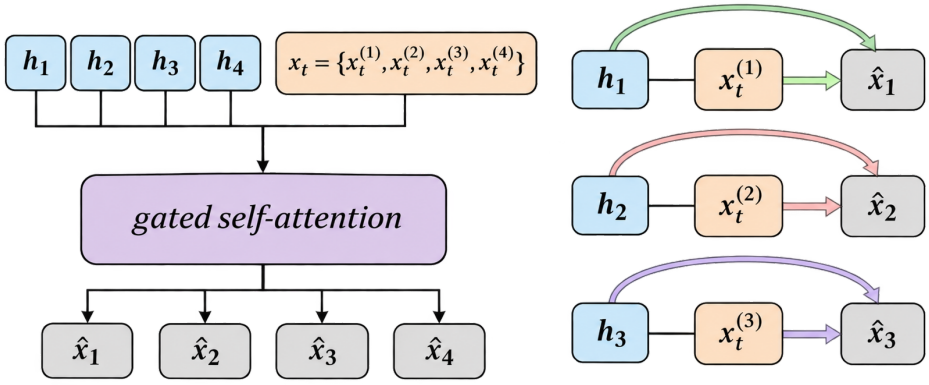

Figure 11: **Set-level context routing versus per-spot paired injection.** The left schematic shows multiple histology tokens and intermediate target states jointly processed by gated self-attention to provide shared context. The right schematic shows direct per-spot pairing between histology and intermediate target states.

D.2 Context Size and Spatial Sampling

Table 14 varies the number of spots M in a token sequence on IDC, with shuffled (global) or spatially local groups.

Table 14: Context-size analysis on IDC (PCC-200). CM-Regression uses early stopping with patience 10.

Sampling	M	Ours	CM-Regression
Global	64	0.658	0.640
Global	128	0.660	0.643
Local	128	0.686	0.626

D.3 Oracle-State Alignment Probe

The alignment plot and its interpretation are included in the main text as Figure 5. The probes use measured test expression at $t = 0.95$ and should be interpreted only as diagnostic interventions.

D.4 Additional Inference-Time Route Intervention

For completeness, the combined `no_cross` mask in the IDC inference-time intervention study yielded the same reported PCC, MSE, and MAE as `no_ig` (0.645, 1.615, and 0.921, respectively). We omit this redundant row from the main-text summary for compactness.

Table 15: Additional combined-route intervention on IDC.

Intervention	PCC	MSE	MAE
`no_cross`	0.645	1.615	0.921

D.5 Layer-Wise Attention Organization

Figure 12 provides the layer-wise attention analysis summarized in Section 3. Gating changes the distribution of image–image, image–ST, ST–image, and ST–ST interactions across depth. This is descriptive evidence of interaction organization; the inference-time route masks in Section 4.5 provide the complementary functional intervention.

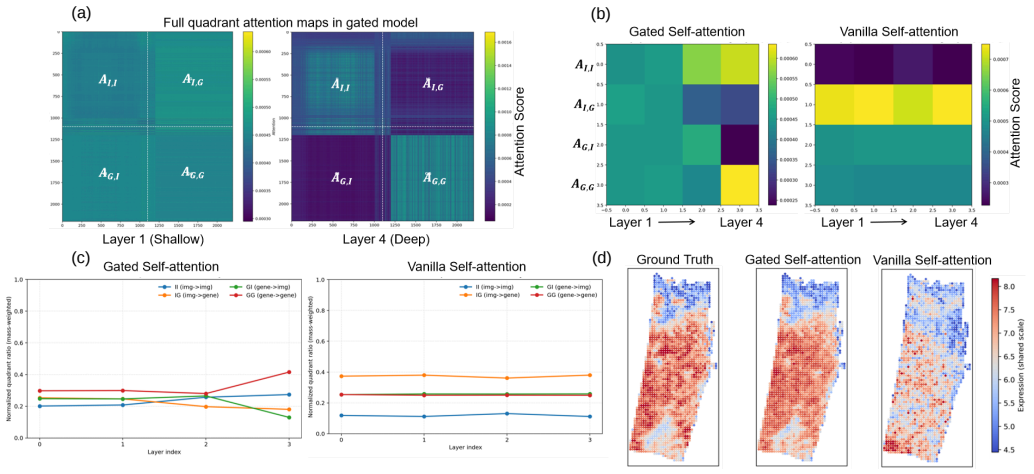

Figure 12: **Gating changes the layer-wise organization of token interactions.** (a) Full quadrant-decomposed attention maps from the gated model. (b) Layer-wise quadrant attention scores comparing gated and vanilla self-attention. (c) Mass-normalized quadrant ratios across layers. (d) Spatial prediction example for MALAT1.

D.6 Spatial Visualization of Quadrant-Wise Attention

Figure 13 provides a spatial view of the quadrant-wise attention scores for MALAT1, the representative gene shown in Figure 12d. With gating, attention patterns become differentiated across tissue locations, network layers, and interaction blocks. In contrast, vanilla self-attention produces more homogeneous maps, suggesting a less adaptive global mixing pattern. This spatial visualization is consistent with the layer-wise analysis in Figure 12: gated attention does not simply inject noisy ST-state tokens into the sequence, but organizes them into structured interaction signals across depth and modality.

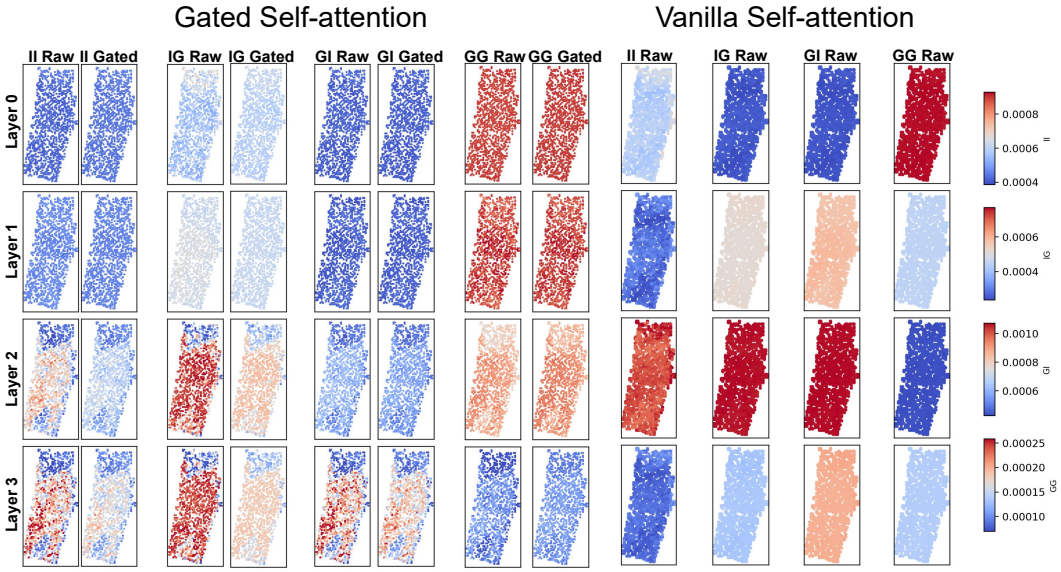

Figure 13: We project the attention scores of MALAT1 from the four interaction blocks back to tissue locations across layers. Gated self-attention produces location-, layer-, and block-dependent patterns, whereas vanilla self-attention remains largely homogeneous, indicating that gating turns noisy intermediate ST states into structured interaction signals rather than uniform token mixing.

D.7 Repeated-Sampling Output Variability

The main text reports the maximum output discrepancy across repeated initial-noise draws. Figure 14 shows the per-gene discrepancies underlying this summary.

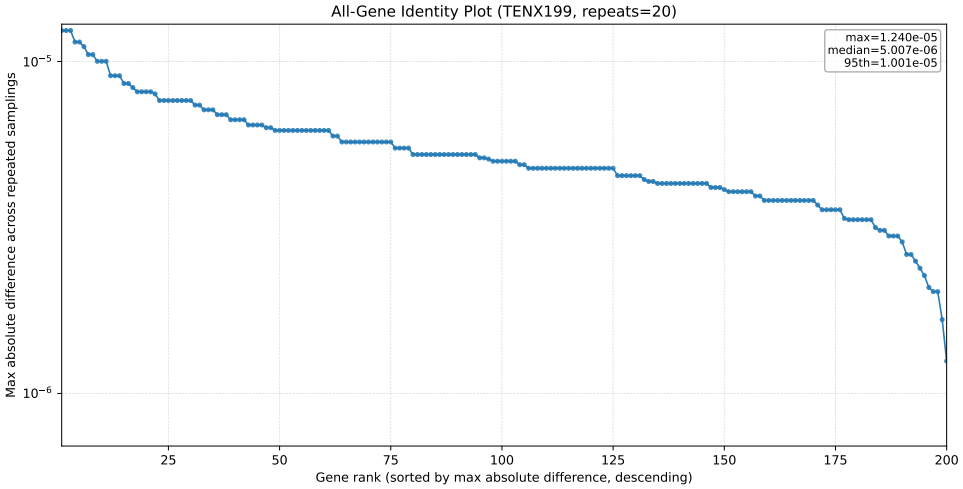

Figure 14: Repeated sampling produces nearly identical predictions under fixed histology inputs. For each gene, we rank the maximum output discrepancy across repeated samplings over all spatial locations.

D.8 Repeated-Sampling Trajectories

To complement the output-level discrepancy analysis in the main paper, we visualize repeated sampling trajectories in a shared PCA space. For each held-out sample, we fix the histology condition and vary only the initial noise. As shown in Figure 15, vanilla flow matching produces more dispersed trajectories, whereas STFlow and our formulation converge to nearly the same endpoint across repeated samplings. This supports the conclusion that, in deterministic histology-conditioned ST prediction, the flow model behaves as a near-deterministic predictor rather than relying on sampling diversity.

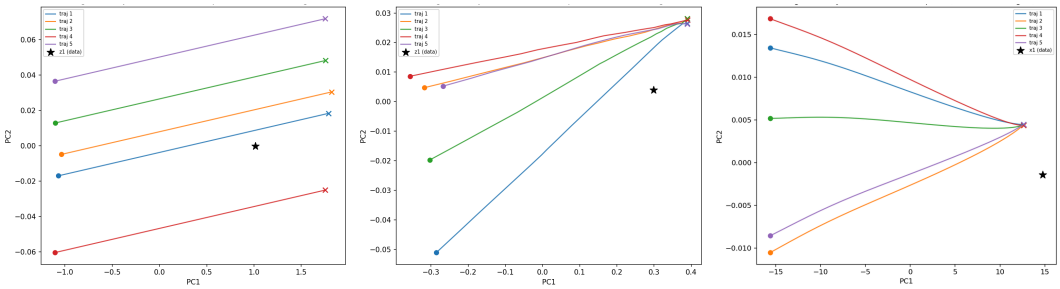

Figure 15: Repeated-sampling trajectories under fixed histology condition and different initial noises. Trajectories are projected into a shared PCA space. Vanilla flow matching (Left) shows more dispersed trajectories, while STFlow (Middle) and our formulation (Right) converge to nearly the same endpoint across repeated samplings.

E Biological and Spatial Visualizations

E.1 Representative Biomarker Examples

We further present representative biomarker examples from a held-out IDC sample to assess whether predicted ST preserves spatially resolved gene expression patterns. We selected five breast-cancer-relevant genes from the measured panel, including *BIRC5*, *BLVRA*, *CDH3*, *CENPF*, and *CXXC5*. These genes are associated with tumor proliferation, epithelial/basal-like programs, and subtype-related transcriptional variation. As shown in Fig. 16, our method produces expression maps that are visually aligned with the measured ST and achieves the best or comparable quantitative performance across these representative genes. Compared with the regression baseline, our predictions better

preserve spatial heterogeneity and avoid overly smoothed expression patterns. Compared with STFlow, our method shows improved reconstruction accuracy, suggesting that the proposed modeling strategy better captures localized biomarker expression in IDC tissue.

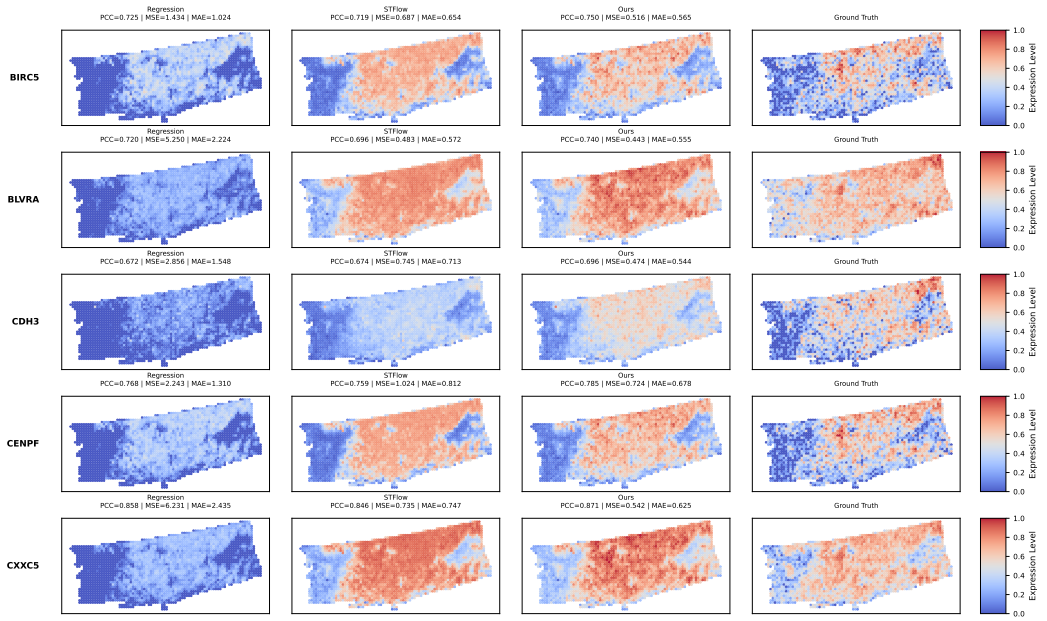

Figure 16: **Representative biomarker gene visualization on an IDC sample.** We visualize spatial expression maps of five breast-cancer-relevant genes, including *BIRC5*, *BLVRA*, *CDH3*, *CENPF*, and *CXXC5*, on a held-out IDC sample. Each row corresponds to one gene, and columns show predictions from Regression, STFlow, our method, and the measured ground-truth ST. Our method achieves higher PCC and lower reconstruction error on these representative genes, while producing spatial expression patterns that better align with the ground truth. This visualization complements the pathway-level analysis by showing that our predicted transcriptomes preserve localized biomarker expression patterns.

E.2 Pathway-Level Biological Program Reconstruction

Beyond evaluating individual gene-level prediction accuracy, we further examined whether the predicted spatial transcriptomes preserve higher-order biological programs that are relevant to invasive ductal carcinoma (IDC). We curated six breast-cancer-related gene programs from the measured 200-gene panel, including epithelial, luminal, HER2-associated, basal-like, proliferation, and invasion programs. These programs were selected to cover major tumor epithelial states, molecular subtype-associated axes, proliferative activity, and invasion-related remodeling.

For each spot and each program, we computed a pathway activity score by averaging the log-transformed non-negative expression values of genes belonging to the program:

$$s_{i,p} = \frac{1}{|\mathcal{G}_p|} \sum_{g \in \mathcal{G}_p} \log(1 + \max(x_{i,g}, 0)),$$

where $x_{i,g}$ denotes the expression of gene g at spot i, and $\mathcal{G}_p$ denotes the gene set of pathway p. The same gene sets and scoring protocol were applied to the measured ST, regression baseline, STFlow, and our method, ensuring a controlled comparison in the same expression domain used during model training.

We evaluated pathway-level reconstruction using spot-wise Pearson correlation coefficient (PCC), Spearman correlation, and mean squared error (MSE) between predicted and measured pathway scores within each slide. This analysis directly tests whether predicted ST recovers spatially varying biological programs rather than only matching isolated marker genes. As shown in Supplementary Table 16, our method achieves the highest average pathway-score PCC and Spearman correlation

across the six programs. In particular, our method improves the reconstruction of proliferation and invasion programs, which are closely associated with tumor aggressiveness and spatially heterogeneous tissue remodeling.

We further evaluated spatial organization using Global Moran's I computed on the spot-level pathway score maps. Moran's I measures whether a spatial feature is clustered, dispersed, or randomly distributed across tissue coordinates. We constructed a KNN graph on the spot coordinates and computed the absolute difference between the Moran's I of predicted and measured pathway scores. Our method achieves lower Moran-I error than STFlow while maintaining the highest pathway-score correlation, suggesting that the predicted transcriptomes preserve biologically meaningful spatial organization at the pathway level.

Overall, these results indicate that our model reconstructs pathway-level biological programs from H&E morphology, including clinically relevant IDC-associated tumor states such as luminal signaling, HER2-associated activity, basal-like programs, proliferation, and invasion.

Table 16: **Pathway-level biological program reconstruction on the IDC test slide.** For each spot, we computed activity scores for six breast-cancer-related gene programs from measured and predicted ST profiles using the same 200-gene panel. We report the spot-wise Pearson correlation coefficient (PCC), Spearman correlation, and mean squared error (MSE) between predicted and measured pathway scores. Our method achieves the highest average PCC and Spearman correlation across programs, indicating better recovery of relative spatial activity patterns. Spatial autocorrelation was further assessed by Global Moran's I on the pathway score maps.

Method	Pathway PCC ↑	Pathway Spearman ↑	Pathway MSE ↓	Moran-I Error ↓
Regression	0.789	0.806	0.453	0.081
STFlow	0.860	0.818	0.044	0.145
Ours	**0.863**	**0.830**	**0.043**	0.120

Table 17: **Curated breast-cancer-related gene programs used for pathway-level validation.** All genes were selected from the measured 200-gene panel and evaluated using the same scoring protocol across measured ST and all prediction methods.

Program	Genes
Epithelial	EPCAM, KRT19, CLDN4, TACSTD2, CD24
Luminal	ESR1, PGR, FOXA1, BCL2, PIP, SCGB2A2
HER2-associated	ERBB2, GRB7, GRB14
Basal-like	KRT5, KRT14, KRT17, EGFR, FOXC1, TP63
Proliferation	MKI67, TOP2A, BIRC5, CCNB1, CDC20, CENPF, RRM2, TYMS
Invasion	MMP11, MMP14, MMP2, LAMA3, LAMB3, LAMC2, COL17A1, VEGFA

F Additional Related Work

F.1 Flow Matching: Generalization and Memorization

Recent work has begun to analyze why flow matching generalizes in high-dimensional generative modeling. A central question in this literature is whether the stochasticity of the conditional flow-matching objective is essential for generalization, or whether similar behavior can be obtained from closed-form or more deterministic velocity targets [1]. Several studies suggest that stochastic target estimation is not necessarily the key driver of generative performance, and that learned flow or score fields may instead reflect structured memorization of the empirical data distribution, image patches, or combinations of training examples [15]. Related analyses also study the transition between memorization and generalization, as well as temporal regimes in which trajectories are initially indistinguishable, later separate by modes, and eventually point toward training samples [11, 12].

Our setting differs from this line of work in both task objective and interpretation. Prior analyses mainly focus on open-ended generation, where the concern is whether a model produces novel

diverse samples or memorizes the training set. In deterministic conditional prediction, however, a concentrated output distribution is not necessarily a failure mode: for a fixed histology condition, the desired output may be a stable molecular profile. Therefore, near-deterministic sampling should not by itself be interpreted as collapse or memorization. The relevant question is instead why the flow-matching path improves prediction over endpoint regression when diversity is unnecessary. We show that the useful effect comes from exposing the model to condition-aligned, time-indexed intermediate target states during training, which provide target-side information beyond the condition but must be selectively integrated with other modalities.

F.2 Histology-Conditioned Spatial Transcriptomics Prediction

Predicting spatial gene expression from histology whole-slide images (WSIs) has attracted increasing attention as spatial transcriptomics (ST) technologies continue to improve in resolution and scale [2, 5, 20, 8]. Because ST profiling remains experimentally expensive and requires specialized platforms [16], computational prediction from routinely available histology images provides a practical way to expand molecular annotation beyond measured ST samples. At the same time, the growing availability of high-resolution paired WSI–ST datasets creates an opportunity to learn pathology representations grounded in molecular structure, which may further support joint pretraining of pathology foundation models with image- and transcriptome-level supervision [10].

Early histology-conditioned ST prediction methods commonly formulate the task as spot-wise supervised regression. ST-Net [7] predicts gene expression from local histology patches using CNN-based encoders, while HisToGene [17] introduces positional embeddings and transformer modules to incorporate spatial layout. TRIPLEX [3] further extends this paradigm by using multi-scale tissue patches to aggregate local and broader histological context. These methods demonstrate that morphology contains predictive signal for spatial gene expression, but they largely rely on mapping histology features directly to endpoint expression values. As a result, they provide limited mechanisms for modeling gene-expression interactions across spots or for exploiting target-side molecular structure during training.

Another line of work incorporates spatial relationships by constructing graphs over ST spots. Graph-based methods can explicitly propagate information across neighboring locations and are therefore well suited for modeling local spatial organization. For example, MERGE [6] reduces the cost of message passing by clustering spot features into representative super-nodes. Such strategies improve scalability compared with dense spot-level graphs, but they also introduce a trade-off between computational efficiency and interaction fidelity: clustering may discard fine-grained spatial variation, nearby spots may be assigned to different representatives, and long-range dependencies may remain difficult to capture. As ST platforms move toward higher spatial resolution and larger slide-level spot counts, maintaining faithful interactions across all relevant tissue regions becomes increasingly challenging for graph-based designs.

More recent generative methods attempt to model richer spatial and molecular structure by jointly considering histology and ST representations. STFlow [9] uses pathology foundation model features as histology inputs and applies a flow-matching formulation to predict slide-level ST profiles, suggesting that generative modeling can improve over independent spot-wise regression. However, its local geometric constraints emphasize spatial smoothness within restricted neighborhoods, which may limit the modeling of broader tissue-level or long-range molecular dependencies. STEM [23] introduces a latent diffusion framework over gene and WSI embeddings, allowing broader context to be incorporated through iterative denoising. Yet diffusion-based inference can be computationally expensive because generation requires many denoising steps, making scalability an important concern for high-resolution ST slides.

Overall, existing methods have advanced histology-conditioned ST prediction from local regression toward spatial and generative modeling, but they still face a trade-off among long-range within-modality interaction, cross-modal interaction, computational efficiency, and scalability to modern high-resolution datasets. Our work systematically considers these factors within a unified and lightweight conditional flow framework. By coupling multiple histology–ST pairs as tokens, the model captures broader tissue-level and cross-modal interactions without relying on costly whole-slide graph propagation or iterative latent diffusion inference. More importantly, our analysis of when and why flow matching helps provides design principles for more expressive, stable and scalable joint generative modeling of histology and ST.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the central claim that flow matching helps deterministic multimodal prediction only when noisy intermediate target states are selectively extracted and integrated with histology features. The scope is further clarified through the listed contributions and the limitation section.

   Guidelines:

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

2. **Limitations**

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

   Answer:[Yes]

   Justification: The paper includes a limitation discussion in Section 5.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: The paper defines the conditional flow-matching formulation but does not claim new formal theoretical results; Appendix A documents empirical spatial dependence and slide heterogeneity.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The paper describes the model architecture, flow parameterization, datasets, metrics, baselines, and evaluation protocol in the main text. Appendix B further provides dataset statistics, splits, gene-panel construction, and hyperparameter settings needed to reproduce the reported results.

Guidelines:

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

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We provide open access to the code repository, with the GitHub link included in the abstract. The repository contains the implementation, configuration files, preprocessing scripts, dataset split files, and instructions needed to reproduce the main experimental results. For data, the benchmark datasets are based on publicly available spatial transcriptomics resources, and we provide the exact sample IDs, train/validation/test splits, and preprocessing details in the supplementary material. In addition, we release the trained model weights for the core IDC dataset to facilitate reproducibility and comparison.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 3 describes the comparative baselines and ablations. Appendix B reports dataset statistics, slide-level splits, preprocessing details, and hyperparameters, including the random seed, hidden dimension, number of layers, attention heads, dropout, learning rate, sampling configuration, and flow schedule.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: Appendix B reports paired statistical significance analyses across matched dataset-split units for multiple gene-panel scales. We include one-sided paired t-tests, Wilcoxon signed-rank tests, Benjamini-Hochberg corrected q-values, and paired effect sizes to assess whether the observed PCC improvements are statistically reliable.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix B reports the compute environment used for the experiments, including the GPU type, memory, and training cost.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The study uses existing research datasets and does not involve intervention, recruitment, or decision-making on human participants. We preserve anonymity in the submission and discuss potential risks and limitations of predicted molecular profiles in the broader-impact section.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Section 5 discusses positive impacts for understanding generative formulations in deterministic biological prediction and for computational pathology research. It also states that predicted molecular profiles may contain systematic errors and should not replace measured ST data or be used for clinical decision-making without experimental validation.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release a high-risk generative model, pretrained language model, image generator, or scraped dataset. The method predicts spatial molecular profiles from histology data for research analysis and is not intended for deployment in clinical decision-making.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The paper cites the existing datasets, baselines, and pathology foundation models used in the experiments. Appendix B summarizes the data sources and usage conditions, and all existing assets are used only according to their intended research purposes and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not introduce or release a new dataset or benchmark asset. The experimental data are derived from existing ST datasets, and the paper reports preprocessing, split, and evaluation details for reproducibility.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, participant recruitment, or direct research interactions with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The study uses existing de-identified research datasets and does not involve new human-subject recruitment, intervention, or interaction. Therefore, IRB approval for new human-subject experiments is not applicable.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The core method, experiments, and analysis do not use LLMs as an important, original, or non-standard component. Any use of language tools, if any, was limited to writing or formatting assistance and did not affect the scientific methodology, results, or originality.

    Guidelines:

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