GENEFLOWV2: RNA-CONTROLLABLE HISTOLOGY GENERATION AND COUNTERFACTUAL EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial transcriptomics (ST) aligns gene expression with histopathological morphology, enabling the investigation of how molecular states translate into observable visual phenotypes. However, RNA-conditioned generative models may produce visually realistic tissue images while underutilizing transcriptomic inputs, limiting their conditional fidelity and controllability. We introduce GeneFlowV2, a rectified-flow framework for generating cellular-level histopathological images from single-cell gene expression profiles. GeneFlowV2 integrates gene-aware RNA encoding with global and spatial conditioning, and further introduces a matched-versus-shuffled RNA ranking objective to enhance the model's sensitivity to transcriptomic variation. The model generates high-quality H&E histopathology images and supports RNA-conditioned visual counterfactual generation by editing an observed single-cell RNA profile toward a target RNA profile. Under the same spatially guarded evaluation protocol, across three 100K-scale image–gene paired datasets, GeneFlowV2 consistently outperforms GeneFlow in all experiments, demonstrating improved image generation quality, and stronger responsiveness to RNA conditioning.

1 INTRODUCTION

Spatial transcriptomics (ST) links spatially resolved gene expression with histopathological images, providing a foundation for studying how molecular variation is reflected in cellular and tissue morphology (Chelebian et al., 2025; Chen et al., 2025; Jaume et al., 2024; Zhang et al., 2024; Fu et al., 2025). This paired setting motivates RNA-to-image generation, which aims to synthesize local histopathological morphology conditioned on a cell's transcriptomic state. GeneFlow (Wang et al., 2026) pioneered this task by introducing a rectified-flow framework for high-quality RNA-conditioned histology generation. Yet visual realism alone does not establish RNA-conditional fidelity: a model can produce plausible histology while responding only weakly to its RNA input. This distinction is critical: without RNA-conditional fidelity, generated morphology cannot be considered transcriptomically grounded or reliably used for controlled generation and counterfactual analysis.

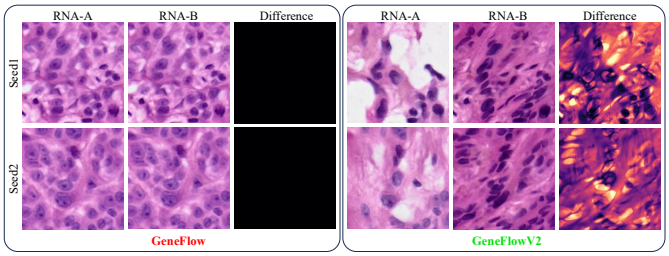

Figure 1: Under fixed-seed RNA swapping, GeneFlow Wang et al. (2026) remains nearly unchanged across RNA profiles but varies markedly across seeds, whereas GeneFlowV2 responds clearly to RNA changes, indicating stronger RNA-conditional sensitivity.

We expose exactly this failure mode in GeneFlow through a controlled RNA-swap experiment. With the initial Gaussian noise tensor fixed, replacing one observed RNA profile with another leaves the generated image nearly unchanged. In contrast, changing the noise seed while keeping the RNA profile fixed produces substantially greater morphological variation (Figure 1). This asymmetry reveals a noise-seed-dominated conditional bypass: the generated morphology responds more strongly to random initialization than to the sample-specific transcriptomic condition.

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

We attribute this behavior to GeneFlow's global-only RNA representation and standard flow-matching objective, which provide insufficient incentive for using sample-specific transcriptomic information. GeneFlowV2 addresses these limitations by retaining contextualized gene-level tokens and routing genes, selected based on their expression levels, to spatial cross-attention modules, enabling local image features to directly access gene-resolved information. We further introduce a matched-versus-shuffled RNA-ranking objective evaluated at the same noisy image state and timestep, requiring the matched RNA condition to yield a lower flow-velocity prediction error than a shuffled condition. Across multiple paired ST datasets evaluated with spatially guarded splits, GeneFlowV2 improves image fidelity and image-to-RNA consistency compared to GeneFlow while producing stronger morphological responses to RNA changes under fixed seeds. Together, these results indicate that expression-routed spatial conditioning and RNA-ranking supervision strengthen the model's dependence on transcriptomic conditions while preserving image-generation quality, thereby supporting RNA-guided generation and *in silico* counterfactual editing.

Contributions. (1) We show GeneFlow's generated morphology is dominated by noise seed rather than the RNA condition, demonstrating that visual realism does not necessarily imply faithful use of the RNA input. (2) We introduce GeneFlowV2, representing RNA as gene-level tokens and injects transcriptomic information via two complementary pathways: block-wise global modulation and expression-routed gated spatial cross-attention. (3) We propose RNA-sensitive flow matching that combines null conditioning with a low-time matched-versus-shuffled RNA-ranking objective, comparing RNA profiles at fixed noisy image state and timestep. (4) We propose an inversion-based framework for RNA-guided counterfactual editing of H&E images, enabling paired morphological comparisons between distinct single-cell transcriptomic states from a shared source inversion.

2 Related Work

Transcriptomics and histopathology. Most paired histology–transcriptomics methods either infer expression from morphology or align the two modalities. BLEEP and OmiCLIP learn shared image–expression representations, while STAMP adds spatial and multi-scale supervision (Xie et al., 2023; Chen et al., 2025; Han et al., 2026). iStar and GHIST predict expression at super-resolved and single-cell resolution; Stem formulates H&E-to-expression inference as conditional generation; and HistoPrism and Path2Space focus on pathway coherence and clinical-scale biomarker discovery (Zhang et al., 2024; Fu et al., 2025; Zhu et al., 2025; Hu et al., 2026; Shulman et al., 2026). The reverse direction remains less explored. GeneVAR generates H&E WSI tiles from case-level bulk RNA-seq under a shared slide-level condition (Zhao et al., 2026). At cellular resolution, GE2Hist reconstructs a local H&E niche from a central cell's Visium HD profile (Cai et al., 2025), whereas GeneFlow maps Xenium profiles to paired local histology patches via rectified flow (Wang et al., 2026; Liu et al., 2023). However, prior work has not systematically tested whether sample-specific RNA controls generated morphology or whether targeted perturbations induce predictable changes. GeneFlowV2 strengthens and evaluates this control through profile swaps and targeted interventions, enabling controlled RNA-guided *in silico* editing.

GeneFlow and RNA-conditional fidelity. GeneFlow is the closest prior work, using conditional rectified flow to generate local H&E images and an auxiliary DAPI channel from RNA profiles (Wang et al., 2026). It compresses each RNA profile into a global embedding that is broadcast through the U-Net, limiting gene-specific control over local image features. Moreover, the standard flow-matching objective can be learned largely from the noisy image, allowing the model to rely only weakly on the RNA condition. Our fixed-noise RNA-swap analysis reveals this conditional bypass: GeneFlow responds more strongly to the noise seed than to the input RNA profile (Figure 1). GeneFlowV2 addresses the global-conditioning bottleneck with gene-resolved spatial cross-attention and counters conditional bypass with a matched-versus-shuffled RNA-ranking objective.

Counterfactual editing in medical imaging. Diffusion models generate counterfactual medical images by modifying pathology-related attributes (Yeganeh et al., 2025; Min et al., 2025; Fontanella et al., 2024). In computational pathology, MoPaDi produces classifier-guided morphological transitions, whereas His-MMDM enables genomic- and transcriptomic-guided editing using sample-level profiles (Žigutytė et al., 2026; Li et al., 2026). Unlike these class- or sample-level approaches, GeneFlowV2 performs single-cell transcriptomic counterfactual editing by holding the real source H&E inversion fixed and replacing only the cell's RNA profile to predict its morphology under an alternative transcriptomic state.

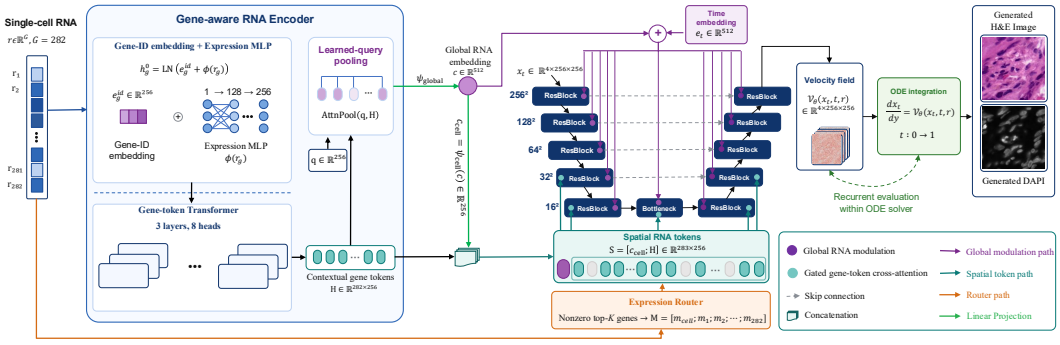

Figure 2: **Overview of GeneFlowV2.** Taking C1 as an example, GeneFlowV2 encodes a 282-gene profile into global and gene-token representations that condition a rectified-flow U-Net via global modulation and routed spatial cross-attention, generating an H&E image with a DAPI channel.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERVIEW

Let $\mathbf{r} \in \mathbb{R}^G$ denote the expression profile of a cell and let $\mathbf{x}_1 \in \mathbb{R}^{C\times H\times W}$ denote its paired target image. Given an intermediate image state $\mathbf{x}_t$ at time $t \in [0,1]$, GeneFlowV2 predicts an RNA-conditioned image-space velocity $\widehat{\mathbf{u}}_t = v_\theta(\mathbf{x}_t, t \mid \mathbf{r})$. At inference, we sample $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and integrate the resulting ordinary differential equation from $t = 0$ to $t = 1$ to obtain a generated image. GeneFlowV2 retains the conditional rectified-flow formulation of GeneFlow (Wang et al., 2026), but redesigns how RNA information is represented, injected, and supervised. As illustrated in Figure 2, GeneFlowV2 comprises four components: (1) a gene-token encoder that preserves gene identity and expression value; (2) a global RNA pathway that conditions residual blocks throughout the U-Net; (3) an expression-routed spatial pathway that exposes selected gene tokens directly to image features; and (4) a matched-versus-shuffled RNA-ranking objective that encourages the predicted velocity to depend on the corresponding expression profile.

3.2 GENE-TOKEN RNA ENCODER

Instead of immediately compressing the complete expression profile into a single vector, GeneFlowV2 constructs one token per gene. The initial representation of the gene g is $\mathbf{h}_g^{(0)} = \mathrm{LN}\left(\mathbf{e}_g^{\mathrm{id}} + \phi(r_g)\right)$, where LN denotes layer normalization, $\mathbf{e}_g^{\mathrm{id}} \in \mathbb{R}^D$ is a learned gene-identity embedding, $\phi : \mathbb{R} \to \mathbb{R}^D$ is a shared expression MLP that maps the scalar expression value r_g into the token space, and the superscript (0) denotes the initial gene-token representation before Transformer contextualization. We stack the initial gene tokens into $\mathbf{H}^{(0)} = [\mathbf{h}_1^{(0)}, \ldots, \mathbf{h}_G^{(0)}]$ and contextualize them using a gene-level Transformer encoder: $\mathbf{H} = \mathcal{T}_{\mathrm{gene}}\left(\mathbf{H}^{(0)}\right) = [\mathbf{h}_1, \ldots, \mathbf{h}_G] \in \mathbb{R}^{G\times D}$. Here, $\mathcal{T}_{\mathrm{gene}}$ denotes the Transformer encoder and $\mathbf{h}_g$ is the contextualized representation of gene g. To obtain a global representation of the complete expression profile, we use a trainable query $\mathbf{q} \in \mathbb{R}^D$ to aggregate the contextualized gene tokens: $\mathbf{p} = \mathrm{AttnPool}(\mathbf{q}, \mathbf{H}), \mathbf{c} = \psi_{\mathrm{global}}(\mathbf{p}) \in \mathbb{R}^{D_c}$. The attention-pooling layer uses $\mathbf{q}$ as its sole query and $\mathbf{H}$ as its key–value memory, while ψ_{global} projects the pooled token $\mathbf{p} \in \mathbb{R}^D$ into the global RNA representation $\mathbf{c}$. We omit positional encodings because gene identity is already encoded by $\mathbf{e}_g^{\mathrm{id}}$ and the conditioning genes has no meaningful sequential order. The encoder therefore produces two complementary representations: contextualized gene tokens $\mathbf{H}$ and a pooled global representation $\mathbf{c}$.

3.3 DUAL-PATH RNA CONDITIONING

Compressing RNA into a single vector provides a stable cell-level condition but does not preserve an explicit interface between individual genes and spatial image features. GeneFlowV2 therefore conditions the image backbone through complementary global and spatial pathways.

Block-wise global modulation. Let $\mathbf{e}_t$ denote the projected sinusoidal time embedding. For the

residual block b, we construct $\mathbf{e}_b = \mathbf{e}_t + \rho\left[P(\mathbf{c}) + \Delta_b(\mathbf{c})\right]$, where P is shared across the U-Net and Δ_b is a block-specific residual adapter. Each residual block maps $\mathbf{e}_b$ to channel-wise scale and shift parameters that modulate its normalized features. The shared projection provides a consistent cell-level condition, specialized by block-specific adapters across the encoder, bottleneck, and decoder.

Expression-routed spatial tokens. To expose both profile-level and gene-level RNA information to image features, we project the global representation $\mathbf{c}$ into the gene-token space and prepend it to the contextualized gene tokens: $\mathbf{c}_{\text{cell}} = \psi_{\text{cell}}(\mathbf{c})$, $\mathbf{S} = [\mathbf{c}_{\text{cell}}; \mathbf{h}_1; \ldots; \mathbf{h}_G] \in \mathbb{R}^{(G+1)\times D}$, where ψ_{cell} is a learned projection and $\mathbf{c}_{\text{cell}}$ is the cell-summary token. For each cell, we select up to $K = 128$ genes with the largest absolute expression values among the nonzero genes, and denote their indices by $\mathcal{I}_K(\mathbf{r})$. The spatial routing mask is defined as $m_{\text{cell}} = 1, m_g = \mathbf{1}\left[g \in \mathcal{I}_K(\mathbf{r})\right], \mathbf{M} = [m_{\text{cell}}; m_1; \ldots; m_G] \in \{0,1\}^{G+1}$, where $\mathbf{1}[\cdot]$ is the indicator function. The cell-summary token is always active, whereas spatial cross-attention uses only the selected gene tokens.

Gated spatial cross-attention. We attach gated cross-attention modules after the U-Net blocks whose outputs have spatial resolutions of 32×32 or 16×16, including blocks in the encoder, bottleneck, and decoder. Let $\mathbf{F}_\ell$ denote the output of one such U-Net block before spatial RNA conditioning. Each spatial feature attends to the cell-summary token and the routed gene tokens in $\mathbf{S}$, while the routing mask $\mathbf{M}$ prevents non-selected gene tokens from participating: $\mathbf{F}'_\ell = \mathbf{F}_\ell + g_\ell \,\mathrm{CrossAttn}_\ell\left(\mathbf{F}_\ell, \mathbf{S}; \mathbf{M}\right)$, where $\mathrm{CrossAttn}_\ell$ uses spatial image features as queries and RNA tokens as keys and values. The scalar $g_\ell \in (0,1)$ is learned independently for each cross-attention module and controls the contribution of its RNA-conditioned residual.

3.4 RNA-Sensitive Flow Matching

Base rectified-flow objective. We retain GeneFlow's sine-interpolated noise–data path with stochastic path perturbation ($\sigma_{\text{path}} = 0.05$), where $t = 0$ denotes noise and $t = 1$ denotes data. Given an RNA profile $\mathbf{r}$, the globally and spatially conditioned network predicts the target velocity $\mathbf{u}_t := \mathrm{d}\mathbf{x}_t/\mathrm{d}t$. For image-space velocity fields $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{C\times H\times W}$, we define the per-example mean squared error as $\ell_{\text{mse}}(\mathbf{a}, \mathbf{b}) = \frac{1}{CHW}\left\|\mathbf{a} - \mathbf{b}\right\|_F^2$, where the mean is taken over channels and spatial locations, but not over minibatch examples. The flow-matching objective is $\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\ell_{\text{mse}}\left(v_\theta(\mathbf{x}_t, t \mid \mathbf{r}), \mathbf{u}_t\right)\right]$. Full definitions of the interpolation path and its target velocity are provided in Appendix A.4. To enable classifier-free guidance, for each training example we jointly replace the global RNA embedding and spatial token bank with their learned null representations with probability $p_{\text{drop}} = 0.15$. At inference, the guided velocity field is

$$v_{\text{CFG}}(\mathbf{x}_t, t \mid \mathbf{r}) = v_\theta(\mathbf{x}_t, t \mid \varnothing) + w_{\text{CFG}}\left[v_\theta(\mathbf{x}_t, t \mid \mathbf{r}) - v_\theta(\mathbf{x}_t, t \mid \varnothing)\right], \tag{1}$$

where $\varnothing$ denotes the joint null condition and w_{CFG} controls the guidance strength.

Matched-versus-shuffled RNA ranking. The base objective evaluates only observed RNA–image pairs and does not compare matched with mismatched RNA. The model may therefore ignore RNA when $\mathbf{x}_t$ already reveals target morphology. We address this shortcut by comparing prediction errors under matched and shuffled RNA for the same flow state. For a minibatch of size $B > 1$, let π be a random derangement of the batch indices, such that $\pi(i) \neq i$ for every i. We sample $t_i \sim \mathcal{U}(0, t_{\max})$ and compute the scalar per-example errors

$$e_i^+ = \ell_{\text{mse}}\left(v_\theta\left(\mathbf{x}_{t_i,i}, t_i \mid \mathbf{r}_i\right), \mathbf{u}_{t_i,i}\right), e_i^- \quad = \ell_{\text{mse}}\left(v_\theta\left(\mathbf{x}_{t_i,i}, t_i \mid \mathbf{r}_{\pi(i)}\right), \mathbf{u}_{t_i,i}\right). \tag{2}$$

The image state $\mathbf{x}_{t_i,i}$, timestep t_i, and target velocity $\mathbf{u}_{t_i,i}$ are held fixed between the two evaluations; only the RNA condition is changed. We encourage the matched condition to satisfy the margin constraint $e_i^+ + \gamma \le e_i^-$ using

$$\mathcal{L}_{\text{rank}} = \frac{1}{B}\sum_{i=1}^{B} \tau\ \mathrm{softplus}\left(\frac{e_i^+ - e_i^- + \gamma}{\tau}\right). \tag{3}$$

We set $t_{\max} = 0.25$ so that ranking is applied in the noise-dominated region, where the flow state contains less information about the target morphology. The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}}. \tag{4}$$

Gradients from $\mathcal{L}_{\text{FM}}$ update the full network, whereas those from $\mathcal{L}_{\text{rank}}$ are restricted to the RNA encoder, global adapters, and spatial cross-attention modules, preventing the margin from being satisfied through the image-only backbone. We fix $\lambda_{\text{rank}} = 0.5$ and $\gamma = \tau = 10^{-3}$ across all datasets and experiments, with γ matching the empirical matched–shuffled per-example MSE gap and τ setting the smoothing scale.

3.5 RNA-GUIDED COUNTERFACTUAL EDITING

Let $\mathrm{Solve}_{a\to b}(\mathbf{x};\mathbf{r})$ denote the state at time b obtained by integrating the guided ODE $\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = v_{\mathrm{CFG}}(\mathbf{x}_t, t \mid \mathbf{r})$ from the initial condition $\mathbf{x}_a = \mathbf{x}$. All backward and forward integrations use the same fixed guidance weight w_{CFG}. Let $\Pi_{[0,1]}$ denote elementwise clipping to the valid image range. Given a real source patch $\mathbf{x}^s$ and its observed RNA profile $\mathbf{r}^s$, we first invert the source from data time to noise time under its matched RNA condition: $\widetilde{\mathbf{x}}_0^s = \mathrm{Solve}_{1\to 0}\left(\mathbf{x}^s;\mathbf{r}^s\right)$. Starting from this shared inverted state, we decode once under the source RNA to quantify inversion error and once under a target RNA profile $\mathbf{r}^t$ to obtain the counterfactual edit:

$$\widetilde{\mathbf{x}}^s = \Pi_{[0,1]}\left[\mathrm{Solve}_{0\to 1}\left(\widetilde{\mathbf{x}}_0^s;\mathbf{r}^s\right)\right], \mathbf{x}^{\mathrm{cf}} \quad = \Pi_{[0,1]}\left[\mathrm{Solve}_{0\to 1}\left(\widetilde{\mathbf{x}}_0^s;\mathbf{r}^t\right)\right]. \tag{5}$$

All backward and forward integrations use 50 explicit-Euler steps. Within each method, reconstruction and editing share the inverted state and guidance strength, differing only in the RNA condition used for forward decoding. The target cell's full RNA profile is provided to the model, whereas its real H&E patch is withheld and used only for distributional evaluation. Here, *counterfactual* denotes the model's response to alternative observed RNA, not an identified biological effect.

4 RESULTS

4.1 DATASETS

We evaluate on three public 10x Genomics Xenium skin-melanoma datasets used by GeneFlow: C1, C2, and P1 (Janesick et al., 2023). After RNA preprocessing and image matching, the datasets contain 100,597, 63,089, and 91,099 image–gene pairs, respectively. Each pair consists of a single-cell expression profile and its registered 256×256 H&E crop with an auxiliary DAPI channel. C1, C2, and P1 use 282, 382, and 5,006 genes, respectively. The public GeneFlow protocol splits cells independently despite substantial overlap between neighboring cell-centered crops. Our C1 audit finds that 98.27% of validation crops overlap at least one training crop (median over 21 random splits; Appendix B.2). To prevent shared tissue content across partitions, we instead construct a spatially guarded 80:20 train–validation split for each dataset: validation cells are selected from a contiguous tissue region, and every candidate training crop that overlaps a validation crop is moved to an excluded guard set. To ensure a fair comparison, all methods are retrained and evaluated on the same guarded partition of each dataset; dataset and split statistics are reported in Appendix B.1. To address potential concerns that the baseline was undertrained, we additionally show representative generations from the official pretrained GeneFlow checkpoint in Appendix B.3.

4.2 QUANTITATIVE AND QUALITATIVE IMAGE ANALYSIS

Evaluation protocol. We evaluate generated H&E images for distributional quality, paired fidelity, and RNA-conditional consistency. We extract pathology-specific features with UNI2-h (Chen et al., 2024). *Distributional quality* is assessed using FID (Heusel et al., 2017) in Inception-v3 (Szegedy et al., 2016) and UNI2-h feature spaces (Inception FID and UNI2-h FID), and the similarity between real and generated gray-level co-occurrence matrix (GLCM) energy distributions (Haralick et al., 1973) (Spatial Energy). *Paired fidelity* is assessed against matched real images using cosine similarity between UNI2-h features, AlexNet-based LPIPS (LPIPS-Alex) (Zhang et al., 2018), and Euclidean distance in Inception-v3 feature space (FeatDist). *RNA-conditional consistency* is assessed with dataset-specific UNI2-h H&E-to-RNA predictors, each fine-tuned on its dataset and fixed during evaluation. For pools of up to 32 candidate profiles, we compare predicted profiles with matched RNA targets using top-5 retrieval accuracy (Top-5@32), matched-target mean reciprocal rank (MRR@32), and Pearson correlations across samples averaged over genes (Gene Pearson). See Appendix C for metric definitions and UNI2-h fine-tuning details.

GeneFlowV2 improves image quality and morphological fidelity. Across the three datasets, GeneFlowV2 more closely matches both the real-image distribution and the image paired with each conditioning profile (Table 1). The clearest cross-dataset signal appears in the pathology-specific representation: relative to GeneFlow, UNI2-h FID decreases by 10.5%, 7.5%, and 2.0% on C1, C2, and P1, respectively. Paired UNI2-h cosine similarity increases by 18.5%, 1.9%, and 4.4% in the same order. Notably, GeneFlowV2 outperforms GeneFlow on eight of the nine reported metrics on P1 despite its substantially larger gene-panel size. Spatial Energy, LPIPS, and FeatDist broadly

Table 1: Quantitative evaluation of GeneFlow and GeneFlowV2 across three complementary axes: distributional image quality, paired morphological fidelity, and image-to-RNA cycle consistency. Cycle metrics are computed using independently trained and frozen H&E-to-RNA predictors on the spatially guarded validation splits. Top-5@32 and MRR@32 summarize retrieval of the paired RNA profile from deterministic candidate pools of up to 32 profiles.

Dataset	Method	**Distributional image quality**			**Paired morphological fidelity**			**Image-to-RNA cycle evaluation**		
		Inception FID ↓	UNI2-h FID ↓	Spatial Energy ↑	UNI2-h Cosine ↑	LPIPS ↓	FeatDist ↓	Top-5@32 (%) ↑	MRR@32 ↑	Gene Pearson ↑
Xenium_{C1}	GeneFlow	31.11	65.71	0.813	0.444	0.409	15.98	15.40	0.125	-0.002
	GeneFlowV2	**27.33**	**58.84**	**0.951**	**0.526**	**0.409**	**14.75**	**53.14**	**0.318**	**0.222**
Xenium_{C2}	GeneFlow	**32.65**	43.50	0.816	0.633	0.453	15.40	37.65	0.246	0.178
	GeneFlowV2	32.90	**40.22**	**0.836**	**0.645**	**0.452**	**15.26**	**48.52**	**0.315**	**0.255**
Xenium_{P1}	GeneFlow	**60.22**	115.04	0.404	0.366	0.508	15.74	15.52	0.126	-0.002
	GeneFlowV2	61.83	**112.79**	**0.459**	**0.382**	**0.506**	**15.69**	**26.73**	**0.186**	**0.096**

support the same conclusion, whereas Inception FID improves on C1 but not on C2 or P1. This agreement across complementary representations matters: distributional metrics measure whether the generated population resembles real H&E, whereas paired metrics ask whether an output preserves the morphology associated with its specific RNA profile. GeneFlowV2 therefore improves population-level realism without sacrificing—and generally strengthening—sample-specific fidelity.

GeneFlowV2 strengthens image-to-RNA correspondence. We next tested whether the improved morphology retained information from the conditioning transcriptome. For each dataset, we trained an independent H&E-to-RNA evaluator on real pairs, froze it, and applied it to both generators. Compared with GeneFlow, GeneFlowV2 increases Top-5@32 by 37.74, 10.87, and 11.21 percentage points on C1, C2, and P1, respectively. MRR@32 reaches $2.56\times$, $1.28\times$, and $1.48\times$ the corresponding GeneFlow values. Gene Pearson also increases from -0.002 to 0.222 on C1 and improves on C2 and P1. Together, these results show that GeneFlowV2 improves alignment with the conditioning transcriptome at both profile and gene levels, beyond visual fidelity.

GeneFlowV2 more closely recapitulates reference morphology. In paired examples, GeneFlowV2 better captures reference-specific variation in nuclear shape, cellularity, and tissue organization (Fig. 3). It more consistently distinguishes elongated from rounded nuclei and sparsely from densely cellular regions. Compared with matched references, GeneFlow more often produces stereotyped images with overly dark, uniformly sized, or crowded nuclei. GeneFlowV2 also better reproduces local nuclear spacing and alignment in regions with organized elongated nuclei.

4.3 RNA Controllability and Counterfactual Editing

High image fidelity does not by itself establish that RNA actively controls generation. We therefore change only the conditioning RNA while holding the image-side starting state fixed. A successful model should redirect morphology toward the requested RNA category without leaving the corresponding real-image distribution.

Matched-noise RNA control. We first generate images from different RNA profiles using eight prespecified shared noise seeds. *RNA Top*-1 measures whether a frozen H&E-to-RNA evaluator assigns each generated patch to the requested category by nearest-centroid classification in predicted-RNA space. *RNA Alignment* measures whether the predicted expression change follows the requested RNA direction, while UNI2-h cFID evaluates target-category realism. Full definitions and uncertainty estimates are provided in Appendix D.

GeneFlowV2 turns RNA into an active control signal. GeneFlowV2 improves all three endpoints on every dataset. At $w_{\text{CFG}} = 2$, *RNA Top*-1 increases from 20.00% to 55.39% on C1, from 41.02% to 51.95% on C2, and from 14.29% to 24.30% on P1 (Table 2). *RNA Alignment* and cFID improve concurrently, showing that the stronger RNA response does not come at the expense of target-category realism. The same ordering holds at $w_{\text{CFG}} = 1$, indicating that classifier-free guidance amplifies an existing controllability advantage rather than creating it. The shared-noise examples make this distinction visually explicit (Figure 4a). GeneFlow produces nearly identical outputs for profiles A and B, yielding weak difference maps despite the substantial morphological differences between the real references. GeneFlowV2 instead changes nuclear shape, cellular density, and tissue organization with the conditioning profile. These structured differences agree with its higher *RNA*

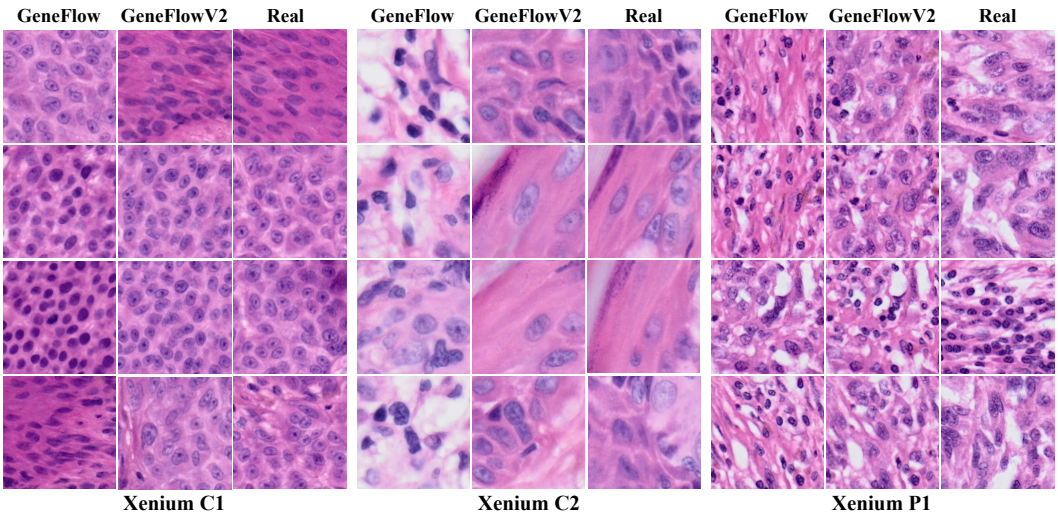

Figure 3: **GeneFlowV2 more closely preserves reference morphology across datasets.** Each triplet contains a held-out real H&E crop followed by GeneFlowV2 and GeneFlow outputs conditioned on the matched RNA profile; four image–RNA pairs are shown per dataset.

Top-1 and are inconsistent with RNA being ignored as an auxiliary input.

Source-anchored counterfactual editing. We next test whether RNA can redirect an observed image. Each method first reconstructs a real source image from its inversion and then decodes the same latent state using a complete target RNA profile. The real target image is used only as an evaluation reference and is never provided to the generator. *RNA Top*-1 retains exactly the same definition as in the matched-noise experiment. *UNI2-h Top*-1 independently tests whether the edited image is assigned to the requested category in pathology-feature space, and cFID evaluates the realism of the target-conditioned distribution. Protocol and aggregation details are provided in Appendix D.

GeneFlowV2 enables effective RNA-guided counterfactual editing. GeneFlowV2 substantially improves counterfactual target matching over GeneFlow on C1 and C2. At $w_{\mathrm{CFG}} = 2$, Predicted-RNA Target Top-1 increases from 7.75% to 35.25% on C1 and from 14.86% to 43.47% on C2, while UNI2-h Target Top-1 rises from 12.75% to 33.00% and from 17.78% to 37.08%, respectively. These gains are accompanied by lower cFID on both datasets and persist at $w_{\mathrm{CFG}} = 1$, indicating that stronger guidance alone does not explain them. P1 shows the same trend: GeneFlowV2 lowers cFID and improves both target-matching metrics. Source-anchored examples support these quantitative results (Fig. 4b). GeneFlow reconstructions and counterfactual outputs are nearly indistinguishable, suggesting that source preservation alone can reflect limited editing. By contrast, GeneFlowV2 clearly changes nuclear morphology and local tissue organization while retaining the source's overall histological character. Difference maps reveal spatially structured changes around cellular regions. Together, the quantitative and qualitative results show that GeneFlowV2 uses RNA to redirect morphology toward the requested state.

4.4 RNA Control Beyond Cell-Type Labels

Full-profile and counterfactual evaluations show that RNA redirects GeneFlowV2 outputs at the category level. We next test two questions separately on C1: whether this control depends on *which genes* are changed and whether it distinguishes RNA profiles within the same annotated cell type.

GeneFlowV2 responds selectively to targeted RNA perturbations. From the same source RNA profile, we construct graded trajectories toward a target gene-program endpoint at doses $d \in \{0, 0.25, 0.5, 0.75, 1\}$, where dose is the interpolation fraction from source to endpoint. For each trajectory, a non-program control perturbs expression-matched genes with the same standardized ℓ_2 magnitude as the target program at each dose. We hold generation noise and all other settings fixed. We define dose specificity as the difference in Spearman correlation between dose and predicted RNA response along the requested direction for target-program versus non-program trajectories. GeneFlow scores -0.194, whereas GeneFlowV2 scores 0.339 at $w_{\mathrm{CFG}} = 1$ and 0.433

Table 2: GeneFlowV2 strengthens matched-noise RNA control and counterfactual target matching. All metrics use complete RGB patches and are macro-averaged over requested target categories. cFID is computed in UNI2-h feature space; *RNA Top*-1 and *UNI2-h Top*-1 measure nearest-centroid target matching in predicted-RNA and image-feature spaces, respectively.

Dataset	Method	w_{CFG}	**Matched-noise RNA control**			**Counterfactual editing**		
			cFID ↓	*RNA Top*-1 (%) ↑	*RNA Alignment* ↑	cFID ↓	*RNA Top*-1 (%) ↑	*UNI2-h Top*-1 (%) ↑
Xenium$_{C1}$	GeneFlow	–	163.79	20.00	-0.001	158.58	7.75	12.75
	GeneFlowV2	1	**134.20**	46.48	0.255	**136.43**	26.50	24.00
	GeneFlowV2	2	140.93	**55.39**	**0.293**	137.55	**35.25**	**33.00**
Xenium$_{C2}$	GeneFlow	–	122.76	41.02	0.195	144.77	14.86	17.78
	GeneFlowV2	1	116.68	47.36	0.267	130.64	31.94	26.53
	GeneFlowV2	2	**113.86**	**51.95**	**0.295**	**121.23**	**43.47**	**37.08**
Xenium$_{P1}$	GeneFlow	–	188.88	14.29	0.001	121.46	13.99	25.10
	GeneFlowV2	1	183.87	22.04	0.081	115.49	14.98	25.79
	GeneFlowV2	2	**179.42**	**24.30**	**0.084**	**115.38**	**15.58**	**26.09**

Table 3: GeneFlowV2 improves targeted perturbation specificity and within-type RNA-profile control. Dose specificity ($\Delta\rho$) is the difference between the dose–response correlations for the target program and an equal-norm non-program control. Swap margin measures preference for the assigned RNA profile over the paired same-type profile. Differences are GeneFlowV2 minus GeneFlow; brackets show 95% confidence intervals.

Task	Primary score ↑	GeneFlowV2 w_{CFG}	GeneFlow	GeneFlowV2	GeneFlowV2 − GeneFlow (95% CI)
Targeted perturbation	Dose specificity ($\Delta\rho$)	1	-0.194	**0.339**	0.533 [0.300, 0.764]
		2	-0.194	**0.433**	0.628 [0.317, 0.994]
Within-type control	Swap margin	1	0.000	**0.119**	0.119 [0.077, 0.149]
		2	0.000	**0.167**	0.167 [0.116, 0.204]

at $w_{\mathrm{CFG}} = 2$, corresponding to improvements of 0.533 and 0.628, respectively (Table 3). Improvements were positive for 11/12 held-out sources and all seeds (Appendix E.1). Thus, GeneFlowV2 responds selectively to perturbed-gene identity rather than merely to RNA-change magnitude.

GeneFlowV2 distinguishes RNA profiles within cell type. We pair two observed RNA profiles from the same annotated cell type and generate an image from each profile using identical noise. A frozen image-to-RNA readout tests whether each generated image is closer to its assigned input profile than to the other paired profile. This correct-versus-swapped margin is approximately zero for GeneFlow. GeneFlowV2 increases it to 0.119 at $w_{\mathrm{CFG}} = 1$ and 0.167 at $w_{\mathrm{CFG}} = 2$ (Table 3), with a positive improvement in every cell type and every leave-one-type-out analysis. The margin disappears when the target profiles are permuted (Appendix E.2). These controls support profile-specific RNA sensitivity within annotated cell types, rather than reliance on RNA solely as a cell-type label.

4.5 Ablation Studies

We ablate GeneFlowV2's key conditioning components on Xenium$_{\mathrm{C1}}$ using the matched protocol (Table 4). The spatial path carries substantially more RNA-conditioning information than the global path alone: MRR@32 increases from 0.149 to 0.267 and Gene Pearson from 0.033 to 0.195. Adding global modulation raises these scores to 0.306 and 0.204, respectively, and adding the cell-summary token further raises them to 0.318 and 0.222. Routing all observed genes rather than the top-K expressed genes reduces the scores to 0.275 and 0.183. Disabling RNA ranking during training and checkpoint selection causes the largest regression, to 0.149 and 0.034, near GeneFlow's 0.125 and -0.002. This identifies RNA ranking as critical for preventing conditional bypass and preserving sensitivity to the conditioning profile. Although individual variants attain the best image-space scores, the full model achieves the strongest RNA consistency, showing that image agreement

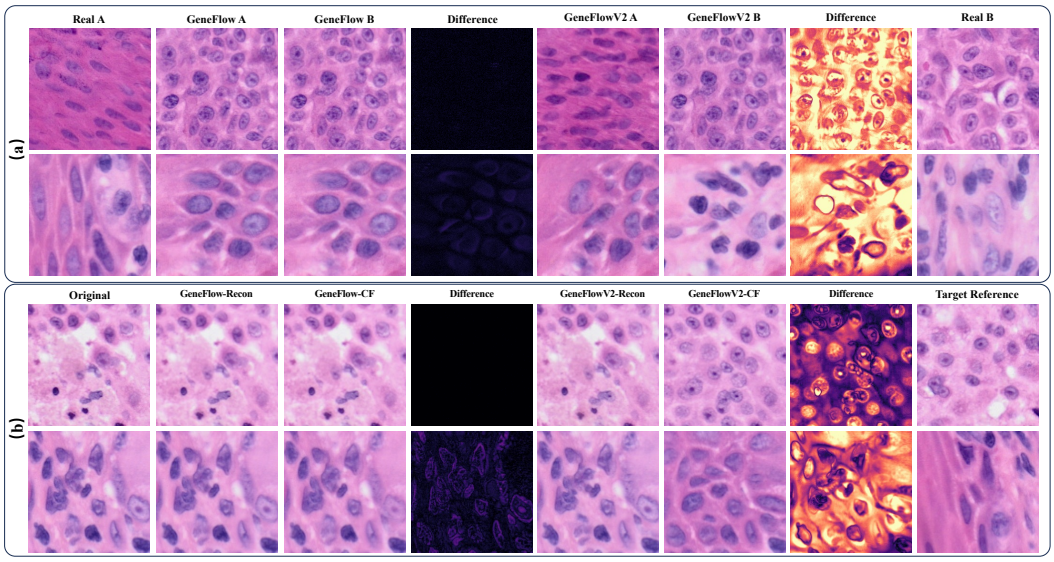

Figure 4: **Changing only RNA redirects generated and source-anchored morphology.** (a) Images generated from RNA profiles A and B with shared initial noise across profiles, methods, and guidance settings; real-image references. (b) Each method decodes source-conditioned reconstructions (Recon) and target-conditioned edits (CF) from one source inversion. Difference maps show A/B contrasts in (a) and Recon/CF contrasts in (b). GeneFlowV2 uses $w_{\mathrm{CFG}} = 2$ throughout. The target reference is neither a generator input nor pixel-level counterfactual ground truth.

alone does not establish adherence to the RNA condition. A sweep over CFG scales 1–4 identifies $w_{\mathrm{CFG}} = 2$ as a favorable balance between RNA consistency and image quality and thus the default operating point (Appendix F.1; Table 19). Complete metrics and variant definitions are in Appendix F. Finally, to test robustness to the H&E-to-RNA predictor, we replaced the fine-tuned UNI2-h evaluator with an independently fine-tuned H0-mini predictor (Filiot et al., 2025). GeneFlowV2 retained its advantage over GeneFlow (Appendix F.2; Table 20).

5 Conclusion

We introduced GeneFlowV2 to address noise-seed-dominated conditional bypass in RNA-conditioned histology generation through gene-level RNA encoding, dual-path conditioning, and RNA-sensitive flow matching. Across three spatially guarded Xenium melanoma datasets, GeneFlowV2 consistently improves histological image quality, image-to-RNA consistency, and RNA controllability. Matched-noise RNA swaps and source-anchored editing demonstrate profile-level control, while graded *in silico* gene-program perturbations and within-cell-type profile swaps reveal sensitivity to gene identity and transcriptomic variation beyond coarse cell-type labels. Together, these results establish conditional fidelity as an evaluation axis distinct from visual realism. GeneFlowV2 provides a unified framework for RNA-guided histology generation, targeted perturbation analysis, and source-anchored counterfactual editing.

Table 4: **Core ablations on C1.** All rows use the same spatially guarded split and evaluation pipeline. GeneFlowV2 variants use $w_{\mathrm{CFG}} = 2$; GeneFlow is included as a contextual reference.

Method	Image agreement		RNA consistency	
	UNI2-h FID ↓	UNI2-h cosine ↑	MRR@32 ↑	Gene Pearson ↑
GeneFlow	65.71	0.444	0.125	-0.002
Global path only	67.82	0.468	0.149	0.033
Spatial path only	57.68	**0.527**	0.267	0.195
All observed genes	**57.46**	0.514	0.275	0.183
w/o cell-summary token	58.28	0.523	0.306	0.204
w/o RNA ranking	64.30	0.459	0.149	0.034
GeneFlowV2	58.84	0.526	**0.318**	**0.222**

AI USE STATEMENT

In this work, we used generative AI tools for language editing, improving portions of the implementation code, and assisting with dataset preprocessing. We did not use generative AI tools to formulate the core research ideas, develop the main methodological contributions, generate experimental data, or draw scientific conclusions. Other required disclosure categories are not applicable to this work.

All AI-assisted text was reviewed and revised by the authors to ensure scientific accuracy and consistency with the underlying experiments. AI-assisted code was manually inspected and tested by the authors, and the dataset preprocessing procedures and their outputs were verified against the original data. The authors independently assessed all experimental results and conclusions. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide an anonymous source code repository at `https://anonymous.4open.science/r/GeneFlowV2-E6F4/`. The repository contains the implementation of our method, configuration files, and scripts for data preprocessing, training, and evaluation. Detailed descriptions of the datasets, preprocessing procedures, model configurations, training settings, hyperparameters, and evaluation protocols are provided in the appendix. We will publicly release the trained model weights upon publication. Together, these resources provide the information needed to reproduce the reported results.

REFERENCES

Hongmin Cai, Boan Ji, Shangyan Cai, Yi Liao, Jiazhou Chen, and Weitian Huang. Ge2hist: Generating histology images from single-cell gene expression via cross-modal generative network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 240–250. Springer, 2025.

Eduard Chelebian, Christophe Avenel, and Carolina Wählby. Combining spatial transcriptomics with tissue morphology. *Nature Communications*, 16(1):4452, 2025.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

Weiqing Chen, Pengzhi Zhang, Tu N Tran, Yiwei Xiao, Shengyu Li, Vrutant V Shah, Hao Cheng, Kristopher W Brannan, Keith Youker, Li Lai, et al. A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nature Methods*, 22(7):1568–1582, 2025.

Alexandre Filiot, Nicolas Dop, Oussama Tchita, Auriane Riou, Rémy Dubois, Thomas Peeters, Daria Valter, Marin Scalbert, Charlie Saillard, Geneviève Robin, et al. Distilling foundation models for robust and efficient models in digital pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 162–172. Springer, 2025.

Alessandro Fontanella, Grant Mair, Joanna Wardlaw, Emanuele Trucco, and Amos Storkey. Diffusion models for counterfactual generation and anomaly detection in brain images. *IEEE transactions on medical imaging*, 2024.

Xiaohang Fu, Yue Cao, Beilei Bian, Chuhan Wang, Dinny Graham, Nirmala Pathmanathan, Ellis Patrick, Jinman Kim, and Jean Yee Hwa Yang. Spatial gene expression at single-cell resolution from histology using deep learning with ghist. *Nature methods*, 22(9):1900–1910, 2025.

Minghao Han, Dingkang Yang, Linhao Qu, Zizhi Chen, Gang Li, Han Wang, Jiacong Wang, and Lihua Zhang. Fusing pixels and genes: Spatially-aware learning in computational pathology. In *International Conference on Learning Representations*, volume 2026, pp. 130377–130410, 2026.

Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Susu Hu, Qinghe Zeng, Nithya Bhasker, Jakob Nikolas Kather, and Stefanie Speidel. Histoprism: Unlocking functional pathway analysis from pan-cancer histology via gene expression prediction. *International Conference on Learning Representations*, 2026.

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

Guillaume Jaume, Paul Doucet, Andrew H Song, Ming Y Lu, Cristina Almagro-Pérez, Sophia J Wagner, Anurag J Vaidya, Richard J Chen, Drew F Williamson, Ahrong Kim, et al. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *Advances in Neural Information Processing Systems*, 37:53798–53833, 2024.

Zhongxiao Li, Tianqi Su, Bin Zhang, Wenkai Han, Sibin Zhang, Guiyin Sun, Yuwei Cong, Xin Chen, Jiping Qi, Yujie Wang, et al. His-mmdm: Multi-domain and multi-omics translation of histopathological images with diffusion models. *Advanced Science*, 13(19):e18066, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.

Hyungi Min, Taeseung You, Hangyeul Lee, Yeongjae Cho, and Sungzoon Cho. Instructx2x: an interpretable local editing model for counterfactual medical image generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 279–288. Springer, 2025.

Eldad D Shulman, Emma M Campagnolo, Roshan Lodha, Youngmin Chung, Amos Stemmer, Thomas Cantore, Beibei Ru, Tian-Gen Chang, Sumona Biswas, Saugato Rahman Dhruba, et al. Ai-predicted spatial transcriptomics unlocks breast cancer biomarkers from pathology. *Cell*, 2026.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Mengbo Wang, Shourya Verma, Aditya Malusare, Luopin Wang, Yiyang Lu, Vaneet Aggarwal, Mario Sola, Ananth Grama, and Nadia Lanman. Geneflow: Translation of single-cell gene expression to histopathological images via rectified flow. *Advances in Neural Information Processing Systems*, 38:45396–45429, 2026.

Ronald Xie, Kuan Pang, Sai Chung, Catia Perciani, Sonya MacParland, Bo Wang, and Gary Bader. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. *Advances in Neural Information Processing Systems*, 36:70626–70637, 2023.

Yousef Yeganeh, Azade Farshad, Ioannis Charisiadis, Marta Hasny, Martin Hartenberger, Björn Ommer, Nassir Navab, and Ehsan Adeli. Latent drifting in diffusion models for counterfactual medical image synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7685–7695, 2025. doi: 10.1109/CVPR52734.2025.00720.

Daiwei Zhang, Amelia Schroeder, Hanying Yan, Haochen Yang, Jian Hu, Michelle YY Lee, Kyung S Cho, Katalin Susztak, George X Xu, Michael D Feldman, et al. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature biotechnology*, 42(9):1372–1377, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Jianwei Zhao, Fan Yang, Xin Li, Qiang Zhai, Ao Luo, Ziqi Ren, Zhicheng Jiao, and Hong Cheng. Genevar: Causal meanflow for autoregressive gene-to-wsi tile synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34116–34125, June 2026.

Sichen Zhu, Yuchen Zhu, Molei Tao, and Peng Qiu. Diffusion generative modeling for spatially resolved gene expression inference from histology images. In *International Conference on Learning Representations*, volume 2025, pp. 19666–19690, 2025.

Laura Žigutytė, Tim Lenz, Tianyu Han, Nic G Reitsam, Sebastian Foersch, Katherine J Hewitt, Moritz Jesinghaus, Zunamys I Carrero, Michaela Unger, Asier Rabasco Meneghetti, et al. Counterfactual diffusion models provide interpretable explanations of artificial intelligence models in pathology. *Cancer Research*, pp. OF1–OF20, 2026.

A Additional Method and Implementation Details

This appendix provides the implementation details omitted from the main text. Unless otherwise stated, architectural dimensions and hyperparameters refer to the C1 experiments. The number of input genes follows the assay panel for experiments on other datasets.

A.1 Input Construction and RNA Preprocessing

RNA preprocessing. Following the preprocessing pipeline of GeneFlow, raw gene counts are normalized independently for each cell to a total count of 10^6 and then transformed using $\log(1+x)$. Let y_{ig} denote the raw count of gene g in cell i. The expression value supplied to the model is

$$r_{ig} = \log\left(1 + \frac{10^6 y_{ig}}{\sum_h y_{ih}}\right). \tag{6}$$

The training loader reads the resulting values directly from `adata.X` without applying additional per-gene standardization. The C1 model uses $G = 282$ genes in the fixed order stored with the checkpoint.

Image representation. Each training example contains a registered 256×256 image tensor with four channels. The first three channels are RGB H&E, and the fourth channel is the aligned DAPI image. Cell and nucleus segmentation masks stored with the preprocessed TIFF files are not used as model inputs or prediction targets. Images are converted to floating point in $[0, 1]$ and resized to 256×256 with antialiasing.

A.2 RNA-Encoder Details

Observation status. The implementation additionally maintains an observation indicator $o_g \in \{0, 1\}$ for each gene. This indicator distinguishes a measured zero, for which $o_g = 1$ and $r_g = 0$, from a gene absent from the assay panel, for which $o_g = 0$. The exact implemented gene token is

$$\mathbf{h}_g^{(0)} = \mathrm{LN}\left(\mathbf{e}_g^{\mathrm{id}} + \phi(o_g r_g) + \mathbf{e}_{o_g}^{\mathrm{obs}}\right), \tag{7}$$

where $\mathbf{e}_{o_g}^{\mathrm{obs}}$ is a learned observation-status embedding. Genes absent from the assay are excluded from Transformer self-attention and attention pooling using a key-padding mask, whereas measured-zero genes remain visible. All retained C1 genes are observed, so $\mathbf{o} = \mathbf{1}$ in the reported C1 experiments. The main text therefore omits $\mathbf{o}$ for clarity.

Encoder dimensions. The gene-token dimension is $D = 256$. The shared expression-value encoder has the form

$$\phi : \mathbb{R} \to \mathbb{R}^{128} \to \mathbb{R}^{256}, \tag{8}$$

with a SiLU nonlinearity between its two affine layers. Both the gene-identity embeddings and observation-status embeddings have dimension 256.

The gene Transformer contains three pre-normalized encoder layers. Each layer uses eight attention heads, a feed-forward dimension of 1024, GELU activations, and zero dropout. No positional encoding is used because each token already contains an explicit gene-identity embedding.

The learned-query pooling module uses one trainable 256-dimensional query and eight-head attention over the contextualized gene tokens. Its pooled output $\mathbf{p} \in \mathbb{R}^{256}$ is mapped to the global RNA representation $\mathbf{c} \in \mathbb{R}^{512}$ by

$$\mathrm{LN}_{256} \to \mathrm{Linear}_{256\to512} \to \mathrm{SiLU} \to \mathrm{Linear}_{512\to512} \to \mathrm{LN}_{512}\,. \tag{9}$$

The cell-summary token is obtained using

$$\mathrm{LN}_{512} \to \mathrm{Linear}_{512\to256} \to \mathrm{SiLU} \to \mathrm{LN}_{256}\,. \tag{10}$$

Consequently, the spatial token bank for C1 has shape $\mathbf{S} \in \mathbb{R}^{B\times283\times256}$.

A.3 Dual-Path Conditioning Details

Global modulation. Both the projected sinusoidal time embedding and the global RNA representation have dimension 512. The shared projection $P : \mathbb{R}^{512} \to \mathbb{R}^{512}$ is implemented as a two-layer MLP with a SiLU nonlinearity. Each stage-specific residual adapter has the form

$$\Delta_b(\mathbf{c}) = W_{b,2}\,\mathrm{SiLU}\left(W_{b,1}\,\mathrm{SiLU}(\mathbf{c})\right), \tag{11}$$

where both affine layers map $\mathbb{R}^{512} \to \mathbb{R}^{512}$. The final layer $W_{b,2}$ is initialized to zero, so training begins from the shared global-conditioning pathway and gradually learns stage-specific corrections. We use $\rho = 1$ in Eq. equation 3.3.

For readability, the main text indexes the adapters by U-Net block b. In the implementation, an adapter is associated with each composite U-Net block container. Residual blocks belonging to the same container therefore share its adapter. Each residual block subsequently maps the resulting conditioning embedding to channel-wise scale and shift parameters for feature normalization.

Hard expression routing. The full GeneFlowV2 model uses hard expression routing with $K = 128$ and $\delta = 0$. A gene is eligible for routing when it is observed and has nonzero expression:

$$o_g = 1, \qquad |r_g| > \delta. \tag{12}$$

Among the eligible genes, at most the K largest values of $|r_g|$ are marked active. If fewer than K genes are eligible, all eligible genes are retained.

Routing changes only the visibility mask supplied to spatial cross-attention. The token matrix is neither gathered nor reordered, and each routed gene retains its fixed panel index. Non-routed observed genes still participate in Transformer contextualization and learned global pooling. The cell-summary token is always visible.

Cross-attention implementation. Let $\mathbf{F}_\ell \in \mathbb{R}^{B\times C_\ell\times H_\ell\times W_\ell}$ denote an image feature map and let $N_\ell = H_\ell W_\ell$. We first define

$$\mathbf{X}_\ell = \mathrm{Flatten}\left(\mathrm{GN}_{32}(\mathbf{F}_\ell)\right) \in \mathbb{R}^{B\times N_\ell\times C_\ell}. \tag{13}$$

The attention projections and routed update are

$$\mathbf{Q}_\ell = \mathbf{X}_\ell W_\ell^Q, \qquad \mathbf{K}_\ell = \mathrm{LN}(\mathbf{S})W_\ell^K, \qquad \mathbf{V}_\ell = \mathrm{LN}(\mathbf{S})W_\ell^V,$$
$$\mathbf{\Pi}_\ell = \mathrm{softmax}\left(\frac{\mathbf{Q}_\ell\mathbf{K}_\ell^\top}{\sqrt{d_h}} + \mathcal{B}(\mathbf{m})\right), \quad \mathbf{F}'_\ell = \mathbf{F}_\ell + \sigma(\alpha_\ell)\,\mathrm{Reshape}\left(\mathbf{\Pi}_\ell\mathbf{V}_\ell W_\ell^O\right). \tag{14}$$

Here, $\mathcal{B}(\mathbf{m})$ is zero for active tokens and $-\infty$ for inactive tokens. Each module uses eight heads with $d_h = 32$, giving an inner attention dimension of 256. The learned scalar α_ℓ is initialized as $\log(0.1/0.9)$, so that $\sigma(\alpha_\ell) = 0.1$ at initialization.

Cross-attention is attached after U-Net blocks whose outputs have spatial resolutions of 32×32 or 16×16. The C1 model contains six such modules in the encoder, one in the bottleneck, and five in the decoder. Cross-attention dropout is zero.

Image backbone. The U-Net maps a tensor in $\mathbb{R}^{B\times 4\times 256\times 256}$ to a velocity tensor of the same shape. It uses 128 base channels, channel multipliers $(1, 2, 2, 2, 2)$, and two residual blocks per resolution. The resulting encoder resolutions and channel widths are

$$(256, 128),\ (128, 256),\ (64, 256),\ (32, 256),\ (16, 256), \tag{15}$$

where each pair denotes spatial resolution and channel width. The decoder uses symmetric skip connections. Native image self-attention is applied at 16×16 and in the bottleneck, while RNA cross-attention is applied at both 32×32 and 16×16. Scale–shift normalization and residual up/downsampling are used throughout the backbone.

For C1, the complete model contains approximately 78.38 million parameters, of which approximately 23.25 million belong to the RNA encoder and conditioning pathways.

A.4 Flow Path and Classifier-Free Guidance

Sine-interpolated training path. Following the inherited GeneFlow formulation, let $t \sim \mathcal{U}(0, 1)$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We define the underlying noise–data path as

$$\overline{\mathbf{x}}_t = s(t)\mathbf{x}_1 + [1 - s(t)]\mathbf{z}, \qquad s(t) = \sin\left(\frac{\pi t}{2}\right). \tag{16}$$

Its target velocity is

$$\mathbf{u}_t = \frac{\mathrm{d}\overline{\mathbf{x}}_t}{\mathrm{d}t} = \frac{\pi}{2}\cos\left(\frac{\pi t}{2}\right)(\mathbf{x}_1 - \mathbf{z}). \tag{17}$$

As in GeneFlow, the state supplied to the network is augmented with a small time-dependent perturbation:

$$\widetilde{\mathbf{x}}_t = \overline{\mathbf{x}}_t + (1 - t)\sigma_{\mathrm{path}}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{18}$$

where $\boldsymbol{\epsilon}$ is independent of $\mathbf{z}$ and $\sigma_{\mathrm{path}} = 0.05$. Thus, the perturbation acts as an augmentation of the network input, while the supervised velocity is defined by the underlying sine-interpolated path. We write $\mathbf{x}_t$ for $\widetilde{\mathbf{x}}_t$ in the main text to avoid additional notation.

At generation time, the model starts from $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and follows the deterministic learned ODE without injecting the training-path perturbation.

Learned null condition. Condition dropout is sampled independently for each training example with probability $p_{\mathrm{drop}} = 0.15$. When an RNA condition is dropped, the pooled representation is replaced by a learned 512-dimensional null global embedding. The spatial token bank is replaced by a learned 256-dimensional null token, and only the first null-token position is exposed to spatial cross-attention. Consequently, the null condition is distinct from a measured all-zero expression profile.

Classifier-free guidance follows Eq. equation 1. When $w_{\mathrm{CFG}} = 1$, the implementation directly returns the ordinary RNA-conditioned velocity without performing an additional null-conditioned forward pass.

A.5 Ranking Optimization and Training Configuration

Ranking branch. The shuffled index mapping is applied jointly to the expression vectors and their observation masks. Both the matched and shuffled branches receive the same image state, time, target velocity, and stochastic path realization. RNA dropout is disabled in this branch, so both conditions are explicitly present. Ranking times are sampled from $t \sim \mathcal{U}(0, 0.25)$.

The flow-matching loss updates all model parameters. Ranking gradients are computed separately and applied only to the conditioning parameter set

$$\theta_{\mathrm{cond}} = \{\text{RNA encoder}, P, \{\Delta_b\}_b, \text{spatial cross-attention}\}. \tag{19}$$

The image-only backbone and time-embedding pathway receive no gradients from $\mathcal{L}_{\mathrm{rank}}$. The combined gradients are clipped to a global norm of 1.0 before the optimizer update.

Table 5: Training configuration for the C1 GeneFlowV2 model.

Configuration	Value
Batch size	8
Maximum epochs	50
Optimizer	AdamW
Initial learning rate	3×10^{-5}
Minimum learning rate	1.5×10^{-6}
Weight decay	0.01
Gradient clipping	1.0
EMA decay	0.999
Condition-drop probability	0.15
Path perturbation scale	0.05
Main-branch shared-noise probability	0.75
Ranking time range	$[0, 0.25]$
Ranking weight λ_{rank}	0.5
Ranking margin γ	0.001
Ranking temperature τ	0.001
Checkpoint-selection ranking weight	5.0
Early-stopping patience	15 epochs
Training seed	42
Numerical precision	bfloat16 AMP

Optimization and model selection. We train with AdamW using an initial learning rate of 3×10^{-5} and weight decay 0.01. A cosine schedule decreases the learning rate to 5% of its initial value. Training uses a batch size of eight, bfloat16 automatic mixed precision, activation checkpointing, and an exponential moving average with decay 0.999.

Validation is performed using the EMA model. The best checkpoint is selected according to

$$\mathcal{J}_{\mathrm{sel}} = \mathcal{L}_{\mathrm{FM}}^{\mathrm{val}} + 5\mathcal{L}_{\mathrm{rank}}^{\mathrm{val}}. \tag{20}$$

The selection weight of 5 is used only for checkpoint selection and is distinct from the training weight $\lambda_{\mathrm{rank}} = 0.5$. Training runs for at most 50 epochs and stops after 15 epochs without improvement in $\mathcal{J}_{\mathrm{sel}}$.

The latest checkpoint is saved after every epoch, with additional periodic checkpoints every five epochs. Each checkpoint stores the online and EMA parameters, optimizer and scheduler states, model and training configurations, gene order, and the spatial-split manifest identifier.

B Dataset and Spatial-Split Details

B.1 Dataset Details

10x Genomics melanoma datasets. C1, C2, and P1 are public Xenium datasets of FFPE human skin melanoma used by GeneFlow (Wang et al., 2026). C1 uses the 282-gene Xenium Human Skin Gene Expression Panel. C2 uses this panel with a 100-gene custom add-on, giving 382 genes. P1 uses a development version of the Xenium 5K Human Pan Tissue and Pathways Panel with 5,006 genes. All three datasets provide spatial expression measurements registered to post-Xenium H&E images. For each retained cell, we use a 256×256 crop centered at its registered coordinate; the generation target contains RGB H&E and an auxiliary DAPI channel.

Following GeneFlow, we remove cells with anomalously low or high transcript counts, further exclude the lowest-count 5%, and normalize and log-transform the remaining expression profiles. We then intersect the processed RNA rows with cells that have valid image crops. This produces 100,597 image–gene pairs for C1, 63,089 for C2, and 91,099 for P1. These are the supervised-pair counts used by our models; they should not be confused with the numbers of cells detected in the original Xenium releases or with the numbers of multi-cell tiles.

Table 6: Paired-data statistics and spatially guarded partitions. Pairs are counted after RNA preprocessing and image matching but before spatial guarding. Guard examples are excluded from both optimization and evaluation.

Dataset	Cohort	Genes	Image–gene pairs	Train	Validation	Guard
C1	Skin melanoma	282	100,597	80,197	20,049	351
C2	Skin melanoma	382	63,089	50,427	12,607	55
P1	Primary dermal melanoma	5,006	91,099	72,879	18,220	0
Total	Three datasets	–	254,785	203,503	50,876	406

Table 7: Spatial-overlap audit on C1. The seed-42 column is a canonical-order reconstruction of the public GeneFlow splitting protocol. Random-split results report the median and range across 21 realizations. The guarded split is used to retrain all compared methods.

Metric	Public protocol seed-42 reconstruction	Random protocol 21 realizations	Guarded C1 split
Train / validation / guard cells	80,477 / 20,120 / 0	80,477 / 20,120 / 0	80,197 / 20,049 / 351
Validation crops overlapping training	19,773 (98.28%)	98.27% [98.11, 98.47]	0 (0.00%)
Overlapping train–validation crop pairs	105,810	105,461 [105,246, 105,966]	0
Median nearest-training overlap area	32.53%	32.45% [32.27, 32.68]	0.00%
Validation crops whose nearest-center training crop overlaps by $\geq 50\%$	8.25%	8.13% [7.71, 8.38]	0.00%
Median nearest-training center distance (pixels)	136.91	136.82 [136.32, 137.18]	4,443.65

The three 10x Genomics datasets are distributed under CC BY 4.0.

B.2 Spatial-Overlap Audit

Source of spatial leakage. The public GeneFlow implementation constructs its training and validation sets using an 80:20 cell-wise `torch.utils.data.random_split`. However, its image examples are 256×256 crops centered on individual cells. Spatially adjacent cells can therefore produce strongly overlapping image crops, and assigning such cells independently to different partitions leaks tissue content from training into validation.

For two axis-aligned crops centered at (x_i, y_i) and (x_j, y_j) with crop width $s = 256$, we define them as overlapping when

$$\mathcal{O}(i, j) = \mathbb{I}\left[|x_i - x_j| < s \ \wedge \ |y_i - y_j| < s\right]. \tag{21}$$

Crops that only touch at an edge are treated as non-overlapping.

Audit protocol. We audit all 100,597 C1 cells that have both a valid expression profile and a corresponding image crop. We first reconstruct the public 80:20, seed-42 splitting protocol using a canonical ordering of cell identifiers. We additionally repeat the same random-split protocol 21 times to determine whether overlap is specific to a particular realization. For every validation crop, we identify all overlapping training crops and measure the number and area of their intersections.

The released GeneFlow checkpoint does not store training and validation cell identifiers. Moreover, the public data loader constructs cell identifiers through a process-dependent set ordering. Consequently, the exact checkpoint membership cannot be recovered from the random seed alone. The

results in Table 7 therefore characterize the released splitting protocol rather than asserting the exact membership of the released checkpoint.

Under the seed-42 reconstruction, each validation crop overlaps a median of five training crops, with a 95th percentile of nine. Considering the maximum overlap across all training crops, 75.55% of validation crops have at least one training crop covering at least 25% of their area, and 9.25% have at least one training crop covering at least 50%. By comparison, the nearest- center training crop covers at least 50% of the validation-crop area for 8.25% of validation crops. The median overlap rate remains 98.27% across 21 random-split realizations, demonstrating that the leakage is a systematic consequence of cell-wise random splitting rather than an isolated random seed. We report this protocol-level leakage analysis only for C1 and do not extrapolate its measured rate to the other datasets.

Spatially guarded splits. For each dataset, we select validation cells from a contiguous spatial region under a seeded projection of the tissue coordinates. Any candidate training cell whose 256×256 crop overlaps a validation crop is placed in an excluded guard set. The validation count is chosen to preserve an 80:20 ratio among retained cells. Table 6 reports the resulting partitions. By construction, they contain no overlapping train–validation crop pairs; this verifies the guarded partitions but does not constitute an audit of the public random-split protocol on C2 or P1.

All GeneFlow and GeneFlowV2 variants are retrained using the same partition for a given dataset, so comparisons cannot be explained by differences in data membership. These partitions provide within-sample spatial holdouts rather than external validation across independent specimens.

B.3 Official-Weight RNA Controllability

To rule out inadequate retraining as a potential explanation for GeneFlow's weak RNA controllability, we additionally perform generation using the officially released pretrained weights. Since official checkpoints are available only for C1 and P1, we evaluate these two datasets. For each dataset, we select RNA profiles from distinct RNA categories and generate images under three fixed noise seeds. Within each row, the seed is fixed and only the RNA profile changes.

These results show that GeneFlow's limited responsiveness to RNA conditioning is also present in the officially released models, rather than being an artifact of our retraining.

C Evaluation Metrics and Dataset-Specific Evaluators

C.1 Notation and evaluation roles

For one dataset, let $\{(x_i, \widetilde{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ denote the matched real image, generated image, and target RNA profile for each held-out sample, and let G be the dataset-specific number of genes. All metrics are computed within a dataset. Two uses of UNI2-h must be distinguished. Image metrics use the original pretrained encoder, frozen without task-specific adaptation. RNA metrics use a separate copy fine-tuned only on real image–RNA pairs from the corresponding dataset and subsequently frozen. No generated image is used to fit or select either evaluator.

C.2 Distributional image quality

Feature extraction. Let f_{I} and f_{U} denote the frozen Inception-v3 and UNI2-h image encoders. We retain the 2048-dimensional output of torchvision Inception-v3 with `IMAGENET1K_V1` weights after replacing its final fully connected layer by the identity. RGB values are clipped to $[0, 1]$, bilinearly resized to 299×299, and mapped to $[-1, 1]$. For UNI2-h, the complete RGB patch is bilinearly resized to 224×224 and normalized by the ImageNet mean and standard deviation; its original foundation-model weights are kept fixed.

Inception FID and UNI2-h FID. For $f \in \{f_{\mathrm{I}}, f_{\mathrm{U}}\}$, let $(\boldsymbol{\mu}_f^{\mathrm{r}}, \boldsymbol{\Sigma}_f^{\mathrm{r}})$ and $(\boldsymbol{\mu}_f^{\mathrm{g}}, \boldsymbol{\Sigma}_f^{\mathrm{g}})$ be the empirical mean and covariance of the complete real and generated feature sets. We compute

$$\mathrm{FD}_f = \left\| \boldsymbol{\mu}_f^{\mathrm{r}} - \boldsymbol{\mu}_f^{\mathrm{g}} \right\|_2^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_f^{\mathrm{r}} + \boldsymbol{\Sigma}_f^{\mathrm{g}} - 2 \left[(\boldsymbol{\Sigma}_f^{\mathrm{r}})^{1/2} \boldsymbol{\Sigma}_f^{\mathrm{g}} (\boldsymbol{\Sigma}_f^{\mathrm{r}})^{1/2} \right]^{1/2} \right). \tag{22}$$

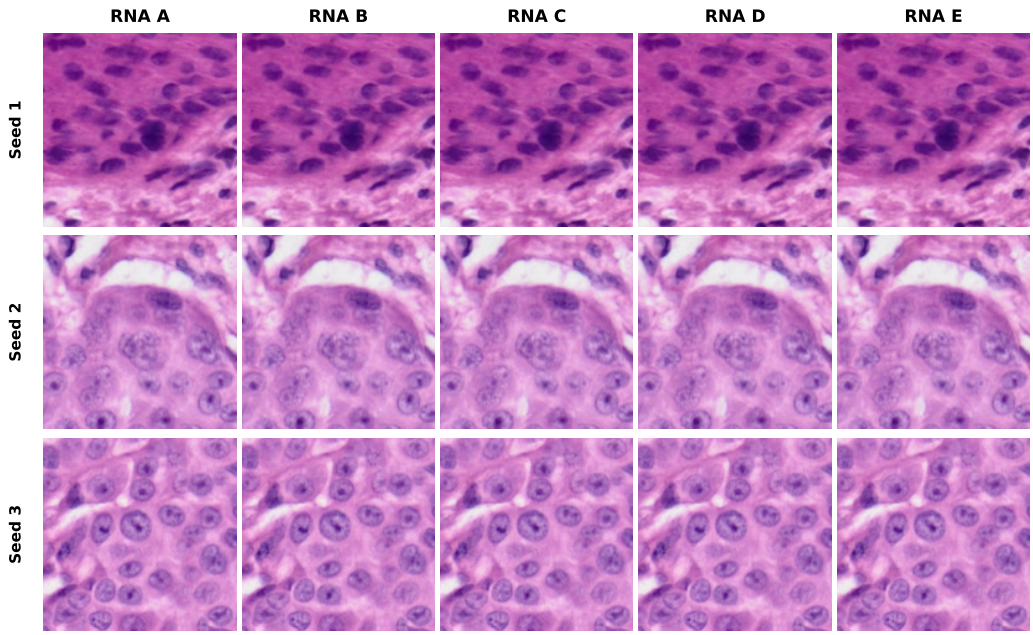

Figure 5: **Official GeneFlow checkpoint on C1.** Columns correspond to five RNA profiles from five distinct RNA categories, and rows correspond to three noise seeds. For a fixed seed, changing the RNA profile produces little visible change, whereas changing the seed leads to substantial morphological variation.

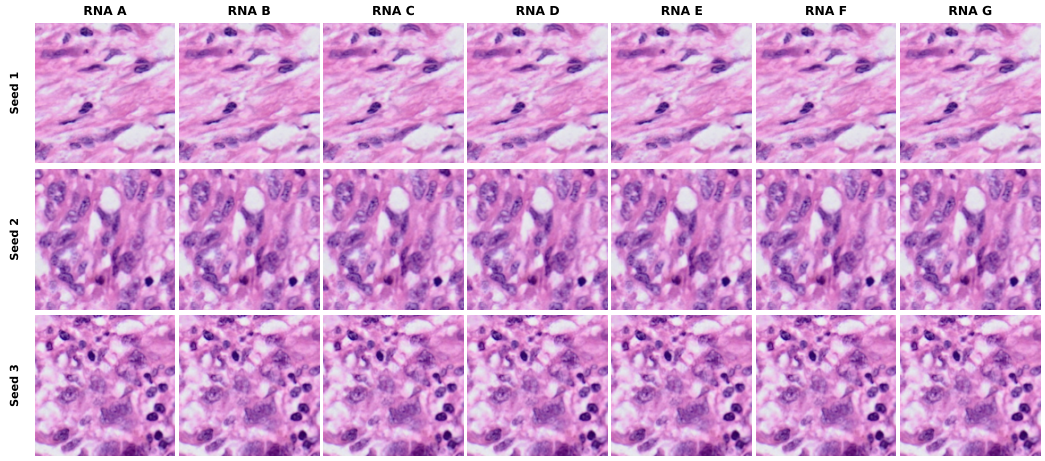

Figure 6: **Official GeneFlow checkpoint on P1.** Columns correspond to seven RNA profiles from seven distinct RNA categories, and rows correspond to three noise seeds. As in C1, outputs remain nearly unchanged across RNA profiles under the same seed, while varying the seed produces clear morphological differences.

We report $\mathrm{FD}_{f_{\mathrm{I}}}$ as Inception FID and $\mathrm{FD}_{f_{\mathrm{U}}}$ as UNI2-h FID. Both are global distributional distances rather than averages of minibatch estimates, and lower values are better.

GLCM-energy distribution similarity (Spatial Energy). Each RGB image is clipped to $[0, 1]$, converted to lossless uint8 RGB by rounding $255x$, and mapped to grayscale with OpenCV `COLOR_RGB2GRAY`. At pixel distance one, we form symmetric normalized 256-level GLCMs for

Table 8: Dataset-specific UNI2-h H&E-to-RNA evaluator configuration. Counts refer to evaluator fitting, inner model selection, and the untouched outer test set, respectively. Batch accumulation gives an effective batch size of 32 for every dataset; "balanced" denotes inverse-frequency cell-type sampling.

Dataset	G	Fit	Dev.	Test	Batch/accum.	Sampling/AMP
Xenium$_{C1}$	282	70,783	7,905	20,049	16×2	balanced / bf16
Xenium$_{C2}$	382	44,927	5,104	12,607	16×2	balanced / bf16
Xenium$_{P1}$	5,006	63,678	7,318	18,220	16×2	natural / bf16

$\Theta = \{0, \pi/4, \pi/2, 3\pi/4\}$ and define

$$E(x) = \frac{1}{4} \sum_{\theta \in \Theta} \left(\sum_{a=0}^{255} \sum_{b=0}^{255} [P_\theta(x)]_{ab}^2 \right)^{1/2}. \tag{23}$$

This matches the mean of scikit-image's GLCM energy property across angles. If $\widehat{F}_{\mathrm{r}}$ and $\widehat{F}_{\mathrm{g}}$ are the empirical CDFs of $E(x_i)$ and $E(\widetilde{x}_i)$, respectively, the reported score is

$$\mathrm{SpatialEnergy} = 1 - \sup_t \left| \widehat{F}_{\mathrm{r}}(t) - \widehat{F}_{\mathrm{g}}(t) \right|. \tag{24}$$

Thus, Spatial Energy is one minus the global two-sample Kolmogorov–Smirnov statistic, not the KS p-value and not the mean image energy. It lies in $[0, 1]$, with higher values indicating closer local-texture distributions.

C.3 Paired image fidelity

The paired metrics preserve the correspondence between x_i and $\widetilde{x}_i$. UNI2-h cosine similarity is

$$\mathrm{Cos}_{\text{UNI2-h}} = \frac{1}{N} \sum_{i=1}^{N} \frac{f_{\mathrm{U}}(x_i)^\top f_{\mathrm{U}}(\widetilde{x}_i)}{\max(\|f_{\mathrm{U}}(x_i)\|_2 \|f_{\mathrm{U}}(\widetilde{x}_i)\|_2, 10^{-12})}, \tag{25}$$

where higher is better. Inception feature distance is the mean unsquared Euclidean distance

$$\mathrm{FeatDist} = \frac{1}{N} \sum_{i=1}^{N} \|f_{\mathrm{I}}(x_i) - f_{\mathrm{I}}(\widetilde{x}_i)\|_2. \tag{26}$$

Finally, LPIPS is the mean AlexNet-based perceptual distance over matched RGB pairs (Zhang et al., 2018). Images are supplied in $[0, 1]$ with the LPIPS normalization option enabled, which maps them to the network's $[-1, 1]$ input range. Lower FeatDist and LPIPS are better.

C.4 Fine-tuning the UNI2-h H&E-to-RNA evaluators

Architecture and input. We train one evaluator h_d per dataset, initialized from the same pre-trained UNI2-h checkpoint. Each model takes the complete 256×256 RGB H&E patch. The RGB patch is resized to 224×224 and ImageNet normalized. A dataset-specific linear head maps the 1536-dimensional UNI2-h representation to all genes in that dataset (282 for C1, 382 for C2, and 5,006 for P1. For each gene, targets are standardized using the mean and standard deviation estimated from the evaluator fit split only, with standard deviations floored at 10^{-3}.

Spatially nested fitting protocol. The evaluator uses the same spatially guarded outer split as generation. The outer test cells—on which generated images are evaluated—are never used for training, early stopping, or checkpoint selection. We further divide the outer training region into a spatially separated fit set and an inner selection set. Table 8 reports the resulting sample counts. Once selected, the same frozen checkpoint, RNA standardization, derangement, and retrieval pools are used for every compared generator on a dataset.

Optimization and selection. We first train only the linear head for five epochs. We then unfreeze the last two UNI2-h transformer blocks while keeping all earlier blocks fixed. Models are optimized with AdamW for at most 20 epochs using learning rates 3×10^{-4} for the head and 3×10^{-6} for the unfrozen backbone, weight decay 10^{-4}, and cosine decay after the head-only warm-up. The training objective is SmoothL1 loss with transition parameter $\beta = 1$ on the standardized gene vector; no program-level, ranking, or correlation term is included. We apply brightness and contrast jitter of 0.08 independently to the RGB channels. Early stopping uses a patience of five epochs, and the checkpoint with the highest inner-set macro gene-wise Pearson correlation is retained. Training uses seed 1701. This supervised adaptation is used only for RNA evaluation; it never replaces the frozen pretrained UNI2-h encoder used by the image metrics above.

C.5 RNA-CONDITIONAL CONSISTENCY METRICS

Let $\widehat{\mathbf{z}}_i = h_d(\widetilde{x}_i) \in \mathbb{R}^G$ and $\mathbf{z}_i \in \mathbb{R}^G$ be the predicted and target profiles after standardization by the fit-set statistics. These metrics quantify agreement as mediated by a fixed H&E-to-RNA predictor; they are not direct molecular measurements of a generated image.

Global top-5 retrieval accuracy (Top-5@32). Using seed 1701, we deterministically shuffle the evaluation indices and partition them into candidate pools of at most 32; a singleton remainder, if present, is excluded. Within a pool $\mathcal{C}$, prediction i scores target j by cosine similarity in standardized RNA space,

$$s_{ij} = \frac{\widehat{\mathbf{z}}_i^\top \mathbf{z}_j}{\|\widehat{\mathbf{z}}_i\|_2 \|\mathbf{z}_j\|_2}. \tag{27}$$

The matched target has rank

$$r_i = 1 + \sum_{j \in \mathcal{C}_i} \mathbf{1}\{s_{ij} > s_{ii}\}, \tag{28}$$

so tied scores do not rank ahead of it. If $\mathcal{I}$ is the set of samples belonging to a non-singleton pool, top-5 retrieval accuracy is

$$\text{Top5@32} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbf{1}\{r_i \le 5\}. \tag{29}$$

Thus, "@32" denotes the maximum candidate-pool size, not a top-32 cutoff over the entire evaluation set. The stored value lies in $[0, 1]$; tables marked Top-5@32 (%) multiply it by 100.

Global retrieval (MRR@32). We deterministically shuffle the evaluation indices with seed 1701 and partition them into candidate pools of at most 32; a singleton remainder, if present, is excluded. Within a pool $\mathcal{C}$, prediction i scores target j by cosine similarity in standardized RNA space,

$$s_{ij} = \frac{\widehat{\mathbf{z}}_i^\top \mathbf{z}_j}{\|\widehat{\mathbf{z}}_i\|_2 \|\mathbf{z}_j\|_2}. \tag{30}$$

The matched target has rank $r_i = 1 + \sum_{j \in \mathcal{C}_i} \mathbf{1}\{s_{ij} > s_{ii}\}$, so tied scores do not rank ahead of it. If $\mathcal{I}$ is the set of samples belonging to a non-singleton pool, then

$$\text{MRR@32} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{1}{r_i}. \tag{31}$$

Thus, "@32" denotes the maximum pool size, not a top-32 cutoff in retrieval over the entire test set. Pair Acc. uses SmoothL1 distance, whereas MRR@32 uses cosine similarity.

Gene-wise Pearson correlation (Gene Pearson). For each gene g, we compute Pearson correlation across held-out samples,

$$\rho_g = \frac{\sum_i (\widehat{z}_{ig} - \overline{\widehat{z}}_g)(z_{ig} - \overline{z}_g)}{\sqrt{\sum_i (\widehat{z}_{ig} - \overline{\widehat{z}}_g)^2} \sqrt{\sum_i (z_{ig} - \overline{z}_g)^2}}, \tag{32}$$

and report the macro-average over genes with finite ρ_g,

$$\text{GenePearson} = \frac{1}{|\mathcal{G}_{\text{finite}}|} \sum_{g \in \mathcal{G}_{\text{finite}}} \rho_g. \tag{33}$$

The implementation also records $|\mathcal{G}_{\text{finite}}|$. This is an average of per-gene correlations across samples, not a correlation after flattening the sample–gene matrix and not an average of within-sample correlations across genes.

D RNA Controllability and Counterfactual Editing Evaluation

D.1 Evaluation Design

Datasets and RNA categories. C1, C2, and P1 use their five, six, and seven native cell categories, respectively. All evaluations use complete observed RNA profiles rather than category centroids as model inputs.

Frozen RNA readout. For each dataset, we train a separate UNI2-h H&E-to-RNA predictor using only real paired data. We fine-tune the last two UNI2-h backbone blocks together with a dataset-specific linear prediction head. Model fitting, checkpoint selection, and outer testing use disjoint spatial partitions (Table 9). After selection, the predictor is frozen and shared across all methods, guidance settings, and evaluation tasks within that dataset. It receives a complete 256×256 RGB patch resized to 224×224. The cFID and UNI2-h Top-1 metrics instead use the original pretrained UNI2-h image encoder, without RNA fine-tuning, and therefore provide an image-space readout independent of the RNA predictor.

Matched-noise RNA control. Each panel contains one observed RNA profile from every category. We use 64 panels and eight prespecified noise seeds $\{1000, 2000, \ldots, 8000\}$ for each dataset. This gives 2,560, 3,072, and 3,584 generated images per method and guidance setting for C1, C2, and P1, respectively. Within a panel and seed, all RNA profiles, methods, and guidance settings receive the same bitwise-identical float32 Gaussian latent. RNA donors are allocated across the outer hold-out's 8×8 spatial blocks. RNA donors, real-image quality references, and evaluator-audit images are disjoint, and all selections are fixed before generation without access to generator or evaluator outputs. RNA-category centroids are estimated using outer-training RNA only. The eight seeds are technical replicates: they are averaged within panel before paired inference and are not treated as independent biological samples.

Source-anchored counterfactual editing. Pair manifests are fixed from held-out data before model evaluation. C1 uses 20 pairs for each of its 10 unordered cross-category combinations and 10 same-category pairs per category, giving 200 cross-category and 50 same-category pairs. C2 uses 24 pairs per cross-category combination and 12 per category, giving 360 and 72 pairs, respectively. P1 uses the analogous seven-category design, giving 504 and 84 pairs. Endpoints are not reused within an experiment. C1 and C2 pairs are matched within the same cell of an 8×8 spatial grid using total expression, detected-gene count, cell area, nuclear area, and spatial position as nuisance variables. Strict same-block matching is infeasible for P1 because the epithelial–melanoma stratum contains only five eligible pairs; P1 therefore uses endpoint-disjoint global nearest-neighbour matching with the same nuisance variables, including spatial coordinates.

Each unordered pair is evaluated in both directions. For a given direction, the model first reconstructs the real source from its inversion and then decodes the same latent state with the complete target RNA profile. The matched real target is used only to construct an unpaired, nuisance-matched evaluation reference; it is never provided to the generator and is not treated as pixel-level counterfactual ground truth. The main results in Table 2 use only cross-category transfers.

D.2 Metrics and Aggregation

Let h denote the frozen dataset-specific H&E-to-RNA predictor, f_{U} the original pretrained UNI2-h image encoder, G the number of measured genes, and K the number of RNA categories. Observed profiles and predicted profiles are compared in the same gene-wise standardized coordinates used to train h. We use the mean smooth-ℓ_1 distance

$$d_{\text{RNA}}(\mathbf{u}, \mathbf{v}) = \frac{1}{G} \sum_{g=1}^{G} \rho(u_g - v_g), \qquad \rho(a) = \begin{cases} a^2/2, & |a| < 1, \\ |a| - 1/2, & |a| \geq 1. \end{cases} \tag{34}$$

Table 9: **Dataset-specific evaluation settings.** Gene Pearson is the mean gene-wise correlation of the frozen H&E-to-RNA predictor on independent real outer-test pairs. cFID entries give the equal real/generated sample count per target category and deterministic resample, not the total number of generated images. UNI2-h centroid references are real images disjoint from the counterfactual endpoints.

Setting	C1	C2	P1
Measured genes	282	382	5,006
RNA categories	5	6	7
Predictor fitting samples	70,783	44,927	63,678
Predictor selection samples	7,905	5,104	7,318
Independent outer-test samples	20,049	12,607	18,220
Real-data mean Gene Pearson	0.3222	0.3039	0.1609
RNA panels / noise seeds	64 / 8	64 / 8	64 / 8
Generated images per condition	2,560	3,072	3,584
RNA-control cFID samples	128	256	256
Counterfactual pairs: cross / same	200 / 50	360 / 72	504 / 84
UNI2-h centroid references	100	256	256
Counterfactual cFID samples	80	120	144

For directional comparisons, we use the zero-safe cosine

$$\cos_0(\mathbf{u}, \mathbf{v}) = \begin{cases} \dfrac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}, & \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{35}$$

The zero convention matches the implementation and assigns zero directional agreement when the predicted change vanishes.

RNA Top-1. Let $\boldsymbol{\mu}_k$ be the mean standardized RNA profile of category k, computed from outer-training cells only. The RNA-predicted category of an image $\mathbf{x}$ is

$$\widehat{k}_{\mathrm{RNA}}(\mathbf{x}) = \arg\min_{\ell \in \{1,\dots,K\}} d_{\mathrm{RNA}}(h(\mathbf{x}), \boldsymbol{\mu}_\ell)\,. \tag{36}$$

For matched-noise generation, let $\mathbf{x}_{p,k,s}$ be the output for panel p, requested category k, and seed s. The reported percentage is

$$\text{RNA-Top1}_{\mathrm{MN}} = \frac{100}{KPS} \sum_{k=1}^{K} \sum_{p=1}^{P} \sum_{s=1}^{S} \mathbb{I}\left[\widehat{k}_{\mathrm{RNA}}(\mathbf{x}_{p,k,s}) = k\right], \tag{37}$$

where $P = 64$ and $S = 8$. Categories receive equal weight, so chance accuracy under balanced prediction is $100/K$ percent.

For counterfactual editing, the same frozen classifier and the same outer-training RNA centroids are applied to the edited image. The score for a source-to-target direction is one when the requested target category is Top-1. The two directions are first averaged within each unordered pair; pairs are then averaged within each unordered cross-category combination, followed by an equal-weight macro-average over category combinations. Thus, the counterfactual RNA Top-1 in Table 2 retains the matched-noise definition while changing only the image-generation protocol and the aggregation unit.

RNA Alignment. For two observed profiles $\mathbf{z}_{p,a}$ and $\mathbf{z}_{p,b}$ in the same matched-noise panel, the corresponding images share the same noise realization. Their alignment is

$$A_{p,s,a,b} = \cos_0(h(\mathbf{x}_{p,b,s}) - h(\mathbf{x}_{p,a,s}), \mathbf{z}_{p,b} - \mathbf{z}_{p,a})\,. \tag{38}$$

We average the eight seeds within each panel and unordered category pair, then average panels within that category pair, and finally macro-average over the $\binom{K}{2}$ category pairs. RNA Alignment measures whether the predicted expression change follows the requested RNA direction; it does not measure the magnitude of the visual change.

Table 10: **Paired improvements in matched-noise RNA control.** Values are GeneFlowV2 ($w_{\mathrm{CFG}} = 2$) minus GeneFlow, with 95% paired-panel bootstrap intervals in brackets. RNA Top-1 differences are in percentage points. All six contrasts have two-sided paired sign-flip $p < 10^{-4}$.

Dataset	Δ RNA Top-1	Δ RNA Alignment
C1	35.39 [33.44, 37.34]	0.2939 [0.2810, 0.3068]
C2	10.94 [8.40, 13.38]	0.0998 [0.0850, 0.1137]
P1	10.02 [8.73, 11.22]	0.0827 [0.0756, 0.0900]

UNI2-h Top-1. This metric is used only for counterfactual editing. Let $\mathcal{R}_k$ be the real-image reference set for category k, disjoint from all counterfactual endpoints, and define

$$\widetilde{f}_{\mathrm{U}}(\mathbf{x}) = \frac{f_{\mathrm{U}}(\mathbf{x})}{\|f_{\mathrm{U}}(\mathbf{x})\|_2}, \qquad \boldsymbol{\nu}_k = \frac{\frac{1}{|\mathcal{R}_k|}\sum_{\mathbf{x}\in\mathcal{R}_k}\widetilde{f}_{\mathrm{U}}(\mathbf{x})}{\left\|\frac{1}{|\mathcal{R}_k|}\sum_{\mathbf{x}\in\mathcal{R}_k}\widetilde{f}_{\mathrm{U}}(\mathbf{x})\right\|_2}. \tag{39}$$

The UNI2-h category prediction is the nearest normalized centroid under cosine similarity,

$$\widehat{k}_{\mathrm{U}}(\mathbf{x}) = \arg\max_{\ell\in\{1,\dots,K\}} \widetilde{f}_{\mathrm{U}}(\mathbf{x})^{\top}\boldsymbol{\nu}_\ell. \tag{40}$$

UNI2-h Top-1 is the percentage of edited images assigned to the requested target category. It uses the same bidirectional pair averaging and category-pair macro-aggregation as counterfactual RNA Top-1. Because it uses the original pretrained image encoder rather than h, it independently tests target matching in pathology-feature space.

Conditional UNI2-h FID. For requested target category k and deterministic balanced resample b, let $\mathcal{R}_{k,b}$ and $\mathcal{G}_{k,b}$ be equally sized real and generated image sets. If $\mathbf{m}$ and $\mathbf{C}$ denote the empirical mean and covariance of their original UNI2-h features, their Fréchet distance is

$$\mathrm{FD}(\mathcal{R}_{k,b},\mathcal{G}_{k,b}) = \|\mathbf{m}_R - \mathbf{m}_G\|_2^2 + \mathrm{tr}\left(\mathbf{C}_R + \mathbf{C}_G - 2\left(\mathbf{C}_R^{1/2}\mathbf{C}_G\mathbf{C}_R^{1/2}\right)^{1/2}\right). \tag{41}$$

We report the equal-weight category macro-average over $B = 100$ deterministic balanced resamples,

$$\mathrm{cFID} = \frac{1}{KB}\sum_{k=1}^{K}\sum_{b=1}^{B}\mathrm{FD}(\mathcal{R}_{k,b},\mathcal{G}_{k,b}). \tag{42}$$

For matched-noise evaluation, generated images are grouped by their requested RNA category. Each resample uses 128 examples per category on C1 and 256 on C2 and P1; when more generated images are available, they are deterministically subsampled by donor and seed. For counterfactual editing, all cross-category directions are grouped by their requested target category, giving 80, 120, and 144 generated images per category on C1, C2, and P1; the real reference set is deterministically subsampled to the same size. The two directions are not averaged for cFID because it is a distributional rather than a pair-level metric.

All UNI2-h metrics use complete RGB patches resized to 224×224. All generated samples are included without success-based filtering. Results are macro-averaged within, and never pooled across, datasets.

D.3 Paired Uncertainty for Matched-Noise RNA Control

For each method contrast, the eight technical seeds are averaged within each of the 64 shared panels before inference. We compute 95% percentile intervals from 10,000 paired bootstrap resamples of panels and two-sided p-values from 100,000 paired sign flips. Table 10 reports the prespecified GeneFlowV2 ($w_{\mathrm{CFG}} = 2$) minus GeneFlow contrast. These intervals quantify variability over the fixed evaluation panels, not over independently trained model instances. Variation across the 100 balanced cFID resamples measures sensitivity to finite reference sampling and is not interpreted as biological uncertainty.

Table 11: **Exploratory direction screen used for case selection.** Entries are mean source-level Spearman correlations between dose and the predicted RNA change along the requested direction, using four sources and one generation seed per row. These sources are excluded from confirmation. Bold marks the case selected for confirmatory evaluation.

Direction	Program ρ	Control ρ	$\Delta\rho$
Epithelial → fibroblast	0.150	−0.225	0.375
Fibroblast → epithelial	0.550	0.350	0.200
Epithelial → myeloid	0.025	−0.150	0.175
Myeloid → epithelial	**0.525**	−0.275	**0.800**

E RNA Control Beyond Cell-Type Labels

E.1 Targeted RNA-Perturbation Validation

Scope and selection disclosure. This experiment asks whether the model response to a sparse, graded RNA perturbation depends on the identities of the perturbed genes rather than only on intervention magnitude. We first conducted an explicitly exploratory GeneFlowV2 $w_{\mathrm{CFG}} = 2$ screen over four directed category–program cases, using four held-out sources and one generation seed per case. The myeloid-to-epithelial direction produced the largest program-minus-control dose–response correlation difference and was frozen for confirmation (Table 11). The four discovery sources were excluded from all confirmatory analyses. This experiment therefore provides a controlled confirmation of the best-supported screened case rather than an unbiased average across programs or directions.

Frozen perturbation and magnitude-matched control. For a held-out myeloid profile $\mathbf{r}^{(0)}$, the target program increased *KRT5*, *KRT17*, *KRT15*, *DSP*, and *TP63* toward their outer-training epithelial 90th percentiles; genes already above their corresponding reference values were left unchanged. Let $\mathbf{r}^{(1)}$ denote the resulting endpoint. We construct five doses, $d \in \{0, 0.25, 0.5, 0.75, 1\}$, by

$$\mathbf{r}^{(d)} = \mathbf{r}^{(0)} + d\big(\mathbf{r}^{(1)} - \mathbf{r}^{(0)}\big). \tag{43}$$

Here, dose d is the interpolation fraction from the source profile to the perturbation endpoint.

The non-program control changes *C11orf96*, *ZNF667-AS1*, *LEPR*, *INSIG1*, and *PLAC9*. Each control gene receives the same outer-training-standardized increment as its paired target-program gene. Consequently, the target-program and control perturbation vectors have identical standardized ℓ_2 norms for every source and dose; the comparison changes gene identity rather than total intervention magnitude.

Matched generation, readout, and inference. We compare the GeneFlow baseline with GeneFlowV2 at $w_{\mathrm{CFG}} = 1$ and $w_{\mathrm{CFG}} = 2$. Every method and guidance condition uses the same source profiles, target-program and control endpoints, dose levels, Gaussian-noise tensors, solver settings, and generation seeds (1101, 2202, and 3303). The frozen C1 UNI2-h H&E-to-RNA readout was fitted and selected only on spatially separated real images and achieved a mean gene-wise Pearson correlation of 0.3222 on an independent real-image test set. It receives the complete 256×256 RGB patch and serves as a fixed directional readout rather than molecular ground truth.

Let $h(\mathbf{x}_d)$ be the predicted standardized RNA vector at dose d, and let

$$\boldsymbol{\delta} = \frac{\mathbf{r}^{(1)} - \mathbf{r}^{(0)}}{\boldsymbol{\sigma}} \tag{44}$$

be the requested target-program change, where $\boldsymbol{\sigma}$ contains the outer-training gene standard deviations. We measure the projected response

$$P_d = \frac{[h(\mathbf{x}_d) - h(\mathbf{x}_0)]^\top \boldsymbol{\delta}}{\max(\|\boldsymbol{\delta}\|_2, 10^{-12})}. \tag{45}$$

The same target-program direction $\boldsymbol{\delta}$ is used to evaluate the target-program and non-program-control trajectories. Dose specificity is

$$S_{\mathrm{dose}} = \rho_{\mathrm{S}}(d, P_d^{\mathrm{program}}) - \rho_{\mathrm{S}}\big(d, P_d^{\mathrm{control}}\big), \tag{46}$$

Table 12: **C1 targeted-program comparison across GeneFlowV2 guidance settings.** Dose specificity is the difference between the target-program and non-program-control dose–response correlations. The $w_{\mathrm{CFG}} = 1$ comparison isolates the model difference, whereas $w_{\mathrm{CFG}} = 2$ evaluates the default guided GeneFlowV2 system against the same GeneFlow baseline. Brackets show 95% paired-source bootstrap confidence intervals. Bold denotes the GeneFlowV2 dose specificity and its improvement over GeneFlow.

GeneFlowV2 w_{CFG}	GeneFlow program	GeneFlow control	GeneFlow specificity	GeneFlowV2 program	GeneFlowV2 control	GeneFlowV2 specificity	GeneFlowV2 − GeneFlow (95% CI)
1	−0.044	0.150	−0.194	0.317	−0.022	**0.339**	**0.533** [0.300, 0.764]
2	−0.044	0.150	−0.194	0.508	0.075	**0.433**	**0.628** [0.317, 0.994]

where ρ_{S} denotes Spearman correlation. The three seeds are technical replicates and are averaged within source before inference. The 12 non-discovery RNA sources are the inferential units. Confidence intervals use 20,000 paired source bootstraps. The prespecified one-sided p-values enumerate all 2^{12} paired sign flips.

GeneFlowV2–GeneFlow comparisons across guidance. At $w_{\mathrm{CFG}} = 1$, GeneFlow did not exhibit a positive program-specific dose response: its dose–response correlation was -0.044 for the epithelial program and 0.150 for the non-program control, giving a dose specificity of -0.194. GeneFlowV2 reached 0.317 and -0.022, respectively, giving a dose specificity of 0.339. The paired GeneFlowV2-minus-GeneFlow improvement was 0.533 (95% CI $[0.300, 0.764]$, exact one-sided $p = 0.0015$), positive for 11 of 12 sources and for every generation seed (Table 12). At full dose, the corresponding GeneFlowV2-minus-GeneFlow improvement in program-minus-control endpoint projection was 0.764 standardized units (95% CI $[0.350, 1.226]$, $p = 0.0022$). At $w_{\mathrm{CFG}} = 2$, GeneFlowV2 dose specificity was 0.433, an improvement of 0.628 over GeneFlow (95% CI $[0.317, 0.994]$, $p = 0.0010$), again positive for 11 of 12 sources and every generation seed.

A representative held-out trajectory illustrates the corresponding selective response at $w_{\mathrm{CFG}} = 1$ (Figure 7).

Interpreting the two guidance settings. The two GeneFlowV2 configurations answer complementary questions: $w_{\mathrm{CFG}} = 1$ isolates the model difference, whereas $w_{\mathrm{CFG}} = 2$ evaluates the default guided system. The direct GeneFlowV2 $w_{\mathrm{CFG}} = 2$ minus $w_{\mathrm{CFG}} = 1$ increment in dose specificity was 0.094 (95% CI $[-0.181, 0.361]$, $p = 0.266$). Thus, although stronger guidance increased the magnitude of pixel-space changes, the additional increase in dose specificity was not resolved, and guidance alone does not account for the GeneFlowV2–GeneFlow separation.

E.2 Within-Type RNA-Profile Control

Purpose and interpretation. The matched-noise experiment changes both the observed RNA profile and the requested category. It therefore cannot exclude a generator that uses RNA only as a coarse category label. This additional C1 experiment holds the annotated cell type fixed and tests whether pairs of generated full patches can be correctly associated with two distinct conditioning profiles drawn from that type. It therefore tests RNA-conditional sensitivity below the annotated cell-type level.

Train-only coordinates and disjoint panels. All transformations were fitted on the spatially guarded C1 outer-training split and then frozen. Within each of the five native cell types, every one of the 282 standardized genes was regressed on log total counts, detected-gene count, log cell area, and log nuclear area. Residuals were centered and scaled using training-only estimates, and a separate 20-component PCA was fitted for each type. We denote the resulting transform by P_c.

For every type, we allocated three mutually disjoint 64-cell outer-test sets: RNA donors, real images for the RNA-readout audit, and real images for UNI2-h quality references. Within each role, the 64 cells formed 32 endpoint-disjoint pairs from the same 8×8 spatial block. Subject to fixed calipers, maximum-weight matching favored large separation in P_c while controlling for total expression,

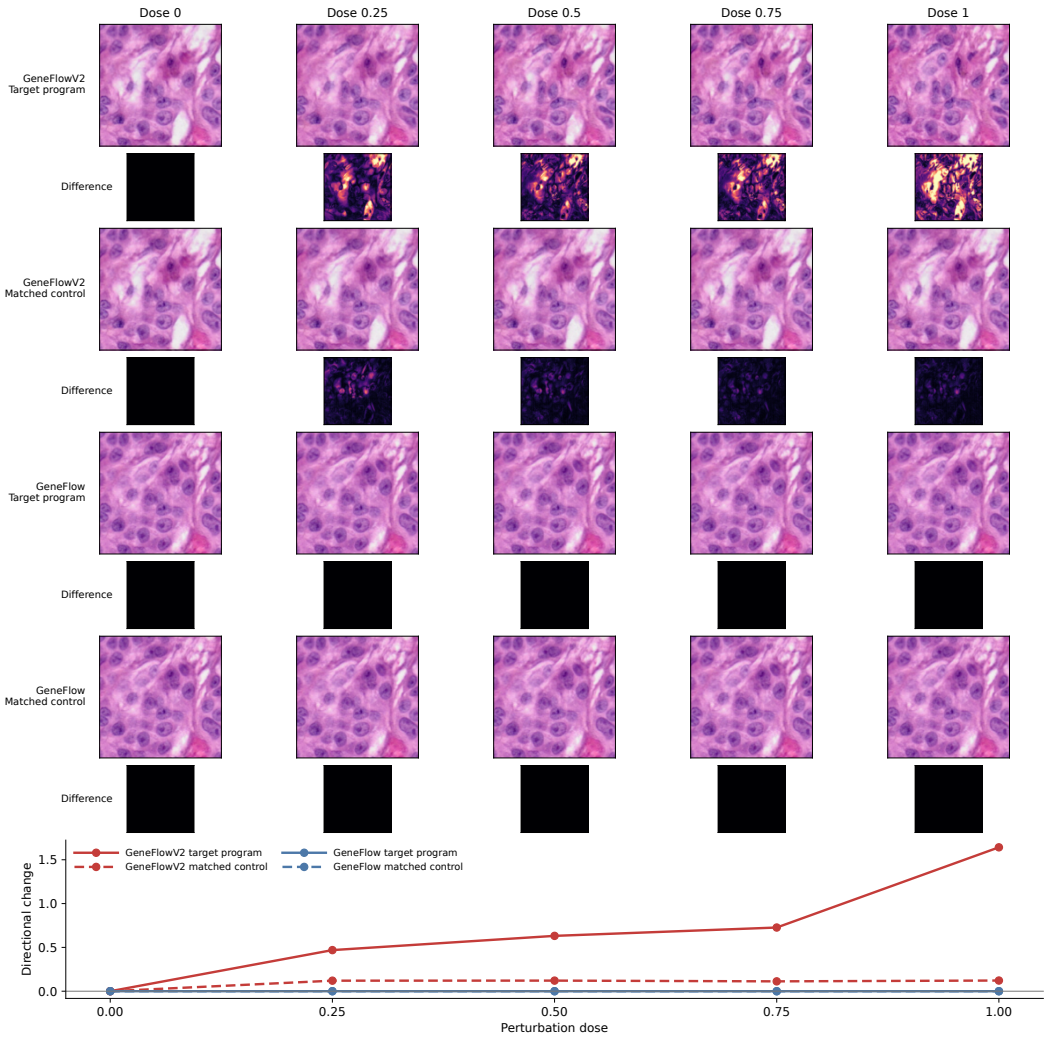

Figure 7: **GeneFlowV2 at $w_{\mathrm{CFG}} = 1$ responds selectively to a graded epithelial-marker program compared with GeneFlow.** Four RGB image trajectories show target-program and equal-norm non-program control perturbations for GeneFlowV2 and GeneFlow at five doses. Each trajectory is followed by its per-pixel mean absolute RGB difference from the corresponding dose-0 image; all difference maps use the same linear scale. The bottom panel shows the predicted change along the requested direction for the displayed source and generation seed. The displayed case was selected from the confirmatory set for visual clarity and then held fixed across methods and controls; all statistical comparisons use every confirmatory source and seed.

detected-gene count, cell area, nuclear area, and within-block position. Selection used no generator output, RNA-readout prediction, or image-quality score (Table 13).

Independent audit of the RNA readout. The primary C1 UNI2-h H&E-to-RNA readout was trained only on real outer-training image–RNA pairs and frozen before this experiment. It receives the complete RGB patch resized to 224×224, without an ROI, mask, or DAPI channel, and predicts all 282 standardized genes. Before generated-image summaries were examined, the same within-type endpoints were evaluated on 160 disjoint real-image pairs. The readout satisfied the prespecified eligibility criteria (Table 14), showing that it can resolve within-type RNA identity on real C1 images.

Table 13: **C1 within-type generation-panel diagnostics.** Every row contains 64 generation donors (32 pairs), 64 disjoint RNA-readout audit cells, and 64 disjoint UNI2-h quality references. Distances and calipers are expressed in standardized matching coordinates.

Cell type	Outer-test cells	Blocks	Max pairs/block	Scalar/position caliper	Median nuisance dist.	Median RNA dist.
Endothelial	776	14	8	0.50/0.75	0.640	12.791
Epithelial/Keratinocyte	4,937	17	8	0.50/0.75	0.608	15.802
Fibroblast/Stromal	2,461	19	8	0.50/0.75	0.576	14.581
Macrophage/Myeloid	207	16	8	1.00/1.25	1.016	7.109
Melanoma	11,668	20	8	0.50/0.75	0.598	15.350

Table 14: **Real-image audit of the within-type RNA readout.** Intervals are 95% spatial-block cluster-bootstrap confidence intervals from 10,000 resamples; p-values use 100,000 two-sided cluster sign flips.

Audit endpoint	Estimate	95% CI	p
Exact-profile swap margin	0.2876	[0.2189, 0.3348]	$< 10^{-4}$
Residual-direction cosine	0.3846	[0.3211, 0.4296]	$< 10^{-4}$
Two-candidate accuracy (%)	63.44	[58.99, 67.46]	–
Target-permutation margin drop	0.3809	[0.3124, 0.4404]	$< 10^{-4}$

Fixed-latent generation. For each of the 160 donor pairs and every seed $s \in \{1000, 2000, \ldots, 8000\}$, the two RNA endpoints were decoded from a bitwise-identical float32 Gaussian tensor. The tensor was also identical across the GeneFlow baseline and GeneFlowV2 at $w_{\mathrm{CFG}} = 1$ and 2; all cross-condition latent hashes were verified to be equal. The three method/guidance conditions produced $5 \times 32 \times 2 \times 8 \times 3 = 7{,}680$ RGB patches, with no result-based filtering. Every endpoint uses the complete 256×256 patch.

Endpoints and inference. For a within-type pair (A, B) of type c, let

$$\mathbf{y}_A = P_c(h(G(\mathbf{z}, \mathbf{r}_A))), \qquad \mathbf{y}_B = P_c(h(G(\mathbf{z}, \mathbf{r}_B))), \tag{47}$$

and

$$\mathbf{t}_A = P_c(\mathbf{r}_A), \qquad \mathbf{t}_B = P_c(\mathbf{r}_B). \tag{48}$$

With mean smooth-ℓ_1 distance D_{RNA}, the primary exact-profile swap margin is

$$M_{\mathrm{swap}} = \frac{1}{2}\left[D_{\mathrm{RNA}}(\mathbf{y}_A, \mathbf{t}_B) - D_{\mathrm{RNA}}(\mathbf{y}_A, \mathbf{t}_A) + D_{\mathrm{RNA}}(\mathbf{y}_B, \mathbf{t}_A) - D_{\mathrm{RNA}}(\mathbf{y}_B, \mathbf{t}_B)\right]. \tag{49}$$

Positive values mean that the two generated patches collectively identify their assigned RNA endpoints better than the swapped assignment. The key directional endpoint is

$$C_{\mathrm{res}} = \cos_0(\mathbf{y}_B - \mathbf{y}_A, \mathbf{t}_B - \mathbf{t}_A), \tag{50}$$

where $\cos_0$ is the zero-safe cosine. We additionally report two-candidate accuracy, a target-permutation control, coarse-type retention, and cell-type-conditional UNI2-h FID and KID.

The eight seeds are technical replicates and are averaged within RNA pair before inference. Cell types receive equal macro weight. Method differences use 10,000 paired spatial-block cluster-bootstrap resamples and 100,000 two-sided cluster sign flips. Holm correction is applied to the two prespecified GeneFlowV2–GeneFlow comparisons on the primary endpoint. The 160 endpoint-disjoint RNA pairs provide the paired observations, and spatial blocks are the resampling and sign-flip clusters; generated images are not treated as independent observations.

Type robustness and target permutation. Both GeneFlowV2–GeneFlow primary differences are positive in every cell type (Table 17). Leaving melanoma out retains positive macro differences of 0.0593 at $w_{\mathrm{CFG}} = 1$ and 0.0922 at $w_{\mathrm{CFG}} = 2$; every other leave-one-type-out estimate is also positive. Permuting targets within cell type and spatial block reduces the GeneFlowV2 margins by 0.1641 (95% CI [0.1224, 0.2097]) and 0.2296 ([0.1771, 0.2908]) at $w_{\mathrm{CFG}} = 1$ and 2, respectively (both $p < 10^{-4}$). The corresponding change for GeneFlow is -0.000001 ([-0.000011, 0.000006]; $p = 0.805$).

Table 15: **Key macro results for C1 within-type RNA-profile control.** M_{perm} is the exact-profile margin after target permutation within cell type and spatial block. Image quality uses the original frozen UNI2-h encoder rather than the RNA-fine-tuned readout. Bold denotes the best value in each column with a specified direction.

Method	w_{CFG}	$M_{\mathrm{swap}} \uparrow$	M_{perm}	$C_{\mathrm{res}} \uparrow$	2AFC (%) ↑	Type retention (%) ↑	cFID ↓	cKID ↓
GeneFlow	–	-0.000001	0.000000	0.0053	50.00	20.00	183.05	0.0999
GeneFlowV2	1	0.1189	-0.0452	0.2076	55.35	37.81	**161.11**	**0.0662**
GeneFlowV2	2	**0.1665**	-0.0631	**0.2233**	**58.16**	**48.48**	170.82	0.0784

Table 16: **Paired within-type contrasts.** Effects are left-minus-right differences with 95% spatial-block cluster-bootstrap confidence intervals. Holm-adjusted p-values apply to the two GeneFlowV2–GeneFlow comparisons on ΔM_{swap}; all other p-values are unadjusted. Effect estimates and intervals are rounded to three decimal places. Bold denotes the larger GeneFlowV2–GeneFlow improvement.

	ΔM_{swap}		ΔC_{res}		Δ 2AFC (percentage points)	
Contrast	Effect [95% CI]	p	Effect [95% CI]	p	Effect [95% CI]	p
GeneFlowV2 ($w_{\mathrm{CFG}} = 1$) − GeneFlow	0.119 [0.077, 0.149]	2.0×10^{-5}	0.202 [0.166, 0.234]	$< 10^{-4}$	5.350 [3.470, 7.160]	$< 10^{-4}$
GeneFlowV2 ($w_{\mathrm{CFG}} = 2$) − GeneFlow	**0.167** [0.116, 0.204]	2.0×10^{-5}	**0.218** [0.181, 0.252]	$< 10^{-4}$	**8.160** [6.030, 10.580]	$< 10^{-4}$
GeneFlowV2 ($w_{\mathrm{CFG}} = 2$) − GeneFlowV2 ($w_{\mathrm{CFG}} = 1$)	0.048 [0.035, 0.059]	$< 10^{-4}$	0.016 [0.001, 0.030]	0.0412	2.810 [1.760, 4.330]	2.0×10^{-5}

Image quality and readout robustness. At $w_{\mathrm{CFG}} = 1$, GeneFlowV2 improves macro cFID from 183.05 to 161.11 and macro cKID from 0.0999 to 0.0662. The $w_{\mathrm{CFG}} = 2$ setting gives the strongest RNA response but a weaker image-quality operating point (cFID 170.82; cKID 0.0784). We therefore report target correctness and response magnitude separately from image quality.

The exact within-type endpoint is mediated by a single frozen H&E-to-RNA readout. This readout passed a disjoint real-image audit, and image quality was measured with an independent frozen encoder, but the exact-profile swap endpoint was not repeated with the H0-mini sensitivity readout used in the broader C1 experiments. Together, these results provide C1-specific evidence that GeneFlowV2 remains sensitive to observed RNA variation below the annotated cell-type level, subject to this readout limitation.

F Complete Component-Ablation Results

Table 18 shows that GeneFlowV2 achieves the strongest RNA consistency, with a global Top-5@32 of 53.14%, MRR@32 of 0.318, and Gene Pearson of 0.222. The spatial pathway carries most of the conditional signal: spatial-only conditioning substantially outperforms global-only conditioning across all three RNA metrics. Crucially, removing RNA ranking reduces these scores to 20.20%, 0.149, and 0.034, respectively—nearly identical to the global-only variant despite retaining the remaining conditioning components. This result identifies RNA ranking as critical for preventing conditional bypass. Removing the cell-summary token or routing all observed genes also weakens RNA consistency, supporting the complementary roles of profile-level context and expression-based token selection. Importantly, these gains do not come at the expense of image quality. GeneFlowV2 achieves the second-best Inception FID (27.33), UNI2-h cosine similarity (0.526), and FeatDist (14.75), while remaining close to the best UNI2-h FID (58.84 versus 57.46) and Spatial Energy (0.951 versus 0.961). Although individual variants optimize particular image-space metrics, the full model provides the strongest overall balance between image fidelity and RNA-conditional consistency.

F.1 Classifier-Free Guidance Sensitivity

Increasing w_{CFG} from 1 to 2 improves global Top-5@32 from 44.74% to 53.14%, MRR@32 from 0.275 to 0.318, and Gene Pearson from 0.183 to 0.222, while UNI2-h FID changes only slightly from 58.23 to 58.84 (Table 19). Stronger guidance yields smaller additional gains in RNA consistency: $w_{\mathrm{CFG}} = 3$ reaches a global Top-5@32 of 56.34% and the highest Gene Pearson (0.231), whereas $w_{\mathrm{CFG}} = 4$ reaches the highest global Top-5@32 (57.59%) and MRR@32 (0.343) without further

Table 17: **Cell-type and leave-one-type-out robustness.** Entries are GeneFlowV2-minus-GeneFlow differences in the primary swap margin. LOTO excludes the indicated cell type and recomputes the equal-type macro effect.

	Type-specific ΔM_{swap}		LOTO ΔM_{swap}	
Cell type	$w_{\text{CFG}} = 1$	$w_{\text{CFG}} = 2$	$w_{\text{CFG}} = 1$	$w_{\text{CFG}} = 2$
Endothelial	0.0518	0.0835	0.1357	0.1873
Epithelial/Keratinocyte	0.0225	0.0531	0.1430	0.1949
Fibroblast/Stromal	0.1249	0.1815	0.1174	0.1628
Macrophage/Myeloid	0.0381	0.0507	0.1391	0.1954
Melanoma	0.3573	0.4637	0.0593	0.0922

Table 18: **Complete component ablations on C1.** Results use generation seed 1000 ($N = 20{,}049$). All GeneFlowV2 variants use $w_{\text{CFG}} = 2$ and the same frozen RNA readout; GeneFlow is included as a contextual reference. Global Top-5@32 is the percentage of matched RNA targets ranked among the top five within deterministic candidate pools of at most 32. Bold denotes the best result in each column.

	Distributional quality			**Paired fidelity**			**RNA consistency**		
Method	Inception FID ↓	UNI2-h FID ↓	Spatial Energy ↑	UNI2-h cosine ↑	LPIPS ↓	FeatDist ↓	Global Top-5@32 (%) ↑	MRR@32 ↑	Gene Pearson ↑
GeneFlow	31.11	65.71	0.813	0.444	0.409	15.98	15.40	0.125	-0.002
Global path only	33.20	67.82	0.871	0.468	0.405	15.70	20.21	0.149	0.033
Spatial path only	30.30	57.68	**0.961**	**0.527**	**0.399**	14.86	43.40	0.267	0.195
All observed genes	27.49	**57.46**	0.914	0.514	0.405	14.86	45.02	0.275	0.183
w/o cell-summary token	**26.19**	58.28	0.953	0.523	0.409	**14.73**	50.84	0.306	0.204
w/o RNA ranking	29.14	64.30	0.893	0.459	0.407	15.74	20.20	0.149	0.034
GeneFlowV2	27.33	58.84	0.951	0.526	0.409	14.75	**53.14**	**0.318**	**0.222**

Table 19: **Classifier-free guidance sensitivity of GeneFlowV2 on C1.** Results use generation seed 1000 ($N = 20{,}049$). The dagger marks the operating point used for de novo generation. Bold denotes the best result in each column.

	Distributional quality			**Paired fidelity**			**RNA consistency**		
w_{CFG}	Inception FID ↓	UNI2-h FID ↓	Spatial Energy ↑	UNI2-h cosine ↑	LPIPS ↓	FeatDist ↓	Global Top-5@32 (%) ↑	MRR@32 ↑	Gene Pearson ↑
1	**24.72**	**58.23**	0.899	0.523	**0.404**	**14.69**	44.74	0.275	0.183
$2^{\dagger}$	27.33	58.84	**0.951**	**0.526**	0.409	14.75	53.14	0.318	0.222
3	30.41	61.91	0.907	0.518	0.420	14.96	56.34	0.334	**0.231**
4	35.13	67.44	0.860	0.504	0.434	15.26	**57.59**	**0.343**	0.231

improving Gene Pearson. Meanwhile, Inception and UNI2-h FID deteriorate steadily under stronger guidance, reaching 35.13 and 67.44, respectively, at $w_{\text{CFG}} = 4$. We therefore use $w_{\text{CFG}} = 2$ as the operating point for de novo generation, providing a favorable balance between image distributional quality and sensitivity to the RNA condition.

F.2 H0-mini Predictor-Swap Control

We additionally trained an H&E-to-RNA evaluator based on the H0-mini histology backbone. The model was trained exclusively on real image–RNA pairs from the evaluator-fitting split, selected using a spatially independent inner development set, and frozen before evaluating the generative models. The final two H0-mini blocks were fine-tuned, and no generated images were used for model fitting or selection. On an independent outer test set of real images, the predictor achieved a mean gene-wise Pearson correlation of 0.351 across all 282 genes. We reused exactly the same generated RGB image sets as in the primary evaluation, changing only the H&E-to-RNA predictor. Table 20 summarizes the three RNA-consistency metrics from the image-to-RNA cycle-evaluation benchmark, together with three matched-noise endpoints and three counterfactual endpoints. The

Table 20: **H0-mini predictor-swap evaluation on C1.** Global Top-5@32, MRR@32, Gene Pearson, target accuracy, target gain, and both RNA-alignment columns are recomputed with the independently fine-tuned H0-mini H&E-to-RNA predictor. All methods reuse the same generated images. Bold denotes the best result in each column.

		Nine-metric benchmark: RNA consistency			Matched-noise RNA control			Counterfactual editing		
Method	w_{CFG}	Global Top-5@32 (%) ↑	MRR@32 ↑	Gene Pearson ↑	cFID ↓	Target acc. (%) ↑	RNA align. ↑	cFID ↓	Target gain ↑	RNA align. ↑
GeneFlow	–	15.31	0.126	-0.002	163.79	20.04	0.000	158.58	0.007	0.056
GeneFlowV2	1	46.29	0.281	0.188	**134.20**	46.02	0.258	**136.43**	0.063	0.228
GeneFlowV2	2	**56.03**	**0.334**	**0.233**	140.93	**55.51**	**0.301**	137.55	**0.070**	**0.255**

two cFID metrics serve as predictor-independent morphological controls and therefore retain their original values from the primary evaluation.