CellFluxMM: Multimodal Generative Modeling of Perturbation-Induced Cellular Responses

Abstract

Understanding cellular responses to genetic and chemical perturbations is essential for uncovering molecular mechanisms and guiding drug discovery. Gene expression and morphology are two of the most widely profiled and complementary readouts: transcriptomics provides molecular insight into regulatory responses, while morphology reflects observable phenotypes. However, most computational approaches treat these modalities in isolation, failing to capture their coupling and limiting both predictive accuracy and interpretability. We introduce CellFluxMM, a multimodal generative framework that simultaneously models perturbation-induced changes in gene expression and morphology. Built upon rectified flow matching, CellFluxMM learns a joint distribution over cellular responses. Given a perturbation and control condition, CellFluxMM can generate consistent transcriptomic and morphological outcomes. Across large-scale datasets spanning both genetic and chemical interventions, the best CellFluxMM achieves an overall 2.5K FID score of 22.26 for morphology generation and a MAE of 0.142 for gene prediction, outperforming unimodal baselines. These results demonstrate the promise of multimodal generative modeling as a scalable and accurate paradigm for in silico perturbation biology, with applications in mechanism-of-action discovery, drug repurposing, and functional genomics.

Keywords: Multimodal generative modeling; Flow matching; Perturbation biology

Data and Code Availability We use public datasets (Haghighi et al., 2022) and will make code publicly available.

Institutional Review Board (IRB) Our research does not require IRB approval.

1. Introduction

Characterizing and predicting how cells respond to genetic and chemical perturbations is a central challenge in cell biology. A wide range of cellular aspects can change upon perturbation, including gene expression (Subramanian et al., 2017b), proteomic composition (Messner et al., 2023), metabolic state (Schuhknecht et al., 2025), and morphology (Chandrasekaran et al., 2024). Among these, gene expression and cell morphology are two of the most widely profiled and complementary readouts: transcriptomics provides direct molecular insight into regulatory responses, whereas morphology reflects observable cellular phenotypes. Together, these modalities provide a more complete view of cellular state transitions than either alone.

Despite their importance, most existing approaches focus on predicting either the transcriptional response or the morphological response to perturbations. This single-modality focus often fails to capture the tight coupling and complementarity between gene expression and morphology, constraining predictive performance and hindering biological insight. Prior studies (Way et al., 2022; Haghighi et al., 2022; Lapins and Spjuth, 2019) have shown that these two readouts share both overlapping and complementary information, suggesting that a unified model that jointly generates both could provide more accurate and generalizable predictions.

Recent advances in generative modeling provide a strong foundation to address these challenges. These models capture complex, high-dimensional distributions and generate coherent samples across domains such as text, images, and videos (Ma et al., 2024; Rombach et al., 2022; Kong et al., 2024). Extending these capabilities to perturbation biology opens the door to in silico simulations of cellular responses.

Recently, a growing number of studies have begun to apply generative models to simulate perturbation effects at the cellular level. IMPA (Palma et al., 2025) and PhenDiff (Bourou et al., 2024) employ a Variational Autoencoder (VAE) and diffusion models, respectively to capture perturbation-affected changes in single-cell states, while CellFlux (Zhang et al., 2025) leverages flow matching to directly model the trajectory from control to perturbed morphology

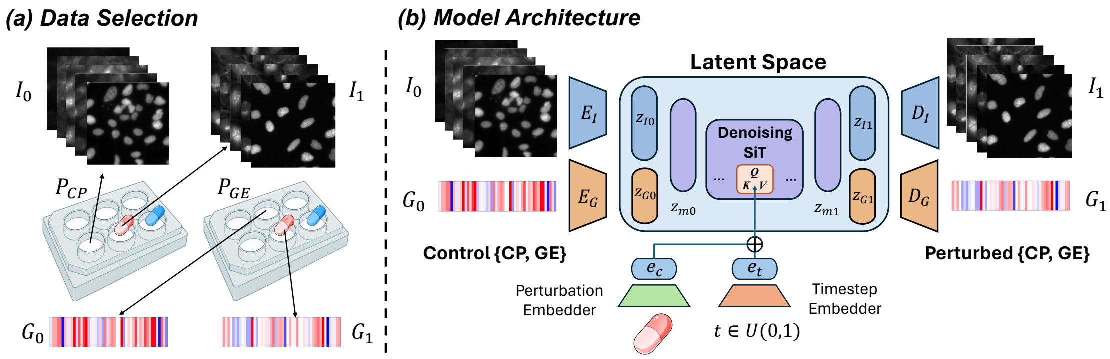

Figure 1: **Overview of CellFluxMM.** (a) *Data Selection.* For a given perturbation, we select wells from the Cell Painting (CP) plate (P_{CP}) treated with this perturbation to obtain the perturbed CP I_1, and randomly choose a control well from the same plate to extract the matched control CP I_0. Similarly, we use the corresponding Gene Expression (GE) plate (P_{GE}) to obtain the perturbed GE G_1 and its control G_0. (b) *Model architecture.* We leverage an image encoder (E_I) and a gene encoder (E_G) to transform control images I_0 and gene expression G_0 into latent representations. A denoising SiT module then integrates the control latent state (z_{m_0}) with a perturbation embedding (e_c) and a stochastic time embedding (e_t, with $t \sim U(0,1)$) to produce the perturbed latent state (z_{m_1}). Finally, decoders (D_I, D_G) reconstruct the perturbed image I_1 and gene expression G_1.

through ODE-based sampling. Compositional Perturbation Autoencoder (Qi et al., 2024) and PerturbNet (Yu et al., 2025) adapt advanced generative models to learn perturbation-dependent changes in gene expression sequences. However, these efforts remain limited to single-modality modeling, focusing exclusively on either morphology or transcriptomics. In addition, some multimodal approaches (Kong et al., 2025; Wang et al., 2025b) condition image generation on perturbed gene expression using diffusion models, but such designs are inherently limited in practical applications since the perturbed transcriptome must be known in advance.

To address this gap, we propose CellFluxMM, a multimodal generative framework that jointly generates perturbed gene expression and morphological changes from control conditions. Built upon rectified flow matching (Liu et al., 2023), the model learns a joint distribution capturing the coupled relationships between transcriptional and morphological responses. During training, it aligns information across modalities to reconstruct consistent outcomes. At inference, given a perturbation, it generates both the perturbed transcriptome and morphology. Compared to prior single-modality approaches (Zhang et al., 2025; Wang et al., 2025a), our framework provides a more comprehensive characterization of perturbation effects and improved cross-modal consistency, enhancing biological interpretability.

We validate CellFluxMM on a large-scale perturbation dataset (Haghighi et al., 2022) spanning both genetic and chemical interventions. Generated outcomes by CellFluxMM capture biologically meaningful relationships between molecular and phenotypic responses, facilitating applications in mechanism of action discovery, drug repurposing and functional genomics.

2. Problem Formulation

In this section, we introduce the objective, data, and mathematical formulation of gene and cellular morphology prediction.

Let $\mathcal{I}$ denotes the cell image space, $\mathcal{G}$ the gene expression space, and $\mathcal{C}$ the perturbation space. Joint modality space $\mathcal{M} = \{\mathcal{I}, \mathcal{G}\}$. Let p_0 represent the joint distribution of untreated cell and gene pairs, and p_1 represent the corresponding distribution after a perturbation $c \in \mathcal{C}$.

We aim to learn a generative model

$$p_\theta : (\mathcal{X} \times \mathcal{M}) \times \mathcal{C} \;\; \rightarrow \;\; \mathcal{P}(\mathcal{X} \times \mathcal{M}),$$

which predicts the conditional joint distribution $p(m_1|m_0, c)$. Here, $m_0 \sim p_0$ denotes the control cell state, and $m_1 \sim p_1$ denotes the perturbed state. Sampling from this conditional distribution enables simultaneous simulation of morphological and transcriptomic responses to perturbations.

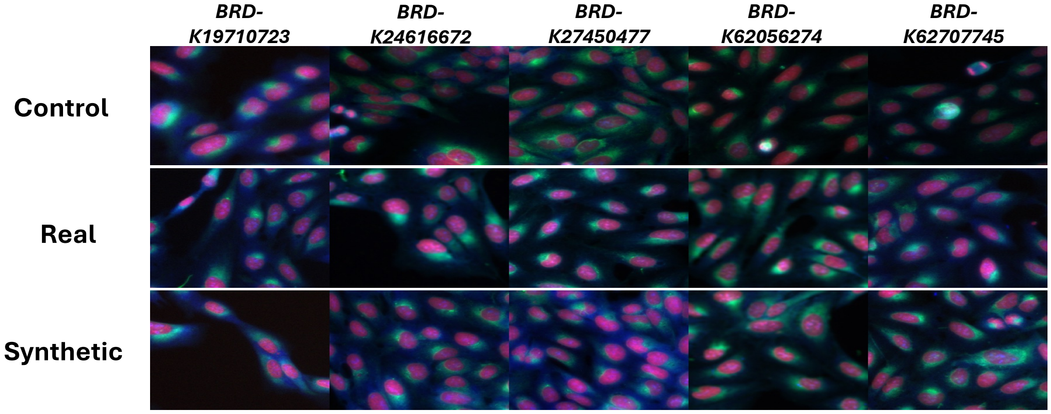

Figure 2: **Image Generation Qualitative Result**. Each column shows a distinct perturbation (BRD ID), with control images, real perturbed images, and synthetic images arranged by row. Perturbations induce clear morphological changes such as elongated cell bodies, thicker cytoplasmic structures, and enhanced fluorescence signals. Synthetic images closely mirror these perturbation-specific patterns, demonstrating the model's ability to capture biologically meaningful phenotypic variations.

2.1. Dataset

We use the public dataset introduced by (Haghighi et al., 2022), a collection of four datasets that systematically profiles over 28,000 chemical and genetic perturbations across both cell morphology which contains 5 channels (DNA, RNA, ER, AGP, and Mito) and gene expression L1000 (Subramanian et al., 2017a) modalities. Providing matches of gene expression and Cell Painting morphology under the same perturbations, this unique dataset enables us to jointly model cellular responses to perturbations across imaging and transcriptomic spaces.

As illustrated in Figure 1a, for each perturbation we select the corresponding perturbed Cell Painting (I_1) and gene expression (G_1) data from their respective experimental plates. To obtain the control counterparts, we then randomly select control samples (I_0, G_0) from the same plates, thereby ensuring that control and perturbed samples share the same batch context. This design allows us to disentangle true perturbation effects from confounding batch effects, and yields paired representations $m_0 = \{G_0, I_0\}$ and $m_1 = \{G_1, I_1\}$, from which we can learn the conditional distribution $p(m_1 \mid m_0, c)$.

In this work, we utilize one of the preprocessed subsets of the dataset, CDRP-BBBC047 (Bray et al., 2017), which contains 21,782 unique chemical perturbations. Altogether, the dataset contains 59,826 matched perturbation–control pairs. Among these, we randomly select 5,000 pairs, corresponding to 5,000 distinct perturbations, to form the test set, while the remaining pairs are used for training.

3. Method

In this work, we consider probability distributions p_0 and p_1 defined over the latent representations z_m, obtained by concatenating the image latent z_I encoded by the image encoder E_I and gene latent z_G encoded by the gene encoder E_G as illustrated in Figure 1b. Given paired samples from these latent distributions, flow matching learns a time-dependent velocity field using a neural network

$$v_\theta : \mathcal{Z} \times [0,1] \to \mathcal{Z},$$

that describes the instantaneous direction and magnitude of change at each point. The transformation process follows the ordinary differential equation:

$$dz_{m_t} = v_\theta(z_{m_t}, t)\, dt, \quad z_{m_0} \sim p_0, \quad z_{m_1} \sim p_1, \quad t \in [0,1].$$

During training, we construct a probability path connecting samples from the source p_0 and target p_1 distributions. We employ the rectified flow formulation (Liu et al., 2023), which yields a simple linear path:

$$z_{m_t} = (1-t) z_{m_0} + t z_{m_1}, \quad t \sim \mathcal{U}[0,1]$$

This linear path has a constant velocity field $v(z_{m_t}, t) = dz_{m_t}/dt = z_{m_1} - z_{m_0}$, which represents the optimal transport direction at each point. The neural network v_θ is trained to approximate this optimal velocity field by minimizing:

$$\mathcal{L}(\theta) = \mathbb{E}_{z_{m_0} \sim p_0, z_{m_1} \sim p_1, t \sim \mathcal{U}[0,1]} \left\| v_\theta(z_{m_t}, t) - v(z_{m_t}, t) \right\|_2^2$$

Table 1: **Main results of CellFluxMM on image generation and gene prediction.** We report FID/KID for generated cell morphology under perturbations and MAE for gene expression prediction. KID values are scaled by 100 for visualization. Variants include removal of modalities (w/o GE, w/o CP) and different weighting factors β to balance the image vs. gene flow-matching loss.

Models	**1K FID**	**2.5K FID**	**5K FID**	**1K KID**	**2.5K KID**	**5K KID**	**5K MAE (Gene)**
CellFluxMM	30.89	**22.26**	17.34	2.05	1.68	1.48	0.142
CellFluxMM w/o GE	31.72	23.76	17.53	2.34	1.72	1.53	–
CellFluxMM w/o CP	–	–	–	–	–	–	0.154
CellFluxMM (β=1)	**30.08**	22.87	**16.78**	**2.03**	**1.62**	**1.44**	0.244
CellFluxMM (β=10)	33.86	24.78	20.05	2.56	1.85	1.65	**0.121**

To further balance the learning of different modalities, we introduce a weighting factor β on the gene channel. Specifically, we decompose the flow-matching loss into an image latent component and a gene latent component as

$$\mathcal{L}_{\text{FM}} = \mathcal{L}_{\text{FM}}^{\text{image}} + \beta\,\mathcal{L}_{\text{FM}}^{\text{gene}},$$

where β controls the relative contribution of the GE modality during training.

At inference time, given a latent sample $m_0 \sim p_0$, we generate m_1 by solving the ODE:

$$z_{m_1} = z_{m_0} + \int_0^1 v_\theta(z_{m_t}, t)dt$$

and reconstruct the image and gene expression by corresponding decoder D_I and D_G. The velocity field v_θ is realized through Scalable Interpolant Transformers (SiT) (Ma et al., 2024). More details about CellFluxMM are presented in Appendix A.

4. Result

In this section, we present detailed results showing CellFluxMM's performance in cellular morphology generation and gene expression under perturbations.

4.1. Evaluation Metrics

We evaluate our framework using both image and gene-level metrics. For image generation under perturbation conditions, we report overall Fréchet Inception Distance (FID) (Heusel et al., 2017) and Kernel Inception Distance (KID) (Bińkowski et al., 2018) to assess visual fidelity and diversity compared with real perturbed images, where both FID evaluation and RGB visualization are conducted on the first three channels (DNA, RNA, ER). For gene prediction, we compute the mean absolute error (MAE) between the generated and ground-truth gene expression vectors to quantify prediction accuracy.

4.2. Qualitative and Quantitative Result

As shown in Fig. 2, the cell images generated by our CellFluxMM are of superior visual quality. Perturbations induce clear morphological changes such as elongated cell bodies, and the synthetic images generated by CellFluxMM closely mirror these perturbation-specific patterns, demonstrating the model's ability to capture biologically meaningful phenotypic variations. This observation is further supported by the quantitative evaluation in Table 1: CellFluxMM achieves an overall FID score (5K) of 17.34 for image generation and a MAE of 0.142 for gene prediction, consistently ranking second-best across both gene prediction and image generation metrics among all model variants, further highlighting the benefit of jointly modeling complementary modalities. Together, these results indicate that CellFluxMM accurately captures perturbation-specific morphological and transcriptional changes. Moreover, the ablation studies confirm that morphology and gene expression provide complementary signals, and leveraging both modalities enables mutual improvement in generation quality and predictive accuracy.

5. Conclusion

In this work, We introduce CellFluxMM, a multimodal generative framework that jointly models perturbation-induced changes in cellular morphology and gene expression. By integrating rectified flow matching in a shared latent space, the model captures both shared and modality-specific responses, achieving strong unimodal performance and improved cross-modal consistency. These results highlight multimodal generative modeling as a promising direction for in silico perturbation biology, with potential applications in mechanism-of-action discovery and drug repurposing.

References

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018.

Anis Bourou, Thomas Boyer, Marzieh Gheisari, Kévin Daupin, Véronique Dubreuil, Aurélie De Thonel, Valérie Mezger, and Auguste Genovesio. PhenDiff: Revealing Subtle Phenotypes with Diffusion Models in Real Images . In *MICCAI*, 2024.

Mark-Anthony Bray, Sigrun M Gustafsdottir, Mohammad H Rohban, Shantanu Singh, Vebjorn Ljosa, Katherine L Sokolnicki, Joshua A Bittker, Nicole E Bodycombe, Vlado Dančík, Thomas P Hasaka, Cindy S Hon, Melissa M Kemp, Kejie Li, Deepika Walpita, Mathias J Wawer, Todd R Golub, Stuart L Schreiber, Paul A Clemons, Alykhan F Shamji, and Anne E Carpenter. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the cell painting assay. *GigaScience*, 2017.

Srinivas Niranj Chandrasekaran, Beth A Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 2024.

Marzieh Haghighi, Juan C Caicedo, Beth A Cimini, Anne E Carpenter, and Shantanu Singh. High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. *Nature Methods*, 2022.

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. In *NeurIPS*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Zhenglun Kong, Mufan Qiu, John Boesen, Xiang Lin, Sukwon Yun, Tianlong Chen, Manolis Kellis, and Marinka Zitnik. SPATIA: Multimodal Model for Prediction and Generation of Spatial Cell Phenotypes. *arXiv preprint arXiv:2507.04704*, 2025.

Maris Lapins and Ola Spjuth. Evaluation of gene expression and phenotypic profiling data as quantitative descriptors for predicting drug targets and mechanisms of action. *Biorxiv*, page 580654, 2019.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *ICLR*, 2023.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring Flow and Diffusion-Based Generative Models with Scalable Interpolant Transformers. In *ECCV*, 2024.

Christoph B Messner, Vadim Demichev, Julia Muenzner, Simran K Aulakh, Natalie Barthel, Annika Röhl, Lucía Herrera-Domínguez, Anna-Sophia Egger, Stephan Kamrad, Jing Hou, et al. The proteomic landscape of genome-wide genetic perturbations. *Cell*, 2023.

Alessandro Palma, Fabian J. Theis, and Mohammad Lotfollahi. Predicting cell morphological responses to perturbations using generative modeling. *Nature Communications*, 2025.

Xiaoning Qi, Lianhe Zhao, Chenyu Tian, Yueyue Li, Zhen-Lin Chen, Peipei Huo, Runsheng Chen, Xiaodong Liu, Baoping Wan, Shengyong Yang, et al. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nature Communications*, 2024.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. *CVPR*, 2022.

Laurentz Schuhknecht, Karin Ortmayr, Jürgen Jänes, Martina Bläsi, Eleni Panoussis, Sebastian Bors, Terézia Dorčáková, Tobias Fuhrer, Pedro Beltrao, and Mattia Zampieri. A human metabolic map of pharmacological perturbations reveals drug modes of action. *Nature Biotechnology*, 2025.

Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong

Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 2017a.

Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 2017b.

Xuesong Wang, Yimin Fan, Yucheng Guo, Chenghao Fu, Kinhei Lee, Khachatur Dallakyan, Yaxuan Li, Qijin Yin, Yu Li, and Le Song. Prediction of cellular morphology changes under perturbations with a transcriptome-guided diffusion model. *Nature Communications*, 2025a.

Xuesong Wang, Yimin Fan, Yucheng Guo, Chenghao Fu, Kinhei Lee, Khachatur Dallakyan, Yaxuan Li, Qijin Yin, Yu Li, and Le Song. Prediction of cellular morphology changes under perturbations with a transcriptome-guided diffusion model. *Nature Communications*, 2025b.

Gregory P Way, Ted Natoli, Adeniyi Adeboye, Lev Litichevskiy, Andrew Yang, Xiaodong Lu, Juan C Caicedo, Beth A Cimini, Kyle Karhohs, David J Logan, et al. Morphology and gene expression profiling provide complementary information for mapping cell state. *Cell systems*, 2022.

Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D Welch. Perturbnet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular Systems Biology*, 2025.

Yuhui Zhang, Yuchang Su, Chenyu Wang, Tianhong Li, Zoe Wefers, Jeffrey J Nirschl, James Burgess, Daisy Ding, Alejandro Lozano, Emma Lundberg, and Serena Yeung-Levy. CellFlux: Simulating Cellular Morphology Changes via Flow Matching. In *ICML*, 2025.

Appendix A. Implementation Details of CellFluxMM

A.1. Image AutoEncoder

We adopt the image VAE from Stable Diffusion (Rombach et al., 2022) as our image autoencoder. Since our Cell Painting data consists of single-channel images, we repeat each channel three times to form pseudo-RGB inputs, which are then processed by the VAE to extract latent features.

A.2. Gene AutoEncoder

We employ a simple MLP-based encoder–decoder architecture as the gene autoencoder consisting of E_G and D_G. This module encodes gene expression profiles into a compact latent representation, facilitating the alignment of gene and image embeddings within a shared latent space.

We employ a simple MLP-based encoder–decoder architecture as the gene autoencoder consisting of E_G and D_G. This module encodes gene expression profiles into a compact latent representation, facilitating the alignment of gene and image embeddings within a shared latent space.

A.3. Perturbation Embedder

In this work, we design a perturbation embedder based on a learnable class embedding table, where each perturbation is assigned a unique embedding vector. The total number of perturbation classes is 21,782, and these embeddings serve as perturbation-specific conditioning signals for the generative model.

A.4. Classifier-Free Guidance

To enhance generation fidelity, we adopt classifier-free guidance (Ho and Salimans, 2022). During training, conditions are randomly dropped with probability p_c by replacing c with a null token $\varnothing$, allowing the model to learn both conditional and unconditional dynamics. At inference, we combine these two predictions through a linear interpolation:

$$v_\theta^{\mathrm{CFG}}(z_{m_t}, t, c) = \alpha\, v_\theta(z_{m_t}, t, c) + (1-\alpha)\, v_\theta(z_{m_t}, t, \varnothing),$$

where $\alpha > 1$ adjusts the strength of the conditioning signal.

A.5. Training and Evaluation Details

We first pretrain the gene autoencoder on all gene expression training datasets using the AdamW optimizer with a learning rate of 1e-4 and a batch size of 256. For image inputs, we randomly select one site per well and crop it to a resolution of 256×256.

The flow matching SiT models are then trained for 30 epochs on 8 L40s GPUs (48 GB) using the AdamW optimizer with a learning rate of 1e-4, a weight decay of 1e-2, and a global batch size of 64. We employ gradient accumulation with a maximum gradient norm of 1. The condition drop probability, classifier-free guidance strength and the gene weighting factor β are set to 0.1 and 1.2, respectively. We adopt the Euler method with 50 sampling steps to generate images and gene expression. Model evaluation is performed using the Exponential Moving Average (EMA) version of the network parameters, and the best checkpoints are selected based on the lowest FID scores on the validation set.