FRESCO: FLOW MATCHING FOR RENDERING EXPRESSION-SPECIFIC CELLULAR MORPHOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting how cells respond to unseen perturbations is central to phenotypic drug discovery and functional genomics. High-throughput screens now measure both cell morphology and single-cell gene expression, but typically in different cells, paired only by treatment. Generating morphologies consistent with a treatment's transcriptional response therefore requires learning a distribution over cells from this group-level pairing. We introduce FRESCO, a conditional flow matching model that generates multi-channel single-cell images from gene expression resampled across a treatment's cells. FRESCO represents expression as transcriptional program tokens, conditions on a batch-matched control cell to account for acquisition effects, and factorizes attention across spatial positions and channels. Because standard image metrics overlook perturbation-specific responses, we also introduce metrics of response direction and retrieval. On two transcriptome–morphology screens (28 compounds; 1,000 gene knockouts), FRESCO reaches a median response cosine of 0.57 and 25.9% top-1 retrieval on 56 held-out knockouts, against 0.28 and 11.1% for the strongest baseline, and roughly halves FID relative to CellFlux on three morphology-only benchmarks. Our implementation is available here.

1 INTRODUCTION

Perturbation experiments allow scientists to identify gene function, compound activity, and mechanism of action by measuring how cells respond to genetic or chemical interventions. Two high throughput readouts provide complementary information. Imaging-based technologies like Cell Painting capture perturbation-induced morphology at single-cell resolution and scale to large libraries (Bray et al., 2016; Cimini et al., 2023; Gustafsdottir et al., 2013; Chandrasekaran et al., 2024), whereas single-cell RNA sequencing measures the gene expression programs induced by the perturbation. Neither modality, however, determines the other fully, since different gene knockouts, and small molecules against different targets, can produce indistinguishable phenotypes(Schneidewind et al., 2020). In matched imaging and scRNA-seq perturbation screens, both modalities resolve local clusters corresponding to individual protein complexes, but transcriptomic profiles separate perturbations at higher phenotypic resolution, whereas image embeddings additionally organize distinct complexes by shared function (Liu et al., 2026; Kudo et al., 2026) . Bridging the two therefore paves the way to connect the structural consequences of a perturbation to the molecular signatures underlying them.

A central obstacle is that these modalities are not observed in the same cell. Imaging and dissociative sequencing are destructive measurements, so multimodal perturbation resources typically apply the same treatment to separate populations of cells and pair the resulting measurements only by treatment (Kudo et al., 2026; Dapello et al., 2025). This setting differs from the two regimes usually assumed in multimodal learning: instance-level pairing, in which every sample contains both views, or fully unpaired data, in which no correspondence is known. Here, correspondence exists only at the group level. Consequently, a transcriptional profile from one treated cell does not uniquely specify the morphology of another treated cell, it constrains a distribution over morphologies compatible with that treatment. This makes cross-modal generation fundamentally distributional rather than pointwise. As a consequence, a deterministic model trained to match independently sampled expression-image pairs can collapse toward a treatment-level average rather than generate realistic individual cells.

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.

Existing approaches do not directly address this setup. Multimodal models of perturbation response have largely focused on representation learning(Agrawal et al., 2025; Kraus et al., 2024), aligning modalities through contrastive or shared-latent objectives (Lu et al., 2025) rather than modeling one conditional distribution from the other. Generative models of cellular morphology have used adversarial (Palma et al., 2025), diffusion (Navidi et al., 2025; Bourou et al., 2024), and flow-matching (Zhang et al., 2025; Jones et al., 2026) objectives, often conditioning on perturbation identity, such as a chemical fingerprint or pretrained gene embedding, often together with a control image. Such representations describe the intervention rather than its measured cellular response. An alternative approach is to condition on experimental measurements. For example, MorphDiff (Wang et al., 2025) conditions on a single population-level transcriptional profile per treatment. Among other measurement-conditioned models, MultiVCDiff (Peng et al., 2026) generates both modalities from a single perturbation representation, and COSMIC (Wen et al., 2026) conditions on the transcriptome of the same cell it generates but produces only the nucleus. Thus, no existing method directly targets generation of full multi-channel morphology while exploiting single-cell transcriptomic gene expression data.

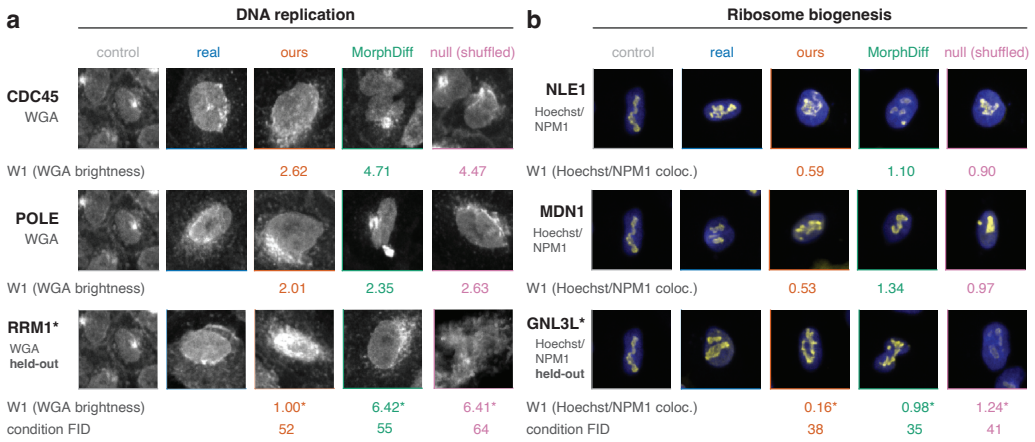

Figure 1: **FRESCO reproduces knockout-specific morphological responses.** Representative crops for non-targeting control, the real knockout, FRESCO, MorphDiff, and a null model in which gene labels are shuffled. Selection and scoring use the same two features throughout: WGA bright-pixel fraction for **(a)** and Hoechst–NPM1 colocalization for **(b)**. The displayed cell is the one nearest the 90th percentile of its own population's response in this feature pair, control included. Numbers are 1D Wasserstein distances (W_1) in control MAD to the real cells of the same knockout (Fig. 6). Condition-level FID for held-out genes is shown at the bottom. Lower is better for both metrics. Channels: Hoechst (blue), NPM1 (yellow). **(a)** DNA replication: *CDC45*, *POLE* and *RRM1*; WGA channel. **(b)** Ribosome biogenesis: *MDN1*, *NLE1* and *GNL3L*; Hoechst and NPM1 channels.

To address these limitations, we introduce **FRESCO**, a conditional generative model for single-cell morphology under group-level multimodal pairing. FRESCO learns a velocity field (Lipman et al., 2023; Esser et al., 2024) in the latent space of a frozen image autoencoder. It generates an image of a treated cell from two inputs: an expression profile estimated from a random subset of that treatment's sequenced cells, and a control cell image sampled independently from the same acquisition batch. The former ensures the model sees the sampling variability of the profile rather than a single fixed estimate, while the latter provides the acquisition background the treatment effect is generated for. Because expression and morphology come from different cells, the model learns the coupling induced by treatment rather than an artificial cell-to-cell correspondence. First, *program pooling* represents expression as tokens corresponding to transcriptional programs rather than collapsing all genes into one dense embedding. Second, *control anchoring* supplies a batch-matched control cell as contextual information to the velocity field rather than using it as the transport source. Third, *channel factorization* encodes fluorescence channels independently and combines them through shared spatial and cross-channel interactions, allowing a single parameterization to accommodate assays with different channel panels.

Our contributions are:

- **Group-level multimodal generation.** We formulate expression-conditioned morphology generation for treatment-matched datasets in which expression and imaging are measured in different cells and paired only through a shared treatment label. We processed two datasets for this setting: a pooled CRISPR screen in A549 cells over 1,000 knockouts (Liu et al., 2026), and scGeneScope (Dapello et al., 2025), 28 chemicals in U2-OS cells.
- **Program pooling** We ablate program pooling against treatment-embedding and one-hot conditioning, and find that retaining program-level structure generalizes better to held-out conditions.
- **Control anchoring and channel factorization** We show that control anchoring recovers acquisition-round shifts in image features, and that cross-channel attention is necessary for realism, with one architecture serving panels of three to seven channels.

2 METHOD

FRESCO models the distribution of single-cell image latents conditioned on a treatment descriptor and an acquisition context. It is a flow transport in the latent space of a frozen image autoencoder, parametrized by a channel-factorized diffusion transformer (Fig. 2).

We deploy FRESCO in two settings. In the *transcriptional* setting, the focus of this work, each treatment is described by single-cell expression profiles measured in separate cells of the same treatment (scGeneScope and the pooled CRISPR screen, OPS Explorer). In the *identity* setting, used for comparison with prior work on the unimodal benchmarks (BBBC021, RxRx1 and JUMP-cpg0000; §3.2), no expression is measured, and each treatment is described only by a fixed property of the perturbation, such as a compound fingerprint or an embedding of the targeted gene (Appendix D.4). Because the identity setting has no expression profiles, program pooling (§2.3) and joint image–gene attention (§2.5) apply only to the transcriptional setting; the transport, control anchoring and channel factorization are shared by both.

2.1 PROBLEM SETTING

An imaging dataset is a set of triples (x_i, y_i, b_i): a single-cell crop $x \in \mathbb{R}^{C\times H\times W}$ with C fluorescence channels, each imaging a distinct subcellular system; a treatment y, such as a drug or genetic perturbation; and the acquisition context b in which the cell was imaged, which captures technical variation such as plate, well position and illumination. Rather than model b explicitly, FRESCO conditions on a control cell from the target's context (§2.4).

Each training example is sampled as

$$(X_1, Y, B) \sim p_{\text{data}}, \quad X^{\text{ctl}} \sim p_{\text{ctl}}(\cdot \mid B), \quad S \sim q_{\text{desc}}(\cdot \mid Y), \tag{1}$$

the target X_1 is a treated cell with treatment Y and acquisition context B; the control crop X^{ctl}, which we call the *anchor*, is drawn from the control cells of the same context; and the descriptor S is drawn from a distribution $q_{\text{desc}}(\cdot \mid Y)$ over descriptors of the treatment (§2.3). X_1 and X^{ctl} are therefore matched by context but are not paired cell to cell (Appendix A.1). Both crops are encoded by the same frozen autoencoder (§2.5) to latents Z_1 and Z^{ctl}, with values z_1 and z^{ctl}.

2.2 CONDITIONAL FLOW MATCHING

With a Gaussian source $Z_0 \sim \mathcal{N}(0, I)$ independent of Z_1, the linear interpolant

$$Z_t = (1-t)\, Z_0 + t\, Z_1, \qquad t \in [0,1], \tag{2}$$

runs from source (t=0) to target (t=1). A velocity network v_θ is trained to regress $Z_1 - Z_0$ (Lipman et al., 2023; Liu et al., 2023),

$$\mathcal{L}(\theta) = \mathbb{E}\left\| v_\theta\big(Z_t, t, S, Z^{\text{ctl}}\big) - \big(Z_1 - Z_0\big) \right\|^2, \tag{3}$$

with the expectation taken over equation 1, Z_0 and a logit-normal t. Samples are generated by integrating $\dot{z} = v_\theta(z, t, s, z^{\text{ctl}})$ from t=0 to t=1 with classifier-free guidance of weight ω, z^{ctl}

being passed unchanged to the conditional and unconditional branches; ω^* is fitted on training data (Appendix G). Details are in Appendix A.2.

The network is conditioned through two pathways. Every input enters through a per-channel modulation vector

$$\mathbf{c}_k = \underbrace{\tau(t)}_{\text{transport time}} + \underbrace{\psi_{\text{pool}}(s)}_{\text{descriptor}} + \underbrace{\epsilon_k}_{\text{channel}} + \underbrace{\psi_{\text{ctl}}(z^{\text{ctl},(k)})}_{\text{control context}} \in \mathbb{R}^D, \tag{4}$$

applied in every block by adaLN-Zero (Peebles & Xie, 2023), where D is the model width, τ a sinusoidal time embedding followed by an MLP, $\psi_{\text{pool}} : \mathbb{R}^{d_c} \to \mathbb{R}^D$ an MLP embedding of the d_c-dimensional descriptor, ϵ_k a learnable embedding of the channel index and ψ_{ctl} the anchor summary of §2.4.

In the transcriptional setting, the expression profile additionally enters as program tokens, constructed in §2.3, which image tokens attend to through joint image–gene attention, detailed in §2.5 (*Joint image–gene attention*) and Appendix A.5.

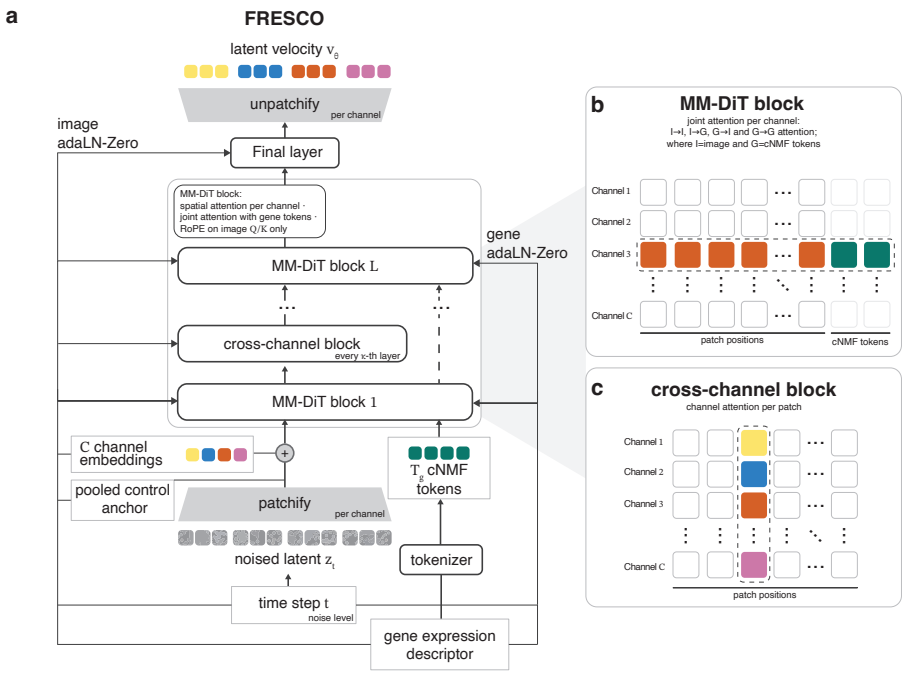

Figure 2: The FRESCO architecture in the transcriptional setting. **(a)** Each channel is encoded by a frozen VAE and the latent is interpolated with Gaussian noise, $z_t = (1-t)z_0 + tz_1$. Image and gene tokens are carried in separate streams, both modulated by the timestep and the descriptor s; the image stream also receives the channel embedding ($\oplus$) and the pooled control latent z^{ctl}. Only image tokens reach the final layer, which predicts the velocity $z_1 - z_0$. **(b)** An MM-DiT block applies one joint attention per channel over image and gene tokens, with RoPE on image tokens only; gene tokens are shared by all channels and re-pooled after each block. **(c)** Every κ-th layer, a cross-channel block attends across the C channels at each patch position.

2.3 Program Pooling

Group-level descriptor. Let $\tilde{e}_i \in \mathbb{R}^G$ be the normalized expression of sequenced cell i over G highly variable genes and $\mathcal{I}_y$ the sequenced cells of treatment y, so that e_y is the mean of $\tilde{e}_i$ over $\mathcal{I}_y$. Rather than e_y itself, the model is given an m-cell estimate of it,

$$S = \frac{1}{\eta m} \sum_{r=1}^{m} \tilde{e}_{J_r}, \qquad J_r \overset{\text{iid}}{\sim} \text{Uniform}(\mathcal{I}_y), \tag{5}$$

redrawn for every training example. Resampling stops the network from memorizing a single fixed input per treatment as a label, encouraging a map that varies smoothly with expression (Bishop, 1995); evaluation draws S at the same m. The constant η fixes the input scale (Appendix A.3).

Program tokens. Rather than compress S into a single vector, we group genes into T_g transcriptional programs by consensus non-negative matrix factorization (cNMF) (Kotliar et al., 2019) of the training cells, giving loadings $M \in \mathbb{R}_{\geq 0}^{T_g \times G}$ with rows summing to one, fitted once and frozen. With frozen per-gene embeddings $E \in \mathbb{R}^{G \times d_e}$ from a single-cell foundation model (scGPT, d_e=512), program j pools expression-weighted gene embeddings,

$$P_j = \sum_{g=1}^{G} M_{jg}\, S_g\, E_{g,:} \in \mathbb{R}^{d_e}, \qquad [\mathbf{g}_0]_j = \mathrm{LN}\big(W_{\mathrm{proj}}\, P_j\big) \in \mathbb{R}^{D}, \tag{6}$$

giving the gene tokens $\mathbf{g}_0 \in \mathbb{R}^{T_g \times D}$ that image tokens attend to (§2.5); only W_{proj} and the LayerNorm are trained. S also enters as a dense summary through ψ_{pool} in equation 4, and we ablate program-only, dense-only and combined conditioning in §3.2.

2.4 Control anchoring

In equation 4, ψ_{ctl} patch-embeds the control latent $z^{\mathrm{ctl},(k)}$ of channel k, mean-pools it over tokens and applies a zero-initialized linear map. Because X^{ctl} is drawn independently of the target given the acquisition context, only its context-level features are informative about the target. Mean-pooling further removes its spatial structure, which provides an inductive bias for z^{ctl} to act as summary of acquisition context rather than a template for the generated cell.

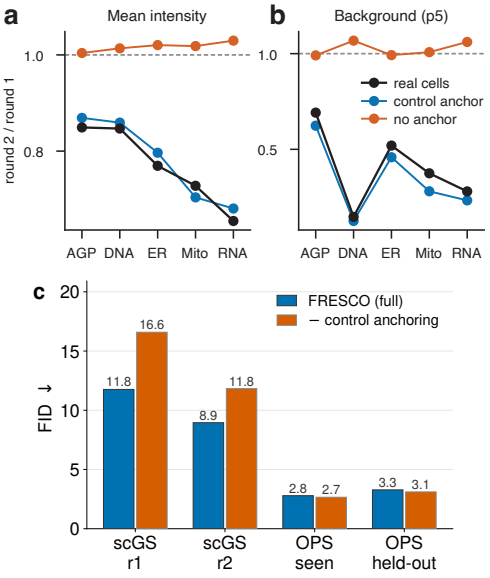

Figure 3: **Control anchoring reproduces imaging-round shifts in scGeneScope (scGS).** Per-channel round-2/round-1 shift computed for **(a)** intensity and **(b)** background over the five Cell Painting channels. **(c)** Comparison between control anchor and without one evaluated with FID. All metrics in Fig. 7.

In scGeneScope, the acquisition context is the imaging round. Fluorescence imaging is subject to batch effects, conventionally handled by including controls in every batch; scGeneScope was imaged over two rounds, and real cells show a round effect in every channel (Fig. 10). Conditioning on a round- and replicate-matched control latent z^{ctl} recovers the direction and approximate magnitude of that shift in mean intensity and background (Fig. 3a,b). A model without ψ_{ctl} receives identical conditioning in both rounds, so the recovered shift comes from z^{ctl}. Removing ψ_{ctl} from training shows what this costs in generation quality. On scGeneScope the evaluated replicates are unseen at training time, whereas on OPS the acquisition context is seen in training. FID rises from 11.76 to 16.59 on round 1 and from 8.95 to 11.83 on round 2 in scGeneScope, while on OPS, it is neutral (2.79 to 2.66 and 3.28 to 3.11; Fig. 3c). All subsequent experiments condition on z^{ctl}.

This differs from control-initialized transport (Zhang et al., 2025), which sets $Z_0 \leftarrow z^{\mathrm{ctl}}$ and transports control images to treated ones. There, the control reaches the output only through the starting point, so sample diversity and control information compete: a deterministic sampler gives one output per control cell, and injecting noise to diversify samples erodes what is carried over from the control. FRESCO keeps a Gaussian source for diversity and passes z^{ctl} as a conditioning input at every step, so the two are controlled separately.

2.5 CHANNEL FACTORIZATION

Per-channel latents. FRESCO operates in the latent space of a frozen pretrained variational autoencoder (Rombach et al., 2022) with downsampling factor f and d_v latent channels. Each fluorescence channel k is encoded independently, replicated to the three RGB channels the encoder expects, giving $z^{(k)} \in \mathbb{R}^{d_v \times h \times w}$ with $h = H/f$ and $w = W/f$. Cross-channel structure is therefore learned by the transformer rather than mixed in a frozen encoder, and one encoder accepts assays with any number of channels. Reconstruction quality is shown in Fig. 8.

Spatial and cross-channel blocks. Fluorescence channels of a cell are strongly dependent: many changes in the AGP channel (actin, Golgi and plasma membrane) are interpretable only relative to the nucleus in the DNA channel. Attending over all CT tokens jointly would capture this but scales poorly in both C and T, so FRESCO factorizes attention into two block types. Each channel latent is patchified by a *shared* embedder with patch size p into tokens $\mathbf{h}^{(k)} \in \mathbb{R}^{T \times D}$, with factored 2D rotary positions and ϵ_k added to the tokens. A *spatial* block is an ordinary transformer block over the T tokens of one channel, with one set of weights for all C channels. Every κ-th layer, a *cross-channel* block attends over the C channel tokens at each spatial position, with per-channel adaLN-Zero modulation (Fig. 9).

To test what the cross-channel block contributes, we replace its attention weights at evaluation time. We compare uniform weights ($A = 1/C$), the identity ($A = I$), and bypassing the block entirely. The trained L-scale model reaches 13.4 $\mathrm{FID_o}$. Uniform weights raise this to 104, bypassing the block to 219 and the identity to 318 (Fig. 4a), so untrained mixing is better than none, and forcing channels to be independent is worse than removing the block. The three substitutions also affect different stages of sampling (Fig. 4b). The identity changes the predicted velocity most near the noise end, with the effect fading as the image forms, whereas the bypass effect grows towards the data end. Uniform weights change it least at every stage. Applied at inference to a model trained with the blocks, these interventions also measure sensitivity to distribution shift. We therefore trained without the blocks, at a matched block count. The verdict is the same: FID rises from 2.79 to 12.26 on OPS and from 11.76 to 20.38 on scGeneScope round 1 (Fig. 4c), and the penalty is largest on the seven-channel panel, where there is most cross-channel structure to model.

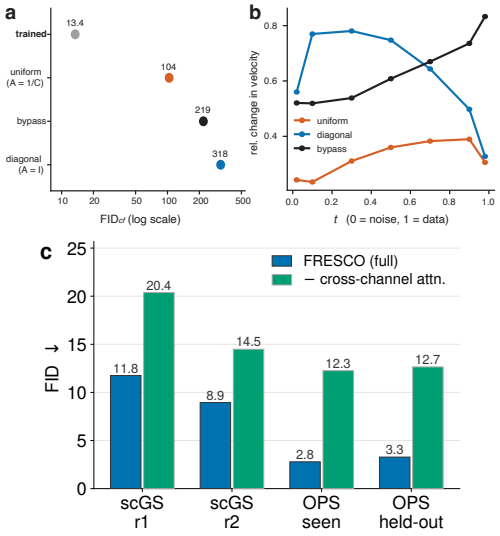

Figure 4: **Channel-DiT blocks. (a–b)** Cross-channel attention interventions trained on RxRx1, applied to all Channel-DiT blocks at once: the learned $C \times C$ attention is replaced by a uniform matrix, or a diagonal one, or the block is bypassed. Shown for **(a)** evaluation $\mathrm{FID_o}$, log axis, and **(b)** relative L_2 change in predicted velocity. **(c)** Training without cross-channel blocks at a matched block count.

Joint image–gene attention. In the transcriptional setting, the spatial block is replaced by an MM-DiT block (Esser et al., 2024). The image tokens of one channel and the gene tokens keep separate projections, normalizations, feed-forwards and modulation, but share one attention over their concatenation, so the gene stream is updated by the image stream as well as read by it. Rotary embeddings apply to image tokens only, the gene stream's modulation omits ϵ_k and the ψ_{ctl} term, and the per-channel copies of the gene tokens are averaged after each block to maintain a single shared representation (Appendix A.5).

Table 1: **Multimodal transcriptome-morphology screens.** $\mathrm{FID}_{\mathrm{inc}}$ is pooled over all perturbations; $\cos(e, d)$ (seed-to-seed noise ≈ 0.03), top-1 and medcos measure per-perturbation conditioning (§3.2). Baselines use their published guidance weights and FRESCO its fitted ω^*. Best per column within each block in bold.

	REALISM	CONDITIONING		
Method	$\mathrm{FID}_{\mathrm{inc}} \downarrow$	$\cos(e, d) \uparrow$	top1 (%) $\uparrow$	medcos $\uparrow$
scGeneScope, round 1	$n = 5{,}600$, $U = 28$, chance top-1 3.6%			
MULTIVCDIFF	34.22	+0.410	18.0	+0.20
CELLFLUX	60.22	+0.560	29.0	+0.44
MORPHDIFF	39.64	+0.279	21.4	+0.29
FRESCO	**11.76**	**+0.640**	**32.1**	**+0.48**
scGeneScope, round 2	$n = 5{,}365$, $U = 27$, chance top-1 3.7%			
MULTIVCDIFF	31.00	+0.150	15.0	+0.07
CELLFLUX	57.13	**+0.770**	**41.0**	**+0.78**
MORPHDIFF	30.88	+0.195	22.2	+0.21
FRESCO	**8.95**	+0.616	37.0	+0.55
OPS Explorer, in-distribution	$n = 19{,}959$, $U = 100$, chance top-1 1.0%			
MULTIVCDIFF	8.41	+0.090	3.1	+0.16
CELLFLUX	13.10	+0.420	3.0	+0.17
MORPHDIFF	3.09	+0.505	3.0	+0.28
FRESCO	**2.79**	**+0.780**	**30.0**	**+0.63**
OPS Explorer, held-out genes	$n = 11{,}183$, chance top-1 1.9% (Appendix G)			
MULTIVCDIFF	7.95	+0.060	4.0	+0.04
CELLFLUX	13.20	+0.270	7.0	+0.09
MORPHDIFF	3.66	+0.402	11.1	+0.28
FRESCO	**3.28**	**+0.680**	**25.9**	**+0.57**

3 EXPERIMENTS

3.1 FRESCO GENERATES PATHWAY-CONSISTENT PHENOTYPES

In the paired OPS Explorer and CROP-seq screen, knockout of three DNA replication genes (*CDC45*, *POLE*, *RRM1*) enlarged cells and redistributed WGA signal from a single compact region beside the nucleus into puncta scattered around it (Fig. 1a), resembling the Golgi remodelling reported under DNA damage (Farber-Katz et al., 2014; Corda et al., 2012). FRESCO reproduced both the spatial pattern and the long-tailed shift in per-cell WGA intensity, including for *RRM1*, which was held out of training ($W_1 = 1.00$ vs. 6.42 for MorphDiff). Three ribosome biogenesis factors (*MDN1*, *NLE1*, *GNL3L*) gave a subtler nuclear phenotype (Fig. 1b): the few large nucleoli of control cells, marked by *NPM1*, fragmented into more numerous, smaller foci. FRESCO captured this for all three, including the held-out *GNL3L* ($W_1 = 0.16$ vs. 0.98), and assigned concordant profiles to *MDN1* and *NLE1*, which act at the same stage of human pre-60S maturation (Thomé et al., 2026). MorphDiff and a shuffled-label null remained near control throughout.

3.2 EXPERIMENTAL SETUP

We evaluate on five high-content screens of crops centered on single nuclei (Table 5), in two groups reported separately. Two *multimodal* screens pair imaging with single-cell RNA-seq of the same treatments, measured in different cells: scGeneScope (Dapello et al., 2025), 28 compounds in U2OS cells, evaluated on acquisition replicates held out from training; and OPS Explorer (Liu et al., 2026), optical pooled screening combined with CROP-seq over 1,000 gene knockouts in A549 cells, evaluated on seen genes and on 56 genes withheld from training. Neither has previously been used for conditional morphology generation, so preprocessing, splits and descriptors are given in Appendix D.1. Three *unimodal* Cell Painting benchmarks describe each perturbation by a label alone: BBBC021 (small molecules, with 8 compounds withheld), RxRx1 (siRNA knockdown) and JUMP-cpg0000 (small molecules, CRISPR knockout and ORF overexpression). For these we inherit preprocessing, splits, metrics and conditioning unchanged from Zhang et al. (2025).

Metrics. We report $\mathrm{FID}_{\mathrm{inc}}$ for realism (Appendix F.1). Conditioning is scored in a space of morphology features per cell, whose number depends on the channel panel, standardised against the pooled real cells of the evaluation split (Appendix F.2). Each perturbation's *response* is its deviation from the mean over perturbations, which removes global offsets and any response shared by all perturbations. $\cos(e, d)$ is the pooled cosine between real and generated responses, **medcos** the median of per-perturbation cosines, and **top-1** the fraction of perturbations whose generated response is closer to its own real response than to any other (chance $1/U$). $\cos(e, d)$ is estimated from unbiased MMD^2 statistics; medcos and top-1 compare mean feature profiles against a reference built from all of a perturbation's cells (Appendix G). Condition-level FID fails to separate FRESCO from MorphDiff on the pathway genes, where the per-feature responses differ several-fold (Appendix E.1).

3.3 RESULTS

Table 2: **Unimodal morphology screens.** $\mathrm{FID}_{\mathrm{o}}$ and $\mathrm{KID}_{\mathrm{o}}$ in Inception feature space (KID $\times 100$); FRESCO is scored with the baselines' released evaluation code. MoA scores generated images with the released classifier, trained on real cells, and reports accuracy and macro-F1. The OOD block withholds eight compounds from training and uses the baseline's OOD checkpoint. Baseline values are quoted from the cited papers (reproduction in Appendix E.3); best generative result per column and block in bold.

	REALISM		MOA CLASSIFICATION	
Method	$\mathrm{FID}_{\mathrm{o}} \downarrow$	$\mathrm{KID}_{\mathrm{o}} \downarrow$	MoA Acc ↑	Macro-F1 ↑
RxRx1				
IMPA	41.6	2.91	–	–
CELLFLUX	33.0	2.38	–	–
FRESCO	**13.7**	**0.42**	–	–
JUMP-cpg0000				
IMPA	14.6	1.08	–	–
CELLFLUX	9.0	0.63	–	–
FRESCO	**4.8**	**0.25**	–	–
BBBC021, in distribution				
Real images *(reference)*	–	–	72.4	69.7
IMPA	33.7	2.60	63.7	40.2
CELLFLUX	18.7	1.62	71.1	49.0
FRESCO	**10.0**	**0.56**	**80.1**	**57.7**
BBBC021, out-of-distribution				
Real images *(reference)*	–	–	88.0	85.0
IMPA	44.5	3.07	16.0	10.0
CELLFLUX	42.0	1.31	43.2	36.6
FRESCO	**25.2**	**1.29**	**54.5**	**40.8**

Morphology generation conditioned on single-cell transcriptome. FRESCO gives the lowest pooled FID in all four evaluation settings of Table 1, by margins of $2.9\times$ on Cell Painting round 1 (34.22 to 11.76) and $1.1\times$ on OPS (3.09 to 2.79), and it leads every column on OPS and scGeneScope on round 1. The gap widens where conditioning is hardest: on the 56 held-out genes it reaches medcos $+0.57$ and top-1 25.9% against $+0.28$ and 11.1% for the strongest baseline, a regime which relies on the descriptor as the perturbation was never seen. On scGeneScope round 2 CellFlux posts higher conditioning scores, but on the easier of the two held-out splits. CellFlux is an image-to-image method that transports a vehicle-control image of the target's batch to a treated one, and round 2's controls are shared across replicates within the round, so its transport starts from control cells already seen in training through round 2/replicate 1, whereas the round-1 test replicate supplies its own (Appendix D.2). We therefore read round 1 as the harder generalization measurement and report both.

Morphology generation conditioned on treatment. On seen perturbations FRESCO attains the lowest $\mathrm{FID}_{\mathrm{o}}$ and $\mathrm{KID}_{\mathrm{o}}$ on all three datasets (Table 2), roughly halving CellFlux's $\mathrm{FID}_{\mathrm{o}}$ ($33.0 \to 13.7$ on RxRx1, $18.7 \to 10.0$ on BBBC021, $9.0 \to 4.8$ on JUMP-cpg0000). BBBC021 MoA accuracy

Table 3: **Conditioning ablations.** Each row represents a separate conditioning type. Columns as in Table 1; each arm is evaluated at its own ω^*.

Model	$\mathrm{FID}_{\mathrm{inc}} \downarrow$	$\cos(e, d) \uparrow$	top1 (%) $\uparrow$	medcos $\uparrow$
scGeneScope, round 1	$n = 5{,}600$, $U = 28$, chance top-1 3.6%			
FRESCO (full)	**11.76**	**+0.640**	32.1	**+0.48**
conditioning → RDKit	16.08	+0.534	35.7	+0.43
conditioning → one-hot	17.01	+0.547	**42.9**	**+0.48**
scGeneScope, round 2	$n = 5{,}365$, $U = 27$, chance top-1 3.7%			
FRESCO (full)	8.95	+0.616	37.0	+0.55
conditioning → RDKit	**8.66**	**+0.737**	**51.9**	+0.67
conditioning → one-hot	9.26	+0.736	**51.9**	**+0.71**
OPS Explorer, in-distribution	$n = 19{,}959$, $U = 100$, chance top-1 1.0%			
FRESCO (full)	**2.79**	+0.780	30.0	+0.63
conditioning → Gene2Vec	3.00	+0.675	20.0	+0.47
conditioning → one-hot	2.89	**+0.837**	**52.0**	**+0.74**
OPS Explorer, held-out genes	$n = 11{,}183$, chance top-1 1.9% (Appendix G)			
FRESCO (full)	**3.28**	**+0.680**	**25.9**	**+0.57**
conditioning → Gene2Vec	3.43	+0.202	1.9	+0.18
conditioning → one-hot	–	–	–	–

reaches 80.1% against 71.1%, above the 72.4% of held-out real cells, which we attribute to the label rather than to fidelity: non-responding cells in real wells inherit their treatment's MoA, whereas guidance concentrates generation on responsive features. On the eight withheld compounds, whose MoA classes all appear in training, FRESCO lowers $\mathrm{FID}_{\mathrm{o}}$ from 42.0 to 25.2 and raises MoA accuracy from 43.2% to 54.5%, though all methods remain far below real cells (88.0%).

Conditioning ablations. Table 3 replaces FRESCO's program-pooled expression descriptor with alternatives while holding the rest of the model fixed. On treatments seen in training, identity descriptors lead the per-perturbation columns: one-hot reaches top-1 52.0% and medcos $+0.74$ on OPS Explorer against 30.0% and $+0.63$ for program pooling, and one-hot and RDKit both lead on scGeneScope round 2. This is expected, since an identity code indexes exactly the treatments it was trained on, and it marks the ceiling reachable when no generalization is required. The ordering reverses out of distribution. One-hot supplies no vector for a gene it never saw and cannot be evaluated, and Gene2Vec, which does supply one, falls to chance ($\cos(e, d) = +0.202$, top-1 1.9%), while program pooling holds at $+0.680$ and 25.9%. Conditioning on the *measured response*, entered as transcriptional programs, therefore transfers where a prior description of the perturbation does not. FRESCO also attains the lowest FID in every block of unseen perturbations, so improved conditioning does not cost image quality.

4 Discussion

FRESCO treats expression-conditioned morphology generation as conditional distribution learning under treatment-level rather than cell-level pairing. Across two transcriptome–morphology screens, it generates realistic cells while recovering perturbation-specific responses, including for gene knockouts absent from training. The morphology-only benchmarks further show that the underlying channel-factorized architecture applies across assays with different perturbation types and fluorescence panels.

The results also separate two aspects of generation that conventional image metrics can conflate. Architectural choices governing transport and cross-channel interactions primarily improve image fidelity, whereas the conditioning representation determines whether generated cells reproduce the measured differences among perturbations. In particular, one-hot and identity-based descriptors can perform well for perturbations seen during training, while expression conditioning provides a route to unseen conditions by representing them through transcriptional programs shared with the training data. Control anchoring complements this representation by adapting generation to the target acquisition context. Together, these findings support group-level multimodal generation as a practical framework for connecting molecular and morphological perturbation responses.

AI USAGE STATEMENT

We used large language models (LLMs) as a general-purpose writing and coding assistant. Their use was limited to organizing the structure of the manuscript, improving clarity and readability, retrieving relevant prior work, and routine software engineering tasks such as boilerplate code, refactoring, and debugging.

All research contributions reported here originate from the authors: the problem formulation and treatment-matched setting, the three method components (program pooling, control anchoring, and the channel-factorized backbone), and the experimental design.

LLMs were not used to generate research ideas, to design the method, or to interpret results. No datasets or mathematical claims were produced with AI assistance.

Every LLM-assisted component was reviewed and verified by the authors. Generated text was checked against our own claims and results, and generated code was read, tested, and validated against the numbers reported in the paper. The authors take full responsibility for all content of this submission.

ETHICS STATEMENT

This work uses publicly available microscopy imaging and single-cell RNA sequencing datasets. Patient data were anonymized by the original data providers, and no new human or animal data were collected. We therefore foresee no ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

Our implementation uses Python 3.12 and PyTorch. We provide the complete training and evaluation pipeline as supplementary material. All experiments were run on up to four NVIDIA H100 GPUs or NVIDIA H200 GPUs (140 GB) with a fixed random seed across every stochastic component. The appendix reports the implementation details, hyperparameters, and data preprocessing steps needed to reproduce our results.

REFERENCES

Vidit Agrawal, John Peters, Tyler N Thompson, Mohammad Vali Sanian, Chau Pham, Nikita Moshkov, Arshad Kazi, Aditya Pillai, Jack Freeman, Byunguk Kang, Samouil L Farhi, Ernest Fraenkel, Ron M Stewart, Lassi Paavolainen, Bryan A Plummer, and Juan C Caicedo. CHAMMI-75: Pre-training multi-channel models with heterogeneous microscopy images. In *The Fourteenth International Conference on Learning Representations*, 2025.

Chris M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995. doi: 10.1162/neco.1995.7.1.108. URL https://doi.org/10.1162/neco.1995.7.1.108.

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 *Lecture Notes in Computer Science*, Lecture Notes in Computer Science, pp. 358–367, Cham, 2024. Springer Nature Switzerland.

Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, September 2016.

Peter D Caie, Rebecca E Walls, Alexandra Ingleston-Orme, Sandeep Daya, Tom Houslay, Rob Eagle, Mark E Roberts, and Neil O Carragher. High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Molecular Cancer Therapeutics*, 9(6):1913–1926, June 2010.

Srinivas Niranj Chandrasekaran, Beth A Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, Alexandr A Kalinin, John Arevalo, Marzieh Haghighi, Juan C Caicedo, Daniel Kuhn, Desiree Hernandez, James Berstler, Hamdah Shafqat-Abbasi, David E Root, Susanne E Swalley, Sakshi Garg, Shantanu Singh, and Anne E Carpenter. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21(6):1114–1121, 2024. doi: 10.1038/s41592-024-02241-6. URL https://www.nature.com/articles/s41592-024-02241-6.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://arxiv.org/abs/1806.07366.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8185–8194. IEEE, 2020. URL https://ieeexplore.ieee.org/document/9157662.

Beth A Cimini, Srinivas Niranj Chandrasekaran, Maria Kost-Alimova, Lisa Miller, Amy Goodale, Briana Fritchman, Patrick Byrne, Sakshi Garg, Nasim Jamali, David J Logan, John B Concannon, Charles-Hugues Lardeau, Elizabeth Mouchet, Shantanu Singh, Hamdah Shafqat Abbasi, Peter Aspesi, Jr, Justin D Boyd, Tamara Gilbert, David Gnutt, Santosh Hariharan, Desiree Hernandez, Gisela Hormel, Karolina Juhani, Michelle Melanson, Lewis H Mervin, Tiziana Monteverde, James E Pilling, Adam Skepner, Susanne E Swalley, Anita Vrcic, Erin Weisbart, Guy Williams, Shan Yu, Bolek Zapiec, and Anne E Carpenter. Optimizing the cell painting assay for image-based profiling. *Nature Protocols*, 18(7):1981–2013, July 2023.

Daniela Corda, Maria Luisa Barretta, Romina Ines Cervigni, and Antonino Colanzi. Golgi complex fragmentation in G2/M transition: An organelle-based cell-cycle checkpoint. *IUBMB Life*, 64(8):661–670, 2012. URL https://iubmb.onlinelibrary.wiley.com/doi/10.1002/iub.1054.

Joel Dapello, Marcel Nassar, Ridvan Eksi, Ban Wang, Jules Gagnon-Marchand, Kenneth Gao, Akram Baharlouei, Kyra Thrush, Nina Riehs, Amy F. Peterson, Aniket Tolpadi, Abhejit Rajagopal, Henry E. Miller, Ashley Mae Conard, David Alvarez-Melis, Rory Stark,

Simone Bianco, Morgan Levine, Ava P. Amini, Alex Xijie Lu, Nicolo Fusi, Ravi Pandya, Valentina Pedoia, and Hana El-Samad. scGeneScope: A treatment-matched single cell imaging and transcriptomics dataset and benchmark for treatment response modeling. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ce02df43d66626bb7087ec699e20c7ea-Abstract-Datasets_and_Benchmarks_Track.html.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2403.03206.

Suzette E Farber-Katz, Holly C Dippold, Matthew D Buschman, Marshall C Peterman, Mengke Xing, Christopher J Noakes, John Tat, Michelle M Ng, Juliati Rahajeng, David M Cowan, Greg J Fuchs, Huilin Zhou, and Seth J Field. DNA damage triggers golgi dispersal via DNA-PK and GOLPH3. *Cell*, 156(3):413–427, January 2014.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Sigrun M Gustafsdottir, Vebjorn Ljosa, Katherine L Sokolnicki, J Anthony Wilson, Deepika Walpita, Melissa M Kemp, Kathleen Petri Seiler, Hyman A Carrel, Todd R Golub, Stuart L Schreiber, Paul A Clemons, Anne E Carpenter, and Alykhan F Shamji. Multiplex cytological profiling assay to measure diverse cellular states. *PLoS One*, 8(12):e80999, December 2013.

Charles Jones, Emmanuel Noutahi, Jason Hartford, and Cian Eastwood. Elucidating the design space of flow matching for cellular microscopy, 2026. URL https://arxiv.org/abs/2603.26790.

Dylan Kotliar, Adrian Veres, M. Aurel Nagy, Shervin Tabrizi, Eran Hodis, Douglas A. Melton, and Pardis C. Sabeti. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-seq. *eLife*, 8:e43803, 2019. doi: 10.7554/eLife.43803. URL https://elifesciences.org/articles/43803.

Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, et al. Masked autoencoders for microscopy are scalable learners of cellular biology. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11757–11768. IEEE, 2024.

Takamasa Kudo, Romain Lopez, Ana M Meireles, Antonio Rios, Paula Coelho, Vineethkrishna Chandrasekar, Dylan C Lam, Jan-Christian Huetter, Mineto Ota, Orit Rozenblatt-Rosen, Levi Garraway, Kathryn Geiger-Shuller, Avtar Singh, Jonathan K Pritchard, and Aviv Regev. Scalable multimodal mapping of macrophage regulatory architecture by integrating optical and transcriptomic pooled screens. *bioRxiv*, pp. 2026.05.27.728345, May 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.02747.

Chad Liu, Alexander Hillsley, Madhurya Sekhar, Cassidy A Jones, Gabriel Sturm, Taihei Fujimori, Diane M Wiener, Karen W Cheng, Talon Chandler, Alexander Lin, Duo Peng, Max Frank, Leah C Dorman, Ilakkiyan Jeyakumar, Ivan E Ivanov, Lin Luan, Greg Courville, Chris B Charlton, Eduardo Hirata-Miyasaki, Sydney Ripsky, Ziwen Liu, Ritvik Vasan, Kyle I Harrington, Kyle Awayan, Trang Le, Yiyun Rao, Giovanni Palla, Vincent Turon-Lagot, Miguel Cid-Rosas, Carolina Arias, Brian DeFelice, Josh Elias, Norma F Neff, Alan R Lowe, Shalin B Mehta, Loïc A Royer, Rafael Gómez-Sjöberg, and Manuel D Leonetti. A multimodal perturbation atlas defines the phenotypic resolution of cellular morphology. *bioRxiv*, pp. 2026.06.01.728087, June 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2209.03003.

MingYu Lu, Ethan Weinberger, Chanwoo Kim, and Su-In Lee. CellCLIP - learning perturbation effects in cell painting via text-guided contrastive learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.06290.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2108.01073.

Zeinab Navidi, Jun Ma, Esteban A. Miglietta, Le Liu, Anne E. Carpenter, Beth A. Cimini, Benjamin Haibe-Kains, and Bo Wang. MorphoDiff: Cellular morphology painting with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=PstM8YfhvI.

Alessandro Palma, Fabian J. Theis, and Mohammad Lotfollahi. Predicting cell morphological responses to perturbations using generative modeling. *Nature Communications*, 16(1):505, 2025. doi: 10.1038/s41467-024-55707-8. URL https://www.nature.com/articles/s41467-024-55707-8.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2212.09748.

Rui Peng, Ziru Liu, Yuanxu Gao, and Jinzhuo Wang. A generative framework for predicting cellular morphological and transcriptomic perturbation responses. *Cell Reports Methods*, 6 (7):101459, 2026. doi: 10.1016/j.crmeth.2026.101459. URL https://www.cell.com/cell-reports-methods/fulltext/S2667-2375(26)00159-1.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022. URL https://arxiv.org/abs/2112.10752.

Tabea Schneidewind, Alexandra Brause, Axel Pahl, Annina Burhop, Tom Mejuch, Sonja Sievers, Herbert Waldmann, and Slava Ziegler. Morphological profiling identifies a common mode of action for small molecules with different targets. *ChemBioChem*, 21(22):3197–3207, 2020. doi: 10.1002/cbic.202000381. URL https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cbic.202000381.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Maciej Sypetkowski, Morteza Rezanejad, Saber Saberian, Oren Kraus, John Urbanik, James Taylor, Ben Mabey, Mason Victors, Jason Yosinski, Alborz Rezazadeh Sereshkeh, Imran Haque, and Berton Earnshaw. Rxrx1: A dataset for evaluating experimental batch correction methods. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4285–4294. IEEE, 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/CVMI/html/Sypetkowski_RxRx1_A_Dataset_for_Evaluating_Experimental_Batch_Correction_Methods_CVPRW_2023_paper.html.

Chairini C Thomé, Nicolas Lemus-Diaz, Katja I Bloch von Blottnitz, Sophie Tagnères, Merle Klein Helmkamp, Mona Honemann-Capito, Philipp Hackert, Sergei Moshkovskii, Christof Lenz, Markus T Bohnsack, Henning Urlaub, and Katherine E Bohnsack. Distinct steps of nuclear maturation of human pre-60S complexes require the activity of GTPases including GNL3L. *Nucleic Acids Research*, 54(5):gkag154, February 2026.

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*, 16:8210, 2025. doi: 10.1038/s41467-025-63478-z. URL https://www.nature.com/articles/s41467-025-63478-z.

Shuo Wen, Ramon Viñas Torné, Johannes Bues, Camille Lucie Lambert, Nadia Grenningloh, Timothée Ferrari, Elisa Bugani, Joern Pezoldt, Jillian Rose Love, Wouter Karthaus, Bart Deplancke, and Maria Brbić. Generative modeling reveals the connection between cellular morphology and gene expression. *bioRxiv*, pp. 2026.01.22.700673, January 2026.

Yuhui Zhang, Yuchang Su, Chenyu Wang, Tianhong Li, Zoe Wefers, Jeffrey Nirschl, James Burgess, Daisy Ding, Alejandro Lozano, Emma Lundberg, and Serena Yeung-Levy. CellFlux: Simulating cellular morphology changes via flow matching. In *International Conference on Machine Learning (ICML)*, 2025. URL https://proceedings.mlr.press/v267/zhang25bt.html.

A METHOD DETAILS

A.1 DATA SAMPLING

The anchor in equation 1 is batch-matched but never cell-paired: target and anchor are drawn independently given the acquisition context. p_{data} is the empirical distribution of treated cells: control cells are used only as anchors and excluded from the training targets, so that capacity is spent on treated cells. The anchor is a single control cell, drawn from the matched pool by the same rule at training and at evaluation (Appendix C), and redrawn every epoch. For identity conditioning S is a deterministic function of the treatment: $q_{\text{desc}}(\cdot \mid y)$ places all its mass on the one fingerprint or gene embedding attached to y, so every draw returns the same vector. For transcriptional conditioning it is the m-cell pooling law equation 5, and a fresh descriptor is drawn for every example. In the transcriptional setting the two assays are run on different physical cells of the same treatment arm, so no cell-level correspondence exists to be recovered.

A.2 FLOW MATCHING AND GUIDANCE

Population minimizer. The minimizer of equation 3 is the conditional expectation

$$v^*(z, t, s, z^{\text{ctl}}) \;=\; \mathbb{E}\big[\, Z_1 - Z_0 \;\big|\; Z_t = z,\, t,\, S = s,\, Z^{\text{ctl}} = z^{\text{ctl}} \,\big]. \tag{7}$$

Under standard regularity conditions, the ODE $\dot{z} = v^*(z, t, s, z^{\text{ctl}})$ realizes the marginal probability path induced by equation 2 and reaches the conditional data distribution at $t{=}1$ (Lipman et al., 2023). The regression target $Z_1 - Z_0$ is constant along each training path. The loss is an elementwise mean-squared error on standardized latents, and training times follow the logit-normal law of Appendix B, where alternative interpolants are also given.

Classifier-free guidance. Conditioning is dropped per sample with probability 0.1, the global descriptor and the program tokens sharing one mask, so a sample is never conditional in one and unconditional in the other. The unconditional signal is learned: a parameter in raw descriptor space pushed through the same ψ_{pool}, and a zero-initialized $T_g \times D$ null token block. The anchor z^{ctl} is passed identically to both branches, so a single doubled-batch evaluation gives

$$v_\theta^{\text{cfg}}(z, t, s, z^{\text{ctl}}) = v_\theta(z, t, \varnothing, z^{\text{ctl}}) + \omega\big[v_\theta(z, t, s, z^{\text{ctl}}) - v_\theta(z, t, \varnothing, z^{\text{ctl}})\big], \tag{8}$$

where $\varnothing$ denotes the learned unconditional signal and ω the guidance weight. Holding the anchor fixed across branches avoids directly scaling an anchor-only contribution, although expression–anchor interactions can still affect the guidance direction.

A.3 EXPRESSION DESCRIPTOR AND PROGRAMS

The expression arm is single-cell RNA-seq, and $\tilde{e}_i$ is a per-cell vector over G genes whose preprocessing is dataset-specific (Appendix D.1, Table 6). On OPS Explorer each cell's vector is additionally centred and scaled to unit norm before pooling: without this, the per-cell counts track sequencing depth, and an m-cell mean is largely a depth average. At evaluation, each generated cell is conditioned on a fresh m-cell draw from equation 5, as in training. S has mean e_y/η and covariance $\Sigma_y/(\eta^2 m)$, with Σ_y the covariance of $\tilde{e}_i$ over $\mathcal{I}_y$: the resampling noise is that of a single scaled cell divided by m. Because subsamples are drawn from the finite pool $\mathcal{I}_y$, they fluctuate around the empirical mean e_y rather than the true treatment mean: resampling conveys the typical size of the estimation error, but cannot remove the error in e_y itself. The constant η is fixed once per m and calibrates the modulation input scale, since the dispersion and sparsity of the pooled vector both depend on m. Because the centring is per cell rather than against controls, S and e_y describe a treatment's expression profile, including basal programs, rather than its response relative to control cells.

The cNMF loadings are fitted on the training-split cells-by-genes matrix before any model training. The split is by acquisition round and replicate, so the fit excludes the held-out round but includes cells of every treatment. In equation 6, W_{proj} is a bias-free map $\mathbb{R}^{d_e} \to \mathbb{R}^D$, and equivalently $P = \big(M \odot (\mathbf{1}_{T_g} S^\top)\big)E$.

Scale invariance of program tokens. Pooling is linear in S and the LayerNorm that follows is scale-invariant: $[\mathbf{g}_0]_j(aS) = [\mathbf{g}_0]_j(S)$ for any $a > 0$. A program token therefore carries which of its genes are elevated relative to one another, not how strongly the program responds overall: a dose that scales a program uniformly leaves its token unchanged, while one that shifts the balance among its genes does not. The norm $\|P_j\|$ cannot supply the missing magnitude either, since gene embeddings within a program can reinforce or cancel. Appendix C therefore lists an explicit program-activity scalar among the tokenizer options.

A.4 Control anchoring

The final linear map of ψ_{ctl} is zero-initialized, so the control term contributes nothing at initialization and the model starts from the same function as one without control conditioning. Conditioning on an observed batch-matched control addresses batch effects at the population optimum, where an informative control can only reduce uncertainty about the target acquisition context; what it contributes to a finite model is measured in §2.4 (all metrics in Fig. 7; examples in Appendix E.7).

A.5 Channel factorization and joint attention

Channel embedding. ϵ_k is added to the token residual as well as entering equation 4. The two injections do separate work: an additive token offset only translates the representation and survives normalization partially, whereas the modulation path acts after normalization and gives each channel its own scale, shift and residual gate. Each cross-channel block applies one projection to each $\mathbf{c}_k$, letting every channel gate the cross-channel update independently.

Joint image–gene attention. Writing $\mathbf{h}_\ell^{(k)}$ for the channel-k image tokens at layer ℓ and $\mathbf{g}_\ell^{(k)}$ for the gene tokens broadcast into that channel, the MM-DiT block computes

$$\big[A_h; A_g\big] = \mathrm{MHA}\big([Q_h; Q_g],\ [K_h; K_g],\ [V_h; V_g]\big), \qquad \begin{aligned} \mathbf{h}_{\ell+1}^{(k)} &= \mathbf{h}_\ell^{(k)} + \lambda_h \odot A_h W_h^o, \\ \mathbf{g}_{\ell+1}^{(k)} &= \mathbf{g}_\ell^{(k)} + \lambda_g \odot A_g W_g^o, \end{aligned} \tag{9}$$

with tokens stored as rows and MHA standard multi-head attention, so A_h and A_g are the head-concatenated outputs entering W_h^o and W_g^o. The bracket denotes concatenation along the token axis: one attention covers image–image, image–gene, gene–image and gene–gene interactions, which distinguishes it from cross-attention conditioning. Rotary embeddings rotate Q_h, K_h only, gene tokens having no spatial position. The feed-forward sublayer that follows is per-stream and omitted here. The per-channel copies are averaged at the end of each block, $\mathbf{g}_{\ell+1} = \frac{1}{C}\sum_k \mathbf{g}_{\ell+1}^{(k)}$. The gene stream's modulation is $\tau(t) + \psi_{\mathrm{pool}}(s)$, omitting ϵ_k and the anchor term, which carry image-side information only. Only image tokens reach the final layer, which projects back to patch space for unpatchification and decoding.

B Transport variants

Alternative interpolants (ablated). Trigonometric (GVP): $\alpha_t = \sin(\frac{\pi t}{2})$, $\sigma_t = \cos(\frac{\pi t}{2})$. Variance-preserving: $\alpha_t = \exp\big(-\frac{1}{2}\int_0^{1-t}\beta_u \mathrm{d}u\big)$, $\sigma_t = \sqrt{1-\alpha_t^2}$ with $\beta_u = \beta_{\min} + (\beta_{\max} - \beta_{\min})u$, $(\beta_{\min}, \beta_{\max}) = (0.1, 20)$; here $\alpha_0 \approx 6.6 \times 10^{-3}$ rather than exactly 0, the usual truncation. For these variants $Z_t = \alpha_t Z_1 + \sigma_t Z_0$, and the regression target in equation 3 becomes $U_t = \dot\alpha_t Z_1 + \dot\sigma_t Z_0$ in place of $Z_1 - Z_0$. Training times follow a logit-normal law, $t = \mathrm{sigmoid}(\varepsilon)$ with $\varepsilon \sim \mathcal{N}(0,1)$, concentrating budget on the mid-schedule where the field is least determined by either endpoint. The flow objective equation 3 is an elementwise mean-squared error.

Relation to score matching. For a variance-preserving path the optimal velocity can be written in terms of the marginal score, which connects flow matching to the probability-flow ODE of Song et al. (2021), itself a neural ODE (Chen et al., 2018): for a Fokker–Planck path with drift f_t and diffusion g_t the transporting field is $f_t - \frac{1}{2}g_t^2 \nabla \log p_t$. The two parameterizations are related by a target transformation and a time-dependent weighting rather than being the same objective; we adopt the velocity form because its regression target is constant in t along a straight-line path.

C IMPLEMENTATION DETAILS

Latent standardization. The mean and standard deviation of each of the $C \times d_v \times h \times w$ latent coordinates are computed once over the training split; the same statistics are applied to target and control latents. Generated latents are unstandardized with those statistics before decoding, which proceeds per channel, averaging the autoencoder's three output channels back to one.

Backbone details. Norms are RMSNorm, feed-forwards SwiGLU, and queries and keys are RMS-normalized per head *before* rotary rotation, so that rotation is the last transform applied before the dot product and the rotary relative-position identity holds. adaLN-Zero predicts a scale γ_i, shift β_i and gate λ_i for each of the block's two sublayers, attention (i=1) and feed-forward (i=2), $(\gamma_i, \beta_i, \lambda_i)_{i=1,2} = \mathrm{MLP}(\mathbf{c}_k)$, and applies them as $\mathbf{h} \leftarrow \mathbf{h} + \lambda_1 \odot \mathrm{Attn}\big(\mathrm{Norm}(\mathbf{h})\,(1+\gamma_1) + \beta_1\big)$, and likewise for the feed-forward with i=2.

Model scale. Scale is varied over three variants. In the single-stream chemical instantiation these are S (12 layers, D=384, 6 heads; 51M parameters), B (12, 768, 12; 199M) and L (24, 1024, 16; 693M), with S approximately parameter-matched to the strongest published baseline ($\approx$ 56M). The transcriptional instantiation carries a second stream in each MM-DiT block; a variant of the same name is therefore larger: its B is $\approx$ 309M. Parameter counts are therefore quoted per instantiation rather than per variant name.

Conditioning and tokenizer. ψ_{pool} is Linear–SiLU–Linear with hidden width $\max(d_c, 4D)$, wider than the width-D conditioning MLP of Peebles & Xie (2023). Genes absent from the foundation model's vocabulary have their rows of E zeroed and so enter no program. The cNMF fit of §2.3 aggregates repeated NMF restarts into consensus programs.

Reproducibility notes. Evaluation configurations are compared under common random numbers, the initial noise for row i being drawn from a generator seeded at $s_0 + i$. The conditioning mask is drawn once, inside the model; drawing it in the training loop as well would compound to an effective rate $p + (1-p)p$. The guidance recombination equation 8 is accumulated in `fp32` under a `bf16` forward pass, since the two velocities are near-equal by construction.

Optimization. AdamW, $\mathrm{lr} = 10^{-4}$, weight decay 0, grad-norm clip 1.0, no LR schedule, 100 epochs. $\beta_2 = 0.999$ for the chemical instantiation, 0.95 for the transcriptional one. OPS Explorer instead uses $\mathrm{lr} = 2 \times 10^{-4}$, 50 epochs and EMA decay 0.9998. Effective global batch 256 (BBBC021: 128×2 accumulation). `bf16` autocast, TF32 matmuls, compiled model, EMA decay 0.9999 raised to the accumulation count per optimiser step. Under data-parallel training the per-rank seed is offset so noise and dropout draws differ across ranks; validation is equation 3 on EMA weights, all-reduced.

Sampling. Fixed-step Heun, 25 steps (adaptive solvers are supported but unused; with a fixed-step method the requested grid *is* the integration grid and solver tolerances are inert). A static timestep shift $s \mapsto \varsigma s/(1 + (\varsigma - 1)s)$ on the noise fraction is available and left at ς=1, leaving the schedule unchanged, unless stated. For the identity-conditioned instantiations, default guidance is $\omega = 3.0$ with sweeps over $\{1.0, 1.5, 2.0, 2.5, 3.0\}$; transcriptional results are reported at the fitted ω^* (Appendix G).

Anchor. Where no vehicle cell matches both the replicate and the batch, the anchor falls back to a batch match alone, then to the global vehicle pool. Anchor dropout is implemented and set to 0 in all reported runs. The anchor is patch-embedded without positional or channel embeddings, since only a pooled summary is consumed, and ψ_{ctl} feeds both the block modulation stream and the final layer's modulation.

Gene tokenizer options (ablated, off by default). Concatenating a per-program activity scalar onto P_j before W_{proj}; a two-layer MLP in place of the bias-free W_{proj}; auxiliary descriptor sources chunked across the T_g token slots and added through zero-initialized projections, each therefore beginning with no contribution.

Table 4: **Multimodal transcriptome-morphology screens.** The two screens used here, where the conditioning descriptor is *measured* single-cell transcriptomics from the same experiment rather than a pretrained embedding. Crop counts are single-cell crops as entered into the splits; image and transcriptome are paired at the perturbation level, not the cell level (§3.2).

	scGeneScope (Dapello et al., 2025)	**OPS Explorer** (Liu et al., 2026)
Cell line	U2OS	A549
Perturbation	Small molecules	CRISPR knockout
Assay	Cell Painting	Optical pooled screening
Channels	5 (AGP, DNA, ER, Mito, RNA)	7 (Hoechst, TOMM20, WGA, Phalloidin, NPM1, Tubulin, ConA)
Perturbations used	29 (28 compounds + DMSO)	1,001 (1,000 targets + non-targeting)
Crops	716,666 at 224×224	539,392 of 683,341 at 128×128
scRNA size	406,412 cells	606,075 cells
Descriptor	2,000 HVG (14 cNMF programs)	2,000 HVG (20 cNMF programs)
Split axis	Replicate 3 / 1 (batch-disjoint)	Two-tier: unseen cells, unseen perturbations

D Experimental configurations

D.1 Datasets

scGeneScope (Dapello et al., 2025) pairs Cell Painting with scRNA-seq on a U2OS cell line treated by a set of small molecules, imaged in five channels over two acquisition rounds. The two assays are run on *distinct* cells of the same treatment arm, so the descriptor is a treatment-level expression profile and no cell-level pairing exists to be recovered. We use 29 conditions (28 compounds and DMSO), 716,666 cells at 224 px crops, with the descriptor read out from 406,412 sequenced cells over 2,000 highly variable genes, summarised as 14 cNMF programs.

OPS Explorer (Liu et al., 2026) is an optical pooled CRISPR knockout screen in A549 lung adenocarcinoma cells (screen `Biohub_OPS0043`), pairing nucleus-centred seven-channel crops to CROP-seq expression of the same screen. Like scGeneScope the descriptor is measured rather than pretrained, but the perturbation vocabulary is two orders of magnitude larger, which makes it the test of whether an expression descriptor still discriminates a thousand perturbations rather than a few dozen. Its nuisance structure is crossed rather than nested, so unlike RxRx1 the perturbation signal exceeds the batch signal and the difficulty is discrimination, not batch correction. We use 1,001 perturbations (1,000 targeted genes and a non-targeting control) and 539,392 cells at 128 px crops, with the descriptor read out from 606,075 sequenced cells over 2,000 highly variable genes as 20 cNMF programs.

BBBC021 (Caie et al., 2010) images MCF-7 breast cancer cells in three channels (nucleus, tubulin, actin) after 24 h exposure to small molecules at multiple doses. A subset of compound–dose pairs carries a curated mechanism-of-action label, which makes it the standard benchmark for asking whether generated images preserve *biological* identity rather than only image statistics: a classifier trained on real perturbed images is applied to generated ones. We use 34 compounds plus DMSO after the baseline's filtering — 26 compounds seen in training and 8 held out entirely — giving 97,504 cells at 96 px crops.

RxRx1 (Sypetkowski et al., 2023) images U2OS cells across many experimental batches after siRNA knockdown of individual genes. Its defining feature is severe batch structure the same perturbation looks different across plates by more than most perturbations differ from each other which makes it the hardest test of the acquisition-context mechanism of §2.4. The descriptor is a pretrained embedding of the targeted gene rather than a chemical fingerprint. We use 1,071 perturbations (1,042 siRNA targets, 29 positive-control siRNAs and an untreated class) over 3 experimental batches spanning 12 plates.

JUMP-cpg0000 (Chandrasekaran et al., 2024) is the pilot release of the JUMP Cell Painting consortium: U2OS cells, five fluorescent channels, profiled under a large perturbation library, including compounds, CRISPR guides and ORF overexpression, with deliberately replicated plate layouts. It is the largest and most heterogeneous of the three, and the only one whose scale tests whether the model benefits from more perturbations rather than more cells per perturbation. We use 747 perturbations: 296 compounds, 296 CRISPR guides and 155 ORFs. The three identity-conditioned screens follow the released preprocessing and splits of the baselines (Palma et al., 2025; Zhang et al., 2025), so that numbers are directly comparable.

Table 5: **Unimodal morphology screens.** Composition of the three identity-conditioned screens as used here, after the baselines' filtering. Crop counts are single-nucleus crops at 96×96; descriptors are those released with the baseline protocol (§3.2).

	BBBC021 (Caie et al., 2010)	**RxRx1** (Sypetkowski et al., 2023)	**JUMP cpg0000** (Chandrasekaran et al., 2024)
Cell line	MCF-7 (breast cancer)	U2OS	U2OS
Perturbation	Small molecules	siRNA knockdown	Compound, CRISPR, ORF
Channels	3	6	5
Perturbations used	34 compounds	1,071 (1,042 siRNA targets, 29 positive controls, untreated)	747 (296 compounds, 296 guides, 155 ORFs)
Crops (96 px)	97,504	170,943	435,161
Descriptor	1024-d Morgan fingerprint	Pretrained gene embedding	Modality-specific, gene-anchored

D.2 Dataset preprocessing

scGeneScope. We take the released 224×224 crops at native resolution and 8-bit depth; a per-channel percentile clip and rescale is applied for consistency but is the identity to within one grey level, so the model sees the released intensities. We drop a cell if any channel saturates over half its pixels or if its mean intensity across channels falls below 0.025, which removes a mitochondrial saturation artefact and blank frames. Our splits are round- and replicate-disjoint rather than random: 303,624 cells from round 1/replicate 3 and round 2/replicate 1 train, round 1/replicate 5 (159,221 cells) is validation, round 1/replicate 4 (120,905 cells) is test, and round 2/replicate 2 (132,916 cells) is a second held-out split. The round-2 holdout is the *easier* of the two despite being nominally further from training: its vehicle controls are shared across replicates within the round, so a model conditioned on a control anchor receives an anchor it has already seen in training via round 2/replicate 1, whereas the round-1 test replicate supplies its own. We therefore treat round 1/replicate 4 as the harder generalisation measurement and report both. For conditioning we library-normalise the counts to 10^4, $\log 1p$-transform, select 2,000 highly variable genes once across both rounds so the gene axis is shared, and fit the 14 cNMF programs on the 115,308 training cells only, so no program sees a held-out batch.

OPS Explorer. We crop our own nucleus-centred 128 px crops from the well mosaics, registering the second imaging cycle onto the first by phase correlation on Hoechst; the duplicated Hoechst channel is the registration anchor and is dropped, leaving seven, and one chunk is rejected for residual misregistration ($r = 0.558$). Per-channel `uint8` quantisation limits are the $[0.05, 99.95]$ percentiles computed once across sampled chunks after registration and then frozen, so shards stay mutually comparable. We exclude cells near the well rim and intersect the remainder with perturbations that have a conditioning profile, leaving 539,392 trainable crops of the 683,341 we extract. Because neighbouring cells are clonal siblings 58% of the time, we split by imaging tile rather than by cell. We first withhold 56 perturbations entirely (27,712 crops), drawn at random as 5% of each of 35 low-resolution transcriptional clusters, so that held-out genes span the transcriptional landscape. The remaining crops are split by $\sim$1.5 mm tile, with tiles assigned at random to train, validation and test (70/15/15; 355,372, 79,854 and 76,454 crops over 944 genes and a non-targeting control), so validation and test share no tile with training. Non-targeting cells are split by tile in the same way, and held-out perturbations are anchored on test-split controls from the same well. In-distribution metrics use 100 seen genes, with up to 200 crops each drawn from all of that gene's crops including training ones, and held-out metrics all 56 genes with up to 200 crops each; medcos and top-1 score 54

of them, the two others falling below the reference-cell floor of Appendix G. Test tiles leave about 80 crops per seen gene, too few for a stable per-gene reference, so the in-distribution real sets pool across splits; because this admits crops seen in training, the held-out-gene block is the primary measure of generalization on OPS Explorer. the GEO per-cell CROP-seq counts, pooled per cell as in equation 5, over 2,000 genes ranked by variance across per-perturbation mean profiles, summarised by consensus cNMF; the held-out genes' cells enter neither the ranking nor the fit.

D.3 Related work

CellFlux (Zhang et al., 2025) formulates perturbation prediction as transport between image distributions. Each perturbed image is paired with an arbitrary control image from its batch: the coupling is independent rather than an optimal transport plan, and batch effects enter only through this source image. A UNet regresses the velocity along the linear interpolant between the pair under a flow matching objective, with all fluorescence channels stacked into one tensor and processed by a shared convolutional stack. The perturbation descriptor, a chemical fingerprint for compounds or a pretrained embedding for genetic perturbations, is embedded, summed with the Fourier time embedding, and injected into the UNet blocks, so conditioning is a global feature-wise signal rather than attention over the descriptor. Training adds Gaussian noise to the source and randomly replaces the descriptor with a null token for classifier-free guidance.

IMPA (Palma et al., 2025) casts perturbation response as image-to-image style transfer, adapting StarGAN v2 (Choi et al., 2020). A content encoder maps the control image to a representation that preserves cell geometry, and a decoder reconstructs the perturbed image from it. The perturbation embedding, concatenated with Gaussian noise, is mapped to a style vector that modulates the decoder through adaptive instance normalization, so conditioning is a spatially uniform scale and shift rather than attention over the embedding. Training is adversarial with a per-perturbation discriminator head, together with a style reconstruction loss requiring the injected style to be recoverable from the generated image and a cycle consistency loss requiring that re-translating the output under the source style reconstructs the input.

MorphDiff (Wang et al., 2025) predicts perturbed morphology with a latent diffusion model (Rombach et al., 2022). Images are encoded by a pretrained autoencoder and the multi-channel latent is flattened into a single tensor whose feature planes are the channels stacked, so cross-channel structure is handled by convolution rather than by any explicit channel mechanism; the tensor is zero-padded inside the network when its spatial size is not a multiple of the downsampling factor. A U-Net regresses the noise under a DDPM with ϵ-prediction and an L_1 objective, $T = 1000$ steps and a sqrt-linear noise schedule, with cross-attention at every resolution. The perturbation descriptor is a bulk-style expression profile, one fixed vector per treatment, z-scored per gene; it is mapped to a block of context tokens by a *frozen random* linear projection, which is a buffer rather than a trained parameter, and enters the U-Net only through cross-attention. Sampling is DDIM, either from noise given the descriptor, or as SDEdit (Meng et al., 2022) from a single control image, which supplies a template at sampling time only: acquisition context is never conditioned on during training, and the published model has no classifier-free guidance, so the descriptor cannot be up-weighted at inference. We report the descriptor-only (noise-initialised) mode throughout, at its published 100 DDIM steps with $\eta = 0$, so that no baseline receives a real image of the target context at sampling time.

MultiVCDiff (Peng et al., 2026) generates morphology and expression jointly. A Cell Painting image is encoded by the SD-VAE (Rombach et al., 2022) and patchified into a token sequence, while the expression profile is encoded by an MLP into a single embedding.[1] One denoising transformer, extending DiT (Peebles & Xie, 2023), operates on the concatenation of the two noisy representations under a shared DDPM with ϵ-prediction and a learned variance; the denoised sequence is split and passed to per-modality decoders. A perturbation encoder maps chemical structures and target-gene identities into a shared space, and this embedding is injected with the timestep by adaptive layer norm; in both cases the descriptor is a single fixed vector per treatment. Expression is therefore an

[1]The paper describes the image encoder as a pretrained vision transformer; the released implementation uses the SD-VAE encoder, which we follow here and in our baseline.

output rather than an input: both modalities are corrupted by the same schedule, so the image attends to an expression embedding whose signal-to-noise falls with t. Batch structure is left unmodelled.

D.4 Model instantiations

FRESCO is instantiated on five screens. Two are *transcriptionally conditioned* (Table 6): the descriptor is a measured expression profile, so the model is told what the perturbation did to the cell's transcriptome and must render the corresponding morphology. Three are *identity-conditioned* (Table 7): the descriptor is a property of the perturbation, known before the experiment is run, and the model is told only which perturbation to render. Model scale is varied over three variants (Appendix C); the S variant is approximately parameter-matched to the strongest published baseline.

Table 6: FRESCO configuration on the multimodal transcriptome–morphology screens.

	scGeneScope	**OPS Explorer**
Autoencoder	SD, f=8, d_v=4	SD, f=8, d_v=4
Per-channel latent	4×28^2	4×16^2
Patch size p	2 (T=196)	1 (T=256)
Cross-channel κ	2	2
Descriptor	per-cell cNMF-normalised counts, G=2000, pooled over m=20 cells	per-cell cNMF counts, centred and unit-normalised per cell, G=2000, pooled over m=200 cells
Conditioning	joint attention	adaLN + joint attention
Gene tokens	T_g=14 cNMF programs, d_e=512	T_g=20 cNMF programs, d_e=512
Anchor	round/replicate-matched vehicle	well-matched non-targeting cell
Backbone	channel-factorized MM-DiT	channel-factorized MM-DiT
Held-out axis	acquisition replicate	56 genes

Table 7: FRESCO configuration on the unimodal morphology benchmarks. All three use adaLN conditioning only, with no gene tokens.

	BBBC021	**RxRx1**	**JUMP-cpg0000**
Autoencoder	FLUX, f=8, d_v=16	FLUX, f=8, d_v=16	FLUX, f=8, d_v=16
Per-channel latent	16×12^2	16×12^2	16×12^2
Patch size p	1 (T=144)	1 (T=144)	1 (T=144)
Cross-channel κ	2	2	2
Descriptor	1024-d ECFP4 count fingerprint	200-d Gene2vec	1224-d combined chemical/genetic embedding
Anchor	batch-matched vehicle	batch-matched untreated well	none
Backbone	channel-factorized DiT	channel-factorized DiT	channel-factorized DiT
Held-out axis	8 compounds	none (held-out cells)	100 perturbations

E Extended ablations

E.1 FRESCO recovers morphological distribution for pathway-level perturbation phenotypes

Figure 6 shows the distributions behind Fig. 1 and Fig. 5. Each arm is centered on its own control median, so the quantity plotted is the shift relative to that arm's controls, and W_1 is the distance from that shifted distribution to the real knockout, in units of the median absolute deviation (MAD) of the real control cells. MorphDiff is 1.2–2.5$\times$ further from the real cells than FRESCO for the seen genes and 6.1–6.4$\times$ further for the held-out genes, and remains within the control interquartile range in all six panels. For *MDN1* and *NLE1* it shifts the distribution away from the real one, so the discrepancy is not only in magnitude. W_1 decreases with sample size and should be compared only within a gene. The six genes were chosen to illustrate two pathway phenotypes and each is scored on the feature with its largest real shift, so these panels are illustrative; aggregate conditioning scores are in Table 1.

Condition-level FID, reported for the held-out genes, does not reproduce this ordering. It separates the shuffled null from the other arms (64 and 41 vs. 35–55) but not FRESCO from MorphDiff, and for *GNL3L* MorphDiff has the lower FID (35 vs. 38) with a six-fold larger W_1. FID is computed over all statistics from all the generated images in a condition, most of which are shared across conditions, whereas the per-feature W_1 is restricted to the feature that responds to the knockout.

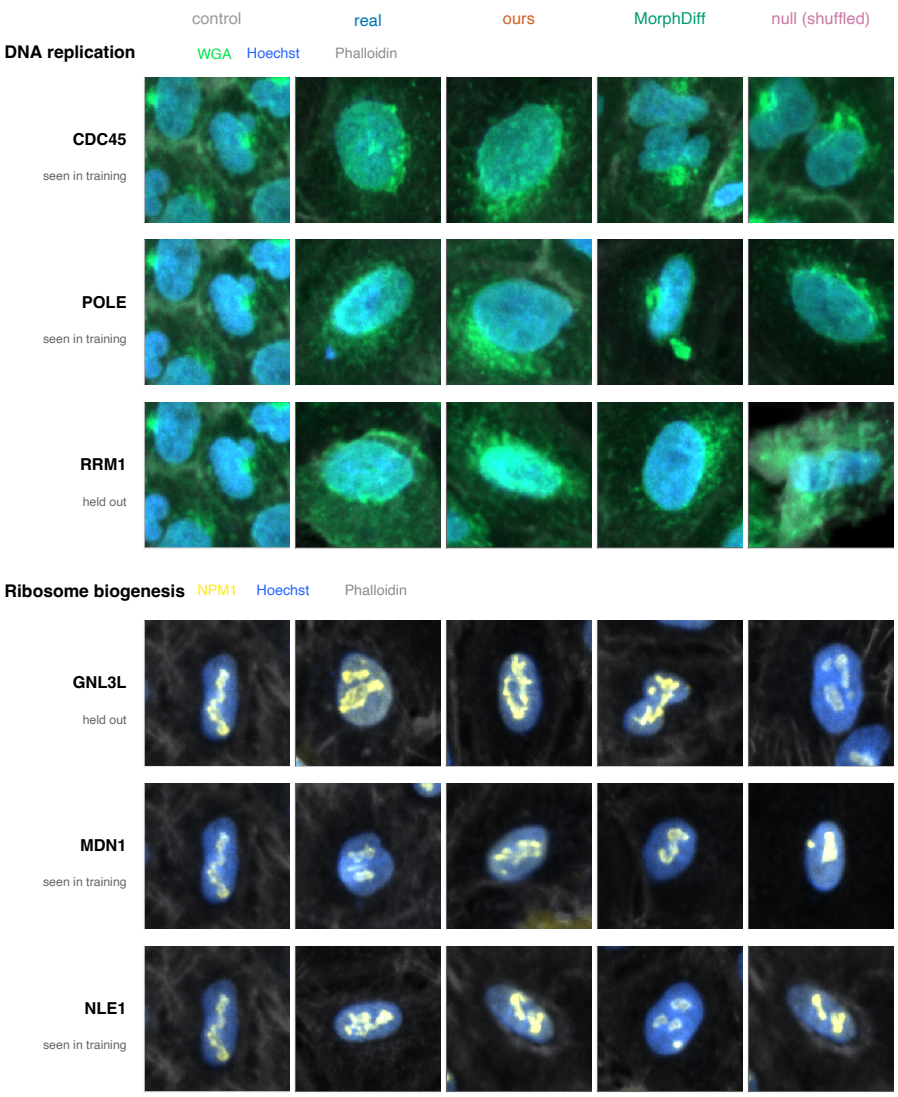

Figure 5: **FRESCO reproduces knockout-specific morphological responses.** Representative multi-channel crops for non-targeting control, the real knockout, FRESCO, MorphDiff, and a null model in which gene labels are shuffled. The displayed cell is the one nearest the 90th percentile of its own population's response in this feature pair, control included. For DNA replication, WGA, Hoechst and Phalloidin channels are overlaid and plotted for *CDC45*, *POLE* and *RRM1* (held-out). For ribosome biogenesis, NPM1, Hoechst and Phalloidin are overlaid and plotted for *GNL3L*, *MDN1* and *NLE1* (held-out).

E.2 Control anchoring: all metrics

Figure 3c reports FID alone. Figure 7 gives the same two arms across all four evaluation metrics, and the split holds throughout: on scGeneScope, where the split is by acquisition round, removing the anchor costs on every metric, while on OPS Explorer, where acquisition context is nearly fixed, the

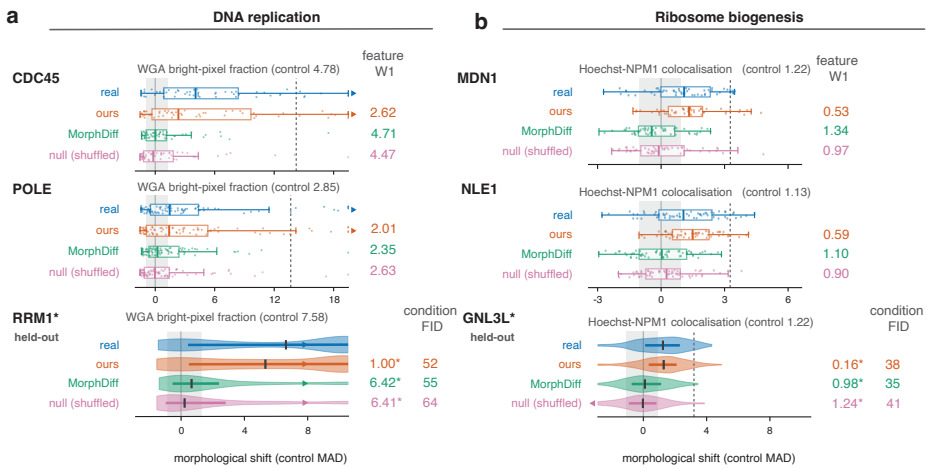

Figure 6: **Per-cell response distributions for the six pathway genes.** Same genes and pathways as Fig. 1. For each gene, the feature with the largest real shift, centered on each arm's own control median and scaled by the real control MAD; the value in brackets is the raw control median. Grey band, control interquartile range; grey line, control median; dashed line, the real 90th percentile. Seen genes are box plots with every cell overlaid ($n = 51$–81), held-out genes (asterisk) violins at full n (183–392). Numbers are W_1 in control MAD to the real cells of the same knockout. Condition-level FID is given for the held-out genes. **(a)** DNA replication, WGA bright-pixel fraction. **(b)** Ribosome biogenesis, Hoechst–NPM1 colocalisation.

two arms sit within noise of each other. The one cell that runs the other way is the $\cos(e, d)$ on OPS held-out, $+0.700$ without the anchor against $+0.680$ with it, a gap inside the seed-to-seed spread of ≈ 0.03 (Appendix G). The component therefore earns its place where batch variation is large enough to compete with the perturbation, and costs nothing where it is not.

E.3 CellFlux benchmark reproducibility

We reran evaluation for CellFlux with their published checkpoints and reproduced them within a small margin, except for the JUMP-cpg0000 dataset, where $|\Delta|$ is 8.45 for FID and 0.70 for KID (Table 8). The values reported in the unimodal morphology benchmark (Table 2) were taken directly from the publication, as these are the best values reported for CellFlux.

Table 8: **CellFlux reproduction.** Published values against values we recovered by running the released code and checkpoints of Zhang et al. (2025) through the same evaluation pipeline used for every number in this paper. FID_o and KID_o ($\times 100$); $|\Delta|$ is the absolute difference between reproduced and published values. IMPA (Palma et al., 2025) is quoted without reproduction, as only the stronger baseline was reproduced. Main-text tables quote published values throughout, so no comparison there depends on this reproduction.

	FID_o			KID_o		
Dataset	Pub.	Repro.	$\lvert\Delta\rvert$	Pub.	Repro.	$\lvert\Delta\rvert$
BBBC021	18.7	18.31	0.39	1.62	1.64	0.02
RxRx1	33.0	32.19	0.81	2.38	2.32	0.06
JUMP-cpg0000	9.0	17.45	8.45	0.63	1.33	0.70

E.4 Guidance scale and model size

Table 9 sweeps the guidance weight against the backbone size on the BBBC021 out-of-distribution split, the eight compounds withheld from training. Each cell generates all $19{,}704$ held-out cells,

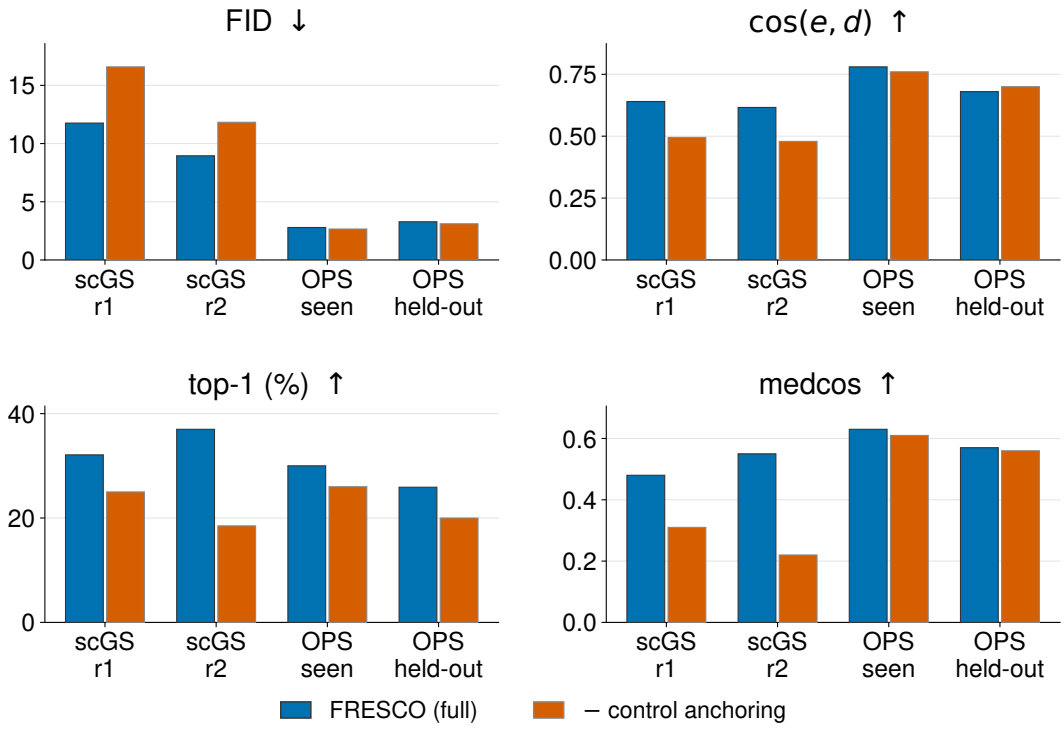

Figure 7: **Control anchoring, all metrics.** As Figure 3c, across the four evaluation metrics. The scGeneScope and OPS split is consistent on FID, top-1 and medcos; on OPS held-out, $\cos(e, d)$ is marginally higher without the anchor ($+0.700$ against $+0.680$).

anchored on the 46-batch vehicle pool, and scores them with the baseline's released mechanism-of-action classifier, which is trained on real perturbed cells and applied to generated ones. No real images enter the comparison, so the number asks only whether a generated cell carries the mechanism its compound induces. Accuracy rises monotonically with guidance at all three sizes, from 45.5–47.0% at $\omega = 1$ to 55.4–58.4% at $\omega = 3$, with macro-F1 following. Capacity contributes little by comparison: the S backbone, approximately parameter-matched to the strongest published baseline, is the best arm at every guidance weight above 1, so the gain is attributable to guidance rather than to size. Every arm exceeds the 43.2% baseline at every setting.

Table 9: Guidance scale and backbone size on BBBC021 out-of-distribution MoA (accuracy / macro-F1). Accuracy rises monotonically with CFG at every backbone size, while size itself contributes little; the baseline is 43.2%.

Backbone	CFG 1.0	CFG 1.5	CFG 2.0	CFG 3.0
S	45.5 / .347	52.9 / .419	55.5 / .433	**58.4 / .472**
B	46.5 / .344	51.0 / .374	54.5 / .408	55.4 / .422
L	47.0 / .360	52.5 / .405	55.2 / .433	56.9 / .460

E.5 Per-channel latent encoding

We generate in a pre-trained VAE latent rather than in pixels, so the encoder must retain the fine structure that carries the morphological signal in microscopy.

We compare two widely used VAEs on RxRx1 without fine-tuning either: the SD-VAE, which assigns four latent channels per microscopy channel (24 × 12 × 12, sixteen-fold compression), and the Flux VAE, which assigns sixteen (96 × 12 × 12, fourfold). The Flux VAE reconstructs more faithfully on all six channels, 0.913 to 0.958 SSIM against 0.652 to 0.848, and the gap is carried by the finer structures: the SD-VAE drops to 0.652 to 0.707 SSIM on ConA, Phalloidin, Syto14 and WGA while

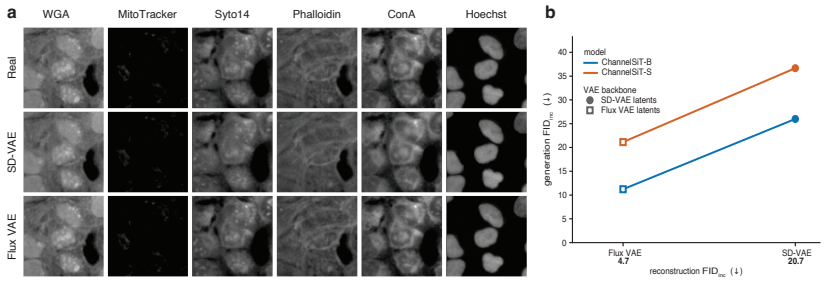

Figure 8: **Zero-shot reconstruction of pre-trained VAE latents for microscopy.** (a) A single field of cells across the six fluorescent channels (columns), comparing the measured image with reconstructions from the SD-VAE and the Flux VAE (rows). (b) Generation $\mathrm{FID}_{\mathrm{inc}}$ of FRESCO trained in each latent space, plotted against the reconstruction $\mathrm{FID}_{\mathrm{rec}}$ of the VAE providing that space. Marker shape indicates the VAE backbone and colour the model scale. Lower is better on both axes.

recovering Hoechst and MitoTracker at 0.848 and 0.818 (Fig. 8a). Reconstruction $\mathrm{FID}_{\mathrm{rec}}$ follows the same trend, 4.7 against 20.7. Because reconstruction sets a ceiling but does not determine generation quality, we train FRESCO in each space: the ordering persists, 21.1 gen. $\mathrm{FID}_{\mathrm{inc}}$ at S scale and 11.2 at B scale on Flux latents against 36.7 and 26.0 on SD-VAE latents (Fig. 8b). Compression and fidelity do not trade off here, since the four times larger Flux latent is better at both model sizes. We therefore use the Flux VAE on RxRx1, and the SD-VAE on scGeneScope, where 224 px images narrow the reconstruction gap to 0.943 against 0.894 SSIM, and we accept that residual loss in exchange for a four times smaller latent.

E.6 Aligning channels with the Channel-DiT block

To measure what the cross-channel attention contributes, we vary the number of Channel-DiT blocks interleaved among twelve spatial blocks from zero to six. Omitting cross-channel attention cost the most here, raising $\mathrm{FID}_{\mathrm{o}}$ from 25.5 to 139.0 and $\mathrm{KID}_{\mathrm{o}}$ roughly sixfold, but a single block recovers almost all of it and further blocks change little: $\mathrm{FID}_{\mathrm{o}}$ spans 25.2 to 27.8 across every allocation containing at least one (Figure 9). Therefore what matters is the presence of cross-channel blocks, not their number. We use one Channel-DiT block per two spatial blocks throughout; the ablation shows this choice has little effect.

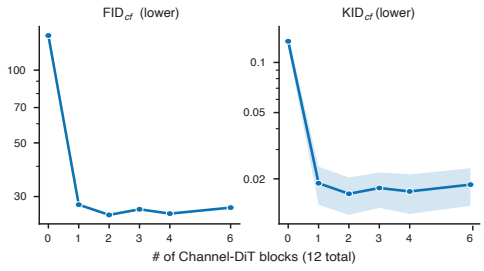

Figure 9: **Channel-DiT to spatial blocks ratio.** Channel-DiT blocks traded against spatial blocks under a fixed budget of twelve alternating blocks (S-scale model). Shading is ± 1 s.d. of $\mathrm{KID}_{\mathrm{o}}$.

E.7 Control

Anchoring: qualitative examples

Section 2.4 shows that control anchoring recovers the shift between imaging rounds. Figure 10 shows representative crops: with control anchoring, generated cells are visually closer to real cells of the same round than without it.

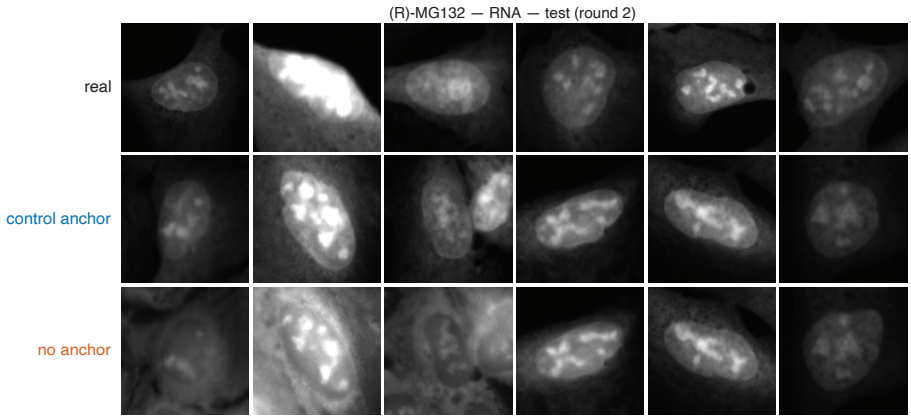

Figure 10: **Generated crops with and without control anchoring, scGeneScope.** Representative crops ordered by control anchor conditions (rows). Images are unpaired; no correspondence is implied

F METRICS

F.1 IMAGE-DISTRIBUTION METRICS

FID and KID compare generated and real images through the 2048-d pool features of an Inception-V3 network trained on natural images, so they measure realism rather than perturbation specificity. Three variants appear in this paper.

- $\mathrm{FID}_{\mathrm{inc}}$, used on the multimodal screens and for internal ablations. A single-channel image is replicated to three channels, and a crop with more than three channels contributes its first three, so on the five- and seven-channel panels this metric sees three channels and is insensitive to failures confined to the others. It is computed over all perturbations pooled; condition-level FID, used in Appendix E.1, applies it to the cells of a single perturbation.
- $\mathrm{FID}_{\mathrm{o}}$ and $\mathrm{KID}_{\mathrm{o}}$, following the notation of Zhang et al. (2025): a verbatim port of their evaluation, computed over all images pooled, and the only numbers comparable with their published values. Crops are mapped to RGB by that work's dataset-specific composite (a weighted colour blend for the six RxRx1 channels, the first three raw channels for JUMP-cpg0000; BBBC021 is already three-channel), quantised to 8 bits, and KID uses subsets of 100.
- $\mathrm{FID}_{\mathrm{rec}}$, the reconstruction FID of Appendix E.5: each channel is replicated to three as a greyscale image, scored separately, and the per-channel values are averaged. It is not comparable with $\mathrm{FID}_{\mathrm{inc}}$ on the same images.

F.2 MORPHOLOGY FEATURES

Every real or generated cell is reduced to a fixed set of scalars, computed identically for both. Sixteen per fluorescence channel gives $16C$:

- **Intensity** (8): mean, standard deviation, median, 10th and 90th percentiles, integrated intensity, median absolute deviation, and the fraction of pixels above mid-range.
- **Texture** (5): contrast, correlation, homogeneity, energy and angular second moment of a grey-level co-occurrence matrix, built at distance 1 on the channel quantized to 16 levels and averaged over the four cardinal and diagonal orientations.
- **Edge** (1): mean Sobel gradient magnitude.
- **Granularity** (2): mean absolute difference between Gaussian blurs of the channel at $\sigma = 1$ and 2, and at $\sigma = 2$ and 4, responding to fine and coarse texture scales respectively.

Five **shape** features are computed once per cell from the Otsu-thresholded sum of all channels, taking the largest connected component: foreground fraction, area, eccentricity, solidity and extent. The remaining $\binom{C}{2}$ are **colocalization** features, the Pearson correlation between the raw pixel vectors of each channel pair (10 at $C = 5$, 21 at $C = 7$); these capture cross-channel spatial relationships that no single-channel statistic sees. The feature count is therefore panel-dependent: $16C + 5 + \binom{C}{2}$, giving 95 at $C = 5$ and 138 on the seven-channel OPS panel.

Features are standardized per dimension against the pooled real cells of the evaluation split, and the same transform is applied to generated cells, so that no feature dominates a distance through its raw units.

G ESTIMATING THE CONDITIONING METRICS

Definitions. Let R_u and G_u be the real and generated cells for perturbation $u = 1, \dots, U$ in the feature space of Appendix F.2, and μ_u^R, μ_u^G their kernel mean embeddings. Each perturbation's response is its deviation from the across-perturbation mean,

$$e_u = \mu_u^R - \tfrac{1}{U}\textstyle\sum_v \mu_v^R, \qquad d_u = \mu_u^G - \tfrac{1}{U}\textstyle\sum_v \mu_v^G. \tag{10}$$

Writing $\mathbb{E}$ for the average over perturbations, the magnitude and direction of the generated responses are

$$\text{gen/real} = \frac{\mathbb{E}\|d\|^2}{\mathbb{E}\|e\|^2}, \qquad \cos(e,d) = \frac{\mathbb{E}\langle e, d\rangle}{\sqrt{\mathbb{E}\|e\|^2\ \mathbb{E}\|d\|^2}}. \tag{11}$$

Classifier-free guidance acts mainly on gen/real, which we use only to fit ω^* (see **Guidance Scale** below); $\cos(e,d)$ is broadly invariant to ω. Because $\cos(e,d)$ is pooled and energy-weighted, it can be carried by a few strong responders, so we also report two per-perturbation scores. These use plain mean feature profiles rather than kernel embeddings: $\bar{e}_u$ and $\bar{d}_u$ are the mean real and generated feature vectors of perturbation u, minus their means over perturbations, and the real profile is estimated from many more cells than are generated, so that reference noise does not attenuate the score. On scGeneScope the reference is the evaluation split itself, about 4,300 cells per compound; on OPS Explorer, whose 1,001 perturbations leave far fewer cells each, it pools every crop of the perturbation across splits, a median of about 500. medcos is the median over u of $\cos(\bar{e}_u, \bar{d}_u)$, and top-1 the fraction of perturbations whose $\bar{d}_u$ is closer by cosine to $\bar{e}_u$ than to any other $\bar{e}_v$. For held-out perturbations, none of whose cells were seen in training, this reference leaks nothing into the model.

Estimating e_u and d_u from sample means and plugging them into equation 11 would inflate every squared norm by the within-perturbation variance divided by the number of cells. What equation 11 needs from the responses (squared norms and inner products) is instead recoverable from pairwise squared distances between mean embeddings, and the unbiased MMD^2 estimator (Gretton et al., 2012) estimates exactly those. This appendix gives that reduction and the properties that follow from it.

Estimator. Cells enter as the standardised morphology features of Appendix F.2. The kernel is a sum of RBF kernels over bandwidth multipliers $\{0.25, 0.5, 1, 2, 4\}\sigma$, with σ the median over perturbations of each perturbation's median pairwise distance, computed on real cells only: fitting σ to generated cells would let a model flatter itself by collapsing onto the kernel's most sensitive scale. We use the unbiased estimator, whose diagonal is excluded, so $\mathbb{E}[\widehat{\mathrm{MMD}}^2(P,Q)] = \|\mu_P - \mu_Q\|^2$ exactly. A perturbation enters $\cos(e,d)$ and gen/real if it has at least 8 real and 8 generated cells, and enters medcos and top-1 if it also has at least 50 reference cells; on the OPS Explorer held-out genes this retains all 56 for the former and 54 for the latter.

Four averages. Collect the pairwise statistics into $U \times U$ tables

$$D_{uv}^{RR} = \widehat{\mathrm{MMD}}^2(R_u, R_v), \qquad D_{uv}^{GG} = \widehat{\mathrm{MMD}}^2(G_u, G_v), \qquad D_{uv}^{RG} = \widehat{\mathrm{MMD}}^2(R_u, G_v), \tag{12}$$

and write $\langle T\rangle_{\neq} = \frac{1}{U(U-1)}\sum_{u\neq v} T_{uv}$ for the off-diagonal mean of a table and $\langle T\rangle_{=} = \frac{1}{U}\sum_u T_{uu}$ for its diagonal mean. The four averages we need are

$$\Delta_{RR} = \langle D^{RR}\rangle_{\neq}, \qquad \Delta_{GG} = \langle D^{GG}\rangle_{\neq}, \qquad \Delta^{\neq} = \langle D^{RG}\rangle_{\neq}, \qquad \Delta^{=} = \langle D^{RG}\rangle_{=}. \tag{13}$$

The first two measure how far apart real, and generated, perturbations are from each other. The last two are the same cross-table read off its diagonal and off it, so their difference isolates the alignment between the two sets of responses.

Reduction. Write $\mu_u^R = \bar{\mu}_R + e_u$ and $\mu_u^G = \bar{\mu}_R + b + d_u$, where $b = \bar{\mu}_G - \bar{\mu}_R$ is a global offset; by construction $\sum_u e_u = \sum_u d_u = 0$. Expanding $\|e_u - e_v\|^2$ and using $\sum_{u \neq v} \langle e_u, e_v \rangle = -U\,\mathbb{E}\|e\|^2$,

$$\Delta_{RR} = \frac{2U}{U-1}\,\mathbb{E}\|e\|^2, \qquad \Delta_{GG} = \frac{2U}{U-1}\,\mathbb{E}\|d\|^2, \tag{14}$$

with b cancelling from Δ_{GG} as it is common to every μ_u^G. For the cross terms, $\mu_u^R - \mu_v^G = e_u - b - d_v$ and $\frac{1}{U-1}\sum_{v \neq u} d_v = -d_u/(U-1)$ give

$$\Delta^{\neq} - \Delta^{=} = \frac{U}{U-1}\,\mathbb{E}\Big[\,2\langle e_u - b, d_u\rangle + \mathbb{E}\|d\|^2 - \|d_u\|^2\Big] = \frac{2U}{U-1}\,\mathbb{E}\langle e, d\rangle, \tag{15}$$

the two nuisance terms vanishing exactly under the average, since $\sum_u d_u = 0$ makes $\mathbb{E}\langle b, d\rangle = 0$ and $\mathbb{E}\|d\|^2$ is by definition the mean of $\|d_u\|^2$. Every factor $2U/(U-1)$ then cancels in the ratios, giving equation 11:

$$\frac{\text{gen}}{\text{real}} = \frac{\Delta_{GG}}{\Delta_{RR}} = \frac{\mathbb{E}\|d\|^2}{\mathbb{E}\|e\|^2} = r^2, \qquad \cos(e,d) = \frac{\Delta^{\neq} - \Delta^{=}}{\sqrt{\Delta_{GG}\,\Delta_{RR}}}. \tag{16}$$

where r is the ratio of generated to real response magnitude. These are identities in the mean embeddings rather than approximations; the estimator is unbiased, so they hold in expectation with finite-sample variance around them. The related quantity $(\Delta^{\neq} - \Delta^{=})/\Delta_{RR} = r\cos(e,d)$ is the slope of a regression of generated on real responses, and $R^2 = 2r\cos(e,d) - r^2$ the corresponding pooled coefficient of determination.

Invariances. The offset b appears nowhere in equation 16: a model whose cells are uniformly displaced from real ones is not penalised by either metric, that being the business of FID. Likewise any response component shared across perturbations – a generic "stressed cell" direction, say – lies in $\bar{\mu}_R$ and is removed by equation 10, so a model is credited only for what distinguishes one perturbation from another.

Uncertainty. The nuisance terms in equation 15 cancel only after averaging, which makes the numerator of $\cos(e,d)$ the noisier ingredient; single perturbations can contribute negatively while the mean is strongly positive (92% of perturbations contribute positively on the OPS in-distribution case). We therefore check values against a permutation null over generated labels and a bootstrap over perturbations. For example, on OPS in distribution the null (500 shuffles) has mean -2.3×10^{-5} and s.d. 5.6×10^{-4}, against an observed $\Delta^{\neq} - \Delta^{=}$ of 2.68×10^{-2}, whose 95% bootstrap interval is $[2.02, 3.48] \times 10^{-2}$. Across seeds $\cos(e,d)$ moves by ≈ 0.03.

Guidance scale. Guidance acts almost entirely on magnitude, so we fit $\text{gen}/\text{real} = a\omega^p$ by least squares in log space through evaluations on the training split at $\omega = 1$ and $\omega = 2$, and set $\omega^* = a^{-1/p}$. The exponent is close to 2 on OPS and flatter on Cell Painting ($p \approx 0.9$–1.1), which is why it is fitted rather than assumed. ω^* is always fitted on the training split and applied unchanged to every evaluation split, in and out of distribution, so no evaluation is scored at a guidance weight tuned on its own cells or perturbations.

H LIMITATION

FRESCO conditions on an expression estimate pooled across treated cells and therefore does not recover the morphology of an individual sequenced cell. Applying it to a new perturbation also requires transcriptomic measurements and suitable control images from the target acquisition context. Moreover, the response metrics operate in a predefined morphology-feature space, which may omit relevant phenotypes, and FID_{inc} sees only three channels of the five- and seven-channel panels. Finally, models are trained separately for each fluorescence panel, and generalization across cell types, doses, time points, laboratories, and unseen panels remains to be established.