FROM SNAPSHOTS TO TRAJECTORIES: BENCHMARKING SENESCENCE-CONDITIONED CELL MORPHOLOGY GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cellular senescence, a state in which cells stop dividing but remain in tissues, drives tissue ageing and is a central readout for the development of new treatments for age-associated diseases. Yet no existing cell morphology benchmark captures this phenotype. We introduce **SENOMORPH**, a *senescence-conditioned* cell morphology benchmark that pairs each single-cell image with a *continuous* senescence score derived from its own transcriptome. SENOMORPH comprises **81,255** real cross-sectional snapshot cells across three tissues and nine cell types. To our knowledge, SENOMORPH is the *first* public benchmark dedicated to this problem. To understand how existing generators perform on SENOMORPH, we assess 9 representative baselines spanning 7 Flow Matching (FM) variants and 2 non-FM generators. Our evaluation shows that these generators, despite attaining competitive distributional scores, reverse the expected morphology trend on individual cell types. We further present **SENOFLOW**, an FM-based generator that enforces two geometric properties under continuous conditions, *batch-invariant pairing* of source-target supervision and *score-Lipschitz, feature-aligned dynamics* at integration. SENOFLOW consistently outperforms all baselines, lowering FID from 14.2 to **11.6** and raising mean $|\rho|$ from 30.7% to **87.7**%. It also follows real trajectories and preserves cell identity on tracked live-cell data. We additionally release a SENOFLOW-generated synthetic companion set of **812,550** score-conditioned cells that lift snapshots into trajectories to support data-scarce downstream tasks like senescent-cell classification with **+3.9** pp ACC and **+1.9** pp AUC gains.

1 INTRODUCTION

Cellular senescence [40, 37] is a stable, non-dividing cell state that accumulates with age and reshapes tissue function [62]. Its morphological signatures [42, 26] have become a central readout in ageing research [11], cancer immunotherapy [5], and senolytic drug development [52, 81]. Yet current understanding remains largely confined to fibroblast *snapshots* [42, 48], and no public benchmark exists for senescence-conditioned cell morphology *trajectories* across cell types and tissues. Existing cell morphology resources focus on discrete drug [39, 9, 10] or gene perturbations [72, 32, 71] at single time points, not on continuous endogenous ageing.

The reason is intrinsic. Senescence is commonly measured with destructive assays such as SA-β-galactosidase staining [23], and even modern in-situ platforms (Xenium [46], MERFISH [15], CosMx [41]) remain cross-sectional, yielding only a single snapshot per cell. Aligned image-transcriptome pairs are abundant, yet no individual cell can be tracked along the senescence axis.

Generative modelling is a natural computational substitute, approximating score-conditioned morphology directly from cross-sectional snapshots. Flow Matching [59, 60] is especially suited to this goal, since it models continuous transformations and generates coherent intermediate states across continuous biological scores. Yet two obstacles prevent its naive application here. ***First***, the standard *minibatch OT* pairing strategy becomes unstable under continuous conditioning, because any given batch sparsely covers the joint space of cell type and senescence score. ***Second***, the standard *single-scale classifier-free guidance* cannot regulate the score-to-morphology mapping at every intermediate score, only at the endpoints.

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

We address these gaps on two fronts. **SENOMORPH** (Fig. 1) is a multi-tissue benchmark of paired DAPI images and transcriptome-derived senescence scores curated from multiple Xenium samples (§2). To extend its coverage along the senescence axis, we further develop **SENOFLOW**, a Flow Matching pipeline that resolves the two obstacles above through *Global Prototype Coupling*, which guarantees batch-invariant pairing via a global prototype memory indexed by cell type and score bin, and *Continuous-score Velocity Correction*, which composes continuity interpolation and bio-direction alignment via weighted gradient on the integrated velocity. Empirically, SENOFLOW improves FID from 16.3 to **11.6** over the OT-CFM baseline, attains the strongest mean $|\rho|$ on every tissue, receives the highest expert ratings on realism and senescence fidelity, and improves downstream minority-cell classification by **+3.9** pp ACC and **+1.9** pp AUC.

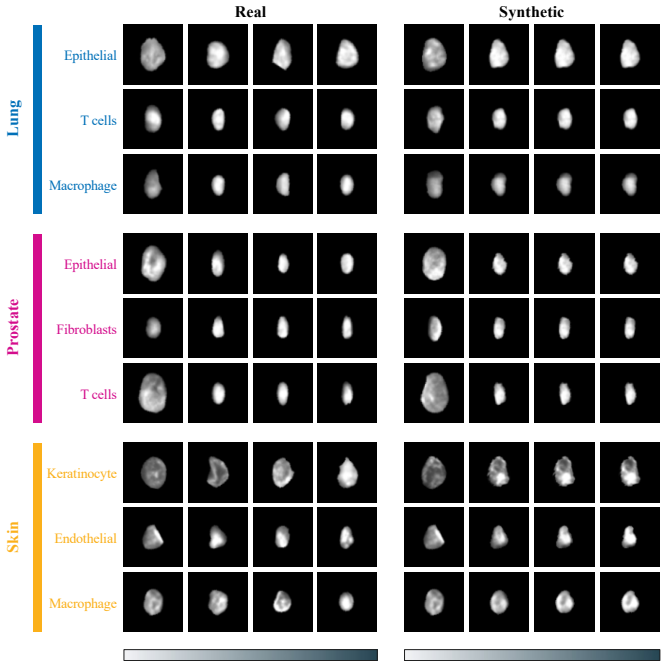

Figure 1: **Cross-sectional snapshots vs. score-conditioned trajectories.** *Left*: real DAPI crops from *distinct* cells, ordered by their transcriptome-derived senescence scores. *Right*: SENOFLOW generates morphology with target scores. Senescence scores span from *LOW* *HIGH*.

Concretely, this work makes three contributions:

- ***Senescence-conditioned dataset*** (§2). We curate the *first* public benchmark of **81,255** real single-cell DAPI crops across nine cell types and three tissues (Fig. 2), each paired with a transcriptome-derived continuous senescence score. Alongside we release an **812,550**-cell SENOFLOW-generated synthetic companion set, providing 10-snapshot trajectory per cell, that supports data-scarce downstream tasks such as senescent-cell classification.
- ***Senescence-conditioned Flow Matching framework*** (§3). We present a Flow Matching framework with two property-driven operators, *Global Prototype Coupling* for batch-invariant pairing and *Continuous-score Velocity Correction* for score-Lipschitz, feature-aligned dynamics, delivering strong gains in both image quality and biological fidelity.
- ***Benchmarking infrastructure for the community*** (§4). We establish a cell-type-aware evaluation protocol, nine reference baselines spanning seven Flow Matching variants (FM) and two non-FM generators, and a public leaderboard. Beyond reporting scores, the benchmark exposes a systematic decoupling between image realism and biological fidelity under continuous conditioning, providing a common testbed for cumulative progress on this problem.

2 SENOMORPH BENCHMARK

SENOMORPH (Fig. 2) is a real cross-sectional snapshot set of `81,255` single-cell DAPI crops from three Xenium tissue sections, each paired with a transcriptome-derived continuous senescence score. Alongside, we release a companion set of `812,550` SENOFLOW-generated synthetic trajectory cells for data augmentation in downstream tasks (§4.1).

2.1 DESIGN PRINCIPLES

We design SENOMORPH around two intrinsic requirements of senescence-conditioned cell morphology generation, *i.e.*, continuous biological annotation and cross-tissue coverage, and additionally release a synthetic companion set to address senescent-cell scarcity in downstream tasks.

- *Continuous biological annotation.* Existing cell morphology datasets usually assign each cell a discrete perturbation label, such as a drug or a knocked-out gene, which reduces generation to translation between a small number of categories. In contrast, each real cell in SENOMORPH

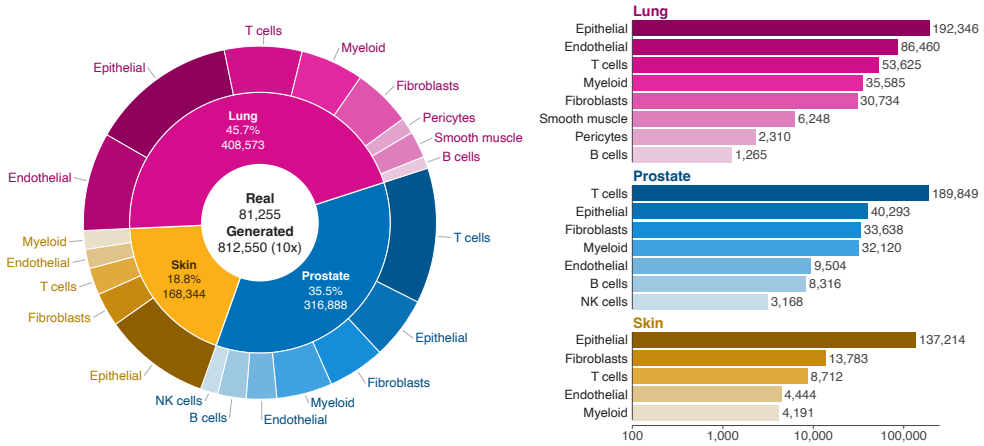

Figure 2: **Dataset composition.** (a) Nested donut showing tissue shares (inner) and per-tissue cell-type breakdown (outer). (b) Per-cell-type counts by tissue on a log scale.

carries a continuous senescence score in $[0, 1]$ computed from its own transcriptome. This better matches the biology, since senescence progresses gradually rather than as an abrupt switch, and it defines a harder task: models must produce smooth morphological change across the full score range, not merely interpolate between two endpoints.

- *Cross-tissue coverage.* A method that performs well on one tissue may rely on tissue-specific shortcuts and fail to transfer. We therefore include three distinct tissues (*i.e.*, lung, prostate, and skin), whose dominant cell-type compositions differ substantially: lung is enriched for epithelial and immune cells, prostate for glandular and stromal cells, and skin for keratinocytes and fibroblasts. Evaluation across all three tissues therefore probes generalization across genuinely different morphologies rather than variants of the same cellular context.
- *Synthetic companion set for data-scarce downstream tasks.* Senescent cells are chronically scarce in ageing research [42, 26]. SENOMORPH alleviates this by releasing the SENOFLOW-generated synthetic companion set: for each real source cell, we provide images at **10** evenly spaced target senescence scores, augmenting cell samples in downstream classification (§4.1).

2.2 DATA ACQUISITION

Acquisition platform. We build the real set of SENOMORPH from three 10x Genomics Xenium tissue sections under one requirement: each cell must provide both a DAPI crop and a senescence score derived from the same physical cell. Xenium is a natural choice because it provides cell-segmented transcript counts aligned to morphology images at subcellular resolution. By contrast, spot-based assays such as Visium v2 [74] measure expression over capture spots rather than individual cells, while high-resolution platforms such as Stereo-seq [14] still require downstream spot-to-cell assignment and segmentation to obtain cell-level profiles. Xenium therefore provides the most direct route to aligned single-cell DAPI crops and transcriptomes for benchmark construction.

Single-cell extraction. We convert the raw Xenium outputs into standardized single-cell DAPI crops that preserve senescence-relevant morphology. For each segmented nucleus, we compute its centroid and bounding box, then crop a square context window centered on the nucleus, with side length scaled by the section-wise mean nucleus size so that cells of different sizes retain comparable local context. Standard quality-control filters remove segmentation artifacts. Crops are extracted at native Xenium resolution, rotated to align the principal nuclear axis vertically, and downsampled to 64×64 pixels for release. We retain only the DAPI channel, since nuclear morphology is the primary visual readout of senescence in our setting, yielding the released set of 64×64 DAPI crops.

2.3 ANNOTATION

Each real cell receives two transcriptome-derived labels, avoiding any additional image-based supervision. Specifically, a discrete cell-type label is assigned by CellTypist [25], which maps single-cell transcriptomes to standardized cell-type ontologies, and a continuous senescence score is computed by DeepScence [70], which predicts cellular senescence from gene expression. Raw DeepScence outputs are min-max normalized to $[0, 1]$ within each tissue so that the full score range is used despite

tissue-specific baseline differences. The companion set inherits tissue and cell-type labels from its source cell and uses the requested target senescence score as its conditioning variable. We validate DeepScence against expert ratings in §4.1 and discuss its rationale in Appendix B. More annotation statistics are in Appendix A.

2.4 Attributes and Monotonic Morphological Trends

To support detailed analysis, each benchmark instance in SenoMorph carries five attributes (see Fig. 3): *tissue source*, *cell type*, *senescence score*, *nuclear area* (NA defined as the pixel count of the nucleus mask), and *mean DAPI intensity* (mean intensity within the mask). Cell type and senescence score are transcriptome-derived for real instances and inherited from the source cell (together with the requested target score) for companion-set ones; NA and DAPI are computed directly from the image and serve as the morphological readouts for biological fidelity.

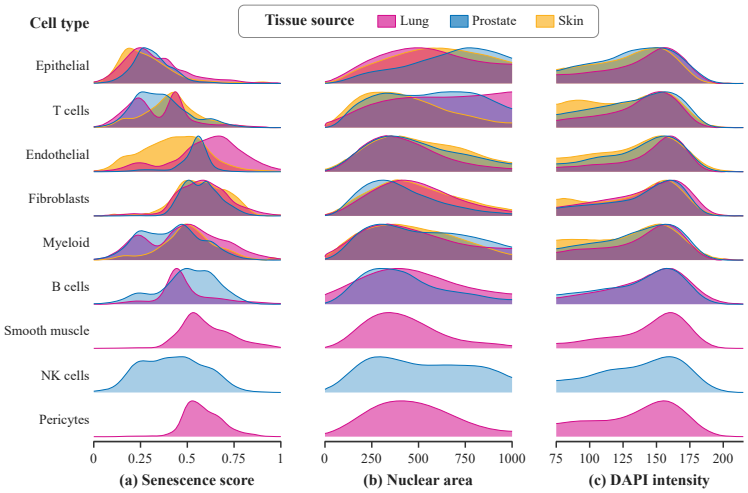

Figure 3: **Per-instance attributes of SenoMorph.** Distribution and composition of the attributes across all benchmark instances.

Monotonic morphological trends. Plotting NA and DAPI against senescence score with LOESS smoothing [20, 19] reveals a systematic aggregate regularity (Fig. 4). *NA decreases and DAPI increases monotonically with senescence*, a pattern that also holds at the per-cell-type level (Appendix A). This empirical structure motivates our method design, with the resulting per-cell-type direction signs $d_i \in \{+1, -1\}$ used by the operators in §3.3 to enforce monotone progression.

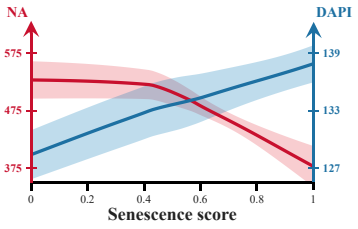

Figure 4: **NA and DAPI vs senescence.** NA ↓, DAPI ↑ with score.

2.5 Dataset Split and Evaluation Protocol

Task. Given a source cell crop, its cell-type label, and a target senescence score, the goal is to generate the corresponding morphology at the requested biological state.

Dataset split. We first split the real set within each tissue into train, validation, and test sets with a $70/10/20$ ratio, stratified by cell type. This preserves the tissue-specific cell-type distribution while ensuring that every cell type appears in all three splits. The companion set inherits the same split from its source cells. The resulting splits $\text{SenoMorph}_{\text{Tra}}$, $\text{SenoMorph}_{\text{Val}}$, and $\text{SenoMorph}_{\text{Tst}}$ are used for training, hyperparameter selection and early stopping, and final evaluation, respectively. Detailed split lists will be released with the dataset.

Evaluation protocol. *All baseline evaluations are conducted on the real test split; the* SenoFlow*-generated synthetic companion set is not used as ground truth for any metric in §4.* At test time, a generator receives a source cell crop, cell-type label, and target senescence score, and produces an image at the requested state. For each source cell in the real test split, we evaluate $K = 10$ uniformly spaced target scores over $[0, 1]$ along the senescence axis. Metrics are detailed in §4.

3 SenoFlow Model

We present SenoFlow (Fig. 5), a Flow Matching framework for senescence-conditioned cell generation. We formalize the task as constructing a velocity field $v_\theta : \mathcal{X} \times [0,1] \times \mathcal{C} \times \mathcal{S} \to \mathbb{R}^d$ that satisfies two geometric properties, (P1) *batch-invariant pairing* that decouples the supervisory signal from minibatch composition, and (P2) *Lipschitz-continuous, feature-aligned dynamics* that respect the biological geometry along the senescence axis. Standard CFM (§3.1) violates both. §3.2

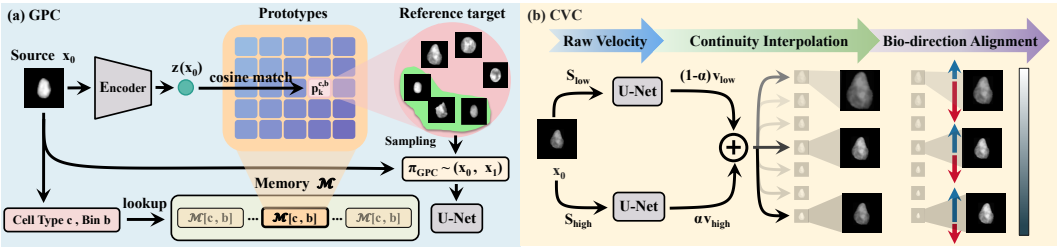

Figure 5: **Overview of SENOFLOW.** (a) **GPC** (§3.2, P1): source x_0 is matched to a prototype $p_k^{c,b}$ in memory $\mathcal{M}[c, b]$ to form batch-invariant pairs π_{GPC}. (b) **CVC** (§3.3, P2): continuity interpolation followed by bio-direction alignment via → NA and → DAPI weighted gradients.

introduces *Global Prototype Coupling* (GPC), the operator on π that satisfies (P1); §3.3 introduces *Continuous-score Velocity Correction* (CVC), the operator on $\mathcal{T}$ that satisfies (P2). GPC and CVC lift CFM into a property-constrained generator, training-free at inference.

3.1 PRELIMINARY

Task definition. Given a source cell image x_0, its cell type c, and a target senescence score s, the goal is to generate $\hat{x}$ that reflects the morphology at s while preserving cell identity. Sweeping s across $[0, 1]$ produces a score-conditioned morphology sequence. All operations below are performed in the latent space of a VAE pretrained on the training split.

Conditional Flow Matching. Conditional Flow Matching [59, 60] learns a velocity field $v_\theta(x_t, t, c, s)$, parameterised over flow time $t \in [0, 1]$ and conditioned on cell type c and target senescence score s, that transports source x_0 to a paired target x_1 along the linear path $x_t = (1-t)x_0 + tx_1$. The pipeline is fully specified by a coupling π that forms training pairs and a velocity transformation $\mathcal{T}$ that post-processes inference-time predictions to produce the generated output $\hat{x}$,

$$\mathcal{L}(\theta; \pi) \;=\; \mathbb{E}_{(x_0,x_1)\sim\pi(\cdot|c,s),\, t\sim\mathcal{U}[0,1]} \big\| \, v_\theta(x_t, t, c, s) - (x_1 - x_0) \, \big\|^2, \quad \hat{x} \;=\; x_0 + \int_0^1 \mathcal{T}(v_\theta)\, dt. \tag{1}$$

Standard CFM [77] sets π to batch-local Sinkhorn OT [21] and $\mathcal{T}$ to the identity. Under continuous scores, batch-local OT is high-variance because any minibatch sparsely covers the joint (c, s) space, and identity transport offers no mechanism to regulate trajectory geometry at intermediate scores.

3.2 GLOBAL PROTOTYPE COUPLING

Obstruction. Batch-local Sinkhorn OT forms source-target pairs within each minibatch, so the conditional pair distribution $\pi(\cdot \mid c, s)$ depends on which other cells happened to be sampled together rather than on the underlying data alone. (P1) requires the opposite, that $\pi(\cdot \mid c, s)$ be a function of the global training distribution, invariant to minibatch composition.

Construction. GPC routes each source cell through a global prototype memory $\mathcal{M}$ indexed by cell type c and quantile-based score bin $b = \text{bin}(s)$, reducing pair formation to a deterministic lookup (Fig. 5(a)). Let $\boldsymbol{z}(x_0)$ denote the VAE latent of x_0, and let $\mathcal{M}[c, b] = \{\boldsymbol{p}_k^{c,b}\}_{k=1}^K$ store K k-means prototypes for the group of cell type c and score bin b. We define:

$$\pi_{\text{GPC}}(x_0, x_1 \mid c, s) \;\propto\; \mathbb{1}\big[\, x_1 \in \mathcal{C}^{k^\star}_{c,\,\text{bin}(s)} \,\big], \qquad k^\star \;=\; \arg\max_k \; \cos\big(\boldsymbol{z}(x_0),\, \boldsymbol{p}_k^{c,\,\text{bin}(s)}\big), \tag{2}$$

where $\mathcal{C}^k_{c,b}$ is the target cluster attached to prototype k, and x_1 is sampled uniformly from this cluster at each training step. Conditioned on (x_0, c, s), the pair distribution depends only on $\mathcal{M}$ and is therefore invariant to minibatch composition, satisfying (P1). $\mathcal{M}$ is built once before training by partitioning training latents into (c, b) groups and reducing each group to K prototypes via k-means; Eq. 2 reduces to an $O(K)$ lookup per sample at training time.

3.3 CONTINUOUS-SCORE VELOCITY CORRECTION

Obstruction. Let $\{m_i\}_{i=1}^M$ be scalar morphological observables (NA and DAPI) differentiable in x, with cell-type-specific directions $d_i \in \{+1, -1\}$ derived from the data and signed biomarker gradients $g_i(x) = d_i \nabla_x m_i(x)$. The per-cell-type monotonicity established in §2.4 sharpens (P2), requiring the integrated velocity be *Lipschitz-continuous in* s and *biologically aligned*, $g_i(x)^\top v \geq 0$ for every i. The raw v_θ violates both. *(i) Score-axis discontinuity*, where $v_\theta(\cdot, s)$ is trained on discrete

bins and unstable at intermediate s. *(ii) Dominant-gradient drag*, where one observable's gradient overshoots and forces a co-varying observable to evolve against its expected direction.

Construction. CVC composes two operators in the order *continuity* then *alignment* (Fig. 5(b)),

$$\mathcal{T}(v_\theta) \;=\; \underbrace{\mathcal{P}_{\text{bio}}}_{\text{alignment}} \circ \underbrace{\mathcal{I}_{\text{smooth}}}_{\text{continuity}} (v_\theta). \tag{3}$$

- **Continuity interpolation** ($\mathcal{I}_{\text{smooth}}$) *enforces Lipschitz continuity in* s, eliminating failure mode (i). We query the velocity network at two boundary-score embeddings $e_{\text{low}}, e_{\text{high}}$ to obtain $v_{\text{low}} = v_\theta(\cdot, e_{\text{low}})$ and $v_{\text{high}} = v_\theta(\cdot, e_{\text{high}})$, and replace $v_\theta(\cdot, s)$ by their linear interpolant:

$$\mathcal{I}_{\text{smooth}}(v_\theta) \;=\; (1-\alpha)\, v_{\text{low}} + \alpha\, v_{\text{high}}, \qquad \alpha = \frac{s - s_{\text{low}}}{s_{\text{high}} - s_{\text{low}}}, \tag{4}$$

  yielding a velocity field that is Lipschitz-continuous in s even at scores absent from training.

- **Bio-direction alignment** ($\mathcal{P}_{\text{bio}}$) *enforces observable-preserving alignment*, addressing failure mode (ii). Let $g_{\text{act}}(x)$ be the gradient of the cell-type-specific active observable (NA or DAPI, derived from §2.4), and stack the gradients of the remaining (preserved) observables as rows of $G_{\text{pres}} \in \mathbb{R}^{(M-1)\times d}$. We project g_{act} onto the null space of G_{pres}, where preserved observables remain unchanged, and scale by β to obtain the *weighted gradient*, which is then added to v,

$$\mathcal{P}_{\text{bio}}(v) \;=\; v \;+\; \underbrace{\beta \left(I - G_{\text{pres}}^\top (G_{\text{pres}} G_{\text{pres}}^\top)^\dagger G_{\text{pres}}\right) g_{\text{act}}(x)}_{\text{weighted gradient}}. \tag{5}$$

  This guarantees $g_j^\top \mathcal{P}_{\text{bio}}(v) = g_j^\top v$ for every preserved j, so advancing the active observable does not perturb the others, an exact invariance independent of v_θ.

The composition $\mathcal{P}_{\text{bio}} \circ \mathcal{I}_{\text{smooth}}$ realises (P2), lifting $\mathcal{T}$ from the identity to a property-constrained correction. More details for SENOFLOW are provided in Appendices C.

4 EXPERIMENTS

In this section, we benchmark SENOFLOW against a comprehensive suite of baselines under the evaluation protocol of §2. §4.1 reports the main benchmark, cross-tissue generalisation, expert evaluation, and downstream utility. §4.2 ablates each component of SENOFLOW.

Baselines. We compare SENOFLOW against nine baselines covering major Flow Matching variants in the literature: I-CFM [59], iMF [2], MeanFlow [36], Rectified Flow [60], Contrastive FM [75], LogitNorm OT-CFM [27], CellFlux [83], and two non-FM generators (cVAE [51], CCDM [24]). All are retrained on SENOMORPH with FiLM [68] conditioning on cell type and senescence score.

Evaluation metrics. Following common practice, we measure *image quality* by FID [43] and KID [7] in InceptionV3 [76] feature space, and *biological fidelity* by mean $|\rho|$, the mean absolute Spearman correlation between the senescence score and the most significant morphological observable per cell type. Six domain experts additionally rate generated trajectories on a 1-to-5 Likert scale [57] along *Senescence Realism* and *Fidelity*.

4.1 EVALUATION RESULTS

Overall Performance. Table 1 reports FID, KID, and mean $|\rho|$ on lung, prostate, and skin against nine baselines. SENOFLOW outperforms every baseline across all three tissues and all three metrics On lung, SENOFLOW reaches **11.6** FID, **1.7** KID, and **87.7%** mean $|\rho|$, surpassing the second-best iMF by 2.6 on FID and 2.0 on KID, and the second-best Contrastive FM by 57.0 pp on $|\rho|$. On prostate, SENOFLOW reaches **11.7** FID, **2.4** KID, and **62.3%** mean $|\rho|$, surpassing the second-best iMF by 1.1 on FID and 0.9 on KID, and the second-best MeanFlow by 4.3 pp on $|\rho|$. On skin, SENOFLOW reaches **11.0** FID, **1.6** KID, and **88.7%** mean $|\rho|$, surpassing the second-best iMF by 0.2 on FID, and the second-best LogitNorm OT-CFM by 70.2 pp on $|\rho|$. The consistent lead across three distinct tissue contexts evidences the cross-tissue generalisation of SENOFLOW. The improvement on FID and KID reflects the effectiveness of GPC, which replaces batch-local OT pairing with a global prototype memory and yields a cleaner supervisory signal for image quality, while the improvement on $|\rho|$ reflects the effectiveness of CVC, which enforces score-Lipschitz, feature-aligned dynamics at integration and translates into stronger biological fidelity. It is worth noting that SENOFLOW is

Table 1: **Benchmark evaluation across three tissues.** All ten methods are trained and evaluated per-tissue. Here presents Non-FM methods, FM methods, and SENOFLOW, respectively.

Method	Venue	Lung			Prostate			Skin		
		FID↓	KID↓	$\lvert\rho\rvert$ ↑	FID↓	KID↓	$\lvert\rho\rvert$ ↑	FID↓	KID↓	$\lvert\rho\rvert$ ↑
cVAE [51]	ICLR'14	61.7	40.0	5.5	49.6	35.4	11.0	78.5	29.5	13.2
CCDM [24]	TMM'26	16.4	4.4	11.8	16.1	4.5	16.3	15.9	4.3	7.6
I-CFM [59]	ICLR'23	85.6	56.8	30.7	59.3	36.6	57.2	97.4	83.7	10.2
iMF [2]	JMLR'25	14.2	3.7	10.8	12.8	3.3	25.8	11.2	1.8	7.3
MeanFlow [36]	NeurIPS'25	18.0	6.1	11.5	17.8	7.0	58.0	16.1	5.9	13.6
Rectified Flow [60]	ICLR'23	61.0	34.7	5.7	30.2	22.8	41.5	97.9	48.9	3.5
Contrastive FM [75]	ICCV'25	26.2	10.8	30.7	29.6	17.7	52.7	19.0	5.2	14.0
LogitNorm OT-CFM [27]	ICML'24	93.0	62.4	7.0	61.8	41.5	22.7	87.4	85.8	18.5
CellFlux [83]	ICML'25	80.2	43.1	1.0	82.3	48.6	12.9	48.1	37.0	10.9
SENOFLOW (ours)	—	**11.6**	**1.7**	**87.7**	**11.7**	**2.4**	**62.3**	**11.0**	**1.6**	**88.7**

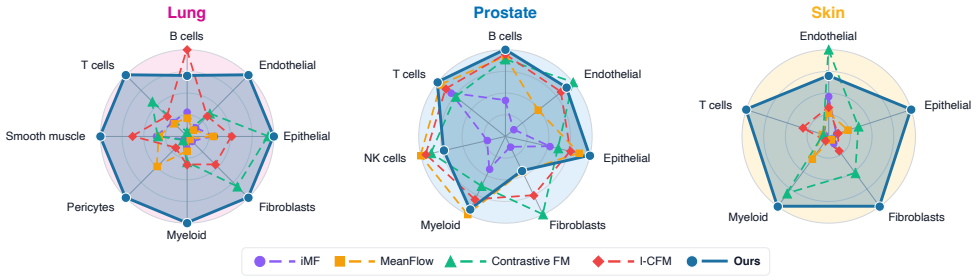

Figure 6: **Per-cell-type $|\rho|$ across tissues (↑).** SENOFLOW yields the largest $|\rho|$ on every tissue.

Table 2: **Expert evaluation across three tissues.** Ten methods rated by six senescence-biology experts on Senescence Realism and Fidelity (1–5 Likert).

Method	Lung		Prostate		Skin	
	Realism↑	Fidelity↑	Realism↑	Fidelity↑	Realism↑	Fidelity↑
cVAE [51]	3.5	2.8	3.2	3.3	2.8	2.0
CCDM [24]	2.8	1.3	2.5	2.7	2.8	2.5
I-CFM [59]	3.3	2.7	2.7	3.7	1.5	1.7
iMF [2]	4.2	2.7	3.3	3.8	3.3	3.0
MeanFlow [36]	4.3	3.0	2.7	3.3	4.2	3.7
Rectified Flow [60]	4.2	3.3	2.8	3.5	2.3	3.3
Contrastive FM [75]	3.3	3.3	2.0	3.3	3.7	4.3
LogitNorm OT-CFM [27]	3.7	3.2	2.2	2.7	2.2	2.5
CellFlux [83]	2.7	2.0	2.5	2.2	3.0	2.8
SENOFLOW (ours)	**4.7**	**4.8**	**4.7**	**5.0**	**4.8**	**4.8**

the only generator with FID below 12 and $|\rho|$ above 50% on every tissue. Side-by-side qualitative trajectories against all baselines on all three tissues are provided in Appendix F.

Expert Evaluation. Six senescence-biology experts further rate generated trajectories along *Realism* and *Senescence Fidelity* on a 1–5 Likert scale, complementing the automatic metrics in Table 1. SENOFLOW receives the highest expert score on every criterion and every tissue in Table 2, reaching cross-tissue means of **4.73** on Realism and **4.87** on Senescence Fidelity. No baseline averages above 4 on both criteria, where the strongest competitors are MeanFlow on Realism (3.73) and Contrastive FM on Senescence Fidelity (3.63), neither reaching 4 on the other axis. This dual-axis lead reflects the two property-driven operators of SENOFLOW, where batch-invariant pairing under GPC delivers snapshot realism and score-Lipschitz, feature-aligned dynamics under CVC deliver trajectory-level senescence fidelity. Inter-rater agreement statistics are reported in Appendix D.

Per-cell-type Performance. To further analyze cross-cell-type robustness, we report per-cell-type $|\rho|$ on lung, prostate, and skin against four representative baselines. As shown in Fig. 6, SENOFLOW attains the higher $|\rho|$ on **15 of 20** cell types. This cross-cell-type robustness traces to GPC's prototype memory $\mathcal{M}[c, b]$ whose cell-type key fixes the target's type to match the source's, correcting the mispairing introduced by batch-local OT whenever a minibatch underrepresents a cell type.

Table 3: **Ablative results on GPC and the two CVC operators.**

Configuration	FID ↓	$\lvert\rho\rvert$ (%) ↑	Configuration	FID ↓	$\lvert\rho\rvert$ (%) ↑
Baseline	16.3	23.7	+ GPC	11.0	34.8
+ CFG	15.8	32.6	+ GPC + CFG	11.2	42.1
+ $\mathcal{I}_{\text{smooth}}$	13.4	55.8	+ GPC + $\mathcal{I}_{\text{smooth}}$	10.9	43.0
+ $\mathcal{P}_{\text{bio}}$	13.1	42.6	+ GPC + $\mathcal{P}_{\text{bio}}$	11.1	36.2
+ $\mathcal{I}_{\text{smooth}} + \mathcal{P}_{\text{bio}}$	12.8	57.1	**Full** (+ GPC + CVC)	**11.6**	**87.7**

(a) GPC and CVC ablation. CVC = $\mathcal{I}_{\text{smooth}} + \mathcal{P}_{\text{bio}}$; CFG: classifier-free guidance [44].

K	FID ↓	$\lvert\rho\rvert$ (%) ↑
8	20.0	54.5
16	18.7	56.5
32	18.9	56.0
64	**11.6**	**87.7**
128	17.3	68.2

(b) GPC cluster count K.

β	FID ↓	$\lvert\rho\rvert$ (%) ↑
0.1	11.5	87.5
0.5	11.5	87.6
1.0	**11.6**	**87.7**
2.0	11.6	87.3
5.0	11.8	85.6

(c) $\mathcal{P}_{\text{bio}}$ guidance strength β.

DeepScence Validation. We further validate DeepScence as the conditioning signal by collecting expert severity ratings on 50 cells per tissue, sampled uniformly along the DeepScence axis and balanced across cell types. Table 4 reports a Spearman correlation of ρ=**0.87** overall, with per-tissue values of **0.86** (lung), **0.88** (prostate), and **0.88** (skin). The tight agreement across tissues confirms that DeepScence tracks expert-perceived senescence, validating its use as the continuous conditioning signal in SENOFLOW.

Table 4: **DeepScence validation.** ρ between DeepScence and expert ratings.

Metric	Lung (N=50)	Prostate (N=50)	Skin (N=50)	Overall (N=150)
ρ ↑	0.86	0.88	0.88	**0.87**

Validation on Tracked Cells. Since the cells are destroyed during senescence measurement, so SENOMORPH holds no real trajectory to score a generated one against. We instead test two properties of the generator on tracked cells. Temporal fidelity measures whether the generated trajectory follows the real one. Identity preservation measures whether it stays with the same cell. We use two additional dataset CellResponse [38] (96-hour drug response) and CancerCellTracker (CCT) [47] (14.5-hour cell death). We remove the original pairing information, reconstructed the pairings solely through coupling, and retrain the SENOFLOW, iMF, and MeanFlow models on the two datasets, respectively. As shown in Tab. 5, SENOFLOW outperforms iMF and MeanFlow on both properties. For temporal fidelity, its generated trajectories track the real frames more closely on CellResponse in both MAD and FID. For identity preservation, its endpoint generated from the 1-hour frame on CCT retrieves the true cell at **53.0**% top-1, whereas iMF and MeanFlow retrieve near chance. This indicates that GPC keeps each generated trajectory anchored to its own source cell. See Appendix G for details.

Table 5: **Validation on tracked cells.**

Method	CellResponse		CCT	
	MAD↓	FID↓	top-1↑	top-5↑
iMF [2]	7.5	100.8	0.5	3.2
MeanFlow [36]	6.6	53.9	0.7	3.3
SENOFLOW (ours)	**5.4**	**19.3**	**53.0**	**64.4**

4.2 ABLATION STUDY

Component on/off (Table 3a). Different from prior conditional flow-matching generators that rely on batch-local OT pairing, SENOFLOW adopts Global Prototype Coupling (GPC) to stabilise training-time pair formation. In addition, SENOFLOW applies two test-time CVC operators, $\mathcal{I}_{\text{smooth}}$ for amplitude interpolation between boundary-score velocities and $\mathcal{P}_{\text{bio}}$ for bio-direction alignment along morphological feature gradients. Table 3a lists the results. The baseline without any component attains 16.3 FID and 23.7% mean $|\rho|$. Enabling GPC alone improves performance to 11.0 FID and 34.8% $|\rho|$, showing the effectiveness of replacing batch-local OT pairing with a global prototype memory. On top of GPC, standard classifier-free guidance (CFG) raises $|\rho|$ to 42.1%. Adding both CVC operators largely boosts the results to **87.7%** $|\rho|$ at **11.6** FID. This validates the efficacy of CVC in enforcing smooth interpolation and bio-aligned dynamics along the senescence axis.

GPC cluster count K (Table 3b). K controls the number of k-means prototypes per (c, b) group in GPC's memory $\mathcal{M}[c, b]$. We observe the performance peaks sharply at K=64, reaching **11.6** FID and **87.7%** $|\rho|$. Smaller $K \in \{8, 16, 32\}$ leaves FID above 18 and $|\rho|$ in the 54-57% range, and

larger $K{=}128$ recovers $|\rho|$ only to **68.2%** at **17.3** FID. Hence, we set $K{=}64$ throughout, balancing per-group prototype coverage against per-prototype sample count.

$\mathcal{P}_{\text{bio}}$ **guidance strength** β **(Table 3c).** The strength β scales the bio-direction projection added to the integrated velocity. We can find the performance is robust across $\beta \in [0.1, 2.0]$, while only $\beta{=}5.0$ shows a degradation to **11.8** FID and **85.6%** $|\rho|$. Thus, we set $\beta{=}1.0$ in this work.

4.3 Downstream Utility

A central application of senescence-conditioned generation is to augment downstream classifiers limited by the scarcity of real senescent cells. We train a DINOv2-backbone [66] binary classifier on a low vs. high senescence task per tissue using real training cells, then retrain the same architecture on real cells augmented with SenoFlow-generated synthetic cells from our companion set. Both variants share the architecture, optimizer, schedule, and real test split (see Appendix H for details). Table 6 reports the results. We observe that augmentation consistently improves both ACC and AUC on every tissue, yielding overall gains of **+3.9** pp ACC and **+1.9** pp AUC. This confirms that SenoFlow's companion set carries senescence-specific morphology beyond what the real training cells alone capture.

Table 6: **Downstream classification with SenoFlow augmentation.**

Setting	Metric	Lung	Prostate	Skin
Real	ACC↑	62.4	64.9	60.3
	AUC↑	72.4	74.4	71.6
+ Aug	ACC↑	**65.0**↑2.6	**66.8**↑1.9	**67.5**↑7.2
	AUC↑	**73.2**↑0.8	**75.9**↑1.5	**74.9**↑3.3

5 Related Work

Datasets for conditional cell morphology. Existing cell morphology datasets condition images on *external* perturbations, either chemical (BBBC021 [61], JUMP Cell Painting [13], the Cell Painting Gallery [79, 8, 12, 64, 39, 9, 10]) or genetic (RxRx3 [31], cpg0021 [13], OPS atlases [32, 71]), and observe each cell only once [72, 8, 12]. No public morphology benchmark, however, conditions on an endogenous biological state. SenoMorph fills this gap. To our knowledge, it is the first cell morphology benchmark dedicated to senescence, an age-accumulating biological state, with a *continuous* score computed from each cell's own transcriptome at single-cell resolution and aligned to its DAPI image (81,255 cells, 9 cell types, 3 tissues). To overcome the inherent scarcity of senescent cells, we additionally release a companion set of 812,550 score-conditioned cells, providing data scale unattainable from cross-sectional measurements alone.

Flow Matching for cell morphology. Flow Matching [59, 60, 2, 27, 63, 82, 36, 75, 69, 45, 4] learns a velocity field that transports samples along straight-line paths between distributions, and has been adopted for cell morphology modelling under discrete drug [83, 67, 65, 84] or gene perturbations [53, 1, 55]. These works share two design choices, source and target cells are paired by minibatch optimal transport [21, 34, 30, 77, 69, 54, 35, 3, 28, 50, 78, 16] at training, and the conditional velocity is steered by classifier-free guidance [22, 44, 6, 17, 49, 18, 73, 56, 85, 33, 29, 80, 58] at inference. Both degenerate under *continuous* biological scores. Batch-local OT becomes high-variance because any minibatch sparsely covers the joint cell-type$\times$score space, and a single guidance scalar regulates trajectory geometry only at the endpoints. SenoFlow replaces both with property-driven alternatives, yielding a generic recipe for FM under continuous biological conditions, which we use to generate the companion set above.

6 Conclusion

SenoMorph establishes the first public benchmark for senescence-conditioned cell morphology generation, addressing a gap that has so far prevented systematic study of an endogenous biological axis at the level of individual cells. The dataset pairs Xenium-derived single-cell DAPI crops with continuous, transcriptome-aligned senescence scores across multiple tissues and cell types, accompanied by a generated companion set that mitigates the scarcity of senescent cells in downstream tasks. Alongside the data, we provide a evaluation protocol, baselines spanning Flow Matching and non-FM models, and SenoFlow, a Flow Matching framework that instantiates the coupling and velocity transformation of CFM with property-driven operators for continuous biological scores. SenoFlow consistently outperforms all baselines, yet still exposes residual gaps on specific cell-type-by-tissue evaluations, which we publicly track on the SenoMorph leaderboard to invite continued progress. We hope SenoMorph supports the broader study of endogenous biological states beyond senescence, and plan to extend it to additional tissues, donors, and imaging modalities as more single-cell data becomes available.

AI USE STATEMENT

Generative AI was used only for grammar correction and sentence-level polishing. It was not involved in research ideation, method design, experiments, data curation, or model implementation. The authors take full responsibility for all content.

ETHICS STATEMENT

We conform to the ICLR Code of Ethics. SENOMORPH is curated from publicly released Xenium tissue sections; no new human-subjects data was collected. The expert ratings in §4.1 were provided voluntarily by six senescence-biology researchers within our group. SENOFLOW's synthetic cells are for research use only and must not be treated as experimental evidence or used for diagnosis. Limitations and broader impact are discussed in Appendix I.

REPRODUCIBILITY STATEMENT

All models are implemented in PyTorch and trained on NVIDIA H200 GPUs. Dataset construction, architecture and pseudo-code, evaluation protocol, and downstream training are detailed in Appendices A, C, D, and H. Code and data will be released upon acceptance.

REFERENCES

[1] Abrar Rahman Abir, Sajib Acharjee Dip, and Liqing Zhang. Cfm-gp: unified conditional flow matching to learn gene perturbation across cell types. *NAR Genomics and Bioinformatics*, 8(3): lqag087, 2026.

[2] Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209): 1–80, 2025.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

[4] Lazar Atanackovic, Xi Nicole Zhang, Brandon Amos, Mathieu Blanchette, Leo J Lee, Yoshua Bengio, Alexander Tong, and Kirill Neklyudov. Meta flow matching: Integrating vector fields on the wasserstein manifold. In *ICLR*, 2025.

[5] Darren J Baker, Bennett G Childs, Matej Durik, Melinde E Wijers, Cynthia J Sieben, Jian Zhong, Rachel A. Saltness, Karthik B Jeganathan, Grace Casaclang Verzosa, Abdulmohammad Pezeshki, et al. Naturally occurring p16ink4a-positive cells shorten healthy lifespan. *Nature*, 530(7589):184–189, 2016.

[6] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *ICLR*, 2024.

[7] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[8] 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, 2016.

[9] Mark-Anthony Bray, Sigrun M Gustafsdottir, Mohammad H Rohban, Shantanu Singh, Vebjorn Ljosa, Katherine L Sokolnicki, Joshua A Bittker, Nicole E Bodycombe, Vlado Dančík, Thomas P Hasaka, et al. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the cell painting assay. *Gigascience*, 6(12):giw014, 2017.

[10] Juan C Caicedo, Claire McQuin, Allen Goodman, Shantanu Singh, and Anne E Carpenter. Weakly supervised learning of single-cell feature embeddings. In *CVPR*, 2018.

[11] Judith Campisi and Fabrizio d'Adda di Fagagna. Cellular senescence: when bad things happen to good cells. *Nature reviews Molecular cell biology*, 8(9):729–740, 2007.

[12] Srinivas Niranj Chandrasekaran, Hugo Ceulemans, Justin D Boyd, and Anne E Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature reviews drug discovery*, 20(2):145–159, 2021.

[13] Srinivas Niranj Chandrasekaran, Beth A Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21(6):1114–1121, 2024.

[14] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792, 2022.

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

[16] Ho Kei Cheng and Alexander Schwing. The curse of conditions: Analyzing and improving optimal transport for conditional flow-based generation. In *ICCV*, 2025.

[17] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

[18] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. In *ICLR*, 2025.

[19] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

[20] William S Cleveland and Susan J Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610, 1988.

[21] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

[22] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

[23] Goberdhan P Dimri, Xinhua Lee, George Basile, Meileen Acosta, Glynis Scott, Calvin Roskelley, Estela E Medrano, Maarten Linskens, Ivica Rubelj, and Olivia Pereira-Smith. A biomarker that identifies senescent human cells in culture and in aging skin in vivo. *Proceedings of the National Academy of Sciences*, 92(20):9363–9367, 1995.

[24] Xin Ding, Yongwei Wang, Kao Zhang, and Z Jane Wang. Ccdm: Continuous conditional diffusion models for image generation. *IEEE Transactions on Multimedia*, 2026.

[25] César Domínguez Conde, Chao Xu, Louie B Jarvis, Daniel B Rainbow, Sara B Wells, Tamir Gomes, Sarah K Howlett, Ondrej Suchanek, Krzysztof Polanski, Hamish W King, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594): eabl5197, 2022.

[26] Imanol Duran, Joaquim Pombo, Bin Sun, Suchira Gallage, Hiromi Kudo, Domhnall McHugh, Laura Bousset, Jose Efren Barragan Avila, Roberta Forlano, Pinelopi Manousou, et al. Detection of senescence using machine learning algorithms based on nuclear features. *Nature communications*, 15(1):1041, 2024.

[27] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[28] Luca Eyring, Dominik Klein, Théo Uscidda, Giovanni Palla, Niki Kilbertus, Zeynep Akata, and Fabian Theis. Unbalancedness in neural monge maps improves unpaired domain translation. In *ICLR*, 2024.

[29] Weichen Fan, Amber Yijia Zheng, Raymond A Yeh, and Ziwei Liu. Cfg-zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*, 2025.

[30] Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein: asymptotic and gradient properties. *arXiv preprint arXiv:1910.04091*, 2019.

[31] Marta M Fay, Oren Kraus, Mason Victors, Lakshmanan Arumugam, Kamal Vuggumudi, John Urbanik, Kyle Hansen, Safiye Celik, Nico Cernek, Ganesh Jagannathan, et al. Rxrx3: Phenomics map of biology. *bioRxiv*, 2023.

[32] David Feldman, Avtar Singh, Jonathan L Schmid-Burgk, Rebecca J Carlson, Anja Mezger, Anthony J Garrity, Feng Zhang, and Paul C Blainey. Optical pooled screens in human cells. *Cell*, 179(3):787–799, 2019.

[33] Ruiqi Feng, Chenglei Yu, Wenhao Deng, Peiyan Hu, and Tailin Wu. On the guidance of flow matching. *arXiv preprint arXiv:2502.02150*, 2025.

[34] Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

[35] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *AISTATS*, 2018.

[36] Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *NeurIPS*, 2026.

[37] Vassilis Gorgoulis, Peter D Adams, Andrea Alimonti, Dorothy C Bennett, Oliver Bischof, Cleo Bishop, Judith Campisi, Manuel Collado, Konstantinos Evangelou, Gerardo Ferbeyre, et al. Cellular senescence: defining a path forward. *Cell*, 179(4):813–827, 2019.

[38] Sean M Gross, Farnaz Mohammadi, Crystal Sanchez-Aguila, Paulina J Zhan, Tiera A Liby, Mark A Dane, Aaron S Meyer, and Laura M Heiser. Analysis and modeling of cancer drug responses using cell cycle phase-specific rate effects. *Nature communications*, 14(1):3450, 2023.

[39] 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, et al. Multiplex cytological profiling assay to measure diverse cellular states. *PloS one*, 8(12):e80999, 2013.

[40] Leonard Hayflick and Paul S Moorhead. The serial cultivation of human diploid cell strains. *Experimental cell research*, 25(3):585–621, 1961.

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

[42] Indra Heckenbach, Garik V Mkrtchyan, Michael Ben Ezra, Daniela Bakula, Jakob Sture Madsen, Malte Hasle Nielsen, Denise Oró, Brenna Osborne, Anthony J Covarrubias, M Laura Idda, et al. Nuclear morphology is a deep learning biomarker of cellular senescence. *Nature Aging*, 2(8): 742–755, 2022.

[43] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

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

[45] Noboru Isobe, Masanori Koyama, Jinzhe Zhang, Kohei Hayashi, and Kenji Fukumizu. Extended flow matching: a method of conditional generation with generalized continuity equation. *arXiv preprint arXiv:2402.18839*, 2024.

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

[47] Qibing Jiang, Praneeth Sudalagunta, Maria C Silva, Rafael R Canevarolo, Xiaohong Zhao, Khandakar Tanvir Ahmed, Raghunandan Reddy Alugubelli, Gabriel DeAvila, Alexandre Tungesvik, Lia Perez, et al. Cancercelltracker: a brightfield time-lapse microscopy framework for cancer drug sensitivity estimation. *Bioinformatics*, 38(16):4002–4010, 2022.

[48] Pratik Kamat, Nico Macaluso, Yukang Li, Anshika Agrawal, Aaron Winston, Lauren Pan, Teasia Stewart, Bartholomew Starich, Nicholas Milcik, Chanhong Min, et al. Single-cell morphology encodes functional subtypes of senescence in aging human dermal fibroblasts. *Science Advances*, 11(17):eads1875, 2025.

[49] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *NeurIPS*, 2024.

[50] Parnian Kassraie, Aram-Alexandre Pooladian, Michal Klein, James Thornton, Jonathan Niles-Weed, and Marco Cuturi. Progressive entropic optimal transport solvers. In *NeurIPS*, 2024.

[51] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[52] James L Kirkland and Tamara Tchkonia. Cellular senescence: a translational perspective. *EBioMedicine*, 21:21–28, 2017.

[53] Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, et al. Cellflow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025.

[54] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. *arXiv preprint arXiv:2201.12220*, 2022.

[55] Oskar Kviman, Kirill Tamogashev, Nicola Branchini, Víctor Elvira, Jens Lagergren, and Nikolay Malkin. Multi-marginal flow matching with adversarially learnt interpolants. In *ICLR*, 2026.

[56] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *NeurIPS*, 2024.

[57] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[58] Haowei Lin, Shanda Li, Haotian Ye, Yiming Yang, Stefano Ermon, Yitao Liang, and Jianzhu Ma. Tfg-flow: training-free guidance in multimodal generative flow. In *ICLR*, 2025.

[59] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[60] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[61] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637, 2012.

[62] Carlos López-Otín, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. Hallmarks of aging: An expanding universe. *Cell*, 186(2):243–278, 2023.

[63] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *ECCV*, 2024.

[64] Nikita Moshkov, Michael Bornholdt, Santiago Benoit, Matthew Smith, Claire McQuin, Allen Goodman, Rebecca A Senft, Yu Han, Mehrtash Babadi, Peter Horvath, et al. Learning representations for image-based profiling of perturbations. *Nature communications*, 15(1):1594, 2024.

[65] Zeinab Navidi, Jun Ma, Esteban Miglietta, Le Liu, Anne Carpenter, Beth Cimini, Benjamin Haibe-Kains, and Bo Wang. Morphodiff: Cellular morphology painting with diffusion models. In *ICLR*, 2025.

[66] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[67] Alessandro Palma, Fabian J Theis, and Mohammad Lotfollahi. Predicting cell morphological responses to perturbations using generative modeling. *Nature Communications*, 16(1):505, 2025.

[68] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[69] Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2304.14772*, 2023.

[70] Yilong Qu, Beijie Ji, Runze Dong, Liangcai Gu, Cliburn Chan, Jichun Xie, Carolyn Glass, Xiao-Fan Wang, Andrew B Nixon, and Zhicheng Ji. Single-cell and spatial detection of senescent cells using deepscence. *Cell Genomics*, 5(12), 2025.

[71] Meraj Ramezani, Erin Weisbart, Julia Bauman, Avtar Singh, John Yong, Maria Lozada, Gregory P Way, Sanam L Kavari, Celeste Diaz, Eddy Leardini, et al. A genome-wide atlas of human cell morphology. *Nature methods*, 22(3):621–633, 2025.

[72] Mohammad Hossein Rohban, Shantanu Singh, Xiaoyun Wu, Julia B Berthet, Mark-Anthony Bray, Yashaswi Shrestha, Xaralabos Varelas, Jesse S Boehm, and Anne E Carpenter. Systematic morphological profiling of human gene and allele function via cell painting. *Elife*, 6:e24060, 2017.

[73] Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *ICLR*, 2024.

[74] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[75] George Stoica, Vivek Ramanujan, Xiang Fan, Ali Farhadi, Ranjay Krishna, and Judy Hoffman. Contrastive flow matching. In *ICCV*, 2025.

[76] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[77] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

[78] Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.

[79] Erin Weisbart, Ankur Kumar, John Arevalo, Anne E Carpenter, Beth A Cimini, and Shantanu Singh. Cell painting gallery: an open resource for image-based profiling. *Nature methods*, 21 (10):1775–1777, 2024.

[80] Mengfei Xia, Nan Xue, Yujun Shen, Ran Yi, Tieliang Gong, and Yong-Jin Liu. Rectified diffusion guidance for conditional generation. In *CVPR*, 2025.

[81] Ming Xu, Tamar Pirtskhalava, Joshua N Farr, Bettina M Weigand, Allyson K Palmer, Megan M Weivoda, Christina L Inman, Mikolaj B Ogrodnik, Christine M Hachfeld, Daniel G Fraser, et al. Senolytics improve physical function and increase lifespan in old age. *Nature medicine*, 24(8): 1246–1256, 2018.

[82] Ling Yang, Zixiang Zhang, Zhilong Zhang, Xingchao Liu, Minkai Xu, Wentao Zhang, Chenlin Meng, Stefano Ermon, and Bin Cui. Consistency flow matching: Defining straight flows with velocity consistency. *arXiv preprint arXiv:2407.02398*, 2024.

[83] Yuhui Zhang, Yuchang Su, Chenyu Wang, Tianhong Li, Zoe Wefers, Jeffrey Nirschl, James Burgess, Daisy Ding, Alejandro Lozano, Emma Lundberg, et al. Cellflux: Simulating cellular morphology changes via flow matching. *arXiv preprint arXiv:2502.09775*, 2025.

[84] Yuhui Zhang, Yuchang Su, Zoe Wefers, Shiye Su, He Li, Tianhong Li, Chenyu Wang, James Burgess, Alejandro Lozano, Linqi Zhou, et al. Cellfluxv2: An image generative foundation model for virtual cell modeling. *bioRxiv*, 2026.

[85] Qinqing Zheng, Matt Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky TQ Chen. Guided flows for generative modeling and decision making. *arXiv preprint arXiv:2311.13443*, 2023.

SUMMARY OF THE APPENDIX

This appendix contains additional details for the ICLR 2027 submission, titled *"From Snapshots to Trajectories: Benchmarking Senescence-Conditioned Cell Morphology Generation."* The appendix is organized as follows:

- §A presents the dataset card of SENOMORPH.
- §B discusses the continuous-proxy assumption for DeepScence.
- §C reports the architecture and pseudo-code for GPC and CVC.
- §D elaborates evaluation metrics and inter-rater agreement.
- §E visualises the FID-vs-$|\rho|$ Pareto frontier.
- §F provides qualitative trajectory comparisons against all baselines.
- §G validates SENOFLOW on tracked cells with ground-truth trajectories.
- §H details the classifier training for downstream task.
- §I discusses extended limitations and broader impact.

A DATASET DETAILS

In this appendix we provide additional details about SENOMORPH.

Composition. The real set contains $81{,}255$ single-cell DAPI crops drawn from three Xenium tissue sections ($37{,}143$ lung, $28{,}808$ prostate, $15{,}304$ skin), each paired with a CellTypist cell-type label and a DeepScence senescence score (per-tissue cell-type shares in Table 7). The companion set provides ten synthetic images per real source cell at evenly spaced target scores, totalling $812{,}550$ score-conditioned crops. Cell-type imbalance is intrinsic to each tissue, with the majority cell type accounting for 47–82% across tissues.

Table 7: **Per-tissue cell-type composition of the real set.** Cell-type labels are produced by CellTypist on the matched transcriptome.

Tissue	Cell type	Share
Lung (N=37,143)	Epithelial cells	47.1%
	Endothelial cells	21.2%
	T cells	13.1%
	Myeloid cells	8.7%
	Fibroblasts	7.5%
	Smooth muscle cells	1.5%
	Pericytes	0.6%
	B cells	0.3%
Prostate (N=28,808)	T cells	59.9%
	Epithelial cells	12.7%
	Fibroblasts	10.6%
	Myeloid cells	10.1%
	Endothelial cells	3.0%
	B cells	2.6%
	NK cells	1.0%
Skin (N=15,304)	Epithelial cells	81.5%
	Fibroblasts	8.2%
	T cells	5.2%
	Endothelial cells	2.6%
	Myeloid cells	2.5%

Per-cell-type monotonicity trends. Fig. 4 in the main text reports the aggregate trend of NA and DAPI against the senescence score across all cell types and tissues. For simplicity, we plot the same LOESS curves for a few representative cell types in Fig. 7. NA and DAPI preserve the same monotone direction observed in the aggregate plot.

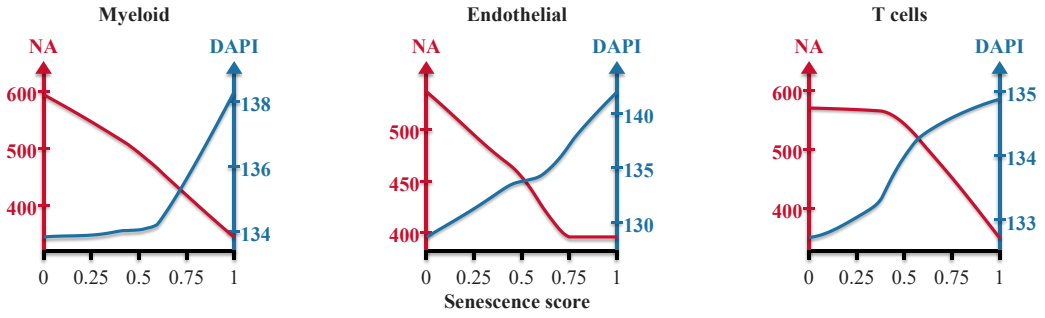

Figure 7: **Per-cell-type monotonicity curves.** LOESS-smoothed NA and DAPI vs senescence score for representative cell types.

B DeepScence as a Continuous Conditioning Signal

We adopt DeepScence [70] as the continuous senescence score in §2.3. Two design choices and one empirical result justify treating its outputs as a continuous proxy for senescence severity.

Multi-gene aggregation mitigates single-gene sparsity. Senescence is not driven by a single canonical marker. Classical markers such as *CDKN1A* (p21), *CDKN2A* (p16), and SASP components are each individually sparse and noisy at single-cell resolution in spatial transcriptomics, where dropout rates are high and panels cover only hundreds of genes. DeepScence instead aggregates information across a curated consensus senescence gene set (SnG), so the score remains well-defined for each cell even when any individual marker drops out.

Nonlinear modelling captures the coordinated programme. Senescence is a coordinated transcriptional programme rather than a linear sum of marker counts: the relevant biology lies in combinatorial co-expression patterns across the SnG. DeepScence is a deep learning-based predictor designed to capture these complex, nonlinear, and combinatorial relationships, and therefore produces a score that better reflects the underlying programme than linear aggregations such as module scores or marker averages.

Empirical validation. Beyond these design choices, our study (§4.1, Table 4) shows Spearman ρ=0.87 between DeepScence scores and expert severity ratings on real cells. The strong agreement provides direct evidence that the predictor tracks expert-perceived senescence and is therefore a meaningful continuous signal to condition on.

C Implementation Details and Pseudo-code

Architecture. All Flow Matching variants share the same U-Net backbone: channel multipliers $(1, 2, 2, 4, 8)$ with base dimension 64, FiLM conditioning with initial embedding dimension 128, and sinusoidal label embedding for the continuous senescence score.

ODE solver. We use an Adams-Bashforth third-order (AB3) solver with 50 uniform steps, initialized via three Euler steps.

All experiments run on NVIDIA H200 GPUs (141 GB).

D Evaluation Metrics and Inter-rater Agreement

This appendix collects the full metric definitions, expert-study protocol, inter-rater agreement, and downstream augmentation setup deferred from §2.5 and §4.1.

Algorithm 1 PyTorch-like pseudo-code for Global Prototype Coupling and Continuous-score Velocity Correction.

```
# ======== Global Prototype Coupling (training-time, pair formation)
    ========
# Z: VAE latents, z0: latent of one source cell
# c, s: cell-type ids and senescence scores
# s_tgt: target senescence score, the condition s in the CFM loss
# K, n_bins: number of prototypes / score bins per (cell_type, bin) group

def build_memory(Z, c, s, K, n_bins):
   # Run once before training, over all training latents.
   M = {} # prototype memory M[c, b]
   for ct in unique(c):
      for b in range(n_bins):
         group = Z[(c == ct) & (score_bin(s) == b)]
         M[ct, b] = kmeans(group, K) # K prototypes per (ct, b)
   return M

def gpc_pair(z0, ct, s_tgt, M):
   # Eq. 2, called per sample at every training step.
   b = score_bin(s_tgt) # target-cell bin, not source bin
   k_star = argmax(cosine_sim(z0, M[ct, b])) # nearest prototype, O(K)
   return sample_from_cluster(ct, b, k_star) # uniform x1 in cluster k*


# ===== Continuous-score Velocity Correction (inference-time, per ODE
    step) =====
# v_theta: velocity network
# e_lo, e_hi, s_lo, s_hi: anchor embeddings/scores at boundary
    percentiles
# beta: bio-direction guidance strength

def cvc_step(x_t, t, dt, score, v_theta, beta):
   # I_smooth: linear interpolation between boundary-score velocities
   v_lo = v_theta(x_t, t, cond=e_lo)
   v_hi = v_theta(x_t, t, cond=e_hi)
   alpha = (score - s_lo) / (s_hi - s_lo)
   v = (1 - alpha) * v_lo + alpha * v_hi

   # P_bio: bio-direction alignment via biomarker gradient, Eq. 5
   g_bio = biomarker_grad(x_t, t)
   v = v + beta * g_bio

   return x_t + dt * v # Euler step
```

D.1 Metric Definitions

Image-quality metrics. FID [43] measures the Fréchet distance between generated and real cell distributions in InceptionV3 [76] feature space, providing a standard measure of visual realism. KID [7] complements FID with an unbiased kernel-based estimator that is more reliable at moderate sample sizes per cell type.

Biological-fidelity metrics. Mean $|\rho|$ is the mean absolute Spearman correlation between the senescence score and the most significant morphological observable for each cell type, measuring whether the generated trajectory preserves the expected rank order of biological change.

Expert Study Metrics. We invite domain experts in senescence and cell biology to rate generated trajectories on two criteria, with method identities anonymised to prevent bias. *Senescence realism* evaluates whether each generated image resembles a real DAPI nuclear crop, independent of fidelity to any specific target state. *Fidelity* evaluates whether the trajectory as a whole reflects biologically plausible senescence progression. Each expert rates a source cell paired with its generated trajectory per method on a 1-to-5 Likert scale [57]; the mean per criterion is reported in Table 2.

Table 8: **Inter-rater agreement.** Fleiss' κ on Likert categories and ICC on raw scores, computed across six experts.

Criterion	Fleiss' κ	ICC
Expert ratings	0.315	0.68

D.2 Inter-rater Agreement

We quantify rating consistency across the six experts using Fleiss' κ on the Likert categories and the intraclass correlation coefficient (ICC) on the raw scores. Fleiss' κ measures categorical agreement after correcting for chance, while ICC measures absolute agreement on the continuous scale. Results are summarized in Table 8.

E FID-vs-$|\rho|$ Pareto Frontier

We visualise the joint FID-$|\rho|$ behaviour reported in Table 1 as a scatter plot in Fig. 8.

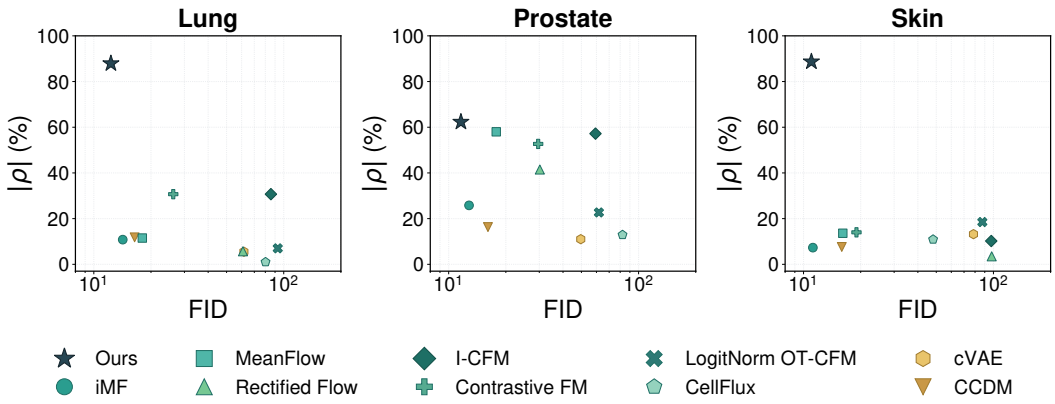

Figure 8: **FID-vs-$|\rho|$ Pareto frontier across lung, prostate, and skin.** Each marker is one method from Table 1; lower FID and higher $|\rho|$ are better, so the top-left corner is Pareto-optimal. Existing FM and non-FM generators either cluster in the low-FID-low-$|\rho|$ region or fail to achieve competitive FID, leaving the top-left corner unoccupied. SenoFlow alone reaches it on every tissue.

F Qualitative Trajectory Comparison Against Baselines

This appendix complements the quantitative results in §4.1 with side-by-side qualitative trajectories on all three tissues. We organise the comparison as a 10×10 grid: columns are the ten methods (SenoFlow + nine baselines), and rows are score steps that sweep the target senescence score from low to high. The top row is the source cell, replicated across all method columns, so every method starts from the same crop.

Figures 9, 11, and 10 present these grids on lung, skin, and prostate respectively. Driven by its bio-guided design (§3, CVC), SenoFlow produces *directionally consistent* trajectories where nuclear area and DAPI intensity evolve monotonically along the senescence axis, whereas baselines yield inconsistent trajectories that flip direction between adjacent steps.

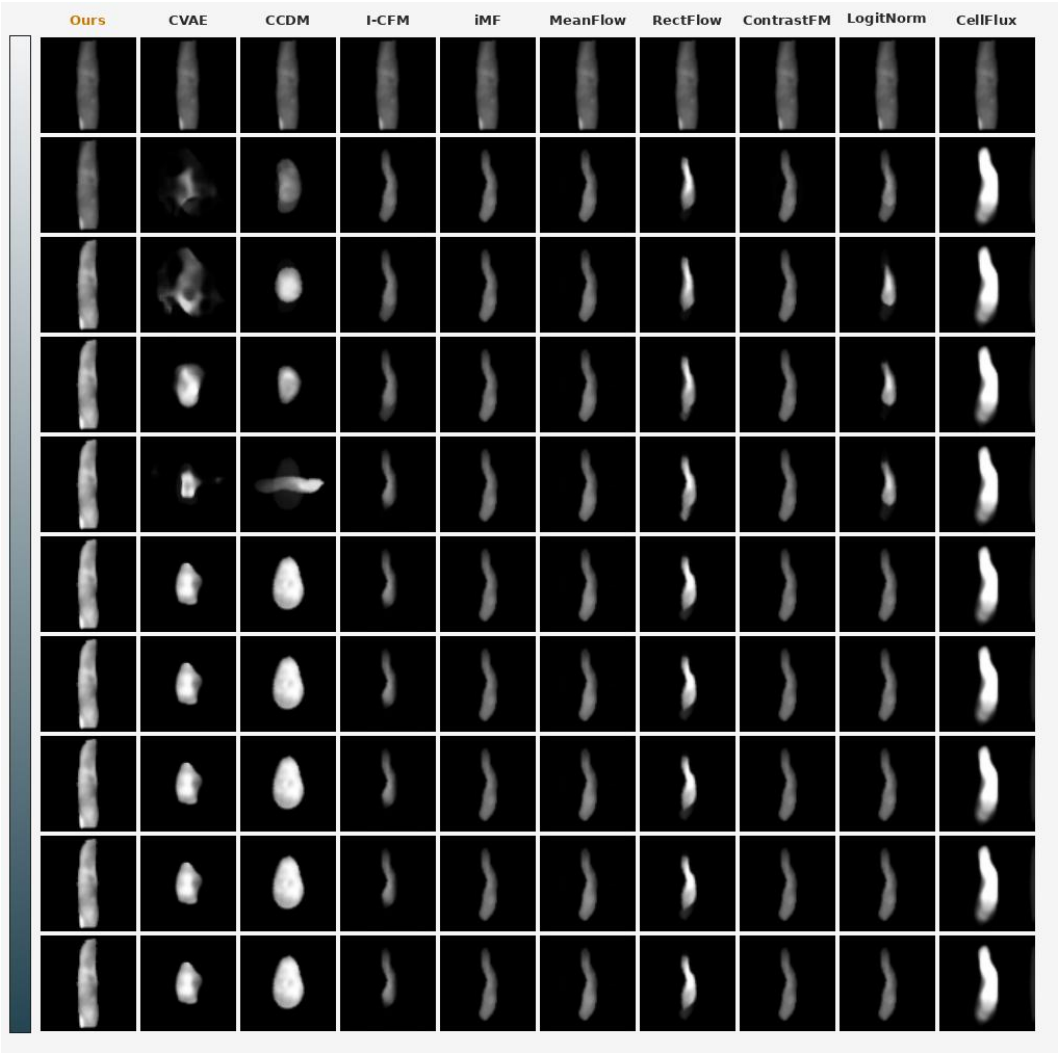

Figure 9: **Qualitative trajectory comparison on lung.** Columns are the ten methods (SENOFLOW in orange, nine baselines in blue); rows sweep the target senescence score from low (top) to high (bottom), starting from the same source cell (top row, identical across columns). Driven by its bio-guided design (§3, CVC), SENOFLOW produces trajectories that align with the biological signature of senescence, whereas baselines do not.

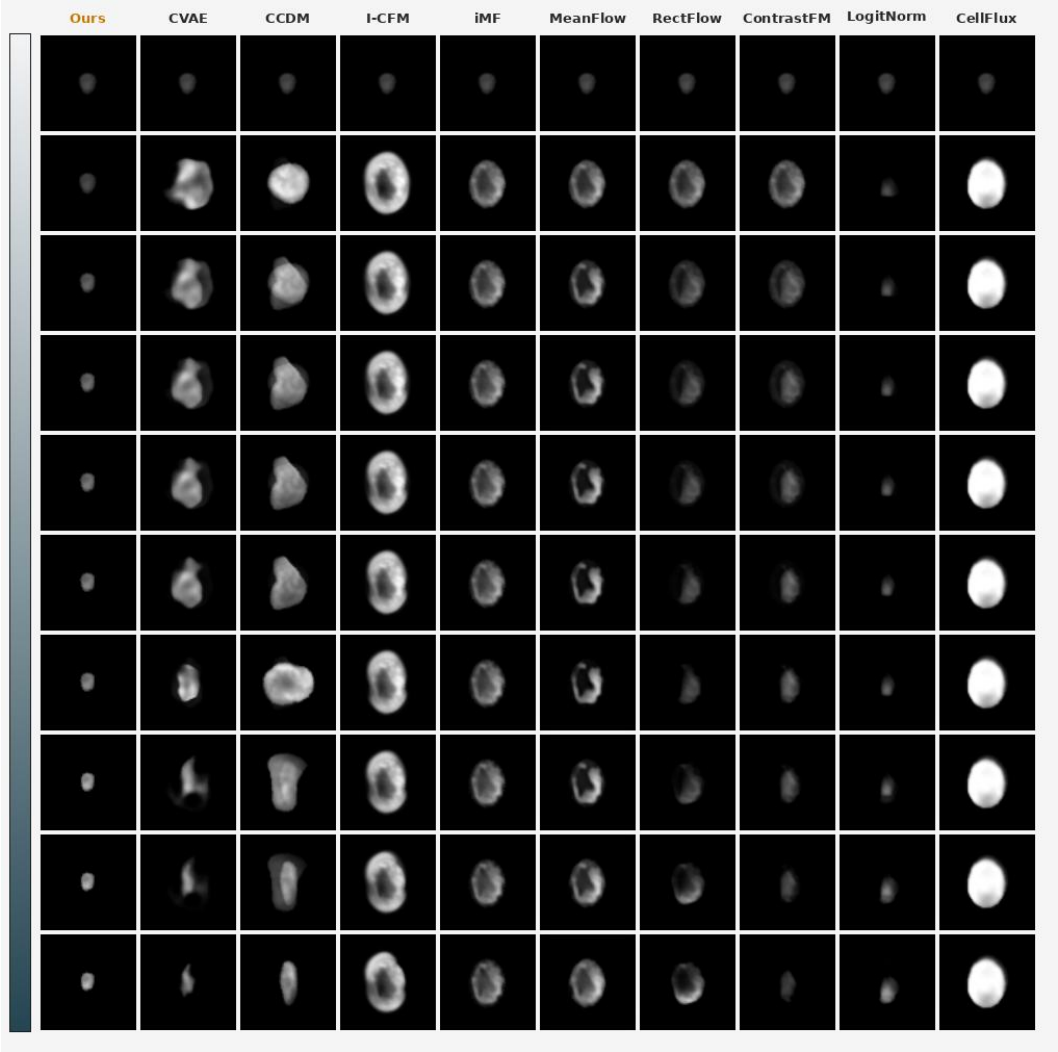

Figure 10: **Qualitative trajectory comparison on prostate.** Layout matches Figure 9.

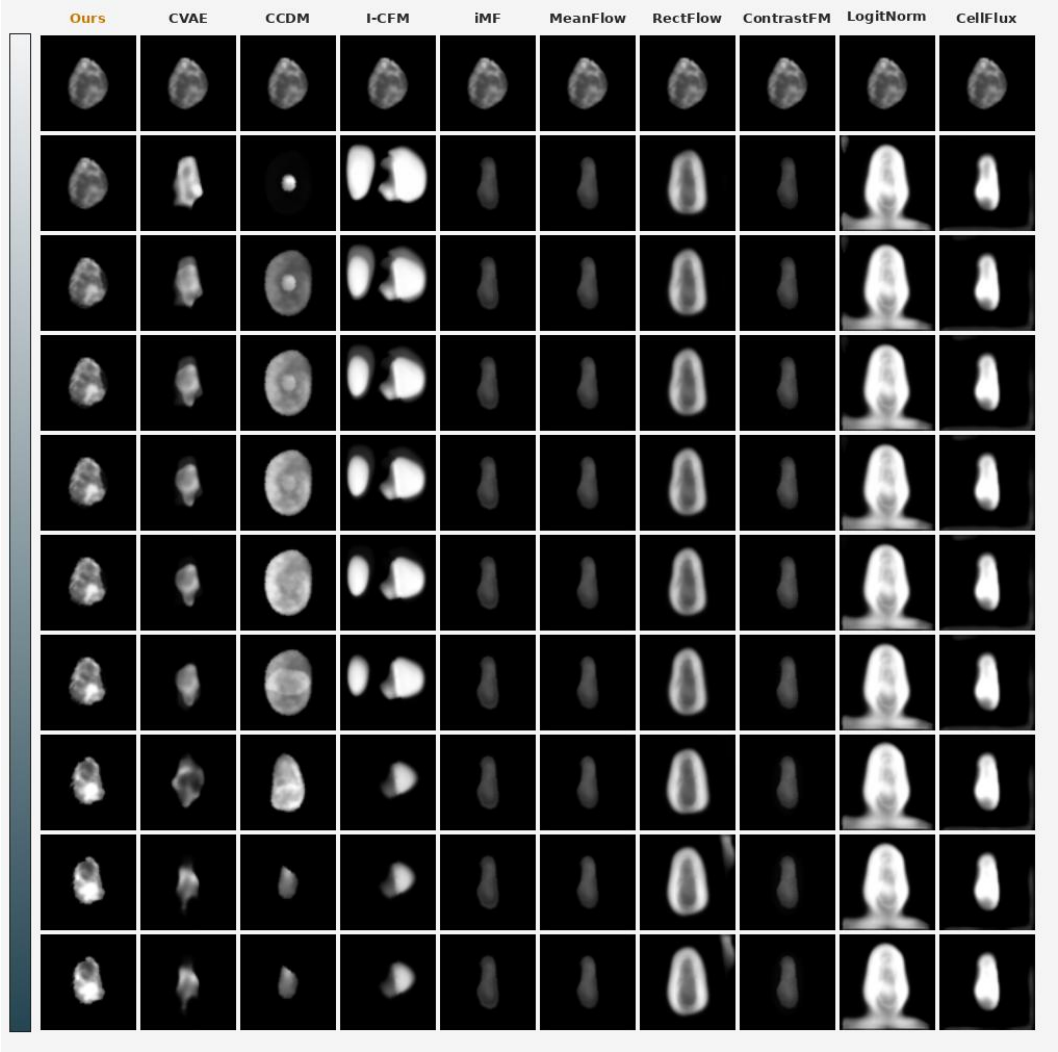

Figure 11: **Qualitative trajectory comparison on skin.** Layout matches Figure 9.

F.1 TEST-TIME SCORE SAMPLING

We sample $N{=}10$ uniformly spaced target senescence scores in $[0, 1]$ for every source cell. $N{=}5$ skips intermediate transitions and $N{=}20$ yields near-identical adjacent frames; $N{=}10$ balances trajectory resolution and evaluation cost (Fig. 12).

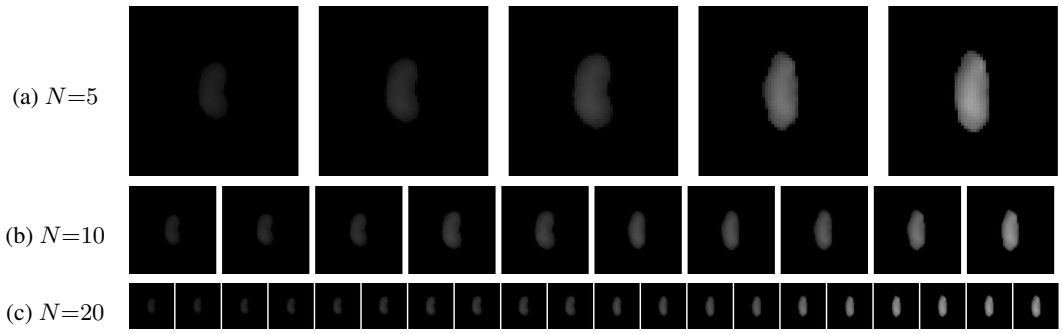

Figure 12: **Effect of N.** SENOFLOW trajectories at $N{=}5$, 10, and 20.

G VALIDATION ON TRACKED CELLS

This section details the experiment in Tab. 5. Every cell in SENOMORPH is imaged once. We therefore validate the generator on two datasets where every cell is tracked over time.

Datasets. *CellResponse* [38] (IDR accession idr0119) images a HER2$^+$ breast cancer line every 30 minutes for 96 hours under six drug arms and 33 dose groups. The baseline frame is taken at dosing. *CCT* [47] is a live-cell death video of 1,461 tracked cells with 29 frames each at 30-minute intervals ($\sim$14.5 hours). Death rate ranges from 5% to 88% along the time axis. We use the public sample released with the code. Tracks are split by cell into 1,023 for training and 438 for test.

Protocol. The true pairing is removed before training and pairs are rebuilt by coupling alone. Elapsed time (and dose on CellResponse) replaces the senescence score as the conditioning variable. $\mathcal{P}_{\text{bio}}$ is disabled in this experiment, as its NA and DAPI directions are specific to senescence. For temporal fidelity, we generate the full trajectory of each test cell from its $t{=}0$ frame. FID is computed between all generated and all real frames. MAD is the median absolute pixel deviation between each generated frame and the real frame of the same cell at the same time. For identity preservation, we generate the final frame of each test cell from its 1-hour frame and retrieve the true endpoint among all 438 candidates.

Results. Fig. 13 shows generated and real trajectories on CellResponse and the retrieval matrix on CCT, where the diagonal is the correct pairing. As seen, SENOFLOW reaches **5.4** MAD and **19.3** FID on CellResponse, and **53.0**% `top-1` and **64.4**% `top-5` on CCT against a chance level of 0.2%.

H DOWNSTREAM CLASSIFIER TRAINING DETAILS

Task setup. For each tissue, we frame senescent-cell detection as a binary classification task by thresholding the tissue-normalised DeepScence score at the median. We sample 10,000 real cells from the training split with a balanced class sampler, optionally augmented with 6,000 SENOFLOW-generated synthetic cells. Evaluation uses a held-out real test split (20% of the real set). Synthetic cells are never used at evaluation time.

Model and training. Both the augmented variant and the real-only baseline use a DINOv2 ViT-S/14 backbone with a two-layer MLP head. DAPI crops ($64{\times}64$) are upsampled to $224{\times}224$, replicated to three channels, and normalized with ImageNet statistics. Training uses AdamW, batch size 48, with early stopping on validation AUC. Implementation details and configurations will be provided in our released code.

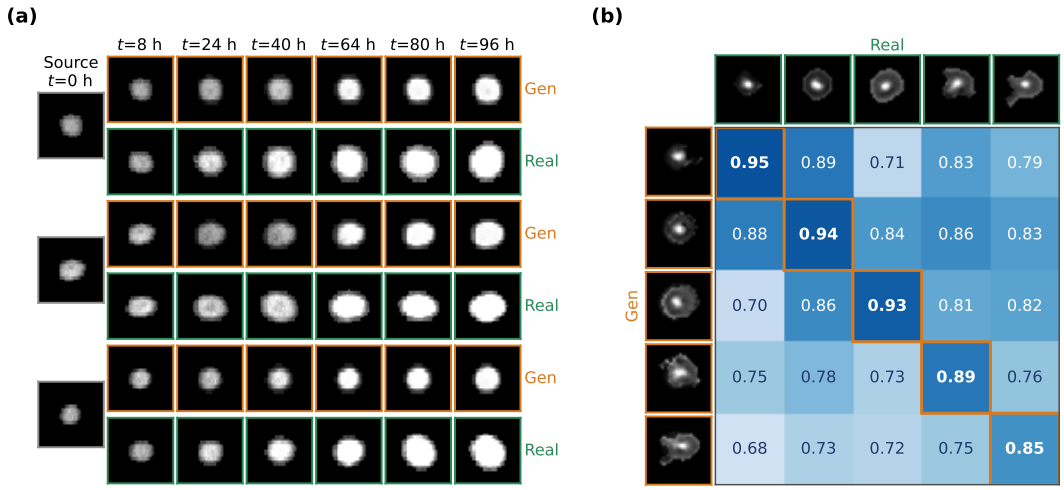

Figure 13: **Validation on tracked cells.** (a) Generated *vs* real trajectories on CellResponse. Each block is one cell, with the source at t=0 h on the left, generated frames in the top row, and real frames of the same cell in the bottom row. (b) Endpoint retrieval on CCT. Each entry is the cosine similarity between a generated endpoint (row) and a real endpoint (column). The diagonal marks the correct pairing.

I Limitations and Broader Impact

Limitations. The current benchmark has two limitations. *Single-channel imaging.* The imaging modality is restricted to single-channel DAPI nuclear staining at 64×64 resolution, which does not resolve cytoplasmic or organelle features. *Donor-level splits.* Each Xenium output contains a single tissue section from a single donor, so donor-level holdout splits are not feasible with the currently available public data. These gaps suggest concrete extensions. Richer multi-channel imaging together with larger gene measurements would enlarge the morphological feature space, and the continued growth of the public Xenium pool will eventually enable donor-level holdouts, extending SenoMorph from an interpolation benchmark to a distributional-shift benchmark.

Broader Impact. SenoMorph is intended as a pre-screening tool for conditional generators before deployment on spatial-transcriptomics atlases. The 64×64 DAPI crops contain no personally identifiable information, and we release the dataset together with the evaluation protocol to mitigate over-interpretation of companion-set samples as biological ground truth. We discourage the use of SenoFlow's synthetic cells in clinical or diagnostic settings without further experimental validation.