SELF-SUPERVISED EDITABLE TRANSPORT FOR COUNTERFACTUAL SPATIOTEMPORAL GENE PERTURBATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting spatiotemporal tissue responses to localized gene edits from destructive, unpaired spatial snapshots is crucial for virtual perturbation screening. However, existing methods isolate temporal dynamics from interventional capacity. Bridging this dichotomy requires targeted edits that remain physically coherent and on-manifold throughout temporal propagation. We introduce **STEP** (**S**elf-supervised **T**ransport of **E**ditable **P**erturbations), a generative framework trained exclusively on unperturbed snapshots to forecast counterfactual spatial dynamics. STEP formalizes gene knockouts as *decoder-consistent proximal projections* within an editable, mask-supervised latent manifold. These interventional states propagate continuously across unpaired snapshots via a shared wild-type path law, governed by mass-calibrated flow matching and unbalanced optimal transport. Extensive benchmarks across static and dynamic spatial CRISPR assays confirm that STEP outperforms existing baselines in counterfactual forecasting and response-gene prioritization. Mechanistic analyses further validate that decoder-consistent proximal projections are crucial for maintaining manifold integrity during latent transport. Applying STEP to a zebrafish embryogenesis atlas uncovers stage-dependent *in silico* knockout heterogeneity, dissecting dynamic developmental mechanisms.

1 INTRODUCTION

Spatial transcriptomics (ST) measures genome-wide expression while preserving tissue architecture, and recent spatial CRISPR screens further reveal how localized genetic perturbations reshape cellular programs and their surrounding microenvironments (Baysoy et al., 2026; Shen et al., 2026). Predicting these perturbation responses *in silico* is an important step toward virtual tissue models that can evaluate how genetic interventions propagate across spatial contexts and developmental stages (Song et al., 2025; Wenckstern et al., 2026). Such models could turn developmental and disease atlases into virtual perturbation systems, reducing the cost and need for exhaustive time-resolved CRISPR experiments.

Current spatial perturbation models primarily operate on static tissue snapshots. They infer counterfactual expression responses from wild-type tissue by propagating interventions through gene regulatory or learned spatial interactions (Kamimoto et al., 2023; Megas et al., 2025; Sun et al., 2025). While capturing spatially resolved perturbation effects, existing methods fail to model temporal propagation of stage-specific interventions, rendering static counterfactual predictions insufficient for evolving developmental systems.

Modeling such temporal responses is challenging because ST measurements are destructive and therefore provide unpaired snapshots across time, often with misaligned coordinates and changing population mass (Stickels et al., 2021; Chen et al., 2022). Optimal transport and stochastic dynamics methods address this setting by connecting disjoint populations and reconstructing continuous wild-type trajectories (Schiebinger et al., 2019; Klein et al., 2025; Zhang et al., 2025; Qiu et al., 2024; Peng et al., 2026). These models, however, remain observational and provide no explicit mechanism for gene-specific intervention. Nascent efforts in temporal perturbation modeling, exemplified by DynPerturb (Qin et al., 2025), incorporate interventions into dynamic representations but rely on latent shifts that lack full-resolution spatial decoding, limiting direct evaluation against *in vivo* spatial CRISPR measurements.

Spatiotemporal counterfactual prediction therefore requires a representation that supports both gene-specific editing and stable temporal evolution, which are difficult to satisfy simultaneously.

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.

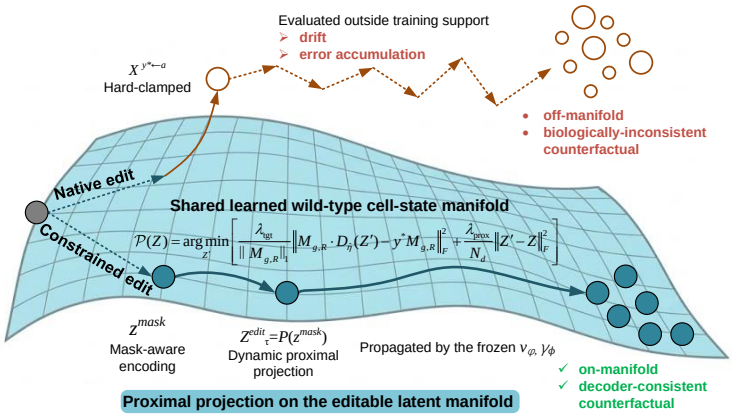


Figure 1: **Motivation of STEP.** Conceptual comparison between naive raw space clamping and constrained latent intervention. Direct one-step editing in native expression space causes counterfactual trajectories to drift off the data manifold into biologically unfeasible regions, whereas mask-supervised latent editing coupled with decoder consistent proximal projection anchors interventional states onto valid biological manifolds throughout temporal transport.

Gene interventions are naturally defined in expression space, where individual genes retain explicit identities, but directly modifying high-dimensional expression profiles can move states away from the biological manifold on which the dynamics are learned. Latent representations provide a more structured space for modeling continuous trajectories, but their abstract coordinates lack direct gene-level mapping, rendering targeted editing non-trivial.

To bridge this gap, we introduce STEP, a generative framework for spatiotemporal *in silico* perturbation. Instead of enforcing brittle feature overrides in ambient count space, STEP formulates gene perturbation as a mask-guided representation learning task. By reconstructing spatial tissue contexts under stochastic gene masking, the spatial autoencoder learns an editable latent manifold where knockouts are processed as manifold-consistent representations rather than destabilizing artifacts. Unpaired temporal dynamics are parameterized within this structured latent space via mass-calibrated flow matching, guided by relation-aware unbalanced optimal transport (Morehead et al., 2026). Consequently, gene edits are executed via a decoder-consistent proximal projection that identifies the minimal latent displacement required for target knockdown, enabling seamless forward trajectory integration through the learned wild-type vector field (Figure 1 and 5). In summary, our main contributions are:

- **Unified Spatiotemporal Framework:** We present STEP, the first generative framework that unifies continuous dynamic trajectory inference across unpaired spatial snapshots with counterfactual *in silico* perturbation, enabling full-resolution spatial expression forecasting beyond static or embedding-only limits.
- **Editable Latent Manifold and proximal projection:** We propose a mask-guided, editable latent space equipped with a decoder-consistent proximal edit operator. This formulation circumvents the off-manifold instability of raw-space gene clamping, ensuring counterfactual states remain biologically coherent before temporal integration.
- **Comprehensive Benchmark and Biological Insight:** Across *in vivo* spatial CRISPR benchmarks and a zebrafish embryogenesis atlas, STEP achieves superior counterfactual perturbation accuracy. Mechanistic analyses confirm that decoder-consistent projections preserve biological coherence, while learned vector fields uncover stage-dependent knockout heterogeneity driven by time-varying gene regulatory network rewiring.

2 Related Work

Snapshot dynamics and spatiotemporal transport. Reconstructing continuous trajectories from unpaired, destructive tissue snapshots relies heavily on optimal transport (OT) (Schiebinger et al.,

2019). Moscot (Klein et al., 2025) extends this via unbalanced fused Gromov-Wasserstein transport to align unaligned spatial sections, while DeepRUOT (Zhang et al., 2025) models stochastic unbalanced dynamics to account for cellular proliferation and apoptosis. To achieve continuous spatiotemporal resolution, frameworks like stVCR (Peng et al., 2026) and Spateo (Qiu et al., 2024) map discrete transport mechanics into continuous-time velocity fields and morphometric ODEs. While these observational models accurately reconstruct wild-type evolution and can technically simulate forward trajectories under forced gene clamps, executing such naive interventions triggers severe off-manifold instability. Because the underlying vector fields are fitted exclusively to unperturbed states, hard-clamping target genes projects cell states into unphysical regions, inducing vector field breakdown and rapid error accumulation during forward trajectory integration.

Spatial *in silico* perturbation. A parallel line of research models counterfactual tissue responses by inferring regulatory mechanisms directly from unperturbed data. Frameworks such as CellOracle (Kamimoto et al., 2023) propagate edits through single-cell gene regulatory networks, whereas generative models like Celcomen (Megas et al., 2025) and SpatialProp (Sun et al., 2025) estimate spatial counterfactuals via causal disentanglement and local context graphs. However, simply feeding dynamic datasets into these static frameworks cannot address dynamic perturbations. Because they assume stationary, time-invariant regulatory rules, they treat dynamic snapshots as isolated static states, ignoring non-equilibrium cellular transitions and non-conserved cell mass. Consequently, they fail to capture how an early intervention alters the continuous, time-varying path law that governs downstream trajectory propagation. While dynamic frameworks like DynPerturb (Qin et al., 2025) attempt temporal interventions, they operate primarily as abstract latent shifts without decoding full-resolution spatial expression profiles. STEP bridges this gap by unifying a mask-guided editable latent manifold for proximal gene edits with a mass-calibrated spatiotemporal flow field, enabling full-resolution counterfactual propagation across unpaired developmental snapshots.

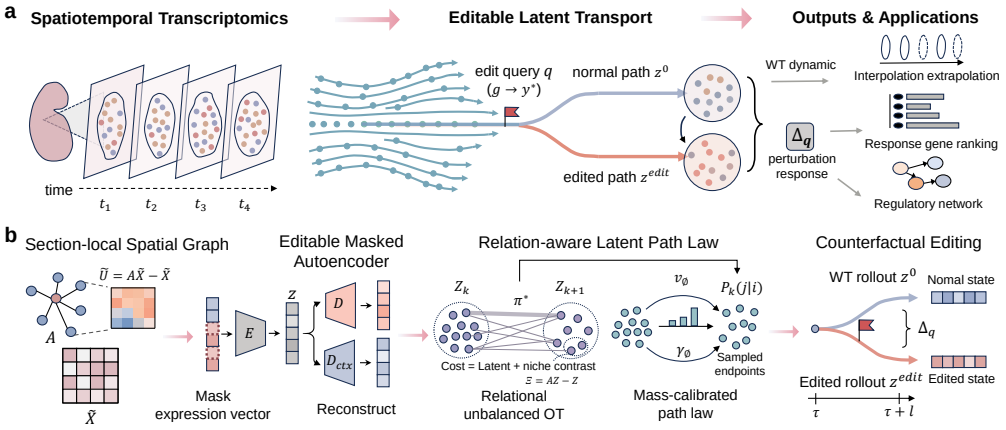


Figure 2: **Overview of STEP. (a)** STEP learns continuous latent dynamic fields from discrete spatiotemporal snapshots, enabling counterfactual perturbation forecasting via targeted node edits. **(b)** STEP combines spatial graphs with editable masked autoencoding to embed cells into an intervention-compatible latent space, driving relation-aware path law learning for spatiotemporal *in silico* perturbations.

3 METHOD

3.1 PROBLEM FORMULATION

Let $\{(\mathbf{X}_k, \mathbf{S}_k)\}_{k=1}^K$ denote spatial transcriptomic (ST) snapshots observed across discrete time points $t_1 < \cdots < t_K$, where stage k measures G genes across N_k spatial locations via an expression matrix $\mathbf{X}_k \in \mathbb{R}_+^{N_k \times G}$ and 2D coordinates $\mathbf{S}_k \in \mathbb{R}^{N_k \times 2}$. Inherently destructive spatial profiling yields temporally unpaired snapshots lacking shared cellular identities or spatial reference frames, with full replicate integration and preprocessing protocols detailed in Appendix H.

Using only wild-type or non-targeting snapshots, we aim to model continuous wild-type tissue evolution and predict its response to a localized gene intervention. Given an onset ST section observed at time τ, we consider an intervention that sets target gene g to expression value $y^\star$ within a spatial region R and maintains this constraint for a prediction horizon ℓ. Let $X^0(t)$ and $X^{\mathrm{edit}}(t)$ denote the wild-type and edited trajectories initialized from the same onset section. The predicted counterfactual response is (Figure 2 and Appendix A)

$$\widehat{\Delta}(\tau, \ell) = X^{\mathrm{edit}}(\tau + \ell) - X^0(\tau + \ell). \tag{1}$$

3.2 Editable Masked Autoencoder

For clarity, we describe representation learning for a single section and suppress section and time indices. This module encodes compact, spatially structured latent states for dynamic modeling while rendering them *editable* to represent gene interventions as controlled local edits rather than out-of-distribution inputs. Given expression matrix $\mathbf{X} \in \mathbb{R}_+^{N\times G}$ and spatial coordinates $\mathbf{S} \in \mathbb{R}^{N\times 2}$, we build a sparse row-normalized weighted kNN graph $\mathbf{A} \in \mathbb{R}^{N\times N}$ (Appendix B.1). After robust gene-wise standardization of $\mathbf{X}$ into $\widetilde{\mathbf{X}}$, we quantify local spatial neighborhood deviation as

$$\widetilde{\mathbf{U}} = \mathbf{A}\widetilde{\mathbf{X}} - \widetilde{\mathbf{X}}. \tag{2}$$

Masking to mimic interventions. To expose the representation to intervention-like missingness, we sample a binary mask $\mathbf{B} \in \{0, 1\}^{N\times G}$ whose nonzero entries indicate masked gene–location pairs. We combine stochastic multi-gene masking with targeted single-gene masking, and apply the same mask consistently to both the expression view and its derived spatial-context view:

$$\widetilde{\mathbf{X}}^B = \widetilde{\mathbf{X}} \odot (\mathbf{1} - \mathbf{B}), \quad \widetilde{\mathbf{U}}^B = \mathbf{A}\widetilde{\mathbf{X}}^B - \widetilde{\mathbf{X}}^B. \tag{3}$$

Encoder with explicit mask awareness. The encoder E_η operates row-wise across the concatenated expression, spatial context, and mask features with shared parameters, mapping an N-unit section to an $N \times d$ latent matrix. Each row of $\mathbf{Z}^B$ (or $\mathbf{Z}^{\mathrm{clean}}$) is the latent state of an individual measured unit; the full matrix is the section-level latent representation

$$\mathbf{Z}^B = E_\eta\left(\mathrm{concat}[\widetilde{\mathbf{X}}^B, \widetilde{\mathbf{U}}^B, \mathbf{B}]\right) \in \mathbb{R}^{N\times d}, \qquad \mathbf{Z}^{\mathrm{clean}} = E_\eta\left(\mathrm{concat}[\widetilde{\mathbf{X}}, \widetilde{\mathbf{U}}, \mathbf{0}]\right) \in \mathbb{R}^{N\times d}. \tag{4}$$

Decoding and objective. Decoders D_η and D_η^{ctx} reconstruct standardized expression and local context, respectively. Model training combines (i) a masked reconstruction objective regularizing unmasked features, (ii) a clean context reconstruction term preserving spatial topology, and (iii) a weak consistency loss aligning masked and clean encodings without enforcing identity.

$$\begin{aligned}\mathcal{L}_{\mathrm{rep}} = \lambda_{\mathrm{mask}} &\frac{\left\|\mathbf{B} \odot \left(D_\eta(\mathbf{Z}^B) - \widetilde{\mathbf{X}}\right)\right\|_F^2}{\|\mathbf{B}\|_1} + \frac{\left\|(\mathbf{1} - \mathbf{B}) \odot \left(D_\eta(\mathbf{Z}^B) - \widetilde{\mathbf{X}}\right)\right\|_F^2}{\|\mathbf{1} - \mathbf{B}\|_1} \\ &+ \lambda_{\mathrm{ctx}} \left\|D_\eta^{\mathrm{ctx}}\left(\mathbf{Z}^{\mathrm{clean}}\right) - \widetilde{\mathbf{U}}\right\|_F^2 + \lambda_{\mathrm{con}} \left\|\mathbf{Z}^B - \mathbf{Z}^{\mathrm{clean}}\right\|_F^2.\end{aligned} \tag{5}$$

After fitting $\hat{\eta}$, the clean encoder is applied independently to every observed stage, yielding $\mathbf{Z}_k \in \mathbb{R}^{N_k\times d}$ for path-law fitting in Section 3.3. The mask-aware branch is retained only for perturbation inference (mask sampling protocols are detailed in Appendix B.2).

3.3 Relation-Aware Latent Path Law

Fixing the pretrained encoder, clean latent snapshots $\mathbf{Z}_k$ define discrete empirical measures over spatial units at stage t_k. We fit a *path law* to transport mass between adjacent stages while preserving both per-unit latent features and spatial relationships encoded by intra-stage graphs $\mathbf{A}_k$. To incorporate spatial relational structure into transport geometry, we define the local latent contrast

$$\mathbf{\Xi}_k = \mathbf{A}_k\mathbf{Z}_k - \mathbf{Z}_k, \tag{6}$$

which measures the deviation of each latent state from its local spatial neighborhood. Let $\mathcal{D}(\mathbf{P}, \mathbf{Q})$ denote the pairwise squared Euclidean-distance matrix between the rows of $\mathbf{P}$ and $\mathbf{Q}$, normalized by its median. For adjacent snapshots, the relation-aware transport cost is

$$\mathbf{C}_k = \underbrace{\mathcal{D}\left(\mathbf{Z}_k, \mathbf{Z}_{k+1}\right)}_{\text{latent-state mismatch}} + \lambda_{\text{rel}} \underbrace{\mathcal{D}\left(\mathbf{\Xi}_k, \mathbf{\Xi}_{k+1}\right)}_{\text{local-context mismatch}} . \tag{7}$$

Each snapshot $\mathbf{Z}_k$ carries an empirical mass vector $\mathbf{m}_k \in \mathbb{R}_+^{N_k}$ representing relative unit abundance (Appendix B.3) to capture stage-to-stage population changes. Given cost matrix $\mathbf{C}_k$ and marginal masses $(\mathbf{m}_k, \mathbf{m}_{k+1})$, we estimate the stage transition coupling via unbalanced optimal transport

$$\begin{aligned} \mathbf{\Pi}_k^\star = \underset{\mathbf{\Pi} \geq 0}{\arg\min} \quad & \underbrace{\langle \mathbf{\Pi}, \mathbf{C}_k \rangle}_{\text{transport cost}} + \underbrace{\epsilon_{\text{OT}} D_{\text{KL}}\left(\mathbf{\Pi} \,\middle\|\, \mathbf{m}_k \mathbf{m}_{k+1}^\top\right)}_{\text{coupling regularization}} \\ & + \lambda_{\text{marg}} \underbrace{\left[D_{\text{KL}}\left(\mathbf{\Pi}\mathbf{1} \,\|\, \mathbf{m}_k\right) + D_{\text{KL}}\left(\mathbf{\Pi}^\top \mathbf{1} \,\middle\|\, \mathbf{m}_{k+1}\right)\right]}_{\text{marginal relaxation}} . \end{aligned} \tag{8}$$

Here, $\mathbf{1}$ denotes an all ones vector of conformable dimension, and $D_{\text{KL}}(\mathbf{a} \,\|\, \mathbf{b}) = \sum_i \left(a_i \log(a_i/b_i) - a_i + b_i\right)$ denotes generalized KL divergence applied elementwise. Product measure $\mathbf{m}_k \mathbf{m}_{k+1}^\top$ serves as the entropic reference, while marginal relaxation terms accommodate mass shifts across stages (Appendix B.3). From the fitted coupling, we extract source marginal $\boldsymbol{\rho}_k$ to compute calibrated per source log-growth target $\boldsymbol{\gamma}_k^\star$

$$\boldsymbol{\rho}_k = \mathbf{\Pi}_k^\star \mathbf{1}, \quad \boldsymbol{\gamma}_k^\star = \frac{1}{\Delta t_k} \log\left[\frac{\sum_{n=1}^{N_{k+1}} m_{n,k+1}}{\sum_{n=1}^{N_k} \rho_{n,k}} \cdot \left(\boldsymbol{\rho}_k \oslash \mathbf{m}_k\right)\right], \quad \Delta t_k = t_{k+1} - t_k. \tag{9}$$

where $\oslash$ and $\log$ operate element-wise. Source–target endpoint pairs $(\mathbf{z}_{\text{src}}, \boldsymbol{\xi}_{\text{src}})$ and $(\mathbf{z}_{\text{tgt}}, \boldsymbol{\xi}_{\text{tgt}})$ are sampled proportionally to the joint probabilities in $\mathbf{\Pi}_k^\star$. Details of the conditional endpoint sampler and mass-calibration identity are provided in Appendices B.4 and B.3. For a uniform time step $u \sim \mathcal{U}(0, 1)$, conditional flow matching defines

$$\mathbf{z}_u = (1-u)\mathbf{z}_{\text{src}} + u\mathbf{z}_{\text{tgt}}, \quad \boldsymbol{\xi}_u = (1-u)\boldsymbol{\xi}_{\text{src}} + u\boldsymbol{\xi}_{\text{tgt}}, \quad \mathbf{v}^\star = \frac{\mathbf{z}_{\text{tgt}} - \mathbf{z}_{\text{src}}}{\Delta t_k}. \tag{10}$$

Each sampled source state inherits its growth target $\gamma^\star \in \boldsymbol{\gamma}_k^\star$. With normalized physical time $\bar{t} = (t - t_1)/(t_K - t_1)$, a single unified network parameterizes both latent velocity and log mass growth rate, simultaneously modeling state evolution and population dynamics

$$\left(\mathbf{v}_\phi\left(\mathbf{z}, \boldsymbol{\xi}, \bar{t}\right), \gamma_\phi\left(\mathbf{z}, \boldsymbol{\xi}, \bar{t}\right)\right) = \text{NN}_\phi\left(\text{concat}\left[\mathbf{z}, \boldsymbol{\xi}, \bar{t}, \sin(\pi\bar{t}), \cos(\pi\bar{t})\right]\right). \tag{11}$$

For endpoint regularization, we define

$$\widehat{\mathbf{z}}_{\text{tgt}} = \mathbf{z}_{\text{src}} + \Delta t_k \mathbf{v}_\phi\left(\mathbf{z}_{\text{src}}, \boldsymbol{\xi}_{\text{src}}, \bar{t}_k\right). \tag{12}$$

The path-law objective is

$$\begin{aligned} \mathcal{L}_{\text{path}} = \mathbb{E}\left[\left\|\mathbf{v}_\phi\left(\mathbf{z}_u, \boldsymbol{\xi}_u, \bar{t}\right) - \mathbf{v}^\star\right\|_2^2\right] + \lambda_{\text{gr}} \mathbb{E}\left[\ell_{\text{H}}\left(\gamma_\phi\left(\mathbf{z}_u, \boldsymbol{\xi}_u, \bar{t}\right), \gamma^\star\right)\right], \\ + \lambda_{\text{end}} \mathbb{E}\left[\left\|\widehat{\mathbf{z}}_{\text{tgt}} - \mathbf{z}_{\text{tgt}}\right\|_2^2\right], \\ + \lambda_{\text{mom}}\left(\left\|\mathbb{E}\left[\widehat{\mathbf{z}}_{\text{tgt}}\right] - \mathbb{E}\left[\mathbf{z}_{\text{tgt}}\right]\right\|_2^2 + \left\|\text{Var}\left[\widehat{\mathbf{z}}_{\text{tgt}}\right] - \text{Var}\left[\mathbf{z}_{\text{tgt}}\right]\right\|_2^2\right). \end{aligned} \tag{13}$$

Where ℓ_{H} denotes the Huber loss. When optimizing ϕ, both encoder parameters and transport couplings remain fixed, implementation details are deferred to Appendix B.5.

3.4 DECODER-CONSISTENT GENE EDITING

Mask-aware anchor at intervention onset. With pretrained representations and wild-type path laws frozen, an intervention setting gene g to $y^\star$ in spatial region R at time τ is defined by binary mask $\mathbf{M}_{g,R} \in \{0,1\}^{N\times G}$. To maintain the distribution features at intervention onset, we reuse the masking channel from Section 3.2, zeroing targeted standardized expression entries and encoding with explicit input mask features

$$\begin{aligned}\widetilde{\mathbf{X}}_\tau^{\mathrm{mask}} &= \widetilde{\mathbf{X}}_\tau \odot (\mathbf{1} - \mathbf{M}_{g,R}),\\ \mathbf{Z}_\tau^{\mathrm{mask}} &= E_{\hat{\eta}}\left(\mathrm{concat}\left[\widetilde{\mathbf{X}}_\tau^{\mathrm{mask}}, \mathbf{A}_\tau\widetilde{\mathbf{X}}_\tau^{\mathrm{mask}} - \widetilde{\mathbf{X}}_\tau^{\mathrm{mask}}, \mathbf{M}_{g,R}\right]\right).\end{aligned} \tag{14}$$

Decoder-consistent proximal projection. Directly clamping target genes risks displacing latent states off the decoder manifold, causing unrealistic rollouts under frozen wild-type vector fields. We define a decoder-consistent proximal projection operator $\mathcal{P}$ enforcing desired edits on the *decoder expression scale* within specified footprints while anchoring latent states near input representations. Taking $\mathbf{Z}^0(\tau) = \mathbf{Z}_\tau^{\mathrm{mask}}$, the edited initial state $\mathbf{Z}^{\mathrm{edit}}(\tau) = \mathcal{P}(\mathbf{Z}^0(\tau))$ is computed as (Appendix C)

$$\mathcal{P}(\mathbf{Z}) = \arg\min_{\mathbf{Z}'}\left[\frac{\lambda_{\mathrm{tgt}}}{\|\mathbf{M}_{g,R}\|_1}\left\|\mathbf{M}_{g,R}\odot\left(D_{\hat{\eta}}(\mathbf{Z}') - y^\star\mathbf{1}\right)\right\|_F^2 + \frac{\lambda_{\mathrm{prox}}}{Nd}\|\mathbf{Z}'-\mathbf{Z}\|_F^2\right], \tag{15}$$

Projected rollout under frozen wild-type dynamics. Let $\Phi_{\hat{\phi}}^{t\to t+h}$ denote one numerical integration step of size h under the frozen wild-type path law with fixed spatial coordinates. For continuous time $t \geq \tau$, wild-type and edited latent trajectories evolve recursively via

$$\begin{aligned}\mathbf{Z}^0(t+h) &= \Phi_{\hat{\phi}}^{t\to t+h}\left(\mathbf{Z}^0(t)\right),\\ \mathbf{Z}^{\mathrm{edit}}(t+h) &= \mathcal{P}\left(\Phi_{\hat{\phi}}^{t\to t+h}\left(\mathbf{Z}^{\mathrm{edit}}(t)\right)\right),\end{aligned} \tag{16}$$

Reapplying $\mathcal{P}$ at each integration step enforces projected decoder-consistent transport, preserving intervention constraints throughout rollout. At evaluation horizon ℓ, the predicted counterfactual response program is decoded as

$$\widehat{\boldsymbol{\Delta}} = D_{\hat{\eta}}\left(\mathbf{Z}^{\mathrm{edit}}(\tau+\ell)\right) - D_{\hat{\eta}}\left(\mathbf{Z}^0(\tau+\ell)\right). \tag{17}$$

Numerical integration schemes and finite-step proximal updates are detailed in Appendix B.6, while the sensitivity readout derived from the frozen path law is provided in Appendix B.7.

4 EXPERIMENTAL SETUP

Tasks and datasets. We benchmark STEP along four complementary axes: (i) **Mechanistic correctness** on synthetic scMultiSim systems with known ground truth (Appendix E). (ii) **Wild-type spatiotemporal dynamics** via trajectory interpolation and forecasting on the multi-stage MOSTA and ARTISTA atlases. (iii) **Counterfactual perturbation prediction** in both (a) single-slice assays (HT29 Perturb-DBiT; KP Perturb-map) and (b) multi-slice temporal snapshots (SPAC-seq). (iv) **Mechanistic analysis** that isolates the contributions of latent-space editing and proximal projection to *in silico* perturbation quality. We further include a zebrafish embryogenesis case study (ZESTA) to probe **heterogeneous spatiotemporal interventions** across developmental stages and lineages. Dataset provenance, preprocessing, and fold splits are provided in Appendix H.

Comparators and metrics. We compare against three representative model classes: (1) wild-type snapshot dynamics models (stVCR, DeepRUOT) (Peng et al., 2026; Zhang et al., 2025); (2) static spatial counterfactual models (SpatialProp, Celcomen) (Sun et al., 2025; Megas et al., 2025); and (3) dynamic counterfactual perturbation models (DynPerturb) (Qin et al., 2025) together with a matched direct-clamp ablation. We evaluate pointwise trajectory accuracy using temporal change correlation, RMSE, temporal DE AUPRC, and Sliced Wasserstein Distance (SWD). For perturbation prediction, response gene AUPRC serves as the primary endpoint, complemented by signed Spearman correlations. Appendices J and K detail baseline implementations and formal metric definitions.

Table 1: Wild-type trajectory benchmark on multi-stage tissue atlases (ARTISTA and MOSTA). Values are means across leave-one-stage interpolation folds (ARTISTA: 3; MOSTA: 6) and expanding-window one-step forecasting folds (ARTISTA adult: 1; MOSTA: 4).

Dataset/task	Method	Pearson ↑	Spearman ↑	RMSE ↓	DE AUPRC ↑	SWD ↓
ARTISTA interpolation (3-fold)	STEP	**0.5505**	**0.4939**	**0.1372**	**0.4431**	0.3430
	stVCR (Nat.Met'26)	0.4998	0.4436	0.2742	0.4013	0.4818
	DeepRUOT (ICLR'25)	0.1701	0.1329	0.3507	0.1602	**0.3426**
ARTISTA adult one-step	STEP	**−0.1058**	**0.0103**	**0.1162**	**0.1855**	0.4176
	stVCR	−0.2811	−0.2889	0.2317	0.0859	0.4353
	DeepRUOT	−0.1816	−0.2575	0.2589	0.0966	**0.3566**
MOSTA interpolation (6-fold)	STEP	0.4486	**0.4629**	**0.0700**	**0.3482**	0.2809
	stVCR	0.4960	0.4318	0.1344	0.2462	0.3586
	DeepRUOT	**0.5408**	0.4255	0.1565	0.1565	**0.2637**
MOSTA rolling one-step (4-fold)	STEP	**0.1326**	**0.2552**	**0.0987**	**0.3222**	0.2985
	stVCR	−0.0261	0.0528	0.2439	0.2336	0.3897
	DeepRUOT	0.0450	0.0023	0.2218	0.1216	**0.2659**

Table 2: Task-specific ablations isolating key components. Top: transport fidelity on stVCR observed snapshots. Bottom: perturbation response prioritization on a controlled regional-knockout simulator.

Ablation	Variant	Readouts			
Flow objective		Pearson ↑	RMSE ↓	SWD ↓	Count err. ↓
	full objective	**0.5385**	**0.0125**	**0.1052**	**0.0231**
	no mass/endpoint terms	0.5381	0.0378	0.1968	0.0825
Masked-gene SSL		Effect Pearson ↑	Gene Spearman ↑	Resp. AUPRC ↑	RMSE ↓
	full objective	**0.5708**	**0.4523**	**0.5911**	**0.2041**
	no masked-gene SSL	0.4568	0.4359	0.3042	0.2111

5 RESULTS

5.1 WILD-TYPE DYNAMICS BENCHMARK

Interpolation Precision. On interpolation tasks, STEP exhibits strong performance in capturing directional expression dynamics and prioritizing temporally dynamic genes (Table 1). Across both ARTISTA and MOSTA atlases, STEP leads across rank-correlation and temporal-DE metrics while halving the trajectory reconstruction error relative to stVCR. This demonstrates its strong capacity to accurately interpolate intermediate cell-state transitions rather than merely smoothing state spaces.

Extrapolation Resilience. The superiority of STEP becomes even more pronounced under temporal extrapolation. While baseline models collapse to near-zero or negative correlations when forecasting unobserved time steps, STEP maintains positive rank alignment and cuts magnitude errors by over half. Crucially, under severe developmental regime shifts (e.g., ARTISTA adult forecasting), where competing baselines suffer from strong anti-correlation, STEP mitigates directional drift, standing as the sole approach to deliver positive rank alignment and top dynamic gene prioritization.

Distributional Trade-offs. Notably, DeepRUOT maintains competitive SWD across several benchmarks, reflecting the efficacy of its direct optimal-transport coupling objective in preserving marginal manifold geometry. However, while DeepRUOT prioritizes global spatial overlap, STEP successfully reconciles comparable distributional alignment with vastly superior dynamic fidelity and differential gene directionality.

Targeted ablations confirm that mass calibration selectively rectifies spatial boundary geometry and population scale, while masked-gene representation learning is essential to prevent off-manifold drift and enable accurate perturbation response prioritization (Table 2 and Appendix F).

5.2 IN VIVO VALIDATION OF STATIC SPATIAL COUNTERFACTUALS

We first evaluate spatial counterfactual inference on single static ST slices, including the HT29 Perturb-DBiT (Baysoy et al., 2026) and mouse KP Perturb-map (Dhainaut et al., 2022) assays.

Macro Response Prioritization. On the HT29 atlas (Figure 3a–d), STEP consistently leads competing baselines and direct-clamp ablations in macro response-gene AUPRC. Target-resolved evaluations confirm that prioritization gains span diverse perturbations, with particularly pronounced recovery for *S100A4* and *ATG4A* targets. Near-zero signed Spearman correlations across all evaluated approaches (Figure 3c) suggest that unperturbed reference tissue architecture primarily governs response salience rather than fine-grained downstream expression directionality.

Target- and section-level recovery. On mouse KP Perturb-map (Figure 3e–h), STEP again achieves the strongest response-gene prioritization across knockout targets. For *Tgfbr2*, it nearly doubles response-gene AUPRC relative to SpatialProp and direct-clamp controls (Figure 3f). Averaged over all targets, STEP maintains a clear lead in mean AUPRC (Figure 3h) and remains comparatively robust across tissue sections (Figure 3g).

Why decoder-consistent latent editing matters. Because the direct-clamp ablation shares an identical frozen backbone with STEP, their performance divergence isolates the specific mechanistic necessity of manifold-constrained latent editing (Figure 3b,f). Direct count clamping forces target genes to zero in ambient expression space, generating off-manifold artifacts that collapse downstream response amplitudes near zero. In contrast, STEP applies decoder-consistent proximal projection at initialization and across rollout steps, iteratively eliminating latent residual drift to recover true response amplitudes and signed effect correlations (Figure 6). This mechanistic validation confirms that maintaining latent manifold consistency is indispensable for faithful counterfactual prioritization from unperturbed spatial references.

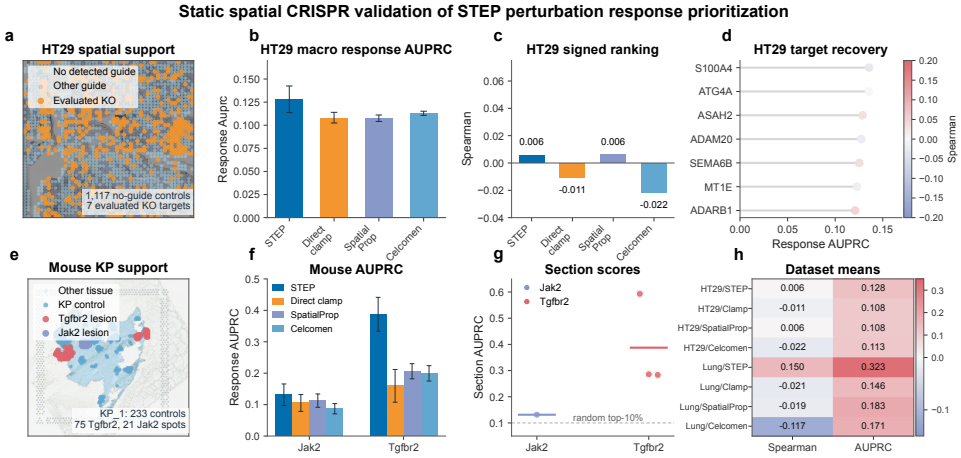


Figure 3: **Static spatial CRISPR benchmark validation. (a–d)** HT29 Perturb DBiT evaluation showing spatial layouts with non-targeting control spots, macro response gene AUPRC across 10 runs, signed Spearman correlations, and target-resolved metrics for STEP versus baselines. **(e–g)** Mouse KP Perturb map sections showing control regions and *Tgfbr2* and *Jak2* lesions, target-level response gene AUPRC over 10 runs, and section-resolved performance with individual section points overlaid on target mean bars. **(h)** Cross-dataset summary comparing mean signed Spearman correlation against response gene AUPRC.

5.3 Spatiotemporal Forecasting of In Vivo Dynamic Interventions

We next evaluate dynamic perturbation forecasting on the spatiotemporal SPAC-seq CRISPR assay (Zhang et al., 2026), predicting Day 7 interventional profiles from unperturbed Day 4 $\rightarrow$ Day 7 transitions without exposure to guide-bearing training data (Figure 4a).

Dynamic counterfactual recovery and mechanistic insight. Across all evaluated targets, STEP substantially outperforms DynPerturb and direct-clamp baselines in mean response-gene AUPRC (Figure 4b,c), with Precision@K metrics confirming robust prioritization of top-decile response targets (Figure 4d). Mechanistic analysis on *Klrd1* knockouts clarifies this performance gap: while direct coordinate clamping induces severe off-manifold trajectory drift and cumulative residual buildup during forward integration, mask-aware encoding and decoder-consistent projection anchor

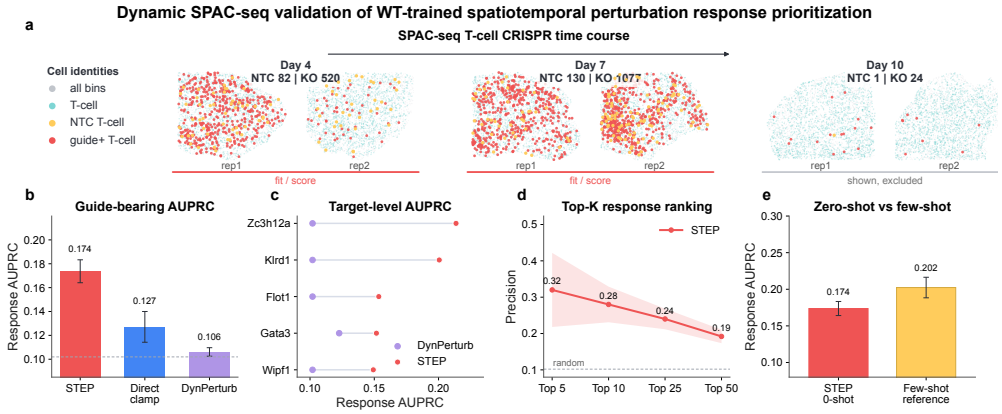


Figure 4: **Spatiotemporal forecasting of dynamic SPAC seq CRISPR interventions. (a)** SPAC-seq section layouts across Days 4, 7, and 10 highlighting control and guide-bearing regions. **(b–d)** Target-level evaluation over 10 runs comparing STEP to DynPerturb and direct clamp baselines, reporting mean response gene AUPRC alongside random baseline, target resolved AUPRC, and mean Precision@K with shaded SEM. **(e)** Zero-shot STEP compared against a perturbation informed reference calibrated on 50% of Day 7 interventional profiles.

perturbed trajectories to the valid cell-state manifold throughout temporal evolution (Figure 7). By suppressing residual accumulation across integration steps, STEP preserves initial edit signals and recovers true response amplitudes alongside signed directional structure.

Comparison with perturbation-supervised references. To contextualize zero-shot capabilities, we evaluate STEP against a semi-supervised reference calibrated on 50% of Day 7 perturbation profiles (Figure 4e). Remarkably, despite receiving zero perturbation-label supervision during representation fitting, zero-shot STEP closely approaches the performance of this perturbation-informed baseline. This outcome confirms that transporting unperturbed initial states through decoder-consistent latent edits effectively recovers dynamic interventional profiles without requiring perturbation labels.

5.4 Spatiotemporal Dynamic GRNs Explain Stage-Specific Interventions

Evaluating STEP on the ZESTA zebrafish atlas demonstrates its practical value in complex biological contexts (Figure 8). *In silico* knockdowns of pioneer factors (*pou5f3*, *nanog*, *sox19b*) reveal pronounced spatiotemporal response heterogeneity, with edit magnitudes attenuating by over an order of magnitude across development. Crucially, local Jacobian sensitivities confirm that this stage-dependent efficacy is driven by dynamic gene regulatory network rewiring and out-degree collapse, directly linking evolving network topology to counterfactual responsiveness (Appendix G).

6 Conclusion and Discussion

Broader impact. STEP unifies spatiotemporal dynamics modeling with *in silico* perturbation through an editable, measure-preserving latent path law. This enables timing-specific counterfactual forecasting across early and late developmental stages, advancing the emerging "world model" paradigm for biomedicine (Noori et al., 2026). By predicting interventional responses from unperturbed wild-type tissues, STEP supports zero-shot identification of time-sensitive therapeutic targets and can narrow the search space for *in vivo* spatial CRISPR screens.

Limitations and Future Directions. Two key avenues remain for future exploration: **(i) Disambiguating perturbation directionality.** Inferring whether interventions activate or repress downstream programs purely from wild-type snapshots is challenging, which may be mitigated by incorporating signed, cell-type-specific GRN topology as soft priors alongside data-driven sensitivity. **(ii) Regime adaptation for long-horizon forecasting.** Abrupt mechanistic transitions amplify multi-step extrapolation errors, necessitating regime shift detection, adaptive path law updates, and calibrated uncertainty quantification for long-horizon intervention design.

AI USE STATEMENT

Generative AI tools were utilized for language editing, grammar refinement, and manuscript formatting. All numerical outputs, analytical results, and scientific claims were independently verified by the authors, who accept full responsibility for the accuracy and integrity of the final work.

ETHICS STATEMENT

This study exclusively analyzes publicly available, fully de-identified spatial transcriptomics datasets. Consequently, no ethical approval or informed consent was required, as no direct human subject or animal experimentation was performed.

REPRODUCIBILITY STATEMENT

Source code is publicly accessible at `https://github.com/Anonymous-2026-coding/STEP`. All experiments were implemented in Python 3.10 using PyTorch 2.2.2, Python Optimal Transport (POT) 0.9.5, and standard scientific computing stacks including NumPy, SciPy, scikit-learn, pandas, and matplotlib. Full dataset provenance, evaluation protocols, and network architectures are detailed in Appendices H and I.

REFERENCES

Alev Baysoy, Xiaolong Tian, Paul Renauer, Feifei Zhang, Zhiliang Bai, Hao Shi, Mingyu Yang, Dingyao Zhang, Miao Liu, Haikuo Li, et al. Large-scale, spatially resolved panoramic crispr screening in native tissue environments using perturb-dbit. *Nature Biotechnology*, pp. 1–14, 2026.

Ozren Bogdanović, Ana Fernandez-Miñán, Juan J Tena, Elisa de la Calle-Mustienes, Carmen Hidalgo, Ila van Kruysbergen, Simon J van Heeringen, Gert Jan C Veenstra, and José Luis Gómez-Skarmeta. Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Research*, 22(10):2043–2053, 2012.

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.

Maxime Dhainaut, Samuel A Rose, Guray Akturk, et al. Spatial crispr genomics identifies regulators of the tumor microenvironment. *Cell*, 185(7):1223–1239, 2022.

Jeffrey A Farrell, Yiqun Wang, Samantha J Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392):eaar3131, 2018.

Kenji Kamimoto, Blerta Stringa, Christy M Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, 2023.

Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Lama Saber, Changying Jing, et al. Mapping cells through time and space with moscot. *Nature*, 638(8052):1065–1075, 2025.

Miler T Lee, Ashley R Bonneau, Carter M Takacs, Ariel A Bazzini, Kate R DiVito, Elizabeth S Fleming, and Antonio J Giraldez. Nanog, pou5f1 and soxb1 activate zygotic gene expression during the maternal-to-zygotic transition. *Nature*, 503(7476):360–364, 2013.

Manuel Leichsenring, Julia Maes, Rebecca Mössner, Wolfgang Driever, and Daria Onichtchouk. Pou5f1 transcription factor controls zygotic gene activation in vertebrates. *Science*, 341(6149): 1005–1009, 2013.

Hechen Li, Ziqi Zhang, Michael Squires, Xi Chen, and Xiuwei Zhang. scmultisim: simulation of single-cell multi-omics and spatial data guided by gene regulatory networks and cell–cell interactions. *Nature Methods*, 22(5):982–993, 2025.

Chang Liu, Rui Li, Young Li, Xiumei Lin, Kaichen Zhao, Qun Liu, Shuowen Wang, Xueqian Yang, Xuyang Shi, Yuting Ma, et al. Spatiotemporal mapping of gene expression landscapes and developmental trajectories during zebrafish embryogenesis. *Developmental Cell*, 57(10): 1284–1298, 2022.

Stathis Megas, Daniel Chen, Krzysztof Polanski, Moshe Eliasof, Carola-Bibiane Schönlieb, and Sarah Teichmann. Estimation of single-cell and tissue perturbation effect in spatial transcriptomics via spatial causal disentanglement. In *International Conference on Learning Representations*, volume 2025, pp. 100192–100217, 2025.

Alex Morehead, Lazar Atanackovic, Akshata Hegde, Yanli Wang, Frimpong Boadu, Joel Selvaraj, Alexander Tong, Aditi Krishnapriyan, and Jianlin Cheng. Flow matching for generative modelling in bioinformatics and computational biology. *Nature Machine Intelligence*, pp. 1–18, 2026.

Ayush Noori, Nic Fishman, Ada Fang, Lukas Fesser, and Marinka Zitnik. World models for biomedicine. *Cell*, 189(19):5845–5869, 2026.

Ertuğrul M Özbudak, Olivier Tassy, and Olivier Pourquié. Spatiotemporal compartmentalization of key physiological processes during muscle precursor differentiation. *Proceedings of the National Academy of Sciences*, 107(9):4224–4229, 2010.

Máté Pálfy, Gunnar Schulze, Eivind Valen, and Nadine L Vastenhouw. Chromatin accessibility established by pou5f3, sox19b and nanog primes genes for activity during zebrafish genome activation. *PLoS Genetics*, 16(1):e1008546, 2020.

Qiangwei Peng, Peijie Zhou, and Tiejun Li. stvcr: spatiotemporal dynamics of single cells. *Nature Methods*, 23(3):542–553, 2026.

Hua Qin, Yilin Zhang, Zhihan Guo, Luni Hu, Qingsong Li, Lei Cao, Tianyi Xia, Ziqing Deng, Yong Zhang, Yuxiang Li, et al. Dynperturb: Dynamic perturbation modeling for spatiotemporal single-cell systems. *bioRxiv*, pp. 2025–09, 2025.

Xiaojie Qiu, Daniel Y Zhu, Yifan Lu, Jiajun Yao, Zehua Jing, Kyung Hoi Min, Mengnan Cheng, Hailin Pan, Lulu Zuo, Samuel King, et al. Spatiotemporal modeling of molecular holograms. *Cell*, 187(26):7351–7373, 2024.

Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, 2019.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Kimberle Shen, Wan Yi Seow, Choong Tat Keng, Michelle Gek Liang Lim, Daryl Shern Lim, Ke Guo, Amine Meliani, Muhammad Irfan Bin Hajis, Bolun Wang, Shyam Prabhakar, et al. Spatial perturb-seq: single-cell functional genomics within intact tissue architecture. *Nature Communications*, 17 (1):3018, 2026.

Xiaoyu Song, Joselyn C Chavez-Fuentes, Weiping Ma, Weijia Fu, Sujung Crystal Shin, Pei Wang, and Guo-Cheng Yuan. sccigen: a high-fidelity spatially resolved transcriptomics data simulator for cell–cell interaction studies. *Genome Biology*, 26(1):354, 2025.

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

Eric D Sun, Alejandro Buendia, Anne Brunet, and James Zou. Spatialprop: tissue perturbation modeling with spatially resolved single-cell transcriptomics. *bioRxiv*, pp. 2025–11, 2025.

Daniel E Wagner, Caleb Weinreb, Zach M Collins, James A Briggs, Sean G Megason, and Allon M Klein. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392):981–987, 2018.

Xiaoyu Wei, Sulei Fu, Hanbo Li, Yang Liu, Shuai Wang, Weimin Feng, Yunzhi Yang, Xiawei Liu, Yan-Yun Zeng, Mengnan Cheng, et al. Single-cell stereo-seq reveals induced progenitor cells involved in axolotl brain regeneration. *Science*, 377(6610):eabp9444, 2022.

Johann Wenckstern, Eeshaan Jain, Benedikt von Querfurth, Yexiang Cheng, Kiril Vasilev, Matteo Pariset, Phil F Cheng, Petros Liakopoulos, Olivier Michielin, Andreas Wicki, et al. The virtual tissues foundation model resolves spatial proteomics across scales. *Nature*, pp. 1–11, 2026.

Haorui Zhang, Zongxu Zhang, Peiyu Wang, Tian Xu, Xiaoyu Chen, Yanping Zhao, Siyu Lin, Wenjie Cai, Pengfei Ren, Ce Luo, et al. Uncovering spatially resolved functional genomics with crispr screen sequencing. *Cell*, 189(15):4594–4618, 2026.

Zhenyi Zhang, Tiejun Li, and Peijie Zhou. Learning stochastic dynamics from snapshots through regularized unbalanced optimal transport. In *International Conference on Learning Representations*, volume 2025, pp. 19888–19919, 2025.

A Algorithm Workflow

Table 3: **Algorithm 1.** STEP training and projected counterfactual rollout.

Require: unpaired snapshots $\{(\mathbf{X}_k, \mathbf{S}_k, \mathbf{m}_k)\}_{k=1}^{K}$, graphs $\{\mathbf{A}_k\}_{k=1}^{K}$, query $q = (g, R, y^\star, \tau, \ell)$, integration step h.

Ensure: frozen $(\widehat{\eta}, \{\mathbf{\Pi}_k^\star\}, \widehat{\phi})$ and counterfactual response $\widehat{\mathbf{\Delta}}_q$.

Training stage

1: $\widetilde{\mathbf{X}}_k \leftarrow \mathrm{RobustScale}(\mathbf{X}_k), \quad \widetilde{\mathbf{U}}_k \leftarrow \mathbf{A}_k\widetilde{\mathbf{X}}_k - \widetilde{\mathbf{X}}_k$ ▷ robust expression and spatial-context inputs.

2: $\widehat{\eta} \leftarrow \arg\min_\eta \mathcal{L}_{\mathrm{rep}}(\eta), \quad \mathbf{Z}_k \leftarrow E_{\widehat{\eta}}(\mathrm{concat}[\widetilde{\mathbf{X}}_k, \widetilde{\mathbf{U}}_k, \mathbf{0}])$ ▷ masked representation learning; retain clean latents.

3: **for** $k = 1, \ldots, K-1$ **do**

4: $\mathbf{\Xi}_k \leftarrow \mathbf{A}_k\mathbf{Z}_k - \mathbf{Z}_k, \quad \mathbf{C}_k \leftarrow \mathcal{D}(\mathbf{Z}_k, \mathbf{Z}_{k+1}) + \lambda_{\mathrm{rel}}\mathcal{D}(\mathbf{\Xi}_k, \mathbf{\Xi}_{k+1})$ ▷ state–relation cost.

5: $\mathbf{\Pi}_k^\star \leftarrow \mathrm{UOT}_{\epsilon_{\mathrm{OT}}, \lambda_{\mathrm{marg}}}(\mathbf{C}_k; \mathbf{m}_k, \mathbf{m}_{k+1})$ ▷ KL-relaxed coupling allows stage-wise mass change.

6: $\boldsymbol{\rho}_k \leftarrow \mathbf{\Pi}_k^\star\mathbf{1}, \quad \boldsymbol{\gamma}_k^\star \leftarrow \Delta t_k^{-1}\log\Big[\frac{\mathbf{1}^\top\mathbf{m}_{k+1}}{\mathbf{1}^\top\boldsymbol{\rho}_k}(\boldsymbol{\rho}_k \oslash \mathbf{m}_k)\Big], \quad w_{i,k} \leftarrow \frac{\rho_{i,k}}{N_k^{-1}\sum_r \rho_{r,k}}$ ▷ calibrate growth and retain row weights.

7: **for each source row** i **with weight** $w_{i,k}$**:** $\quad j \sim \mathsf{P}_k(\cdot \mid i)$, $\mathsf{P}_k(j \mid i) = [\mathbf{\Pi}_k^\star]_{ij}/\rho_{i,k}$; $u \sim \mathcal{U}(0,1)$.

8: $(\mathbf{z}_{\mathrm{src}}, \boldsymbol{\xi}_{\mathrm{src}}, \mathbf{z}_{\mathrm{tgt}}, \boldsymbol{\xi}_{\mathrm{tgt}}, \gamma^\star) \leftarrow (\mathbf{Z}_{k,i}, \mathbf{\Xi}_{k,i}, \mathbf{Z}_{k+1,j}, \mathbf{\Xi}_{k+1,j}, \gamma_{i,k}^\star)$; $\mathbf{z}_u \leftarrow (1-u)\mathbf{z}_{\mathrm{src}} + u\mathbf{z}_{\mathrm{tgt}}$; $\boldsymbol{\xi}_u \leftarrow (1-u)\boldsymbol{\xi}_{\mathrm{src}} + u\boldsymbol{\xi}_{\mathrm{tgt}}$.

9: $\mathbf{v}^\star \leftarrow (\mathbf{z}_{\mathrm{tgt}} - \mathbf{z}_{\mathrm{src}})/\Delta t_k$ ▷ construct the conditional flow-matching targets.

10: **end for**

11: $(v_\phi, \gamma_\phi) \leftarrow \mathrm{NN}_\phi([\mathbf{z}, \boldsymbol{\xi}, \bar{t}, \sin(\pi\bar{t}), \cos(\pi\bar{t})]), \quad \widehat{\phi} \leftarrow \arg\min_\phi \mathcal{L}_{\mathrm{path}}$ ▷ fit velocity, growth, endpoint, and moment terms.

12: $\mathrm{freeze}(\widehat{\eta}, \{\mathbf{\Pi}_k^\star\}, \widehat{\phi})$ ▷ all subsequent edits use the frozen WT path law.

Counterfactual inference

13: $\widetilde{\mathbf{X}}_\tau^{\mathrm{mask}} \leftarrow \widetilde{\mathbf{X}}_\tau \odot (\mathbf{1} - \mathbf{M}_{g,R})$; $\mathbf{Z}_\tau^{\mathrm{mask}} \leftarrow E_{\widehat{\eta}}([\widetilde{\mathbf{X}}_\tau^{\mathrm{mask}}, \mathbf{A}_\tau\widetilde{\mathbf{X}}_\tau^{\mathrm{mask}} - \widetilde{\mathbf{X}}_\tau^{\mathrm{mask}}, \mathbf{M}_{g,R}])$ ▷ encode a location-aware edit anchor.

14: $\mathbf{Z}_\tau^{\mathrm{edit}} \leftarrow \mathcal{P}(\mathbf{Z}_\tau^{\mathrm{mask}})$; $\mathcal{P}(\mathbf{Z}) = \arg\min_{\mathbf{Z}'}\Big[\frac{\lambda_{\mathrm{tgt}}}{\|\mathbf{M}_{g,R}\|_1}\|\mathbf{M}_{g,R} \odot (D_{\widehat{\eta}}(\mathbf{Z}') - y^\star\mathbf{1})\|_F^2 + \frac{\lambda_{\mathrm{prox}}}{Nd}\|\mathbf{Z}' - \mathbf{Z}\|_F^2$ ▷ target fidelity + latent proximity.

15: $\mathbf{Z}^0(\tau) \leftarrow E_{\widehat{\eta}}(\mathrm{concat}[\widetilde{\mathbf{X}}_\tau, \widetilde{\mathbf{U}}_\tau, \mathbf{0}])$ ▷ clean WT state at the same onset.

16: **for** $t = \tau, \tau+h, \ldots, \tau+\ell-h$ **do**

17: $\mathbf{Z}^0(t+h) \leftarrow \Phi_{\widehat{\phi}}^{t\to t+h}(\mathbf{Z}^0(t))$; $\mathbf{Z}^{\mathrm{edit}}(t+h) \leftarrow \mathcal{P}\Big(\Phi_{\widehat{\phi}}^{t\to t+h}(\mathbf{Z}^{\mathrm{edit}}(t))\Big)$ ▷ share WT dynamics and reimpose the edit after each step.

18: **end for**

19: $\widehat{\mathbf{\Delta}}_q \leftarrow D_{\widehat{\eta}}(\mathbf{Z}^{\mathrm{edit}}(\tau+\ell)) - D_{\widehat{\eta}}(\mathbf{Z}^0(\tau+\ell))$ ▷ decoded response at horizon ℓ.

B Additional Method Details

B.1 Spatial Graph Construction

For a stage with N spatial positions, we construct a spatial adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ based on a kNN graph ($k = 8$ by default). Letting $d_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|_2$ and q_d denote the median positive distance across directed kNN edges, unnormalized weights are defined via Gaussian kernel scaling:

$$\widetilde{\mathbf{A}}_{ij} = \exp\left[-\left(\frac{d_{ij}}{\max(q_d, 10^{-8})}\right)^2\right]. \tag{18}$$

We enforce an undirected topology without self-loops by setting $\widetilde{\mathbf{A}} \leftarrow \max(\widetilde{\mathbf{A}}, \widetilde{\mathbf{A}}^\top)$ with a zero diagonal, and row-normalize to obtain $\mathbf{A}$:

$$\mathbf{A}_{ij} = \frac{\widetilde{\mathbf{A}}_{ij}}{\sum_r \widetilde{\mathbf{A}}_{ir}}. \tag{19}$$

Entries outside the symmetrized kNN neighborhood are set to zero. For stages containing multiple tissue replicates, graphs are constructed independently per replicate and combined block-diagonally, preventing spurious cross-replicate spatial edges.

B.2 Representation Preprocessing and Mask Sampling

All normalization statistics are computed on training stages. For each gene g, location c_g and robust scale σ_g are defined as:

$$c_g = \mathrm{median}_i\, \mathbf{X}_{ig}, \qquad \sigma_g = \max\{Q_{0.75}\left(|\mathbf{X}_{ig} - c_g|\right),\, 0.05\}\,, \tag{20}$$

yielding standardized expression $\widetilde{\mathbf{X}}_{ig} = (\mathbf{X}_{ig} - c_g)/\sigma_g$ and spatial context $\widetilde{\mathbf{U}}_{ig} = \mathbf{U}_{ig}/\sigma_g$. The expression decoder in Equation 5 reconstructs preprocessed expression at decoder scale, whereas the context decoder predicts $\widetilde{\mathbf{U}}$.

Training incorporates a two-tier masking protocol. Minibatches are first subjected to independent Bernoulli masking with probability p_{mask}. To train the encoder on targeted knockout queries, a single uniformly selected gene is additionally masked with probability $p_{\mathrm{single}} = \min\{1, \max(0.25, G/32)\}$. In the empirical loss (Equation 5), masked and observed reconstruction terms are normalized by their respective entry counts, while spatial-context and latent-consistency terms are averaged element-wise.

B.3 Empirical Measure Weights

In STEP, "mass" denotes a statistical weight defining the empirical snapshot measure, rather than a physical quantity such as biomass or sequencing depth. Let e index the tissue sections collected at time stage t_k (where $e = 1$ for standard single-section stages). For a raw section e containing $n_{e,k}^{\mathrm{raw}}$ measured units, let $a_{r,e,k}$ denote the design weight of raw unit r ($a_{r,e,k} = 1$ by default, or an explicit sampling multiplicity). Its total raw design mass is

$$\Omega_{e,k}^{\mathrm{raw}} = \sum_{r=1}^{n_{e,k}^{\mathrm{raw}}} a_{r,e,k}. \tag{21}$$

Replicate sections at the same stage contribute additively. To fix a non-dimensional mass unit while preserving temporal mass ratios, raw stage masses within each training fold are normalized by their cross-stage geometric mean, yielding $\Omega_{e,k}$. When quadrature retains $N_{e,k}$ representative rows, each unit receives

$$\omega_{i,e,k} = \frac{\Omega_{e,k}}{N_{e,k}}, \tag{22}$$

thereby preserving the normalized mass of the original section. For static assays without population dynamics, we set $\Omega_{e,k} = 1$.

When multiple sections are observed at stage t_k, their representative units are concatenated consistently with X_k and A_k. The stage-level empirical mass vector $\mathbf{m}_k$ used in the UOT objective is defined as:

$$\mathbf{m}_k = \mathrm{concat}_{e:\, t_e = t_k}\left(\omega_{1,e,k}, \ldots, \omega_{N_{e,k},e,k}\right)^\top, \qquad \Omega_k = \mathbf{1}^\top \mathbf{m}_k. \tag{23}$$

The vector $\mathbf{m}_k$ directly parameterizes the optimal transport coupling and path-law learning without further transformation.

B.4 Coupling-Consistent Sampling and Mass Calibration

For transition $k \to k{+}1$, let $\mathbf{m}_k$ and $\mathbf{m}_{k+1}$ be the stage particle-mass vectors defined in Appendix B.3. The coupling in Equation 8 uses these vectors as both the source/target marginal reference and the product reference measure. Define the transported source-row mass $\varrho_{i,k} = \sum_j [\mathbf{\Pi}_k^\star]_{ij}$ and conditional endpoint distribution $\mathsf{P}_k(j \mid i) = [\mathbf{\Pi}_k^\star]_{ij}/\varrho_{i,k}$. Sampling S endpoints per source row gives the normalized path weight

$$w_{i,k} = \frac{\varrho_{i,k}}{N_k^{-1} \sum_{r=1}^{N_k} \varrho_{r,k}}. \tag{24}$$

Proposition 1 (Coupling-consistent conditional sampling)**.** *For every integrable function f,*

$$\sum_i \varrho_{i,k}\, \mathbb{E}_{j\sim \mathsf{P}_k(\cdot|i)}[f(i,j)] = \sum_{i,j} [\mathbf{\Pi}^\star_k]_{ij} f(i,j). \tag{25}$$

Proof. Substitute $\mathsf{P}_k(j \mid i) = [\mathbf{\Pi}^\star_k]_{ij}/\varrho_{i,k}$ and cancel $\varrho_{i,k}$. □

Thus conditional endpoint sampling preserves the UOT measure only when the row mass is retained.

Proposition 2 (Observed target-mass consistency)**.** *The scalar calibration in Equation 9 preserves the relative UOT row allocation and makes the calibrated source-row masses sum to the observed target mass Ω_{k+1}.*

Proof. The rescaling multiplies every entry of $\boldsymbol{\rho}_k$ by $\Omega_{k+1}/(\mathbf{1}^\top \boldsymbol{\rho}_k)$; summing the result gives Ω_{k+1}. □

B.5 Path-Law Sampling, Validation, and Forecasting

For each source state (k, i), S independently sampled endpoints inherit its transported weight $w_{i,k}$ and growth label $\gamma^\star_{i,k}$, which scale the velocity, growth, and endpoint loss terms in Equation 13. The moment-matching term is evaluated over minibatch statistics of the one-step surrogates and sampled targets. Validation splits are defined strictly at the level of unique source rows (k, i) across all S endpoint copies, retaining the checkpoint with minimum validation loss. Upstream representation parameters and UOT couplings remain frozen throughout path-law optimization.

Given an initial latent state $\mathbf{Z}(t_0)$, spatial adjacency $\mathbf{A}$, and mass vector $\mathbf{m}(t_0)$, trajectories are forecasted by integrating the coupled neural ODE system:

$$\frac{\mathrm{d}\mathbf{z}_i(t)}{\mathrm{d}t} = v_{\hat\phi}(\mathbf{z}_i(t), (\mathbf{A}\mathbf{Z}(t) - \mathbf{Z}(t))_i, \bar t)\,, \qquad \frac{\mathrm{d}\log m_i(t)}{\mathrm{d}t} = \gamma_{\hat\phi}(\mathbf{z}_i(t), (\mathbf{A}\mathbf{Z}(t) - \mathbf{Z}(t))_i, \bar t)\,, \tag{26}$$

using a second-order midpoint solver.

B.6 Decoder-Consistent Gene Editing

For a counterfactual query $\mathsf{q} = (g, R, y^\star, \tau, \ell)$, let $\mathbf{M}_{g,R}$ denote the indicator matrix for targeted gene–row pairs. The mask-aware anchor (Equation 14) is generated from the onset section following the identical preprocessing, spatial graph, and masking channels as representation training. The latent projection (Equation 15) optimizes from $\mathbf{Z}^{\mathrm{mask}}_\tau$ over n_{proj} gradient steps with step size γ_{proj}. The target $y^\star$ is specified on the decoder scale, with $y^\star = 0$ for gene knockout experiments.

During counterfactual rollout, STEP advances the edited state under the frozen velocity field, reapplying the latent projection after each numerical integration step up to time horizon ℓ. The final response is decoded and paired with the unperturbed wild-type trajectory from the same onset stage (Equation 17).

B.7 Dynamic Regulatory Sensitivity

For a spatial domain specified by expression matrix $\mathbf{X}$ and graph $\mathbf{A}$, let $\mathbf{Z}_{\mathbf{X}} = E_{\tilde\eta}(\widetilde{\mathbf{X}}, \widetilde{\mathbf{U}}, \mathbf{0})$ be the unperturbed latent encoding. We define the decoded one-step map as:

$$\Psi_{t,\Delta_{\mathrm{sens}}}(\mathbf{X}; \mathbf{A}) = D_{\tilde\eta}\Big(\mathbf{Z}_{\mathbf{X}} + \Delta_{\mathrm{sens}}\, v_{\hat\phi}(\mathbf{Z}_{\mathbf{X}}, \mathbf{A}\mathbf{Z}_{\mathbf{X}} - \mathbf{Z}_{\mathbf{X}}, \bar t)\Big)\,. \tag{27}$$

Letting $\overline{\Psi}_{t,\Delta_{\mathrm{sens}},h}(\mathbf{X}; \mathbf{A}) = N^{-1}\sum_{i=1}^N \Psi_{t,\Delta_{\mathrm{sens}}}(\mathbf{X}; \mathbf{A})_{ih}$ denote the domain-averaged prediction for target gene h, the sensitivity of gene h to a domain-wide infinitesimal shift in regulator g is formulated as:

$$J_{hg}(\mathbf{X}, t; \mathbf{A}) = \lim_{\varepsilon_{\mathrm{fd}}\to 0} \frac{\overline{\Psi}_{t,\Delta_{\mathrm{sens}},h}(\mathbf{X} + \varepsilon_{\mathrm{fd}}\mathbf{1}_N \mathbf{e}_g^\top; \mathbf{A}) - \overline{\Psi}_{t,\Delta_{\mathrm{sens}},h}(\mathbf{X}; \mathbf{A})}{\varepsilon_{\mathrm{fd}}}, \tag{28}$$

where $\mathbf{e}_g$ is the g-th canonical basis vector. In practice, J_{hg} is evaluated via finite differences. Assembling J_{hg} yields a dynamic sensitivity matrix $\mathbf{J}(\mathbf{X}, t; \mathbf{A})$, providing a spatially resolved regulatory readout when restricted to targeted tissue domains.

C Theoretical Analysis of Decoder-Consistent Gene Editing

To motivate the editable-latent-manifold design of Section 3.4, we discuss how a decoder-consistent proximal edit remains identifiable under the learned wild-type dynamics, whereas a direct ambient-space clamp may not. Let

$$\mathcal{M}_{\hat{\eta}} = \left\{ D_{\hat{\eta}}(\mathbf{Z}) : \mathbf{Z} \in \mathbb{R}^{N \times d} \right\} \tag{29}$$

denote the learned decoder manifold in the preprocessed expression space. For a query $q = (g, R, y^\star, \tau, \ell)$, write $\mathbf{Z}_a = \mathbf{Z}_\tau^{\text{mask}}$ and $\mathbf{M} = \mathbf{M}_{g,R}$. The direct clamp and the decoder-consistent edit are

$$\widetilde{\mathbf{X}}^{\text{clamp}} = (\mathbf{1} - \mathbf{M}) \odot \widetilde{\mathbf{X}}_\tau + y^\star \mathbf{M}, \qquad \widetilde{\mathbf{X}}^\star = D_{\hat{\eta}}(\mathbf{Z}^\star), \quad \mathbf{Z}^\star = \mathcal{P}(\mathbf{Z}_a), \tag{30}$$

where $\mathcal{P}$ is the proximal operator in Equation 15. In the matched direct-clamp ablation, $\widetilde{\mathbf{X}}^{\text{clamp}}$ is encoded once and the frozen path law is integrated without reapplying $\mathcal{P}$, so the comparison isolates the edit operator rather than the learned dynamics.

Proposition 3 (Manifold support and off-manifold non-identifiability)**.** *Consider the proximal regime $\lambda_{\text{tgt}} > 0$, $\lambda_{\text{prox}} > 0$, with $D_{\hat{\eta}}$ L_D-Lipschitz on $\mathcal{B}_r(\mathbf{Z}_a)$, a decoder-feasible state $\mathbf{Z}_f \in \mathcal{B}_r(\mathbf{Z}_a)$ satisfying $\mathbf{M} \odot \left(D_{\hat{\eta}}(\mathbf{Z}_f) - y^\star \mathbf{1}\right) = \mathbf{0}$, and a global minimizer $\mathbf{Z}^\star$ of Equation 15. Then:*

- *(a)* ***Manifold membership and bounded displacement.*** *$\widetilde{\mathbf{X}}^\star \in \mathcal{M}_{\hat{\eta}}$ and $\|\mathbf{Z}^\star - \mathbf{Z}_a\|_F \le \|\mathbf{Z}_f - \mathbf{Z}_a\|_F$.*
- *(b)* ***Fidelity and decoder-distance bounds.***

$$\left\|\mathbf{M} \odot \left(D_{\hat{\eta}}(\mathbf{Z}^\star) - y^\star \mathbf{1}\right)\right\|_F \le \sqrt{\frac{\lambda_{\text{prox}} \|\mathbf{M}\|_1}{\lambda_{\text{tgt}} N d}}\, \|\mathbf{Z}_f - \mathbf{Z}_a\|_F, \tag{31}$$

$$\|\widetilde{\mathbf{X}}^\star - D_{\hat{\eta}}(\mathbf{Z}_a)\|_F \le L_D \|\mathbf{Z}_f - \mathbf{Z}_a\|_F. \tag{32}$$

- *(c)* ***Rollout consistency.*** *Under the projected rollout in Equation 16, every decoded edited state belongs to $\mathcal{M}_{\hat{\eta}}$, and, if $v_{\hat{\phi}}$ is L_v-Lipschitz near $\mathbf{Z}_a$, $\left\| v_{\hat{\phi}}(\mathbf{Z}^\star, \boldsymbol{\xi}, \bar{t}) - v_{\hat{\phi}}(\mathbf{Z}_a, \boldsymbol{\xi}, \bar{t}) \right\|_F \le L_v \|\mathbf{Z}_f - \mathbf{Z}_a\|_F$.*
- *(d)* ***Non-identifiability of the direct clamp.*** *If $\delta_{\text{clamp}} = \text{dist}(\widetilde{\mathbf{X}}^{\text{clamp}}, \mathcal{M}_{\hat{\eta}}) > 0$, no finite stability bound for the ambient vector field at $\widetilde{\mathbf{X}}^{\text{clamp}}$ is identifiable from wild-type observations supported on $\mathcal{M}_{\hat{\eta}}$.*

Proof. Let $J(\mathbf{Z})$ denote the objective in Equation 15. Since $\mathbf{Z}_f$ is decoder-feasible, its target term vanishes, so optimality gives $J(\mathbf{Z}^\star) \le J(\mathbf{Z}_f) = \frac{\lambda_{\text{prox}}}{Nd} \|\mathbf{Z}_f - \mathbf{Z}_a\|_F^2$. Because both terms of J are nonnegative, $\frac{\lambda_{\text{prox}}}{Nd} \|\mathbf{Z}^\star - \mathbf{Z}_a\|_F^2 \le J(\mathbf{Z}^\star) \le \frac{\lambda_{\text{prox}}}{Nd} \|\mathbf{Z}_f - \mathbf{Z}_a\|_F^2$, proving (a). This places $\mathbf{Z}^\star \in \mathcal{B}_r(\mathbf{Z}_a)$, so the local Lipschitz bound on $D_{\hat{\eta}}$ applies; the same comparison inequality bounds the target term, giving the first part of (b), and combining with the Lipschitz bound gives the second. Manifold membership follows from $\widetilde{\mathbf{X}}^\star = D_{\hat{\eta}}(\mathbf{Z}^\star)$; since every projected rollout state is again of the form $D_{\hat{\eta}}(\mathbf{Z}^{\text{edit}}(t))$, induction gives (c), together with the stated velocity bound from the local Lipschitz property of $v_{\hat{\phi}}$.

For (d), let F_0 be any continuous extension of the on-manifold wild-type vector field to a neighborhood of $\widetilde{\mathbf{X}}^{\text{clamp}}$, and define $\psi(\mathbf{X}) = \min\{1, \text{dist}(\mathbf{X}, \mathcal{M}_{\hat{\eta}})/\delta_{\text{clamp}}\}$. For any Frobenius-unit $\mathbf{U}$ and $a > 0$, $F_a = F_0 + a\psi\,\mathbf{U}$ agrees with F_0 on every wild-type training state, since ψ vanishes on $\mathcal{M}_{\hat{\eta}}$, yet $\|F_a(\widetilde{\mathbf{X}}^{\text{clamp}}) - F_0(\widetilde{\mathbf{X}}^{\text{clamp}})\|_F = a$ is unbounded as $a \to \infty$. Wild-type data restricted to $\mathcal{M}_{\hat{\eta}}$ therefore cannot identify a finite stability bound for the ambient field at $\widetilde{\mathbf{X}}^{\text{clamp}}$. □

Proposition 3 shows that whenever a nearby decoder-feasible target exists, decoder-consistent projection guarantees a finite, quantifiable edit that stays supported on the learned manifold throughout rollout (parts a–c), while an ambient clamp that leaves the manifold cannot be assigned any comparable stability guarantee from wild-type data alone (part d). This asymmetry is consistent with our mechanism analyses: direct clamping collapses response amplitudes toward zero and shows

no analogous error-correction behavior (Figures 6c, 7c), whereas repeated projection recovers both the magnitude and sign of the downstream response (Figures 6b, 7b), and the manifold-distance distributions in Figure 5 directly visualize the geometric separation established in part (a).

D ENTENDED DATA FIGURES

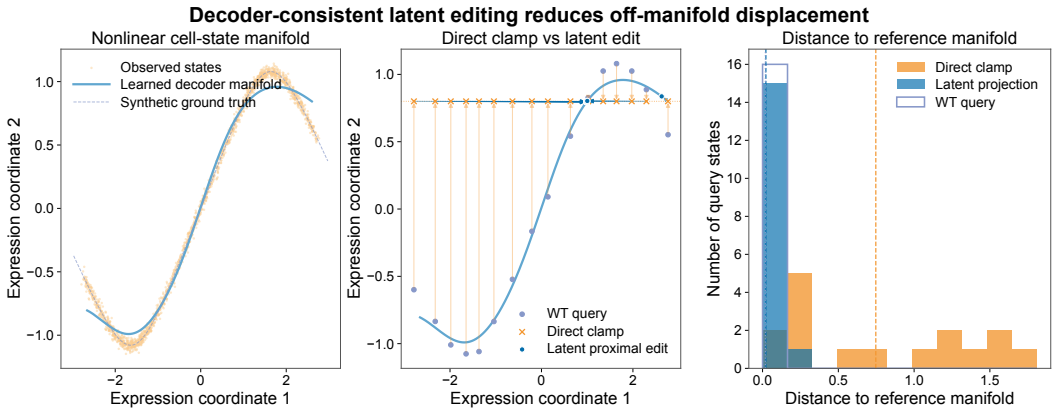


Figure 5: **Decoder-consistent latent editing maps ambient clamps back to the valid state manifold. Left** Synthetic query states with the ground-truth reference manifold and the manifold induced by the trained autoencoder decoder. **Middle** Editing trajectories in a shared coordinate space: orange arrows show ambient coordinate clamping that moves wild-type queries to target values ($y = 0.8$), whereas dark-blue arrows apply proximal latent projection to return the clamped anchors to decoder-consistent (on-manifold) states while satisfying the target constraint. **Right** Distributions of manifold distance for wild-type baselines, direct clamps, and latent projections, showing that proximal editing keeps edited states within standard autoencoder reconstruction tolerances.

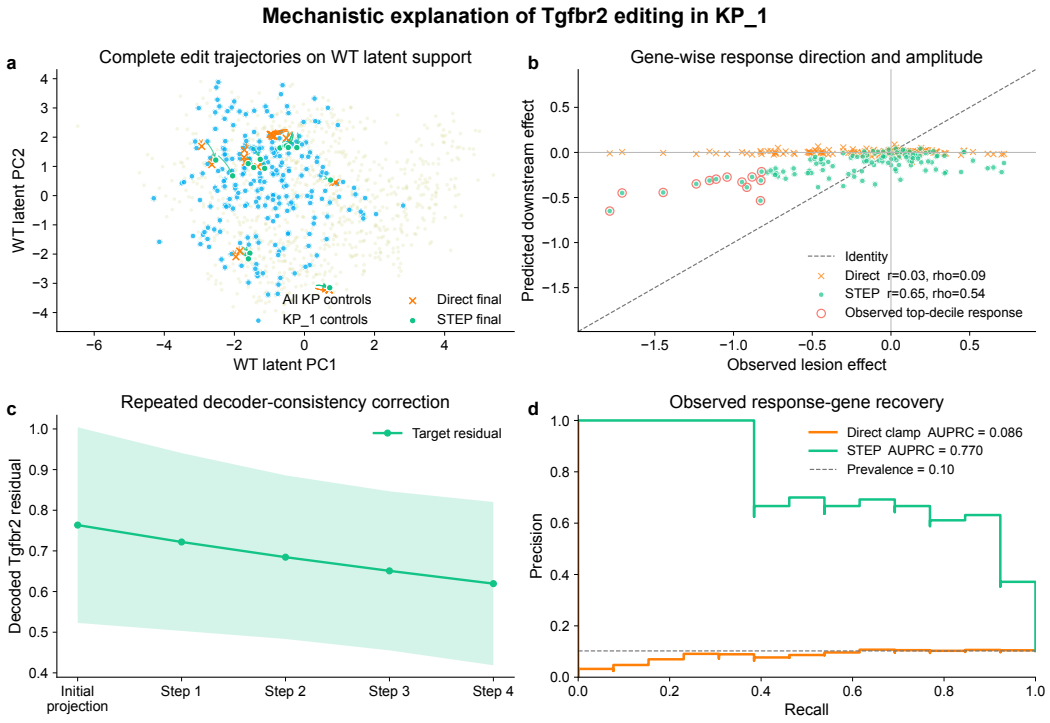


Figure 6: **Decoder-consistent latent editing improves downstream response recovery for *Tgfbr2* lesions in mouse KP. (a)** Latent training support from control regions across four KP sections (controls highlighted). Orange and teal arrows show direct-clamp and STEP rollout trajectories from onset (crosses) to the final state (filled circles). STEP applies decoder-consistent projection at initialization and after each integration step. **(b)** Gene-wise predicted versus observed downstream *Tgfbr2* effects. Direct clamp yields near-zero responses (Pearson $= 0.027$, Spearman $= 0.085$, norm ratio $= 0.047$), whereas STEP recovers both effect magnitude and signed structure (Pearson $= 0.654$, Spearman $= 0.541$, norm ratio $= 0.374$). Outlined points denote observed top-decile response genes. **(c)** Mean decoded *Tgfbr2* residual across initialization and four rollout steps (shaded: SD over 20 queries), illustrating iterative decoder-consistency corrections. **(d)** Precision–recall for downstream response-gene recovery (top decile of absolute lesion-versus-matched-control effects, excluding *Tgfbr2*). Ranking by absolute mean predicted effect over 20 queries, STEP achieves AUPRC $= 0.770$ versus 0.086 for direct clamp.

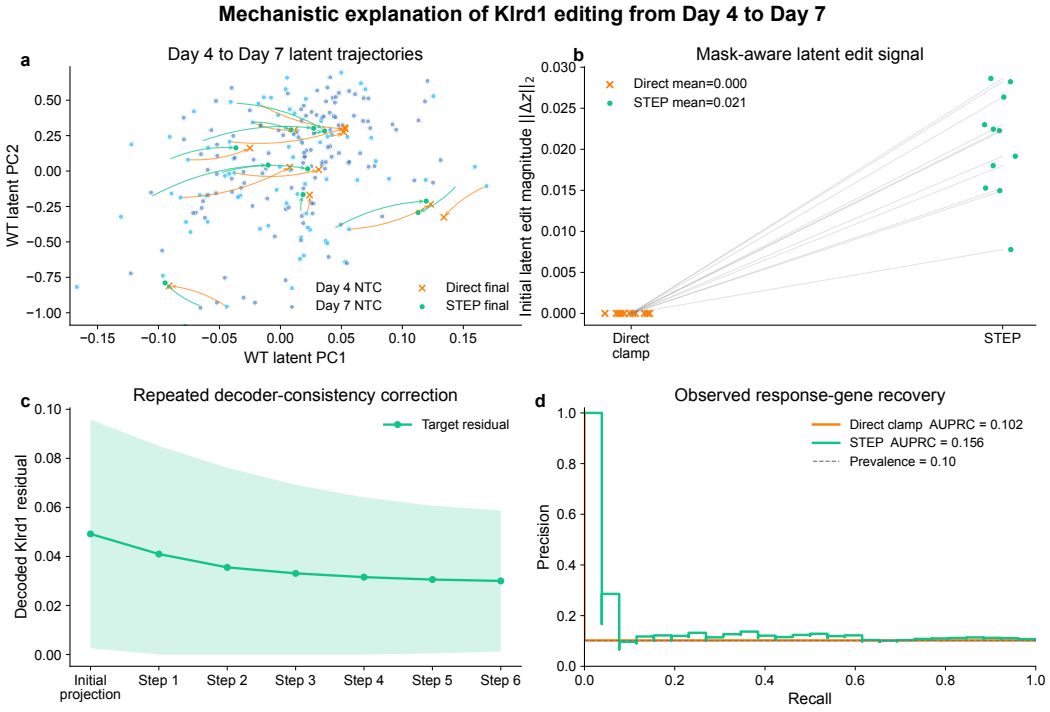


Figure 7: **Mask-aware latent editing improves recovery of the *Klrd1* response program under temporal rollout.** We evaluate *Klrd1* knockout responses with the wild-type path law fitted strictly on non-targeting control bins before intervention testing. **(a)** Day 4 and Day 7 control latent support with overlaid query trajectories; orange and teal vectors denote direct-clamp and STEP rollouts from onset (crosses) to final states (circles). **(b)** Initial latent edit magnitude after applying direct clamping versus STEP's mask-aware encoding and initialization projection, isolating the immediate edit signal before propagation. **(c)** Decoded *Klrd1* residual over initialization and six half-day integration steps (shaded: SD across query scenes), illustrating drift suppression by repeated projection. **(d)** Precision–recall for downstream response-gene recovery (top decile by absolute observed effect, excluding *Klrd1*). STEP attains AUPRC $= 0.156$ and signed Spearman 0.147, outperforming direct clamping (AUPRC $= 0.102$, signed Spearman 0.000) and the prevalence baseline (0.10).

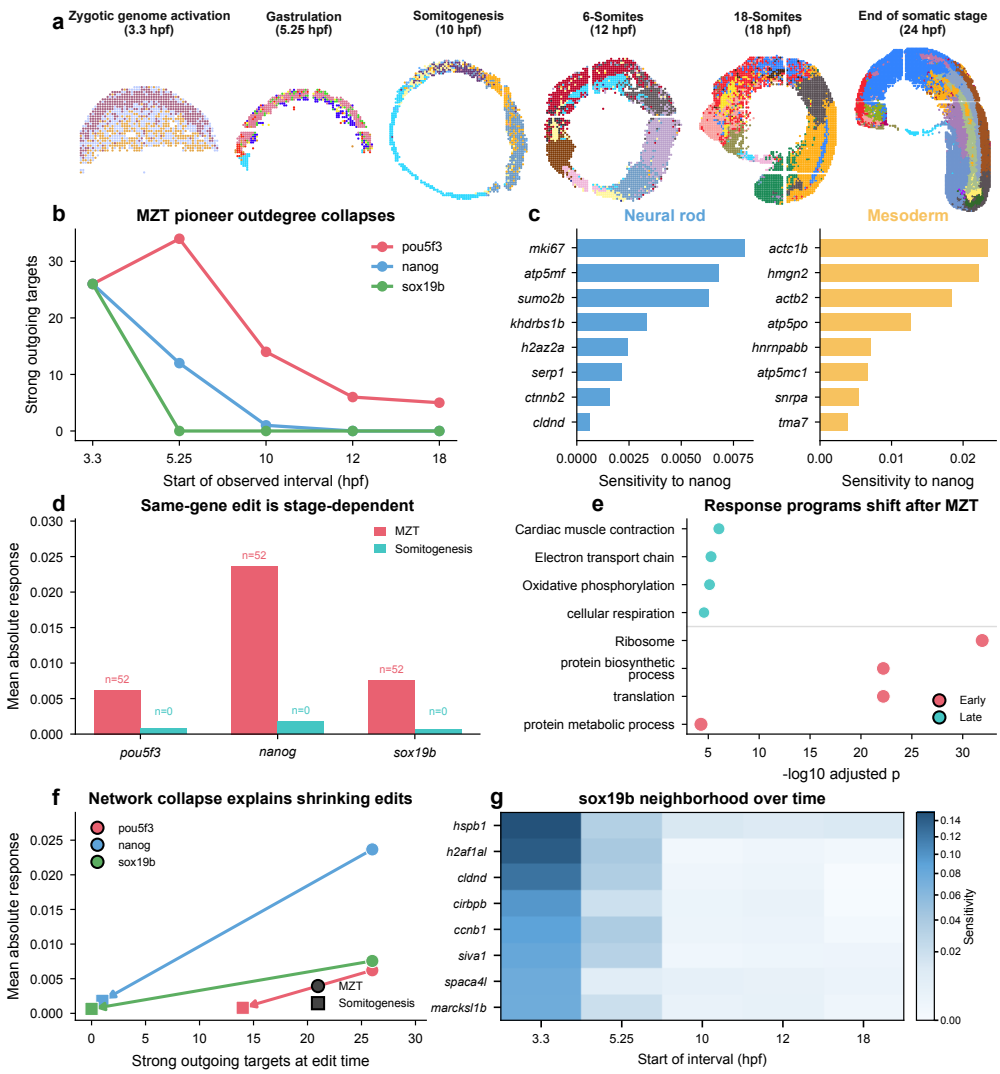


Figure 8: **Spatiotemporal GRN dynamics explain stage-specific *in silico* intervention efficacy on the ZESTA atlas. (a)** Stereo-seq spatial bin layouts across six developmental stages (3.3–24 hpf). **(b)** Developmental collapse in out-degree connectivity for maternal-to-zygotic transition (MZT) pioneer factors (*pou5f3*, *nanog*, *sox19b*). **(c)** Spatially restricted *nanog* regulatory neighborhoods at 12 hpf, contrasting neural rod and somitic mesoderm domains. **(d)** Predicted response magnitudes for identical one-step knockdowns applied at MZT versus somitogenesis; annotations indicate target counts above the MZT threshold. **(e)** Pathway enrichment highlights a shift from early programs (translation, ribosome biogenesis) to late programs (muscle contraction, oxidative phosphorylation). **(f)** Coupled decrease in pioneer out-degree and predicted edit magnitude across development. **(g)** Temporal decay of outgoing Jacobian sensitivity from *sox19b* to its core MZT neighborhood.

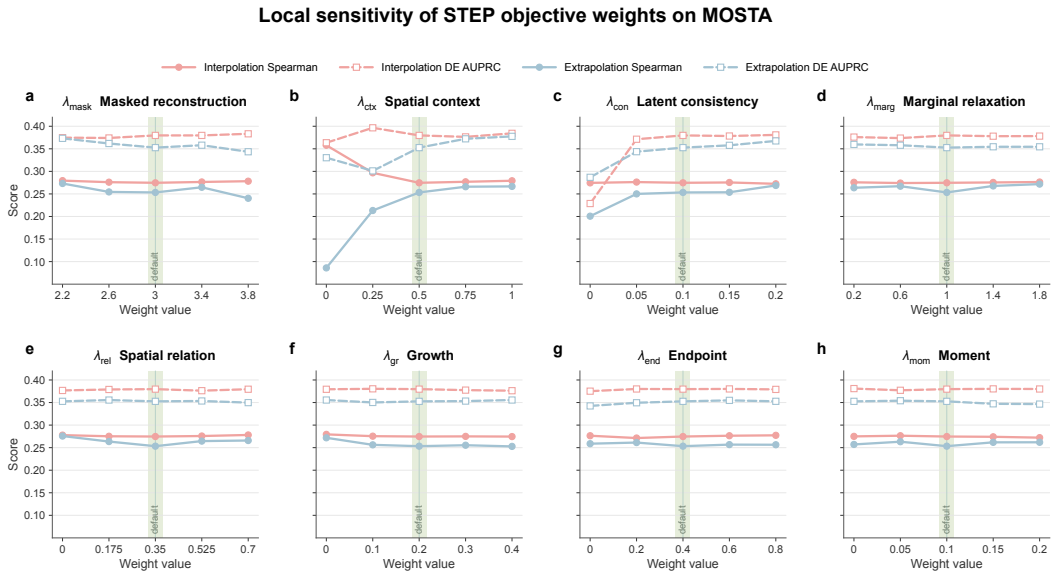


Figure 9: **Hyperparameter sensitivity.** Coral and blue curves correspond to the E12.5→E13.5 interpolation fold and the E15.5→E16.5 extrapolation fold, respectively. Solid lines (filled circles) report temporal-change Spearman correlation; dashed lines (open squares) report temporal-DE AUPRC. Pale-green bands indicate default settings. Panels sweep individual loss weights: **(a)** masked reconstruction (λ_{mask}), **(b)** spatial context (λ_{ctx}), **(c)** latent consistency (λ_{con}), **(d)** UOT marginal relaxation (λ_{marg}), **(e)** spatial relation (λ_{rel}), **(f)** path growth (λ_{gr}), **(g)** endpoint consistency (λ_{end}), and **(h)** moment consistency (λ_{mom}); all panels share the same y-axis. Performance is most sensitive to λ_{ctx} and λ_{con}. Setting $\lambda_{\mathrm{ctx}} = 0$ sharply degrades extrapolation Spearman ($0.253 \rightarrow 0.086$), underscoring the importance of microenvironmental contrast for future-stage transfer. Removing λ_{con} consistently reduces temporal-DE AUPRC in both interpolation and extrapolation, indicating that aligning masked and clean encodings is critical for prioritizing dynamic response genes. In contrast, $\lambda_{\mathrm{marg}}, \lambda_{\mathrm{rel}}, \lambda_{\mathrm{gr}}, \lambda_{\mathrm{end}}$, and λ_{mom} exhibit comparatively flat profiles, suggesting the default configuration lies in a robust region.

E SIMULATED MECHANISM BENCHMARK RESULTS

To evaluate STEP's capacity for mechanistic recovery under known ground truth, we construct a synthetic benchmark comprising 96 genes across 144 cells per stage over five time points ($t \in \{0, \ldots, 4\}$) using `scMultiSim` (Li et al., 2025). A Hill reaction–diffusion system generates wild-type (WT) and regional knockout (KO) trajectories, where a central hub gene governs seven exclusive downstream targets. Models are trained on unperturbed WT snapshots ($t \leq 3$), reserving KO trajectories, spatial masks, and ground-truth GRNs exclusively for evaluation. Performance is assessed via two complementary readouts: (1) **Latent Editing**, applying decoder-consistent projection in a target region at $t = 1$ to forecast downstream KO trajectories ($t \in \{2, 3, 4\}$); and (2) **Local Sensitivity Analysis**, evaluating the hub gene's outgoing Jacobian column $J_{hg}(\mathbf{X}, t)$ against the true network topology (Table 4).

Table 4: Mechanistic evaluation on the simulated GRN benchmark. WT metrics report gene-wise temporal-change correlations. KO metrics are macro-averaged over forecast stages $t \in \{2, 3, 4\}$. Sensitivity metrics average the predicted outgoing Jacobian column of the hub gene over WT stages $t \in \{0, 1, 2, 3\}$. *Ambient clamp* denotes direct expression-space clamping evaluated under the same learned path law ("–": not applicable).

Readout	Pearson ↑	Response AUPRC ↑	Recall@7 ↑	RMSE ↓
WT one-step mean	0.9870	–	–	0.0161
WT $t = 2 \to 4$	0.9574	–	–	0.0319
Latent edit	0.5708	0.5911	0.5714	0.2041
Local sensitivity	**0.7865**	**1.0000**	**1.0000**	**0.0586**
Ambient clamp	0.1216	0.0868	0.0000	0.0748

Key insights from the simulated benchmark include:

- **High-fidelity wild-type dynamics:** STEP accurately captures baseline spatiotemporal dynamics, achieving a single-step Pearson correlation of 0.9870 and maintaining 0.9574 over two-step unrollings ($t = 2 \to 4$).
- **Perfect modular sensitivity recovery:** The local Jacobian readout achieves perfect target identification ($\mathrm{AUPRC} = 1.0000$, $\mathrm{Recall@7} = 1.0000$), confirming that the learned vector field isolates localized, high-impact regulatory modules.
- **Manifold-constrained projection superiority:** For counterfactual forecasting, latent-space editing recovers four of the seven exclusive targets ($\mathrm{AUPRC} = 0.5911$), dramatically outperforming ambient expression-space clamping ($\mathrm{AUPRC} = 0.0868$).
- **Necessity of self-supervised pretraining:** Ablating self-supervised masking severely degrades latent-edit AUPRC from 0.5911 to 0.3042 (Table 2), validating the indispensable role of manifold-constrained pretraining for counterfactual editing.

F Ablation Studies

To validate our core methodological choices, we conduct targeted ablations evaluating: (i) mass-calibrated flow dynamics, and (ii) masked-gene self-supervised pretraining (SSL) (Table 2).

Mass-Calibrated Flow Dynamics. Ablating row-weighting, mass calibration, and endpoint regularization (Propositions 1 and 2) leaves directional correlation virtually intact, but severely impairs spatial density alignment and population scale. Re-introducing the complete objective resolves these boundary and scale distortions, yielding a 47% to 72% reduction across trajectory RMSE, SWD, and population count error. This confirms that mass calibration selectively rectifies endpoint boundary geometry and cell scale without altering velocity directionality.

Masked-Gene Representation Learning. Evaluating stochastic and single-gene masking highlights their necessity for interventional representation learning. Omitting masked-gene pretraining severely degrades interventional prediction fidelity, cutting response-gene AUPRC by nearly half and impairing effect correlation. Training exclusively on unmasked wild-type profiles confines the autoencoder to passive co-expression modeling. In contrast, training with targeted zero-out masks forces counterfactual knockouts onto in-distribution latent trajectories. Crucially, local sensitivity analysis recovers all ground-truth hub targets (Appendix E), confirming mechanism-level perturbation responsiveness.

G Case study of ZESTA zebrafish embryo atlas

To understand why counterfactual responsiveness varies across development, we interpret the local Jacobian sensitivity $J_{hg}(X, t)$ (Equation 28) as a *spatiotemporal* gene regulatory network (GRN) on the ZESTA zebrafish embryo atlas (Liu et al., 2022) (Figure 8a). We focus on three canonical maternal-to-zygotic transition (MZT) pioneer transcription factors—*pou5f3*, *nanog*, and *sox19b* (Lee et al., 2013; Leichsenring et al., 2013)—fitted across six wild-type Stereo-seq snapshots spanning 3.3–$24\ \mathrm{hpf}$.

Temporal collapse and spatial specialization. At early MZT ($3.3\ \mathrm{hpf}$), each pioneer exhibits high out-degree, regulating 26 strong downstream targets. By somitogenesis (10–$12\ \mathrm{hpf}$), out-degree drops to 14, 1, and 0 for *pou5f3*, *nanog*, and *sox19b*, respectively (Figure 8b). This stage-dependent rewiring is not visible to static GRN baselines and provides a direct mechanistic explanation for why the same edit can have very different effects across developmental time.

Spatially, conditioning $J_{hg}(X, t)$ on $12\ \mathrm{hpf}$ subregions reveals domain-specific neighborhoods: *nanog* connects to distinct targets in the neural rod (*ctnnb2*, *mki67*, *h2az2a*) versus the somitic mesoderm (*actc1b*, *actb2*) (Figure 8c), consistent with post-gastrulation lineage specialization (Wagner et al., 2018; Farrell et al., 2018).

Mechanistic basis of stage-specific editing. This evolving topology translates into stage-dependent perturbation sensitivity. Executing identical one-step *in silico* latent knockdowns at MZT ($3.3 \rightarrow 5.25\ \mathrm{hpf}$) versus somitogenesis ($10 \rightarrow 12\ \mathrm{hpf}$) yields an order-of-magnitude attenuation in response magnitude: predicted effects drop by $13\times$, $7.7\times$, and $11.6\times$ for *nanog*, *pou5f3*, and *sox19b*, respectively (Figure 8d). Under a fixed response threshold, early MZT edits perturb 52 downstream targets enriched for ribosome biogenesis and translation, reflecting pioneer-driven zygotic genome activation (Pálfy et al., 2020). In contrast, late-stage edits produce few primary targets and shift residual signatures toward oxidative phosphorylation and cardiac muscle contraction (Figure 8e), consistent with organogenesis during somitogenesis (Özbudak et al., 2010). As pioneer out-degree collapses alongside epigenetic restriction of cell fate (Bogdanović et al., 2012), predicted edit magnitudes contract in tandem (Figure 8f,g), supporting the conclusion that local GRN state governs counterfactual edit strength across both space and time.

H Dataset Provenance and Preprocessing

H.1 Common Preprocessing Principles

To preclude data leakage, all gene selection and feature scaling parameters are fit within training folds. Raw count matrices are normalized to 10^4 counts per cell, $\log(1 + x)$-transformed, and filtered to exclude mitochondrial and ribosomal genes before intersecting candidate highly variable

Table 5: Summary of datasets evaluated in this study. "Rows" indicates the computational sample size per snapshot.

Dataset	Times / stages	Genes	Rows	Role
Simulated replay	5	96	144 / time	WT dynamics, GRN, KO
stVCR rectangle	6 observed snapshots	3	512 / time	Flow-objective ablation
MOSTA	E9.5–E16.5	1,000	180 / time	WT dynamics
ARTISTA	5 ordered stages	1,000	180 / time	WT dynamics
HT29 Perturb-DBiT	Static	256	1,117 controls	Static CRISPR
Mouse lung Perturb-map	Static	128	1,061 KP controls	Static CRISPR
SPAC-seq	Day 4/7	256	Section dependent	Dynamic CRISPR
ZESTA	3.3–24 hpf	512	13,166 total	Case study

genes across developmental stages. For MOSTA and ARTISTA, section-wise dispersion ranks are aggregated across training slices, with expression profiles scaled by the 0.99 training quantile and clipped to the analysis range. Spatial topology is encoded via row-normalized $k = 8$ nearest-neighbor graphs restricted strictly to intra-section edges. Finally, representative spatial locations are sampled via computational quadrature, preserving empirical-measure weights during unbalanced optimal transport (UOT) and mass growth evaluation, with reproducible state trajectories enforced through deterministic inverse-CDF sampling.

H.2 stVCR Rectangle Simulation

The stVCR rectangle simulation (Peng et al., 2026) provides six discrete spatial snapshots ($t \in \{0, 0.5, \ldots, 2.5\}$) used for the flow-objective ablations in Table 2. The benchmark models dynamic population growth from $1{,}500$ to $2{,}410$ cells across 3 gene expression dimensions and 2 spatial coordinates, evaluated against 60 continuous ground-truth frames ($t \in [0.05, 2.95]$ at step size $\Delta t = 0.05$). Each observed stage is discretized into 512 deterministic particles.

H.3 Mouse Organogenesis Benchmark (MOSTA)

The MOSTA dataset (Chen et al., 2022) captures mouse organogenesis across eight embryonic stages (E9.5–E16.5). We evaluate models on the intersected gene set under two protocols that match Table 1:

- **Leave-one-stage interpolation (6-fold).** Each internal age E10.5–E15.5 is held out in turn. The model is trained on the remaining stages, and the held-out stage is predicted by a one-step advance from the immediately preceding observed age.
- **Expanding-window one-step extrapolation (4-fold).** With a minimum history of four ages, the four forecasts are E9.5–E12.5→E13.5, training through E13.5→E14.5, training through E14.5→E15.5, and training through E15.5→E16.5. Preprocessing, UOT, and model parameters for target k are derived strictly from preceding stages.

Each fold independently selects 1,000 HVGs and 180 quadrature particles per stage, and uses 15 representation epochs and 30 flow epochs. Reported interpolation and extrapolation scores are means over these folds.

H.4 Axolotl Regeneration Benchmark (ARTISTA)

The ARTISTA dataset provides Stereo-seq spatial transcriptomics of axolotl telencephalon regeneration (Wei et al., 2022). We evaluate the natural regeneration sequence across five ordered stages: stage 44, stage 54, stage 57, juvenile, and adult, excluding the metamorphosed branch. Ordinal indices $t \in \{0, 1, 2, 3, 4\}$ encode stage order rather than equal physical durations. Evaluation covers two protocols that match Table 1:

- **Leave-one-stage interpolation (3-fold).** Each internal stage (stage 54, stage 57, juvenile) is held out in turn. The model is trained on the remaining stages, and the held-out stage is predicted by a one-step advance from the immediately preceding observed stage.

- **Adult one-step extrapolation (1-fold).** The model is trained through juvenile and predicts the immediately next adult stage; adult profiles are excluded from training, preprocessing, and coupling.

Each fold independently selects 1,000 HVGs and 180 quadrature particles per stage. Reported interpolation scores are means over the three internal folds; adult extrapolation is a single biologically available transition.

H.5 Static Human Metastasis Benchmark (HT29 Perturb-DBiT)

The HT29-Cas9 lung metastasis section (Bru.LC.6) (Baysoy et al., 2026) comprises 2,166 spatial spots over a 256-gene panel selected from 1,117 non-targeting (NT) control spots. Seven perturbation targets meet the minimum support threshold (≥ 20 assigned spots): *ATG4A*, *MT1E*, *S100A4*, *SEMA6B*, *ADARB1*, *ASAH2*, and *ADAM20*. All models are trained exclusively on unperturbed NT control sections, reserving perturbation profiles for downstream evaluation.

To model counterfactuals on a single static tissue section without temporal dynamics, STEP operates in its stationary limit (zero temporal flow) using targeted zero-out masks at central query spots. Predicted downstream gene-effect vectors are evaluated against ground-truth Mixscape log-fold-change (LFC) profiles and macro-averaged across targets after excluding the knocked-down gene. All baselines (SpatialProp, Celcomen) and the direct-clamp ablation share identical NT controls, gene panels, query contexts, and target contrasts, ensuring strict isolation of perturbation data during representation fitting.

H.6 Static Mouse Lung Cancer Benchmark (Perturb-map)

The mouse KP Perturb-map dataset (GSE193460; four 10x Visium sections) (Dhainaut et al., 2022) leverages Hyperion-linked protein annotations to establish an unperturbed KP control reference across sections, reserving *Tgfbr2* and *Jak2* lesion spots for counterfactual evaluation. Normalized spots (≥ 100 genes) yield a 128-gene panel based on pooled control variance.

To evaluate counterfactual predictions, up to 20 lesion spots per target and section are paired with spatially matched control contexts in native coordinate space. Queries are initialized from control patches with targeted zero-out masks at central spots, and predicted effects are benchmarked against empirical lesion contrasts (mean lesion minus matched-control expression). Metrics are calculated per section excluding the target gene and macro-averaged across eligible sections (three for *Tgfbr2*, one for *Jak2*). All competing approaches (SpatialProp, Celcomen, and direct-clamp) share the identical control reference, gene panel, target subsets, and evaluation contrasts, ensuring isolation of knockout profiles during training.

H.7 Dynamic In Vivo CRISPR Benchmark (SPAC-seq)

The spatiotemporal SPAC-seq spatial CRISPR assay (Zhang et al., 2026) evaluates dynamic perturbation responses across spatial bins aligned with sgRNA matrices. Guide assignments follow the TARDIS protocol (Zhang et al., 2026), retaining unique sgRNA calls per bin. Within annotated T-cell regions, non-targeting control (NTC) spots across duplicate sections provide baseline spatial reference at Day 4 and Day 7, whereas Day 10 sections are excluded due to insufficient NTC support. Five perturbation targets meet query and evaluation thresholds: *Flot1*, *Gata3*, *Klrd1*, *Wipf1*, and *Zc3h12a*.

Because sgRNA vectors are introduced prior to Day 4, the task is formulated as a single-interval conditional forecast across the Day 4 $\rightarrow$ Day 7 horizon. Queries are initialized from metadata-matched Day 4 NTC spatial patches. Predicted effects are scored exclusively on cell-autonomous, guide-bearing Day 7 bins. Perturbation-bearing profiles remain unblinded only after prediction, ensuring zero perturbation labels enter model fitting.

H.8 Zebrafish Embryogenesis Case Study (ZESTA)

The ZESTA Stereo-seq atlas (Liu et al., 2022) spans six zebrafish developmental stages ($t \in \{3.3, 5.25, 10, 12, 18, 24\}$ hpf), comprising 13,166 spatial bins across 26,628 genes. Raw counts

are library-size normalized and $\log(1+x)$-transformed to derive a 512-gene panel based on pooled variance across snapshots. Model fitting on unperturbed wild-type trajectories utilizes 240 spatially stratified representative bins per stage, with empirical bin counts specifying spatial measure weights during unbalanced optimal transport (UOT).

Following representation fitting, outgoing Jacobian sensitivity $J_{hg}(X,t)$ is evaluated across adjacent developmental intervals, defining strong downstream targets as genes exceeding the 95th-percentile sensitivity threshold at $3.3\ \mathrm{hpf}$. At $12\ \mathrm{hpf}$, domain-restricted *nanog* Jacobians are averaged separately over neural-rod and somitic-mesoderm domains. ***In silico*** knockdowns of pioneer factors are executed at MZT ($3.3 \to 5.25\ \mathrm{hpf}$) and somitogenesis ($10 \to 12\ \mathrm{hpf}$) by applying targeted zero-out masks across representative bins. Active response programs, defined by genes exceeding the 90th-percentile response magnitude, are benchmarked via g:Profiler (Raudvere et al., 2019) against zebrafish Gene Ontology, KEGG, and WikiPathways databases using the top 50 response genes per interval.

I Architecture and Optimization Details

Neural Network Architecture. The representation encoder E_η is parameterized as a three-layer Multi-Layer Perceptron (MLP) operating on concatenated normalized gene expression, spatial context vectors, and binary perturbation masks, stabilized by LayerNorm and SiLU activations. Dual single-hidden-layer decoders D_η and D_η^{ctx} reconstruct expression profiles and standardized spatial context. The dynamic path law incorporates a shared two-layer backbone with separate latent-velocity v_ϕ and log-mass-growth γ_ϕ heads, conditioned on latent states, latent niche contrast, normalized time t, and harmonic time features.

Objective Formulation and Hyperparameters. Model training proceeds via a multi-task objective balancing self-supervised representation learning and dynamic transport alignment. Self-supervised pretraining employs masking probability $p_{\mathrm{mask}} = 0.12$, with masked reconstruction, spatial-context, and latent-consistency weights $\lambda_{\mathrm{mask}} = 3.0$, $\lambda_{\mathrm{ctx}} = 0.5$, and $\lambda_{\mathrm{con}} = 0.1$. Unbalanced optimal transport uses entropic regularization $\epsilon_{\mathrm{OT}} = 0.08$, marginal relaxation $\lambda_{\mathrm{marg}} = 1.0$, and spatial relation weight $\lambda_{\mathrm{rel}} = 0.35$. Path-law growth, endpoint, and moment weights are $\lambda_{\mathrm{gr}} = 0.2$, $\lambda_{\mathrm{end}} = 0.4$, and $\lambda_{\mathrm{mom}} = 0.1$. The core parameter sensitivity analysis is presented in Figure 9.

J Baseline Methods and Scope

stVCR. Evaluated using its official implementation, spatial alignment, and Wasserstein–Fisher–Rao (WFR) Neural ODE formulation (Peng et al., 2026). As a wild-type snapshot-dynamics baseline, it models temporal distribution flows without a counterfactual gene-editing operator.

DeepRUOT. Implemented following its published unbalanced neural optimal transport (Neural-OT) formulation (Zhang et al., 2025), and evaluated on the same fold-specific 1,000-gene matrices and 180-particle sections as STEP. It serves as a pure wild-type temporal dynamics baseline without perturbation capability.

SpatialProp. Evaluated on static spatial CRISPR benchmarks using its spatial-graph architecture and matched gene panels (Sun et al., 2025). It models static spatial perturbation propagation within single tissue sections but lacks a temporal path law.

Celcomen. Evaluated on static HT29 and mouse lung Perturb-map benchmarks following its spatial protocol (Megas et al., 2025). It generates static spatial counterfactuals trained exclusively on unperturbed reference tissue.

DynPerturb. DynPerturb is a dynamic inference framework for single-cell and spatial transcriptomics, which models counterfactual perturbation dynamics via temporal attention and recurrent spatial representations (Qin et al., 2025). We use this framework to analyze dynamic SPAC-seq data to enable an equitable benchmark comparison.

Direct clamp. Serves as an ablation benchmark isolating the contribution of the latent edit operator. Utilizing STEP's frozen wild-type encoder, UOT couplings, and latent velocity field, it clamps target genes to zero directly in expression space, encodes the modified profile, and integrates the trajectory without proximal decoder projection. This isolates the specific gain of decoder-consistent latent editing over shared wild-type dynamics.

Few-shot reference. Constructed for SPAC-seq by utilizing 50% of target-specific Day-7 guide-bearing spatial bins for profile supervision and evaluating on the remaining held-out bins. It serves as an empirical upper bound for label-efficient perturbation regimes rather than a wild-type-only model.

K Evaluation Metrics

Temporal distribution dynamics. Let $\bar{\mathbf{x}}_s, \bar{\mathbf{x}}_t, \bar{\mathbf{x}}_p \in \mathbb{R}^G$ denote the population-mean expression vectors for the source, observed target, and predicted target distributions, respectively.

- **Temporal-change correlation and AUPRC:** Trajectory directional agreement is quantified by the Pearson correlation between predicted changes $(\bar{\mathbf{x}}_p - \bar{\mathbf{x}}_s)$ and true changes $(\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_s)$. The Area Under the Precision-Recall Curve (AUPRC) evaluates ranking performance on salient temporally differentially expressed genes, defined as top target changes selected independently of model training.
- **Population-mean and change RMSE:** Measures expression amplitude and trajectory scale calibration at the population-mean level.

Distributional and mass metrics.

- **Sliced Wasserstein Distance (SWD):** Evaluates high-dimensional distributional alignment across unpaired snapshots by projecting expression matrices X_P and X_Q onto L uniformly sampled unit hyperspheres $\theta_l \in \mathbb{S}^{d-1}$:

$$\mathrm{SWD}(P, Q) = \frac{1}{L} \sum_{l=1}^{L} W_2(\theta_l^\top X_P, \theta_l^\top X_Q). \tag{33}$$

  SWD complements gene-level metrics by quantifying global population manifold overlap.
- **Mass growth error:** Population-scale accuracy is evaluated by the relative count error $\frac{|\widehat{N}-N|}{N}$, as reported in the flow-objective ablation (Table 2).

Counterfactual perturbation response.

- **Signed Spearman correlation and response AUPRC:** Downstream effect rankings across non-edited genes are evaluated via signed Spearman rank correlation. Response AUPRC tests response-gene prioritization, treating the top decile of observed absolute effect magnitudes as ground-truth positives and ranking genes by predicted absolute effects. On SPAC-seq, this endpoint is computed exclusively on Day 7 guide-bearing bins.
- **Precision@K:** Precision@K measures recovery of the top 10% observed response genes at K equal to that support size.
- **Response calibration:** Effect RMSE and norm ratio $\frac{\|\boldsymbol{\Delta}\bar{\mathbf{x}}_p\|_2}{\|\boldsymbol{\Delta}\bar{\mathbf{x}}_t\|_2}$ quantify response amplitude calibration.