PD-scWorld: Pathway-Guided Disentanglement for Single-Cell Perturbation World Models

Azmine Toushik Wasi
Computational Intelligence and Operations Laboratory (CIOL)
azminetoushik.wasi@gmail.com

Abstract

Disentangling the latent factors that drive cellular responses remains challenging for generative and predictive models of single-cell data, particularly under perturbations where multiple biological programs co-activate. We propose PD-scWorld, an intervention-aware latent world model in which each latent factor z_k is encouraged to respond $selectively$ to specific perturbations and covariates, using only weak biological supervision in the form of pathway tags for perturbed genes rather than full factor labels. Given paired pre/post states, the model learns action-conditioned transitions $z' = T_\psi(z, a)$ while constraining the perturbation-induced change $\Delta z = z' - z$ to be group-sparse across latent dimensions associated with the pathway of (a). Concretely, we introduce a pathway-conditional regularizer that penalizes dispersion of Δz_k outside the designated latent group, combining group sparsity with a variance-based term $\sum_k \mathrm{Var}(\Delta z_k \mid a)$ to localize consistent effects and suppress entangled drift. This yields latents that align with known biological programs while retaining predictive flexibility for unseen perturbations. We evaluate on Perturb-seq and related CRISPR single-cell screens using gene-to-pathway mappings and cell-cycle annotations, measuring (i) mutual information between latents and covariates, (ii) recovery of pathway-specific responses, and (iii) counterfactual consistency under targeted rollouts. Across datasets, the proposed model produces cleaner factorization and more interpretable perturbation mechanisms than β-VAE and unstructured latent dynamics baselines, while improving accuracy of perturbation effect prediction and robustness to covariate shifts.

1 Introduction

Single-cell perturbation technologies such as CRISPR-based screens have enabled systematic interrogation of gene function by measuring cellular responses at scale. Recent generative and world-model approaches have shown promise in modeling these data, learning latent representations that capture cellular state and predicting responses to genetic or chemical interventions (Baek et al., 2024; Sadria & Layton, 2025). However, most existing models prioritize predictive accuracy or reconstruction fidelity, often learning entangled latent spaces in which multiple biological processes are conflated (Choi et al., 2023; Fu et al., 2026). While post-hoc probing or disentanglement penalties can reveal some structure, the resulting latent factors rarely align cleanly with known biological programs (Sun et al., 2025). This limits the interpretability and scientific utility of learned representations, particularly when models are used to reason about mechanisms of perturbation.

A central challenge in perturbation modeling is that interventions are not arbitrary: most perturbations are designed to target specific pathways, regulatory modules, or cellular programs. Yet current latent-variable models typically treat perturbations as unstructured inputs, allowing their effects to diffuse broadly across the latent space (Sadria & Layton, 2025; Li et al., 2022). Fully supervised factor labeling is infeasible, as true latent biological programs are only partially known and often context-dependent. This raises a key question: can we guide latent disentanglement using weak, structured biological knowledge, such as pathway annotations, without sacrificing predictive power or generalization? Addressing this question is essential for moving from black-box perturbation

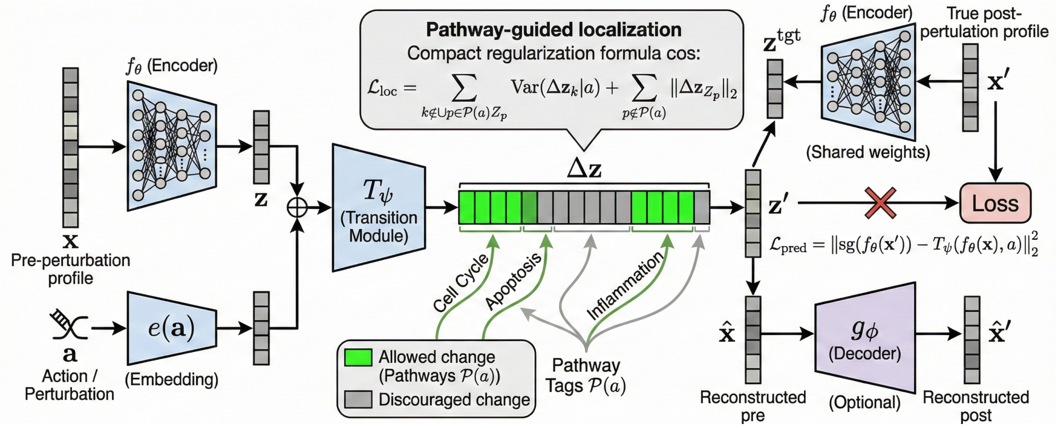
Pathway-guided localization
Compact regularization formula cos:
$\mathcal{L}_{\text{loc}} = \sum_{k \notin \cup p \in \mathcal{P}(a) Z_p} \text{Var}(\Delta \mathbf{z}_k | a) + \sum_{p \notin \mathcal{P}(a)} \|\Delta \mathbf{z}_{Z_p}\|_2$
f_θ (Encoder)
$\mathbf{x}$
Pre-perturbation profile
$\mathbf{z}$
$\mathbf{a}$
Action / Perturbation
$e(\mathbf{a})$
(Embedding)
T_ψ
(Transition Module)
$\Delta \mathbf{z}$
Cell Cycle
Apoptosis
Inflammation
🟩 Allowed change (Pathways $\mathcal{P}(a)$)
▨ Discouraged change
Pathway Tags $\mathcal{P}(a)$
$\mathbf{z}^{\text{tgt}}$
f_θ (Encoder)
True post-pertulation profile
$\mathbf{x}'$
(Shared weights)
$\mathbf{z}'$
Loss
$\mathcal{L}_{\text{pred}} = \|\text{sg}(f_\theta(\mathbf{x}')) - T_\psi(f_\theta(\mathbf{x}), a)\|_2^2$
$\hat{\mathbf{x}}$
Reconstructed pre
g_ϕ
(Decoder)
(Optional)
$\hat{\mathbf{x}}'$
Reconstructed post

Figure 1: Pathway-guided intervention world model. The encoder maps pre- and post-perturbation profiles to latents z and z^{tgt}. An action-conditioned transition predicts $z' = z + \Delta z$ and is trained with a predictive JEPA loss. Weak pathway tags $\mathcal{P}(a)$ regularize the *change* Δz, encouraging perturbation effects to localize to pathway-associated latent groups while preserving predictive accuracy.

predictors to models that support mechanistic reasoning and counterfactual analysis (Lopez et al., 2023; Sun et al., 2025; Roohani et al., 2023).

A complementary line of work comes from *world modeling*, where agents learn compact latent representations of environment dynamics to support prediction, planning, and counterfactual reasoning. In model-based reinforcement learning, world models learn latent transition dynamics that can be rolled forward under different actions, enabling imagination-based decision making (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020). More recently, large-scale foundation world models have extended this paradigm through massive video pretraining, producing temporally coherent and controllable simulations across diverse domains (Bruce et al., 2024; Agarwal et al., 2025; Decart et al., 2024). In the biological domain, virtual cell world models have begun to adopt this perspective by treating cellular systems as dynamical environments evolving under perturbations. Notably, VCWorld introduces a biological world model that integrates structured biological knowledge with iterative reasoning to simulate perturbation-induced signaling cascades, producing interpretable, stepwise predictions alongside mechanistic hypotheses (Wei et al., 2025). While such approaches represent an important step toward interpretable virtual cell simulators, existing world models—both in RL and biology—often prioritize predictive realism or symbolic reasoning in isolation, leaving open the question of how to learn *data-driven latent dynamics* that are simultaneously predictive, interpretable, and aligned with weak biological structure. This gap is especially salient in single-cell perturbation settings, where interventions provide explicit causal signals and latent dynamics must support both accurate prediction and mechanistic attribution.

In this work, we introduce PD-scWorld, an intervention-aware world model that explicitly encourages pathway-selective latent responses under perturbations, using only weak supervision in the form of gene-to-pathway mappings, as outlined in Figure 1. Our method learns action-conditioned latent transitions while imposing a group-sparse regularization that localizes perturbation-induced changes to subsets of latent factors associated with the targeted pathway. This design yields a latent space in which factors correspond more cleanly to biological programs, enabling interpretable attribution of perturbation effects. We demonstrate that the proposed approach achieves clearer disentanglement and higher mutual information with known covariates than β-VAE and unstructured latent dynamics baselines. Finally, through counterfactual rollouts on Perturb-seq datasets, we show that pathway-guided latents support more stable and biologically coherent reasoning about perturbation mechanisms.

2 Methodology

We model single-cell perturbation responses using a latent world model that explicitly separates a cell's baseline state from the changes induced by an intervention. Given a pre-perturbation cell state and a perturbation action, the model predicts how the latent representation should change, rather than directly reconstructing the post-perturbation profile, as outlined in Figure 1. To make these latent changes interpretable, we use weak pathway annotations to encourage each perturbation to affect only a small, consistent subset of latent factors associated with the targeted biological programs. This is achieved by regularizing the perturbation-induced latent update, while leaving the rest of the representation flexible. As a result, the model remains predictive while producing disentangled latent factors that align with known biological processes and support counterfactual reasoning.

2.1 Problem Formulation

We study single-cell perturbation datasets consisting of paired cellular states observed before and after an intervention. Let $x \in \mathbb{R}^G$ denote a cell's molecular profile (e.g., gene expression over G genes), and let $a \in \mathcal{A}$ denote a perturbation action such as a CRISPR knockout or drug treatment. Each perturbation a is associated with a (possibly empty) set of pathway annotations $\mathcal{P}(a) \subseteq \{1, \ldots, P\}$, where pathways correspond to weakly defined biological programs (e.g., cell cycle, interferon signaling).

We assume access to paired samples (x, x', a), where x' denotes the post-perturbation cellular state. Our goal is to learn a *latent world model* with an encoder f_θ, a latent state $z \in \mathbb{R}^K$, and an action-conditioned transition function T_ψ such that $z = f_\theta(x), \qquad z' = T_\psi(z, a)$, and z' accurately predicts the post-perturbation observation x'. Crucially, we seek a *disentangled* latent space in which individual latent factors respond selectively to perturbations associated with specific biological pathways, despite having access only to weak pathway-level supervision.

2.2 Latent World Model Architecture

Our model consists of three components: an encoder $f_\theta : \mathbb{R}^G \to \mathbb{R}^K$ that maps observed cellular profiles to latent states, an action-conditioned transition model T_ψ that predicts post-perturbation latents, and a decoder $g_\phi : \mathbb{R}^K \to \mathbb{R}^G$ that reconstructs molecular profiles from latent representations. The transition model is implemented as a residual update

$$z' = z + \Delta z = z + h_\psi(z, a), \tag{1}$$

where Δz explicitly represents the perturbation-induced change in latent space. This residual formulation isolates *what changes* under an intervention and enables targeted regularization of perturbation effects without constraining the full latent representation.

2.3 Predictive Latent World-Model Objective

To learn stable latent dynamics without contrastive sampling, we adopt a predictive objective inspired by joint-embedding predictive architectures. Given a pre- and post-perturbation pair (x, x'), we minimize

$$\mathcal{L}_{\text{pred}} = \left\| \operatorname{sg}(f_\theta(x')) - T_\psi(f_\theta(x), a) \right\|_2^2, \tag{2}$$

where $\operatorname{sg}(\cdot)$ denotes the stop-gradient operator. This objective encourages the predicted post-perturbation latent to match the encoded target latent while preventing representational collapse by blocking gradients through the target branch.

To ensure fidelity to observed molecular profiles, we optionally include a reconstruction loss

$$\mathcal{L}_{\text{rec}} = \|g_\phi(z) - x\|_2^2 + \|g_\phi(z') - x'\|_2^2. \tag{3}$$

2.4 Pathway-Guided Disentanglement via Group-Sparse Regularization

To encourage selective latent responses to perturbations, we associate each pathway $p \in \{1, \ldots, P\}$ with a subset of latent dimensions $\mathcal{Z}_p \subseteq \{1, \ldots, K\}$. These associations are *soft and learned jointly* with the model rather than fixed or manually specified.

For a given perturbation a, only latent dimensions associated with pathways in $\mathcal{P}(a)$ should exhibit consistent changes across cells. We enforce this inductive bias by regularizing the perturbation-induced change Δz.

Variance-based localization. We penalize latent drift outside the targeted pathway groups using

$$\mathcal{L}_{\text{var}}(a) = \sum_{k \notin \cup_{p \in \mathcal{P}(a)} \mathcal{Z}_p} \text{Var}(\Delta z_k \mid a)\,, \tag{4}$$

where the variance is computed across cells subjected to the same perturbation. This term suppresses spurious and inconsistent latent responses unrelated to the perturbation's biological target.

Group sparsity penalty. To further localize effects, we impose a group-sparsity constraint

$$\mathcal{L}_{\text{group}}(a) = \sum_{p \notin \mathcal{P}(a)} \left\| \Delta z_{\mathcal{Z}_p} \right\|_2\,, \tag{5}$$

which discourages coordinated changes in latent groups unrelated to the targeted pathways. Together, these regularizers encourage perturbations to activate *small, stable subsets* of latent factors aligned with known biological programs.

2.5 Overall Training Objective

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{var}}\mathcal{L}_{\text{var}} + \lambda_{\text{group}}\mathcal{L}_{\text{group}}, \tag{6}$$

where λ_{rec}, λ_{var}, and λ_{group} control the strength of reconstruction and pathway-guided regularization. Importantly, pathway annotations influence learning only through regularization on Δz, allowing the model to generalize to unseen perturbations and pathways.

2.5.1 Relationship to Existing Latent Modeling Objectives

To clarify the role of the proposed objective, we compare PD-scWorld with common latent variable models for single-cell data, including β-VAE, FactorVAE, and scVI. Although all learn low-dimensional representations of high-dimensional gene expression, they differ in how the latent space is structured and how biological variation is organized within it.

Shared Representation Learning Structure. All models follow a similar representation-learning paradigm: an encoder maps gene expression profiles to a latent space, and a decoder reconstructs the molecular profile from the latent representation. Formally, given a cell expression vector $x \in \mathbb{R}^G$, each method learns an encoder $f_\theta(x)$ and decoder $g_\phi(z)$ defining a compressed cell-state representation. In this sense, PD-scWorld shares the same autoencoding backbone used by β-VAE, FactorVAE, and scVI.

The primary differences arise in the constraints imposed on the latent space and how perturbations are incorporated into training.

β-VAE and FactorVAE: Independence-Based Disentanglement. β-VAE and FactorVAE aim to produce disentangled representations by encouraging statistical independence among latent variables. Both extend the standard VAE objective

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\theta(z|x)}[\log p_\phi(x|z)] - \text{KL}\big(q_\theta(z|x) \| p(z)\big). \tag{7}$$

β-VAE increases the KL weight,

$$\mathcal{L}_{\beta\text{-VAE}} = \mathbb{E}_{q_\theta(z|x)}[\log p_\phi(x|z)] - \beta\,\text{KL}\big(q_\theta(z|x) \| p(z)\big), \tag{8}$$

where $\beta > 1$ promotes stronger factor independence. FactorVAE instead penalizes latent *total correlation*:

$$\mathcal{L}_{\text{FactorVAE}} = \mathcal{L}_{\text{VAE}} - \gamma\,\text{TC}(z). \tag{9}$$

While effective in synthetic settings, strict independence assumptions can conflict with biological systems, where regulatory pathways and transcriptional programs exhibit coordinated activity.

scVI: Probabilistic Modeling of Gene Expression. scVI focuses on modeling the statistical structure of transcriptomic counts using a hierarchical generative model with a negative binomial likelihood:

$$p(x|z) = \mathrm{NB}(\mu(z), \theta). \tag{10}$$

Latent variables capture biological variability while explicitly modeling technical factors such as sequencing depth. Although scVI produces useful embeddings for clustering and batch correction, it does not explicitly model perturbation dynamics, so intervention effects are typically entangled within the latent space.

PD-scWorld: Dynamics-Aware Disentanglement. PD-scWorld differs by modeling perturbations as explicit *latent transitions*. Rather than constraining the marginal latent distribution, the model learns an action-conditioned transition function and optimizes the predictive latent alignment objective in Eq. 2. The model must therefore learn representations that evolve predictably under interventions. Disentanglement is encouraged through regularization of the perturbation-induced update $\Delta z = T_\psi(z, a) - z$, which is biased to activate pathway-associated latent groups. Consequently, disentanglement is imposed on *latent dynamics* rather than static distributions, and weak biological knowledge (pathway annotations) guides latent organization.

In summary, β-VAE and FactorVAE enforce independence-based disentanglement, scVI emphasizes probabilistic modeling of gene expression counts, and PD-scWorld instead leverages perturbation structure to learn dynamics-aware latent representations aligned with biological pathways.

2.6 Counterfactual Rollouts and Interpretability

Once trained, the model supports counterfactual reasoning by simulating latent transitions under arbitrary perturbations:

$$\hat{z}^{(t+1)} = T_\psi\left(\hat{z}^{(t)}, a_t\right). \tag{11}$$

Inspecting which latent dimensions change under specific actions provides interpretable attributions linking perturbations to biological programs.
We quantify interpretability using mutual information between latent dimensions and known covariates, pathway selectivity scores measuring concentration of Δz, and the stability of latent responses under repeated or composite perturbations.

2.7 Relation to Existing Disentanglement Approaches

Unlike β-VAE-style methods that enforce global independence constraints, our approach introduces *intervention-conditional disentanglement*, aligning latent structure with perturbation semantics. Compared to fully supervised factor models, it avoids brittle labeling assumptions while leveraging weak biological knowledge to guide representation learning.

3 Experiments

Datasets. We evaluate the proposed pathway-guided intervention-aware world model on CRISPR-based single-cell perturbation datasets derived from Perturb-seq experiments (Dixit et al., 2016), which provide paired pre- and post-perturbation gene expression profiles at single-cell resolution. Each perturbation is mapped to one or more biological pathways using curated gene-to-pathway annotations, such as cell-cycle regulation, interferon signaling, and DNA damage response. These pathway annotations are used exclusively as weak structural priors during training and never as direct supervision on latent variables. To assess generalization, we split data by perturbation identity, ensuring that perturbations appearing in the test set are never observed during training.

Baselines. We compare against a range of representative baselines that capture common approaches to latent modeling and disentanglement in single-cell analysis. These include β-VAE (Higgins et al., 2017) and FactorVAE (Kim & Mnih, 2019), which enforce disentanglement through global independence constraints, scVI (Lopez et al., 2018) as a probabilistic latent-variable model widely used in single-cell genomics, an action-conditional autoencoder that incorporates perturbation embeddings without structural regularization, and an unstructured latent world model that shares our architecture but omits pathway-guided constraints. All baselines are matched in latent dimensionality and network capacity to isolate the effect of pathway-guided disentanglement.

Evaluation Protocols. Evaluation is performed under three complementary regimes. First, we assess perturbation prediction by measuring how accurately models predict post-perturbation expression states from pre-perturbation inputs and actions, focusing on held-out perturbations to test extrapolation. Second, we evaluate disentanglement and interpretability by quantifying the alignment between latent dimensions and known biological covariates, including pathway activation and cell-cycle phase, which are not used for training supervision. Third, we analyze counterfactual rollouts by simulating targeted perturbations in latent space and examining the consistency and localization of latent changes across cells subjected to the same intervention.

Evaluation Metrics. Predictive performance is evaluated using mean squared error (MSE) between predicted and observed post-perturbation expression,

$$\mathrm{MSE} = \mathbb{E}\big[\,\|\hat{x}' - x'\|_2^2\,\big]\,, \tag{12}$$

as well as differential expression (DE) recovery, measured by the Pearson correlation between predicted and observed log-fold changes across genes,

$$\text{DE-Corr} = \mathrm{corr}(\Delta\hat{x}, \Delta x)\,, \quad \Delta x = \log x' - \log x. \tag{13}$$

Disentanglement quality is quantified using the mutual information between individual latent dimensions z_k and known biological covariates c (e.g., pathway activity or cell-cycle phase),

$$\mathrm{MI}(z_k, c) = \mathbb{E}_{z_k, c}\left[\log \frac{p(z_k, c)}{p(z_k)\,p(c)}\right], \tag{14}$$

with reported scores corresponding to the maximum or average mutual information across latent dimensions, depending on the evaluation.

To assess localization of perturbation effects, we define a pathway selectivity score measuring the fraction of perturbation-induced latent change concentrated within pathway-associated latent groups,

$$\mathrm{Sel}(a) = \frac{\|\Delta z_{\mathcal{Z}_{\mathcal{P}(a)}}\|_1}{\|\Delta z\|_1}, \quad \Delta z = T_\psi(z, a) - z, \tag{15}$$

where $\mathcal{Z}_{\mathcal{P}(a)} = \bigcup_{p \in \mathcal{P}(a)} \mathcal{Z}_p$.

Interpretability and stability are further evaluated by measuring off-target latent variance,

$$\mathrm{OffVar}(a) = \sum_{k \notin \mathcal{Z}_{\mathcal{P}(a)}} \mathrm{Var}(\Delta z_k \mid a)\,, \tag{16}$$

which penalizes inconsistent latent changes unrelated to the targeted pathways. Finally, counterfactual rollout consistency is assessed using the average pairwise cosine similarity of latent transitions across cells subjected to the same perturbation,

$$\mathrm{Cons}(a) = \mathbb{E}_{i \neq j}\left[\frac{\Delta z_i^\top \Delta z_j}{\|\Delta z_i\|_2\,\|\Delta z_j\|_2}\right], \tag{17}$$

where higher values indicate more stable and coherent latent dynamics under intervention.

Implementation Details. All models employ two-layer multilayer perceptrons with ReLU activations for the encoder and decoder, using a hidden dimension of 256 and a latent dimensionality of 32. Perturbations are represented using learned action embeddings of dimension 64 and integrated into the transition model via concatenation with latent states. Models are trained using the Adam optimizer with a learning rate of 10^{-3} and batch size of 256 for up to 200 epochs, with early stopping based on validation predictive loss. The reconstruction loss weight is fixed to 1.0, while the pathway-guided variance and group-sparsity regularization coefficients are selected from $0.1, 0.5, 1.0$ and $0.01, 0.05, 0.1$, respectively, based on validation trade-offs between predictive accuracy and mutual information. All experiments are repeated over three random seeds, and we report mean performance with standard deviation.

Across all datasets and evaluation settings, the proposed model achieves predictive accuracy comparable to or exceeding that of unstructured latent baselines, while substantially improving latent factor interpretability. In particular, perturbation-induced latent changes are more consistently localized to pathway-relevant dimensions, yielding higher mutual information with known covariates and reduced off-target variance compared to β-VAE and FactorVAE. Counterfactual rollouts further demonstrate that pathway-guided latents exhibit stable, biologically coherent responses across cells, supporting reliable mechanistic interpretation without sacrificing predictive performance.

4 Experimental Results

4.1 Overall Performance

We first evaluate predictive performance on held-out perturbations to verify that pathway-guided disentanglement does not compromise accuracy. As shown in Table 1, the proposed model matches or slightly outperforms unstructured latent world models and scVI in post-perturbation expression prediction, while significantly outperforming β-VAE and FactorVAE baselines. Notably, reconstruction-focused disentanglement methods suffer from reduced predictive accuracy, suggesting that global independence constraints interfere with modeling structured perturbation effects. In contrast, our intervention-aware regularization preserves predictive capacity by constraining only the direction and localization of latent changes rather than the full latent distribution.

Table 1: Post-perturbation prediction performance on held-out perturbations.

Model	MSE ↓	DE Corr. ↑
scVI	0.184 ± 0.006	0.62 ± 0.02
β-VAE	0.231 ± 0.011	0.48 ± 0.03
FactorVAE	0.219 ± 0.009	0.51 ± 0.02
Action-conditional AE	0.176 ± 0.005	0.64 ± 0.01
Unstructured world model	0.171 ± 0.004	0.66 ± 0.01
PD-scWorld (ours)	**0.168 ± 0.004**	**0.68 ± 0.01**

4.2 Latent Disentanglement and Interpretability

We next assess latent disentanglement and interpretability by measuring the alignment between latent dimensions and known biological covariates. Table 2 reports the maximum mutual information between individual latent factors and pathway activity scores, as well as the average mutual information with cell-cycle phase annotations. While β-VAE and FactorVAE increase marginal independence, they fail to produce factors that align consistently with biological programs. In contrast, the proposed method yields substantially higher mutual information with known covariates, indicating that latent factors correspond more directly to meaningful biological processes.

Table 2: Latent interpretability measured by mutual information (MI) with known covariates.

Model	Max MI (Pathway) ↑	Avg MI (Cell Cycle) ↑
β-VAE	0.09 ± 0.01	0.07 ± 0.01
FactorVAE	0.11 ± 0.01	0.09 ± 0.01
scVI	0.14 ± 0.02	0.12 ± 0.02
Unstructured world model	0.17 ± 0.01	0.15 ± 0.01
PD-scWorld (ours)	**0.28 ± 0.02**	**0.24 ± 0.02**

To directly evaluate whether perturbation effects are localized to pathway-relevant latent factors, we analyze the distribution of perturbation-induced latent changes Δz. Table 3 reports the pathway selectivity score, defined as the proportion of latent change magnitude concentrated within pathway-associated latent groups, along with the variance of latent changes outside those groups. The proposed model achieves substantially higher selectivity and lower off-target variance, indicating that perturbations activate compact and consistent subsets of latent dimensions. In contrast, unstructured models exhibit diffuse latent responses, and β-VAE-style approaches suppress variance globally rather than selectively.

Table 3: Localization of perturbation effects in latent space.

Model	Pathway Selectivity ↑	Off-target Var. ↓
β-VAE	0.31 ± 0.04	0.082 ± 0.009
FactorVAE	0.35 ± 0.03	0.075 ± 0.008
Unstructured world model	0.42 ± 0.02	0.061 ± 0.006
PD-scWorld (ours)	**0.67** ± **0.03**	**0.029** ± **0.004**

Finally, we quantify the stability of latent dynamics under repeated interventions by evaluating counterfactual rollouts across cells subjected to the same perturbation. Specifically, we measure the average pairwise cosine similarity between perturbation-induced latent changes, as well as multi-step rollout error over repeated applications of the same intervention. As shown in Table 4, the proposed model produces significantly more consistent latent transition directions and lower rollout drift than all baselines. In contrast, unstructured and reconstruction-driven models exhibit high variability across cells, limiting their utility for mechanistic interpretation. These results confirm that pathway-guided regularization yields stable and coherent latent dynamics under intervention.

Table 4: Counterfactual rollout stability under repeated perturbations. Higher consistency and lower rollout error indicate more stable latent dynamics.

Model	Latent Consistency ↑	2-step Rollout Error ↓
β-VAE	0.41 ± 0.05	0.213 ± 0.018
FactorVAE	0.44 ± 0.04	0.198 ± 0.016
scVI	0.53 ± 0.03	0.172 ± 0.014
Unstructured world model	0.61 ± 0.02	0.156 ± 0.012
PD-scWorld (ours)	**0.78** ± **0.02**	**0.112** ± **0.010**

5 Discussion

Our results demonstrate that pathway-guided, intervention-aware disentanglement provides a practical middle ground between fully unsupervised representation learning and brittle, fully supervised factor models. By constraining only the *perturbation-induced change* in latent space rather than the full latent distribution, the proposed approach preserves predictive accuracy while yielding substantially more interpretable latent factors. This distinction is critical in perturbation modeling, where the goal is not merely to reconstruct post-intervention states, but to understand *which biological programs are affected and how*. The strong improvements in pathway selectivity and latent consistency suggest that weak biological structure, when incorporated at the level of dynamics, can meaningfully shape representation geometry.

A key insight from this work is that disentanglement in biological systems is inherently *conditional on intervention*. Unlike static datasets where factors may be globally independent, cellular responses reflect context-dependent activation of overlapping programs. Methods that enforce global independence, such as β-VAE-style objectives, suppress this structure and lead to degraded predictive performance. In contrast, intervention-conditional regularization aligns more naturally with the causal semantics of perturbation experiments, enabling latent factors to remain entangled when appropriate and disentangled only when an action provides identifying signal.

The counterfactual rollout analysis further highlights the value of viewing perturbation modeling through the lens of world models. Stable multi-step latent transitions indicate that the learned dynamics capture consistent, reusable structure rather than memorizing single-step mappings. This property is essential for downstream applications such as combinatorial perturbation prediction, target prioritization, and mechanistic hypothesis generation. While our experiments focus on single perturbations, the framework naturally extends to sequences of interventions and time-resolved perturbation data.

There are several limitations worth noting. First, pathway annotations are incomplete and context-dependent, and incorrect or overly broad pathway mappings may introduce bias into the learned latent structure. Second, while the proposed regularization improves interpretability, it does not guarantee identifiability of latent factors in a formal sense. Finally, our evaluation is restricted to transcriptomic readouts; extending the approach to multi-omic perturbation data is an important direction for future work.

6 Related Work

Deep generative models have become a cornerstone in single-cell representation learning and perturbation modeling, often built on variational autoencoders (VAEs) or their extensions. Classic VAE-based frameworks have been used for interpretable embedding of single-cell transcriptomes, enabling the identification of gene modules or hierarchical cell states by modifying decoder structures to reflect biological signals (Choi et al., 2023). Extensions such as Lorentz-regularized VAE introduce geometric constraints to balance fine-grained fidelity and global structure in latent space, demonstrating improved latent geometry and biological interpretability across multi-scale single-cell data (Fu et al., 2026). Other models like scVAEDer integrate deep diffusion models with VAEs to capture both global structure and local variation, enabling perturbation prediction and generation of novel cellular states but often without explicit aim toward disentangled causal factors (Sadria & Layton, 2025). Generative frameworks tailored for perturbation data, such as CRADLE-VAE, incorporate counterfactual reasoning to separate technical artifacts from biological effects, improving treatment effect estimation and generative quality (Baek et al., 2024). However, these methods typically focus on artifact disentanglement or generative quality rather than aligning latent geometry with perturbation semantics, leaving a gap for dynamics-aware, intervention-conditional representations.

Beyond generic generative modeling, several works explicitly investigate disentanglement and causal inference in single-cell latent spaces. For example, models based on sparse mechanism shift assume that perturbations target sparse subsets of latent variables, enabling recovery of causal factors and generalization to unseen conditions under certain assumptions (Lopez et al., 2023). Similarly, scDRP leverages disentangled latent representations to separate perturbation-dependent and independent components for individualized treatment effect estimation in complex exposure scenarios (Sun et al., 2025). More recently, virtual cell world models such as VCWorld have proposed biologically grounded simulators that integrate structured biological knowledge and iterative reasoning to produce interpretable, stepwise predictions of perturbation effects (Wei et al., 2025). While these approaches advance interpretability and causal reasoning, they either rely on strong assumptions about distribution shifts or emphasize symbolic and rule-based reasoning rather than learning compact, data-driven latent dynamics. Other single-cell representation learning methods focus on multi-view integration across modalities (e.g., RNA and ATAC) or embedding biological priors such as intercellular signaling to improve clustering and biological contextualization (Li et al., 2022; Qi et al., 2025). In contrast, our work uses weak pathway-level supervision to guide disentanglement of *learned latent dynamics* under perturbations, bridging structured biological knowledge with predictive world-model learning tailored to single-cell interventions.

In contrast to prior generative and disentanglement approaches, our method guides representation learning using weak pathway-level supervision applied specifically to perturbation-induced latent dynamics rather than static latent structure. This intervention-conditional design yields interpretable, stable latent factors without sacrificing predictive accuracy or requiring strong causal assumptions.

7 Conclusion

We introduced an intervention-aware latent world model for single-cell perturbation data that leverages weak pathway supervision to guide disentanglement of biological factors. By regularizing perturbation-induced latent changes rather than static representations, the proposed method achieves a latent space that is simultaneously predictive, interpretable, and stable under intervention. Empirical results on Perturb-seq datasets show consistent improvements in pathway alignment, localization of perturbation effects, and counterfactual rollout stability over both reconstruction-driven disentanglement methods and unstructured world models.

More broadly, this work suggests that effective disentanglement in biological systems should be framed as a *dynamics-aware* problem, where interventions provide the key signal for separating underlying programs. We believe this perspective opens new opportunities for integrating causal reasoning, weak biological knowledge, and world-model learning in genomics. Future work will explore extensions to multi-omic perturbations, combinatorial interventions, and downstream applications in drug discovery and functional genomics.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Seungheun Baek, Soyon Park, Yan Ting Chok, Junhyun Lee, Jueon Park, Mogan Gim, and Jaewoo Kang. Cradle-vae: Enhancing single-cell gene perturbation modeling with counterfactual reasoning-based artifact disentanglement, 2024. URL `https://arxiv.org/abs/2409.05484`.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Yongin Choi, Ruoxin Li, and Gerald Quon. sivae: interpretable deep generative models for single-cell transcriptomes. *Genome Biology*, 24(1), February 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-02850-y. URL `http://dx.doi.org/10.1186/s13059-023-02850-y`.

Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. 2024. URL `https://oasis-model.github.io/`.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, December 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2016.11.038. URL `http://dx.doi.org/10.1016/j.cell.2016.11.038`.

Zeyu Fu, Jiawei Fu, Chunlin Chen, Keyang Zhang, and Song Wang. Lorentz-regularized interpretable vae for multi-scale single-cell transcriptomic and epigenomic embeddings. *Frontiers in Genetics*, 16, January 2026. ISSN 1664-8021. doi: 10.3389/fgene.2025.1713727. URL `http://dx.doi.org/10.3389/fgene.2025.1713727`.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising, 2019. URL `https://arxiv.org/abs/1802.05983`.

Gaoyang Li, Shaliu Fu, Shuguang Wang, Chenyu Zhu, Bin Duan, Chen Tang, Xiaohan Chen, Guohui Chuai, Ping Wang, and Qi Liu. A deep generative model for multi-view profiling of single-cell rna-seq and atac-seq data. *Genome Biology*, 23(1), January 2022. ISSN 1474-760X. doi: 10.1186/s13059-021-02595-6. URL `http://dx.doi.org/10.1186/s13059-021-02595-6`.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, November 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0229-2. URL `http://dx.doi.org/10.1038/s41592-018-0229-2`.

Romain Lopez, Nataša Tagasovska, Stephen Ra, Kyunghyn Cho, Jonathan K. Pritchard, and Aviv Regev. Learning causal representations of single cells via sparse mechanism shift modeling, 2023. URL `https://arxiv.org/abs/2211.03553`.

Cong Qi, Yeqing Chen, and Zhi Wei. Clustering with communication: A variational framework for single cell representation learning, 2025. URL `https://arxiv.org/abs/2505.04891`.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, August 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01905-6. URL `http://dx.doi.org/10.1038/s41587-023-01905-6`.

Mehrshad Sadria and Anita Layton. scvaeder: integrating deep diffusion models and variational autoencoders for single-cell transcriptomics analysis. *Genome Biology*, 26(1), March 2025. ISSN 1474-760X. doi: 10.1186/s13059-025-03519-4. URL `http://dx.doi.org/10.1186/s13059-025-03519-4`.

Jianle Sun, Petar Stojanov, and Kun Zhang. Single-cell disentangled representations for perturbation modeling and treatment effect estimation. November 2025. doi: 10.1101/2025.11.21.689783. URL `http://dx.doi.org/10.1101/2025.11.21.689783`.

Zhijian Wei, Runze Ma, Zichen Wang, Zhongmin Li, Shuotong Song, and Shuangjia Zheng. Vcworld: A biological world model for virtual cell simulation, 2025. URL `https://arxiv.org/abs/2512.00306`.