A Scalable Latent Diffusion Model for Single-Cell Gene Expression Data

Giovanni Palla
Chan Zuckerberg Initiative
Redwood City, CA
gpalla@chanzuckerberg.com

Sudarshan Babu
CZ Biohub
Chicago, IL
sudarshan.babu@czbiohub.org

Payam Dibaeinia
CZ Biohub
Chicago, IL
payam.dibaeinia@czbiohub.org

Donghui Li
Chan Zuckerberg Initiative
Redwood City, CA
dli@chanzuckerberg.com

Aly A. Khan
CZ Biohub
Chicago, IL
aakhan@czbiohub.org

Theofanis Karaletsos*
Chan Zuckerberg Initiative
Redwood City, CA
tkaraletsos@chanzuckerberg.com

Jakub M. Tomczak*
Chan Zuckerberg Initiative
Redwood City, CA
jtomczak@chanzuckerberg.com

Abstract

Computational modeling of single-cell gene expression is crucial for understanding cellular processes, but generating realistic expression profiles remains a major challenge. This difficulty arises from unbounded counts of expression data and complex dependencies within gene sets. Existing generative models often impose artificial gene orderings, reducing both their flexibility and biological relevance. We introduce a scalable latent diffusion model for single-cell gene expression that respects the fundamental exchangeability property of gene measurements. Unlike existing approaches requiring artificial orderings or complex hierarchies, we propose a streamlined VAE using fixed-size latent variables with permutation-invariant and permutation-equivariant components. Our unified Multi-head Cross-Attention Block (MCAB) serves dual roles: permutation-invariant pooling in the encoder and permutation-equivariant unpooling in the decoder, eliminating separate mechanisms for varying gene sets. We enhance this framework by replacing the Gaussian prior with a latent diffusion model using Diffusion Transformers and linear interpolants, enabling high-quality generation with multi-conditional classifier-free guidance. Our approach naturally handles single-cell data challenges like high-dimensional sparsity that is demonstrated by its superior performance in unconditional and conditional cell generation benchmarks.

*Co-last authors

The NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development (AI4D3 2025), San Diego, California, USA, 2025.

1 Introduction

Single-cell transcriptomics has revolutionized our understanding of cellular heterogeneity and biological processes at unprecedented resolution [26], enabling high-throughput gene expression profiling across thousands/millions of cells [30], and providing valuable insights into cellular differentiation [7], disease progression [31], and responses to drug perturbations [2, 3, 32]. However, modeling the complex, high-dimensional gene expression data from single cells presents significant computational and methodological challenges [12, 17, 20].

Deep generative modeling [28] offers a powerful framework to formulate expressive probability distributions. In the context of single-cell data, multiple methods have been proposed. In particular, Variational Auto-Encoders (VAEs) have been extensively utilized for representation learning (single-cell Variational Inference; scVI) [14], perturbation modeling [15, 22], trajectory inference [5], among others [4]. Additionally, Generative Adversarial Networks (GANs) have also been proposed, both for generating realistic cell populations (scGAN; [19]) and for inferring cellular trajectories [24]. Recently, diffusion-based models have also been adopted for single-cell gene expression [18]. An interesting research line was proposed in [21] that combines scVI with a flow matching in the latent space (CellFlow for Generation; CFGen).

However, two key challenges limit existing methods. First, and most fundamentally, they often require a fixed ordering of genes or operate on a restricted subset of highly variable genes (HGVs). This assumption directly clashes with the biological reality that gene expression profiles are **exchangeable** sets, where the order of genes carries no meaning. Second, approaches based on GANs inherit well-known training instabilities and risks of mode collapse. These limitations make current models inflexible, difficult to scale, and unable to properly handle the unordered nature of single-cell data.

This paper introduces a novel approach that combines the flexibility of VAEs with the power of latent diffusion models (see Figure 1), specifically designed to handle the exchangeable nature of gene expression data. The key insight is that careful architectural choices, particularly in the parameterization of permutation-invariant and permutation-equivariant components, results in a scalable, deep, and exchangeable generative model. The contributions of the paper are the following:

- We propose a novel fully transformer-based VAE architecture for exchangeable data that uses a single set of fixed-size, permutation-invariant latent variables. The model introduces a Multi-head Cross-Attention Block (MCAB) that serves dual purposes: It acts as a permutation-invariant pooling operator in the encoder, and functions as a permutation-equivariant unpooling operator in the decoder. This unified approach eliminates the need for separate architectural components for handling varying set sizes.
- We replace the standard Gaussian prior with a latent diffusion model trained using linear interpolants using SiT (Scalable Interpolant Transformers) and a denoiser parameterized by Diffusion Transformers (DiT). This allows for better modeling of the complex distribution of cellular states and enables controlled generation through classifier-free guidance.
- The proposed framework supports generation conditioned on multiple attributes simultaneously through an extended classifier-free guidance mechanism, enabling fine-grained control over generated cell states, as demonstrated on multiple benchmark datasets.

2 Methodology

2.1 Problem formulation

Let us consider M random variables, $\mathbf{x}$, where each $\mathbf{x}_i \in \mathbb{X}^D$, e.g., $\mathbb{X} = \mathbb{N}$. A set of indices of M random variables is denoted as $\mathcal{I}$, namely, $\mathcal{I} = \pi(\{1, 2, \ldots, M\})$, where $\pi(\cdot)$ is a permutation[2]. Further, we denote a specific order of variables in $\mathbf{x}$ determined by $\mathcal{I}$ as $\mathbf{x}_\mathcal{I}$. We assume that for a given $\mathcal{I}$, an object $\mathbf{x}_\mathcal{I}$ is equivalent to an object defined by $\pi(\mathcal{I})$, namely, $\mathbf{x}_\mathcal{I} = \mathbf{x}_{\pi(\mathcal{I})}$. An example of such a setting is gene expression data where $\{1, 2, \ldots, M\}$ corresponds to gene IDs and the order of gene IDs does not change the state of a cell.

[2]We denote a permutation either as a function $\pi(\cdot)$ or, equivalently, as a matrix $\mathbf{P}$.

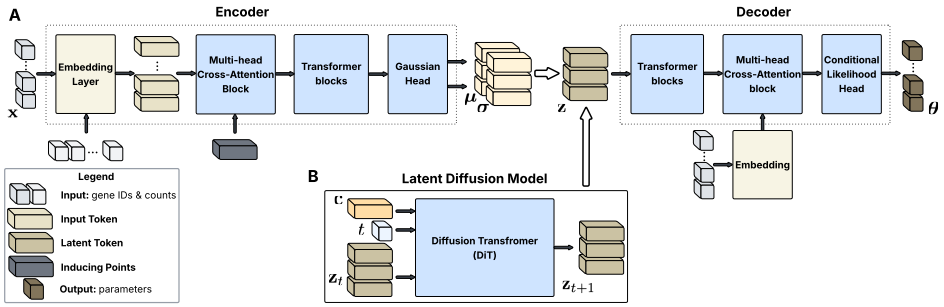

Figure 1: Our deep generative model for single-cell gene expression data. **A**: A fully transformer-based architecture for processing gene expressions. The encoder network results in permutation-invariant latent variables represented as tokens. The decoder network returns permutation-equivariant counts for given gene IDs. **B**: At the second stage, a vanilla prior is replaced by a latent diffusion model. We model latent tokens using Diffusion Transformers (DiT), and train the resulting LDM using linear interpolants and the flow matching loss. Sampling is carried out by applying the Scalable Interpolant Transformers (SiT) library.

Further, we assume a *true* conditional distribution model $p(\mathbf{x}_\mathcal{I}|\mathcal{I})$ that for given indices $\mathcal{I}$ allows sampling $\mathbf{x}_\mathcal{I}$ and representing $\mathbf{x}_\mathcal{I}$ either in the form of the likelihood or embeddings. We access this *true* distribution through observed *iid* data $\mathcal{D} = \{(\mathbf{x}_{\mathcal{I}_n}, \mathcal{I}_n)\}_{n=1}^N$. We look for a model $p_\theta(\mathbf{x}_\mathcal{I}|\mathcal{I})$ with parameters θ that fits best the *true* distribution approximated by the empirical distribution for given data $\mathcal{D}$, $p_\mathcal{D}(\mathbf{x}_\mathcal{I}|\mathcal{I})$ in the sense of the log-likelihood function:

$$\ell(\theta; \mathcal{D}) = \sum_{n=1}^{N} \ln p_\theta(\mathbf{x}_{\mathcal{I}_n}|\mathcal{I}_n). \tag{1}$$

Moreover, we are interested in finding a single model that for given indices $\mathcal{I}$ generates corresponding $\mathbf{x}_\mathcal{I}$. For instance, we would like to have a model that can generate gene expression for given different orders of gene IDs. Formally, we require the model to be *exchangeable*, namely, $p(\mathbf{x}_\mathcal{I}|\mathcal{I}) = p(\mathbf{x}_{\pi(\mathcal{I})}|\pi(\mathcal{I}))$.

2.2 scLDM: Our approach

2.2.1 Variational Auto-Encoder for exchangeable random variables

To model an exchangeable probabilistic model $p(\mathbf{x}_\mathcal{I}|\mathcal{I})$, we introduce m latent variables (i.e., the number of latents is fixed for all subsets $\mathcal{I}$), $\mathbf{Z} \in \mathbb{R}^{m \times D}$. By using the family of variational posteriors of the form $q_\phi(\mathbf{Z}|\mathbf{x}_\mathcal{I})$, the Evidence Lower BOund (ELBO) is the following:

$$\ln p(\mathbf{x}_\mathcal{I}|\mathcal{I}) \geq \mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{Z}|\mathbf{x}_\mathcal{I})} \left[\ln p_\theta(\mathbf{x}_\mathcal{I}|\mathbf{Z}, \mathcal{I}) + \ln p_\psi(\mathbf{Z}) - \ln q_\phi(\mathbf{Z}|\mathbf{x}_\mathcal{I})\right]. \tag{2}$$

Since our assumption is that the model must be exchangable, we propose to parameterize the distributions in a way that: (i) $\mathbf{Z}$ is permutation-invariant, namely, we aim for defining variational posteriors as Gaussian distributions with permutation-invariant neural networks parameterizing means and variances: $\{\mu, \sigma^2\} = \mathrm{NN}(\mathbf{x}_\mathcal{I}) = \mathrm{NN}(\mathbf{x}_{\pi(\mathcal{I})})$, (ii) the conditional likelihood is defined as $p_\theta(\mathbf{x}_\mathcal{I}|\mathbf{Z}, \mathcal{I}) = \prod_{i\in I} p_\theta(\mathbf{x}_i|\mathbf{Z})$, hence, we must ensure that for permuting $\mathcal{I}$, we permute the parameters accordingly; in other words, we must have a permutation-equivariant neural net: $\mathbf{P}_\theta = \mathrm{NN}(\mathbf{Z}, \pi(\mathcal{I}))$.

2.2.2 Fully-transformer-based Variational Auto-Encoder

Permutation-invariant/equivariant Cross-Attention Our VAE is parameterized by transformer-based architecture; however, we do not use causal masks. Moreover, we require a layer that allows pooling/unpooling because, otherwise, attention layers would deal with tens of thousends or more

tokens. We introduce the following multi-head cross-attention block (MCAB):

$$\mathrm{MCAB}_{\mathbf{S}}(\mathbf{X}) = F(\mathbf{X}, \mathbf{S}) + \mathrm{MLP}(\mathrm{LN}_F(F(\mathbf{X}, \mathbf{S})) \tag{3}$$

$$F(\mathbf{X}, \mathbf{S}) = \mathbf{Q} + \mathrm{Att}_K\left(\mathrm{LN}_Q(\mathbf{Q}), \mathbf{K}, \mathbf{V}\right) \tag{4}$$

$$\mathbf{Q} = \mathrm{Linear}_S(\mathbf{S}) \tag{5}$$

$$\mathbf{K} = \mathrm{Linear}_K(\mathrm{LN}_K(\mathbf{X})) \tag{6}$$

$$\mathbf{V} = \mathrm{Linear}_V(\mathrm{LN}_V(\mathbf{X})), \tag{7}$$

where Linear is a linear layer, $\mathrm{LN}(\cdot)$ denotes a layer norm, and $\mathrm{MLP}(\cdot)$ is a small fully-connected neural network, e.g., $\mathrm{MLP}(\mathbf{X}) = (\mathrm{Linear} \circ (\mathrm{Linear} \odot (\mathrm{silu} \circ \mathrm{Linear}))(\mathbf{X})$.[3] This block is defined similarly to a block used in a successor of SetTransformer called Perceiver [10, 11].

MCAB is either permutation-invariant or permutation-equivariant. Since it relies on the attention mechanism, if we permute $\mathbf{X}$ but do not permute $\mathbf{S}$, then MCAB is permutation-invariant (see Property 3). However, if we process $\mathbf{Z}$ by a permutation-invariant function and we permute $\mathbf{S}$ accordingly to the permuted indices, then MCAB becomes permutation-equivariant (see Property 4). As a result, we use MCAB as a permutation-invariant pooling operator in the encoder network, and as a permutation-equivariant unpooling operator in the decoder network.

Encoder (Variational Posterior) We define the family of variational posteriors as Gaussians, $q_\phi(\mathbf{Z}|\mathbf{x}_\mathcal{I}) = \mathcal{N}(\mathbf{Z}|\mu(\mathbf{x}_\mathcal{I}), \sigma(\mathbf{x}_\mathcal{I}))$. We need $\mathbf{Z}$ to be of fixed size and invariant to permutations of $\mathbf{x}_\mathcal{I}$, hence, we propose the following architecture of the encoder network:

$$\mathrm{NN}_{enc}(\mathbf{x}_\mathcal{I}, \mathcal{I}) = \left(\mathrm{T}_L \circ \mathrm{T}_{L-1} \circ \ldots \circ \mathrm{T}_1 \circ \mathrm{MSCAB}_{\mathbf{S}} \circ \mathrm{Embedding}\right)(\mathbf{x}_\mathcal{I}, \mathcal{I}), \tag{8}$$

where $\mathrm{T}_l(\cdot)$ denotes a transformer block, e.g., $\mathrm{T}_l(\mathbf{X}) = ((\mathrm{Id} \oplus (\mathrm{MLP} \circ \mathrm{LN}_2)) \circ (\mathrm{Id} \oplus (\mathrm{Att}_K \circ \mathrm{LN}_1)))(\mathbf{X})$, and $\mathrm{Embedding}(\cdot, \cdot)$ is an embedding layer. Since inputs $\mathbf{x}_\mathcal{I}$ form a (column) vector of counts, and $\mathcal{I}$ are IDs, we propose to use the following embedding layer:

$$\mathrm{Embedding}(\mathbf{x}_\mathcal{I}, \mathcal{I}) = \mathrm{Linear} \circ (\mathrm{repeat}_D(\mathbf{x}_\mathcal{I}) \boxplus \mathbf{E}_\mathcal{I}), \tag{9}$$

where repeat_D repeats the counts D-times resulting in a matrix $M \times D$, Linear projects the concatenated $2D$-dimensional space to the D-dimensional space, and $\mathbf{E} \in \mathbb{R}^{M \times D}$ is the embedding matrix. The rationale behind this way of embedding both counts and indices is to mix the information and be able to learn the mixing through a projection (linear) layer.

The last transformer block duplicates the embedding dimension such that both the means μ and the variances σ^2 of a Gaussian are modeled. Note that all transformer blocks are permutation-equivariant but our multi-head cross-attention block ($\mathrm{MCAB}_{\mathbf{S}}$) is permutation-invariant. As a result, the proposed parameterization NN_{enc} results in permutation-invariant variational posteriors.

Decoder (Conditional Likelihood) The decoder network parameterizes the conditional likelihood function $p(\mathbf{x}_\mathcal{I}|\mathbf{Z}, \mathcal{I})$ for given latents $\mathbf{Z}$ and indices $\mathcal{I}$. The conditional likelihood could be a Gaussian if $\mathbf{x}$'s are continuous, or Poisson or Negative Binomial for counts. To fulfill the requirement on modeling exchangeable distributions, we need to ensure the conditional likelihood is exchangeable. In other words, for a given permutation π, the following holds true: $p_\theta(\mathbf{x}_\mathcal{I}|\mathbf{Z}, \mathcal{I}) = p_\theta(\mathbf{x}_{\pi(\mathcal{I})}|\mathbf{Z}, \pi(\mathcal{I}))$. First, we assume that for given $\mathbf{Z}$, the conditional likelihood is fully factorized: $p_\theta(\mathbf{x}_\mathcal{I}|\mathbf{Z}, \mathcal{I}) = \prod_{i \in I} p_\theta(\mathbf{x}_i|\mathbf{Z})$. Next, we make the parameterization of $p_\theta(\mathbf{x}_\mathcal{I}|\mathbf{Z}, \mathcal{I})$ permutation equivariant, because, otherwise, transforming $\mathbf{Z}$ would result in incorrect parameters for each component $p_\theta(\mathbf{x}_i|\mathbf{Z})$. Keeping in mind that $\mathbf{Z}$ is permutation-invariant to permutations of $\mathbf{x}_\mathcal{I}$, we propose the following decoder network:

$$\mathrm{NN}_{dec}(\mathbf{Z}, \mathcal{I}) = (\mathrm{MCAB}_{\mathbf{E}_\mathcal{I}} \circ \mathrm{T}_L \circ \ldots \circ \mathrm{T}_1)(\mathbf{Z}, \mathcal{I}). \tag{10}$$

and then parameterize $p_\theta(\mathbf{x}_\mathcal{I}|\mathbf{Z}, \mathcal{I}) = p_\theta(\mathbf{x}_\mathcal{I}|\mathrm{NN}_{dec}(\mathbf{Z}, \mathcal{I}))$, where $\mathrm{NN}_{dec}(\mathbf{Z}, \mathcal{I})$ regresses the parameters of the count likelihoods chosen per dimension.

In our decoder network, we use $\mathrm{MCAB}_{\mathbf{E}_\mathcal{I}}$ as our final block that outputs the parameters of the conditional likelihood. To make sure the model is permutation-equivariant, we define pseudoinputs in the multi-head cross-attention block selecting embedding vectors specified by $\mathcal{I}$, $\mathbf{S} = \mathbf{E}_\mathcal{I}$, where . This way, we ensure permutation-equivariance since permuting indices is equivalent to permuting embedding vectors, $\mathbf{E}_{\pi(\mathcal{I})} = \mathbf{E}_\mathcal{I}$, see Property 4 in Appendix. Eventually, we obtain a family of exchangeable conditional likelihood functions.

[3] We use the following notation for function composition: $(f \circ g)(x) \stackrel{df}{=} f(g(x))$, $(f \cdot g)(x) \stackrel{df}{=} f(x)g(x)$, $(f \oplus g)(x) \stackrel{df}{=} f(x) + g(x)$, and $(f \boxplus g)(x) \stackrel{df}{=} \mathrm{concatenate}(f(x), g(x))$.

Prior (Marginal over Latents) In this paper, we advocate to use a Latent Diffusion Model (LDM) [25], namely, for a pre-trained VAE, we fit a diffusion-based model in the latent space to replace a *simpler* prior like $\mathcal{N}(\mathbf{Z}|\mathbf{0}, \mathbf{I})$. Using LDMs not only results in a better match with the aggregated posterior [27, 28], but allows the application of controlled sampling using techniques such as classifier-free guidance [9]. In particular, we focus on linear interpolants and the flow matching (FM) loss [13, 29], and the following version of the classifier-free guidance for FM:

$$\tilde{v}_{t,\epsilon}(\mathbf{Z}, y) = v_{t,\epsilon}(\mathbf{Z}; \text{Null}) + \omega \left[v_{t,\epsilon}(\mathbf{Z}; y) - v_{t,\epsilon}(\mathbf{Z}; \text{Null}) \right], \tag{11}$$

where $v_{t,\epsilon}(\mathbf{Z}; \cdot)$ is a parameterized vector field, and ω is the guidance strenght for attributes $\mathbf{y} \in \{0, 1\}^J$, where any combination of attributes is possible (we refer to it as *joint conditioning*); the Null attribute corresponds to no conditioning. In CFGen [21], a different classifier-free guidance was used, namely, $\tilde{v}_{t,\epsilon}(\mathbf{Z}, y) = v_{t,\epsilon}(\mathbf{Z}; \text{Null}) + \sum_{j=1}^{J} \omega_j \left[v_{t,\epsilon}(\mathbf{Z}; y_j) - v_{t,\epsilon}(\mathbf{Z}; \text{Null}) \right]$, that assumes *additive conditioning* s.t. $\sum_j y_j = 1$.

We parameterize the vector field (score) model using Diffusion Transformer (DiT) blocks [23]. The network is a composition of DiT and perfectly fits our modeling scenario since latents $\mathbf{Z}$ are tokens.

2.2.3 Training

We train our model (VAE + LDM) using the two-stage approach: (1) A VAE is trained to learn a permutation-invariant latent space by reconstructing subsets of variables; and (2) An LDM is trained to generate new samples from this latent space which can be controlled by classifier-free guidance [9] with multiple conditions [21].

Stage 1: VAE We train our VAE with a standard Gaussian prior by optimizing the ELBO in (2). However, to encourage better reconstruction capabilities, we introduce β-weighting of the KL-term like in [8]. Moreover, since we want our model to *generalize* across various subsets of variables, we modify our training. We assume that for all datapoints, we have access to all variables; hence, for each element in a minibatch, we sample indices $\mathcal{I}$ using a uniform distribution, and then pass $\mathbf{x}_\mathcal{I}$ to our VAE. In this way, we enforce our model to implicitly determine *relevant* information in the input object and encode them in $\mathbf{Z}$. Since $\mathbf{Z}$ is invariant to permutations of variables, we expect to obtain robust data representations.

Stage 2: LDM In the second stage, we freeze the VAE and replace the standard Gaussian prior with a score-based (diffusion) model parameterized by a DiT network trained with linear interpolants and the flow matching loss. Additionally, to encourage controlled sampling, for each element of a mini-batch, we sample from the Bernoulli distribution with probability ρ to determine whether conditioning is used or not. This is a typical approach to teaching the model to use classifier-free guidance.

2.2.4 (Un)conditional Generation

In our model, sampling $\mathbf{x}$'s determined by the indices $\mathcal{I}$ is defined by the following generative process: (i) $\mathbf{Z} \sim p(\mathbf{Z})$, (ii) $\mathbf{x}_\mathcal{I} \sim p_\theta(\mathbf{x}_\mathcal{I}|\mathbf{Z}, \mathcal{I})$. However, we can also sample *conditionally* by applying the classifier-free guided sampling technique, following the vector field defined in (11).

3 Experiments

3.1 (Un)conditional Cell Generation on a Single Attribute

Details For the first experiment, we used single-cell RNA-sequencing data from the benchmark datasets used in [21]. Here, we are interested in evaluating the reconstructive and generative capabilities of our scLDM. For generations, we train our scLDM to synthesize gene expression profiles conditioned on a single attribute. At inference time, we query the model with specific labels to generate new synthetic cells that match the desired cellular identity. In the case of unconditional generation, we sample from the vector field without conditioning on the cell type label (i.e., $y = \text{Null}$). We compare our approach to scVI [14], scDiffusion [18], and the current SOTA generative model CFGen [21].

Results and discussion Our proposed scLDM model demonstrates substantial improvements over existing approaches across all evaluated datasets and metrics, see Table 1. scLDM consistently achieves the lowest reconstruction error values, with particularly notable improvements on Tabula Muris (4569.6 vs. 5547.6 for CFGen) and HLCA (4102.1 vs. 5428.7 for CFGen) datasets. The Pearson correlation coefficients show dramatic improvements, with scLDM achieving 0.391 on Tabula Muris compared to 0.221 for scVI and 0.136 for CFGen—nearly doubling the correlation with ground truth. Similarly, MSE is consistently reduced, with scLDM achieving 0.069 on HLCA compared to 0.117 for CFGen and 0.238 for scVI. These results suggest that our fully transformer-based VAE is able to more effectively capture the complex structure of single-cell gene expression data compared to traditional VAE-based methods (scVI, CFGen). The consistent improvements across diverse tissue types (brain, entire organism, and lung) indicate the generalizability of our approach, namely, a parameterization of the VAE using the proposed transformer-based architectures.

Table 1: Model performance comparison on cell reconstruction task.

Dataset	Model	RE ↓	PCC ↑	MSE ↓
Dentate Gyrus	scVI	**5193.2** ± 0.1	0.058 ± 0.000	0.378 ± 0.000
	CFGen	5468.8 ± N/A	0.076 ± N/A	0.253 ± N/A
	scLDM	**5232.9** ± 43.1	**0.103** ± 0.005	**0.249** ± 0.002
Tabula Muris	scVI	5588.2 ± 1.7	0.221 ± 0.000	0.132 ± 0.000
	CFGen	5547.6 ± N/A	0.136 ± N/A	0.127 ± N/A
	scLDM	**4569.6** ± 105.1	**0.391** ± 0.021	**0.092** ± 0.004
HLCA	scVI	5659.2 ± 0.5	0.125 ± 0.000	0.238 ± 0.000
	CFGen	5428.7 ± N/A	0.146 ± N/A	0.117 ± N/A
	scLDM	**4102.1** ± 41.1	**0.421** ± 0.013	**0.069** ± 0.001

Table 2 presents the generation benchmarks, where scLDM demonstrates superior performance across both unconditional and conditional generation sampling. In the unconditional setting, our model achieves the lowest Wasserstein-2 distance across all datasets, with improvements ranging from 14% on Dentate Gyrus to 12% on Tabula Muris. While CFGen shows competitive performance on MMD^2 RBF, our approach matches or outperforms it, achieving identical scores on HLCA and superior results on Tabula Muris. In terms of the Fréchet Distance (FD), scLDM still shows superior performance, with particularly striking improvements on Tabula Muris, where it achieves a nearly three-fold reduction compared to the second-best baseline. For conditional generation, scLDM maintains its performance edge with consistent improvements in W2, MMD^2 RBF, and FD scores across all datasets. We report further generation results on all genes in Appendix. In Figure 2 we report qualitative evaluations of generation results for the three datasets for all three models. Our model shows qualitatively a better coverage of the cell state variation on UMAP coordinates, showcasing how it is able to recapitulate high resolution cell states in highly heterogenous tissues like the human lung. These results demonstrate that our latent diffusion approach not only generates more realistic single-cell expression profiles but also maintains superior performance when conditioning on cell state information, a crucial capability for practical applications in single-cell genomics.

Table 2: Model performance comparison on (un)conditional cell generation benchmarks on highly variable genes.

Setting	Model	W2 ↓	MMD^2 RBF ↓	FD ↓
Dentate Gyrus				
Uncond	scDiffusion	17.443 ± 0.028	0.258 ± 0.002	256.630 ± 0.357
	CFGen	12.617 ± 0.034	**0.022** ± 0.001	**28.105** ± 0.332
	scLDM	**10.817** ± 0.065	**0.023** ± 0.000	28.403 ± 0.099
Cond	scDiffusion	17.321 ± 0.041	0.689 ± 0.000	261.217 ± 1.856
	CFGen	11.608 ± 0.066	**0.075** ± 0.000	41.425 ± 1.612
	scLDM	**10.615** ± 0.028	0.102 ± 0.003	**34.388** ± 1.014
Tabula Muris				
Uncond	scDiffusion	14.143 ± 0.007	0.144 ± 0.001	158.977 ± 1.070
	CFGen	11.658 ± 0.127	0.008 ± 0.000	36.373 ± 1.165
	scLDM	**10.295** ± 0.110	**0.004** ± 0.000	**13.130** ± 0.318
Cond	scDiffusion	14.143 ± 0.007	0.144 ± 0.001	158.977 ± 1.070
	CFGen	8.921 ± 0.034	0.026 ± 0.000	21.517 ± 0.596
	scLDM	**7.717** ± 0.030	**0.016** ± 0.000	**11.008** ± 0.716
HLCA				
Uncond	scDiffusion	15.886 ± 0.038	0.163 ± 0.001	210.853 ± 1.165
	CFGen	12.433 ± 0.045	**0.007** ± 0.000	24.639 ± 0.738
	scLDM	**10.419** ± 0.079	**0.007** ± 0.000	**18.024** ± 0.372
Cond	scDiffusion	15.886 ± 0.038	0.163 ± 0.001	210.853 ± 1.165
	CFGen	9.757 ± 0.078	0.090 ± 0.006	33.900 ± 5.116
	scLDM	**8.445** ± 0.045	**0.074** ± 0.002	**20.974** ± 1.504

3.2 Conditional Generation on Multiple Attributes

Details For the second experiment, we used perturbational single-cell RNA-sequencing data from human peripheral blood mononuclear cells (PBMCs) generated by Parse Biosciences [1]. We selected the donor with the largest number of profiled cells, comprising 1,267,690 single cells subjected to one of 90 cytokine perturbations or a control condition, spanning 18 annotated cell types. To reduce dimensionality and focus on the most informative features, we restricted the analysis to the top 2,000 highly variable genes (HVGs), as previously identified for this dataset [2].

In this experiment, we train our model to generate gene expression when conditioned on cell type and cytokine perturbation. During training, the model is exposed to cells from diverse cell types under multiple perturbations, allowing it to capture the joint structure across these axes of variation. At inference time, we can query the model with specific combinations of cell type and perturbation to

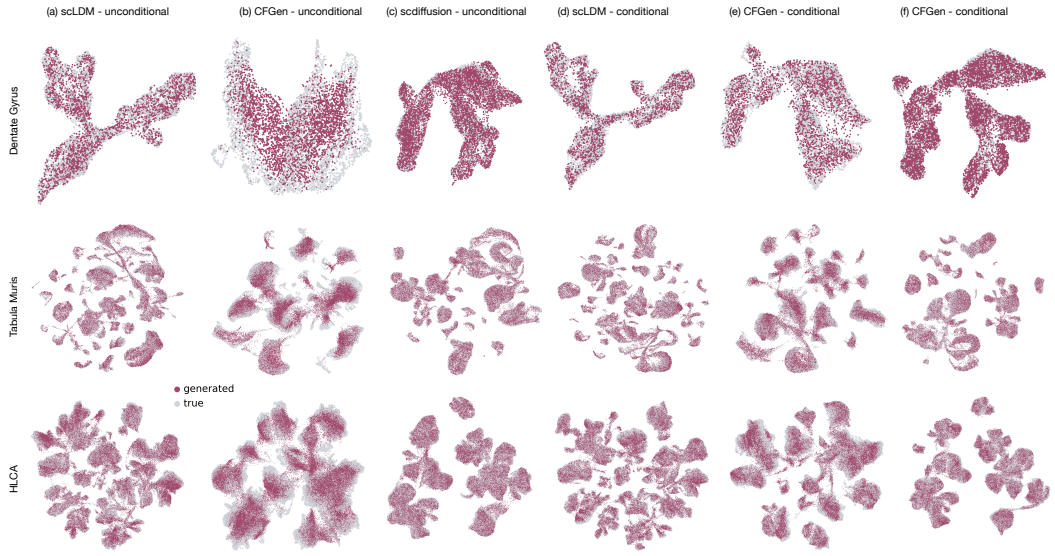

Figure 2: Conditional and unconditional generation across methods. (**a**) Generated vs. true cells for our methods in unconditional settings. (**b**) Generated vs. true cells for CFGen in unconditional settings. (**c**) Generated vs. true cells for our methods in conditional settings. (**d**) Generated vs. true cells for CFGen in conditional settings. **Our method reports qualitatively better results in both conditional and unconditional generation settings.**

generate new samples. We compare our model against established generative baselines: CPA [16] and scVI [14].

Results and Discussion The results presented in Table 3 demonstrate that our proposed approach significantly outperforms the baselines in the Parse 1M dataset (cytokine perturbation). Our model scLDM is substantially better across all metrics, improving up to $\sim$90% for MMD^2 RBF and FD for the Parse 1M dataset. This showcases how our model is superior in covering the full cellular variation in perturbation response in unseen combinations of cell contexts and perturbations.

Table 3: Model performance comparison on conditional cell generation on Parse1M.

Dataset	Model	W2 ↓	MMD^2 RBF ↓	FD ↓
Parse 1M	scVI	35.508 ± 0.014	1.372 ± 0.002	1233.109 ± 2.762
	CPA	13.534 ± 0.010	1.117 ± 0.020	181.324 ± 0.302
	scLDM	$\mathbf{12.455} \pm 0.001$	$\mathbf{0.027} \pm 0.000$	$\mathbf{18.145} \pm 0.068$

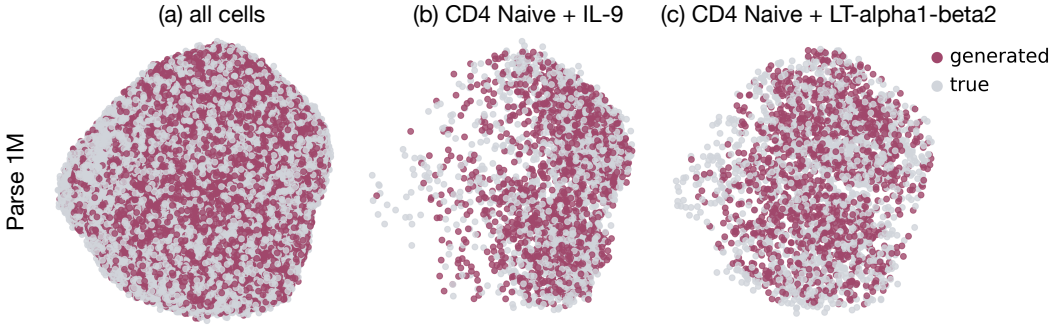

Figure 3: Conditional generation across multiple attributes: cell type and perturbation. (a) Generated vs. true cells across all cell types in the Parse 1M dataset show close alignment. (b–c) For CD4 Naive cells, conditioning on cytokine perturbations (IL-9, LT-alpha1-beta2) produces perturbation-specific shifts consistent with the true test distributions.

In Figure 3, we report a qualitative evaluation of our model generative performances for the Parse 1M dataset for unseen combinations of CD4-Naive cells with various cytokine perturbations such as IL-9

and LT-alpha1-beta2. These results demonstrate our model's ability to be conditioned on multiple attributes, such as cell type and cytokine, simultaneously.

4 Conclusion

In this paper, we demonstrated that enforcing the inductive bias of exchangeability is critical for generative modeling of single-cell data. We introduced a scalable architecture combining a permutation-invariant encoder and a permutation-equivariant decoder within a fully transformer-based VAE with a latent diffusion model parameterized with DiTs, achieving state-of-the-art performance on cell generation benchmarks, on both observational and perturbational data.

Our work moves beyond imposing artificial structure on gene expression data and instead provides a principled framework for learning from unordered sets. This approach is not limited to transcriptomics and lays the groundwork for developing foundational models for other exchangeable biological data, such as proteomics and epigenomics, as well as multi-omics and multi-modal data, thereby enabling more faithful and powerful virtual models of cellular biology.

Author Contributions

GP: Conteptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – review & editing, SB: Investigation, Software, Validation, Visualization, Writing – review & editing, PD: Investigation, Software, Validation, Visualization, Writing – review & editing, DL: Project administration, AK: Supervision, Writing – review & editing, TK: Conceptualization, Project administration, Supervision, Writing – review & editing, JT: Conceptualization, Investigation, Methodology, Software, Supervision, Validation, Writing – original draft, Writing – review & editing

Acknowledgments

The authors would like to kindly thank Isaac Virshup, and Lakshmi Krishnan for insightful discussions.

References

[1] 10 Million Human PBMCs in a Single Experiment. https://www.parsebiosciences.com/datasets/10-million-human-pbmcs-in-a-single-experiment/. [Online; accessed on September 5, 2025].

[2] Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, et al. Predicting cellular responses to perturbation across diverse contexts with STATE. *bioRxiv*, pages 2025–06, 2025.

[3] Michael Bereket and Theofanis Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. In *Neural Information Processing Systems*, 2023.

[4] Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mariano Gabitto, Mohammad Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Vitalii Kleshchevnikov, Carlos Talavera-López, Lior Pachter, Fabian J. Theis, Aaron Streets, Michael I. Jordan, Jeffrey Regier, and Nir Yosef. A python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, Feb 2022.

[5] Adam Gayoso, Philipp Weiler, Mohammad Lotfollahi, Dominik Klein, Justin Hong, Aaron Streets, Fabian J Theis, and Nir Yosef. Deep generative modeling of transcriptional dynamics for rna velocity analysis in single cells. *Nature Methods*, 21(1):50–59, 2024.

[6] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

[7] Gunsagar S Gulati, Shaheen S Sikandar, Daniel J Wesche, Anoop Manjunath, Anjan Bharadwaj, Mark J Berger, Francisco Ilagan, Angera H Kuo, Robert W Hsieh, Shang Cai, et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*, 367(6476):405, 2020.

[8] 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.

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

[10] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver IO: A General Architecture for Structured Inputs & Outputs. In *International Conference on Learning Representations*, 2022.

[11] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[12] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21:1–35, 2020.

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

[14] 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, 2018.

[15] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):e11517, 2023.

[16] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, Jose L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023.

[17] Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

[18] Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. scdiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40(9), 2024.

[19] Mohamed Marouf, Pierre Machart, Vikas Bansal, Christoph Kilian, Daniel S Magruder, Christian F Krebs, and Stefan Bonn. Realistic in silico generation and augmentation of single-cell rna-seq data using generative adversarial networks. *Nature communications*, 11(1):166, 2020.

[20] Karlynn E Neu, Qingming Tang, Patrick C Wilson, and Aly A Khan. Single-cell genomics: approaches and utility in immunology. *Trends in immunology*, 38(2):140–149, 2017.

[21] Alessandro Palma, Till Richter, Hanyi Zhang, Manuel Lubetzki, Alexander Tong, Andrea Dittadi, and Fabian J Theis. Multi-modal and multi-attribute generation of single cells with cfgen. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

[23] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[24] Derek Reiman, Godhev Kumar Manakkat Vijay, Heping Xu, Andrew Sonin, Dianyu Chen, Nathan Salomonis, Harinder Singh, and Aly A Khan. Pseudocell tracer—a method for inferring dynamic trajectories using scrnaseq and its application to b cells undergoing immunoglobulin class switch recombination. *PLoS computational biology*, 17(5):e1008094, 2021.

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[26] Orit Rozenblatt-Rosen, Michael JT Stubbington, Aviv Regev, and Sarah A Teichmann. The Human Cell Atlas: from vision to reality. *Nature*, 550(7677):451–453, 2017.

[27] Jakub Tomczak and Max Welling. Vae with a vampprior. In *International conference on artificial intelligence and statistics*, pages 1214–1223. PMLR, 2018.

[28] Jakub M. Tomczak. *Deep Generative Modeling*. Springer International Publishing, 2024.

[29] Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional flow matching: Simulation-free dynamic optimal transport. *arXiv preprint arXiv:2302.00482*, 2(3), 2023.

[30] Isaac Virshup, Danila Bredikhin, Lukas Heumos, Giovanni Palla, Gregor Sturm, Adam Gayoso, Ilia Kats, Mikaela Koutrouli, Bonnie Berger, et al. The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature biotechnology*, 41(5):604–606, 2023.

[31] Tao Zeng and Hao Dai. Single-cell rna sequencing-based computational analysis to describe disease heterogeneity. *Frontiers in Genetics*, 10:629, 2019.

[32] Jesse Zhang, Airol A Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G Jones, et al. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *BioRxiv*, pages 2025–02, 2025.