SINGLE-CELL GENE EXPRESSION GENERATION WITH HYBRID DISCRETE-CONTINUOUS DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-cell gene expression data is an important pathway to understanding the fundamentals of various cellular applications, like disease tracking and drug targeting. Obtaining high-quality single-cell data is challenging due to the high cost compared to bulk sequencing. This has led to the use of generative models to create synthetic data. The goal is to efficiently generate realistic cells, as that is vital for stronger downstream analysis for applications like cell trajectories and differentiations. Current directions either use purely continuous or purely discrete diffusion to tackle this challenge. Using a purely continuous method allows the model to learn the joint expression levels through a continuous score function. This lets the model learn any potential cellular connections between each gene expression. Using a discrete method respects the sparse nature of the data, removing any representational mismatch that occurs from working in a continuous to discrete space. We propose to leverage the best of both worlds through **scCANDI**: single-cell continuous and discrete diffusion. We highlight the differences between continuous and discrete diffusion through a case study, pinpointing the exact differences in the generated cell profiles under discrete, continuous, and hybrid settings. We demonstrate that scCANDI achieves strong performance against current state-of-the-art masked discrete diffusion methods while requiring **5-10**$\times$ fewer function evaluations.

1 INTRODUCTION

Single-cell RNA sequencing provides high-resolution measurements of gene expression within individual cells, enabling a detailed study of cell differentiation, disease progression, and other biological processes that are missed when cells are measured in bulk (Gulati et al., 2019; Zeng & Dai, 2019; Heumos et al., 2023). Despite the benefits of measuring single-cell data, the real-world experimental cost of collecting single-cell profiles is extremely high. Generative modeling offers a practical alternative. By learning the distribution of real cell states, we can synthesize realistic single-cell profiles to augment downstream tasks such as perturbation prediction, trajectory inference, and cell type annotation (Lotfollahi et al., 2019; Yeo et al., 2021).

In single-cell gene expression, each cell is discretely represented as the gene expression level, or a measurement of active genes within individual cells. The majority of genes end up with low gene expressions, making vector representations sparse. While there are various generative modeling techniques applied to single cell data like using variational autoencoders and generative adversarial networks (Lopez et al., 2018; Marouf et al., 2020). We focus on generative diffusion models, which have broadly followed two paradigms: continuous and discrete diffusion processes (Luo et al., 2024; Palla et al., 2025; Bhattacharya et al., 2026).

Continuous diffusion approaches first learn an encoder-decoder setup to map the discrete integer counts into a learned continuous space, where standard continuous diffusion techniques can be applied (Luo et al., 2024). By training a continuous diffusion model, it is possible to model how positions should jointly evolve through a learned continuous geometry. While continuous approaches benefit from the extensive literature on continuous diffusion, these approaches are at odds with the inherently discrete and sparse nature of single-cell data. The continuous relaxations introduce a representation gap. The continuous model will have to learn the discrete boundaries rather than solely optimize for the real distribution (Maddison et al., 2017).

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.

This leads us to discrete diffusion approaches for single-cell data (Wang et al., 2026; Bhattacharya et al., 2026). We avoid the representation mismatch between discrete data and continuous diffusion. However, pure discrete approaches also have their limitations: these methods suffer in the low-NFE (number of function evaluations) regime, where factorization error of independent predictions accumulate over the trajectory. Given that the motivation behind modeling single-cell gene expression is as a cheap and reliable form of generating realistic synthetic data, this limitation is undesirable. The ideal generative models would be able to generate diverse, realistic samples at low NFE. This would allow for an efficient synthetic data generator while providing realistic synthetic data for downstream tasks. Thus we need a generative modeling framework that respects the discrete nature of the data, but also allows for modeling how positions should evolve simultaneously for efficient generation.

This motivates our proposed method, *scCANDI*, single-cell continuous and discrete diffusion. Instead of choosing between discrete or continuous diffusion, hybrid diffusion allows us to leverage the best of both worlds. Building on the single-cell pipeline (Bhattacharya et al., 2026), scCANDI replaces the continuous and discrete diffusion processes with a hybrid diffusion process from Pynadath et al. (2026). The discrete component of the hybrid trajectory respects the sparse nature of single-cell data, while the continuous component jointly evolves multiple positions in parallel for efficient generation.

We summarize our contributions as follows:

- We introduce scCANDI, a hybrid discrete-continuous diffusion framework, leveraging properties of discrete and continuous diffusion for single-cell gene expression. We utilize the efficiency of the joint continuous updates and the representational match in the discrete process to create a framework that can efficiently generate realistic samples.
- We analyze the differences between discrete, continuous, and hybrid generative processes by isolating each process in scCANDI. We discover a unique trade-off for single-cell data regarding standard metrics in maximum mean distance (MMD) and waserstein distance (WD). We note the discrete diffusion generally leads to a better MMD score, while continuous diffusion leads to a better WD score. scCANDI offers a unique trade-off balancing between the MMD and WD based on the strength of the continuous updates.
- We show that scCANDI matches or outperforms state-of-the-art methods across multiple datasets in terms of distribution fidelity while requiring 5–10$\times$ fewer function evaluations.

2 Related Works

2.1 Diffusion for Single-Cell Gene Expression

There has been various approaches to apply generative models for single-cell gene expression. Luo et al. (2024) introduces an encoder-decoder to feed latent embeddings directly into a diffusion model; Palma et al. (2025) introduces a flow-based conditional generative model to preserve the discreteness of single-cell data; Palla et al. (2025) utilizes a special latent representation to capture the discreteness of single-cell data to properly apply the forward and reverse process in a continuous representation. More recently, Bhattacharya et al. (2026) introduces Discrete Cell Models (DCM), a masked discrete diffusion method utilizing a discrete representation in the forward and reverse process to match the discrete nature of the data. We introduce a new hybrid diffusion framework to leverage the benefits from discrete and continuous diffusion for single-cell gene expression.

2.2 Hybrid Discrete-Continuous Diffusion

Hybrid Diffusion refers to some combination of utilizing continuous and discrete information for the diffusion process. Many works have presented various approaches, such as working in the one-hot space, embedding space, or even latent space (Pynadath et al., 2026; Zheng et al., 2025; Zhou et al., 2025; Shariatian et al., 2025). More recently, Jutras-Dubé et al. (2026) demonstrates that hybrid, discrete, and uniform methods can be unified under sticky stochastic diffusion processes. Other works have demonstrated that hybrid methods can take advantage of the adaptive ordering techniques commonly used in masked diffusion Agarwal et al. (2026); Azangulov et al. (2026). We use the framework introduced in Pynadath et al. (2026), the framework utilizes a one-hot representation where single-cell data can be easily adapted.

3 PRELIMINARIES

3.1 PROBLEM FORMULATION

We represent a cell's gene expression profile as a discrete sequence. Let $\mathbf{x} \in \mathcal{X}^M$ denote the expression profile across M genes, where $\mathcal{X} = \{0, 1, \ldots, K\}$ is the discrete vocabulary of raw integer counts and K is the maximum expression count in the dataset. Each element $x_i \in \mathcal{X}$ represents the expression level of gene i. This formulation treats single-cell gene expression identically to a text sequence, where each gene plays the role of a token position. For conditional generation, we aim to learn $p_\theta(\mathbf{x} \mid \mathbf{c})$ where $\mathbf{c} = (\text{perturbation}, \text{cell type})$, maximizing the conditional log-likelihood:

$$\ell(\theta; \mathcal{D}) = \sum_{n=1}^{N} \log p_\theta(\mathbf{x}_n \mid \mathbf{c}_n) \tag{1}$$

over observed data $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{c}_n)\}_{n=1}^N$.

3.2 MASKED DISCRETE DIFFUSION

Masked discrete diffusion defines a forward process that gradually interpolates between the data distribution $P(X_0)$ and a degenerate distribution in which all positions are replaced by a special mask token M (Lou et al., 2024; Sahoo et al., 2024; Shi et al., 2025):

$$q(X_t \mid X_0) = \mathrm{Cat}(X_t;\, (1 - \alpha(t))\, X_0 + \alpha(t)\, M) \tag{2}$$

where $\alpha(t)$ is a monotonically decreasing function controlling the masking rate. The reverse transition factorizes cleanly:

$$q(X_s \mid X_t, X_0) = \begin{cases} \mathrm{Cat}(X_s;\, X_t) & X_t \neq M \\ \mathrm{Cat}\left(X_s;\, \dfrac{(1-\alpha_s)\, M + (\alpha_s - \alpha_t)\, X_0}{1 - \alpha_t}\right) & X_t = M \end{cases} \tag{3}$$

When a position is masked, the reverse distribution depends on X_0 only through the expectation $\mathbb{E}[X_0 \mid X_t]$, allowing us to use mean parameterization. The neural network learns to predict $\mathbb{E}[X_0 \mid X_t]$ directly. The training objective reduces to a weighted cross-entropy loss over masked positions.

3.3 CONTINUOUS DIFFUSION

We define a continuous diffusion with a forward SDE from the data distribution $p(\tilde{X}_0)$ toward a Gaussian prior. This is a Variance-Exploding (VE) SDE (Song et al., 2021), hence we have:

$$d\tilde{X}_t = g(t)\, dW_t \tag{4}$$

with reverse SDE:

$$d\tilde{X}_t = -g(t)^2\, \nabla_{\tilde{X}_t} \log p_t(\tilde{X}_t)\, dt + g(t)\, d\tilde{W}_t. \tag{5}$$

The score function can be expressed in terms of the conditional expectation and cumulative noise $\sigma(t)^2 = \int_0^t g(s)^2\, ds$:

$$\nabla_{\tilde{X}_t} \log p_t(\tilde{X}_t) = -\frac{\tilde{X}_t - \mathbb{E}[\tilde{X}_0 \mid \tilde{X}_t]}{\sigma(t)^2}. \tag{6}$$

Thus, learning $\mathbb{E}[\tilde{X}_0 \mid \tilde{X}_t]$ is sufficient to simulate the reverse process. A key advantage of continuous diffusion is that the score function enables *joint* updates across all positions simultaneously, in contrast to the independent per-position sampling of masked discrete diffusion.

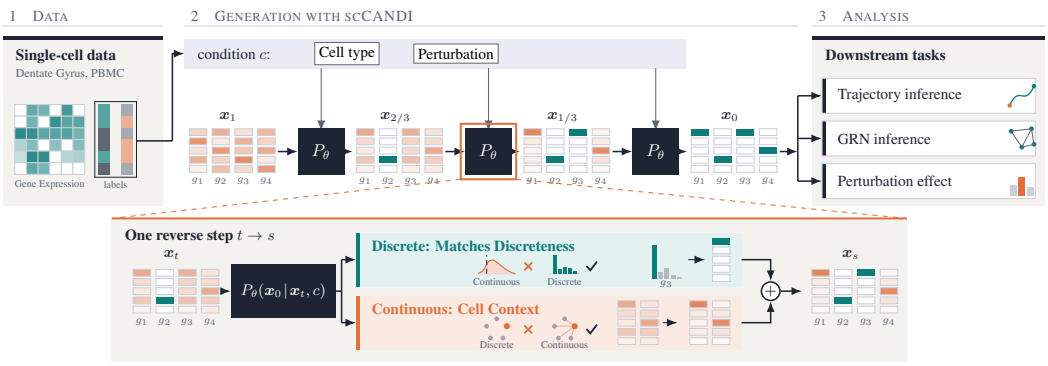


Figure 1: **scCANDI overview**: After training the model, we take the single-cell data and apply the desired conditioning c. We show the reverse step utilizing the discrete anchors with the continuous update information. Repeating that process, we are able to generate realistic cell samples for downstream tasks. This allows us to represent the generated cell in the discrete space while applying the joint information from the continuous process.

4 Methodology

4.1 Hybrid Framework

We introduce the main adaptations scCANDI makes from Discrete Cell Models (DCM) (Bhattacharya et al., 2026) and Continuous and Discrete Diffusion (CANDI) (Pynadath et al., 2026). We adapt a simple yet effective tokenization framework from DCM to represent gene expression profiles. Each gene count $x_i \in \mathcal{X} = \{0, \dots, K\}$ is represented as a binned vector, instead of using the raw counts. We further adapt the representation for CANDI which utilizes a one-hot vector $\tilde{x}_i \in \{0,1\}^{|\mathcal{X}|}$, giving a full expression profile $\tilde{X}_0 \in \mathbb{R}^{M \times |\mathcal{X}|}$. We distinguish between the discrete token $x_i \in \mathcal{X}$ and its one-hot representation $\tilde{x}_i$.

We utilize this one-hot representation to effectively run the hybrid noising kernel from CANDI. A position-wise mask $M_t \in \{0,1\}^M$ determines which gene positions remain as clean discrete anchors and which receive Gaussian noise:

$$X_t = \tilde{X}_0 \odot M_t + \tilde{X}_t \odot (1 - M_t) \tag{7}$$

where $M_t[i] \sim \text{Bernoulli}(\alpha(t))$, $\alpha(t) = 1 - t$, and $\tilde{X}_t \sim \mathcal{N}(\tilde{X}_0, \sigma(t)^2 I)$. The noise level $\sigma(t)$ is chosen so that continuous rank degradation, increases linearly with time.

$$r(t) = \Phi\left(-\frac{1}{\sigma(t)\sqrt{2}}\right) \tag{8}$$

This maintains alignment between discrete and continuous corruption throughout the forward process. Clean positions provide discrete conditioning signals and corrupted positions carry continuous gradient information enabling joint refinement during inference.

4.2 Model Parameterization

DCM utilizes Score Entropy Discrete Diffusion (SEDD) which avoids direct parameterization to get gene expression samples, and uses an estimate instead (Lou et al., 2024). CANDI allows for the model to learn the conditional distribution $P_\theta(X_0 \mid X_t, \mathbf{c})$ over discrete gene counts. We follow the CANDI formulation from which both the discrete reverse transitions and the continuous score function are derived:

$$\nabla_{\tilde{X}_t} \log p_t(\tilde{X}_t) = -\frac{\tilde{X}_t - \mathbb{E}[\tilde{X}_0 \mid \tilde{X}_t]}{\sigma(t)^2} \tag{9}$$

We follow CANDI and add a learnable bias vector $\mathbf{b} \in \mathbb{R}^{|X|}$ to the input representation of corrupted positions, to help the model distinguish clean positions from corrupted ones:

$$e_i = \begin{cases} \tilde{X}_t[i] & M_t[i] = 1 \quad \text{(clean)} \\ (1-\lambda) \cdot \tilde{X}_t[i] + \lambda \cdot \mathbf{b} & M_t[i] = 0 \quad \text{(corrupted)} \end{cases} \tag{10}$$

where $\lambda \in [0, 1]$ is a learnable interpolation weight. This bias provides a learnable signal that allows the model to identify which positions require denoising. This allows us to not rely on a gene-specific embedding table which is used in DCM.

4.3 Training Objective

DCM (Bhattacharya et al., 2026) utilizes a simplified weighted cross-entropy in denoising cross entropy (Ou et al., 2025). We utilize a weighted cross-entropy loss in CANDI (Pynadath et al., 2026), which weights the loss according to the number of masked positions to balance the loss across time steps. This also corresponds to the standard discrete diffusion ELBO (Sahoo et al., 2024). We use:

$$\mathcal{L}(\theta, \mathcal{D}) = \mathbb{E}_{\substack{X_0 \sim \mathcal{D} \\ t \sim [0,1] \\ X_t \sim q(\cdot|X_0)}} \left[-\frac{1}{1-\alpha(t)} \log P_\theta(X_0 \mid X_t, \mathbf{c}) \right] \tag{11}$$

4.4 Hybrid Inference

scCANDI performs hybrid inference leveraging both discrete conditional structure and the continuous score function. Given the current state (X_t, M_t) and the next time step $s < t$. Sample which corrupted positions transition to clean:

$$M'_s[i] \sim \text{Bernoulli}\left(\frac{\alpha(s) - \alpha(t)}{1 - \alpha(t)} \right) \tag{12}$$

For these positions, sample discrete gene counts X'_s from $P_\theta(\cdot \mid \tilde{X}_t, \mathbf{c})$ and convert it to one-hots. For positions remaining corrupted ($M_t[i] = M'_s[i] = 0$), apply the probability flow ODE update:

$$\tilde{X}''_s = \tilde{X}_t - \tfrac{1}{2}\big(\sigma(t)^2 - \sigma(s)^2\big) \cdot \nabla \log p_t(\tilde{X}_t) \tag{13}$$

Clean positions from M_t are unchanged. We can derive the final update rule as follows:

$$\tilde{X}_s = M_t \odot \tilde{X}_t + \neg M_t \odot M'_s \odot \tilde{X}'_s + \neg M_t \odot \neg M'_s \odot \tilde{X}''_s \tag{14}$$

This joint update is the key advantage over discrete diffusion methods, where positions are sampled independently. The continuous score function allows scCANDI to refine many gene positions simultaneously, which is particularly beneficial at low NFE where a large fraction of positions must be resolved in each step.

4.5 Advantages

From these adaptations from DCM (Bhattacharya et al., 2026) and CANDI(Pynadath et al., 2026), scCANDI is able to be applied to the single-cell gene expression. We expect scCANDI to leverage the advantages of purely discrete and purely continuous diffusion. Discrete diffusion has a weakness in the low-NFE space, but can generate high quality synthetic cells. Continuous diffusion has joint updates, but loses generation quality due to the representation gap. scCANDI should be able to generate quality realistic samples while utilizing the continuous joint updates be efficient at low-NFE.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate scCANDI with DCM (Bhattacharya et al., 2026), since DCM has shown strong results compared to other state-of-the-art methods like CFGen and scLDM (Palma et al., 2025; Palla et al., 2025)

Datasets. We compare the performance between models with the PBMC3K (2638 cells, 8573 genes, 8 cell types)(Zheng et al., 2017), Dentate Gyrus (18,213 cells, 17,002 genes, 14 cell types) (Hochgerner et al., 2018), and PBMC COVID (422,220 cells, 2000 genes, 29 cell types) (Arunachalam et al., 2020). All datasets can be found together in Palma et al. (2025)

Metrics. Given a set of ground truth cells $X = \{x_1, \dots x_n\}$ and a set of generated cells $Y = \{y_1, \dots y_n\}$. We utilize RBF-Kernel Maximum Mean Discrepancy (MMD) and 2-Wasserstein Distance (WD). Both are standard evaluation metrics between various single-cell works. MMD is able to capture the overall distribution structure between X and Y, while WD is able to capture any shifts between the distributions.

Experimental Design. For each experiment (both the tables and the figures), we utilize three randomly seeded runs. We generate 100 cells to compare with a subset of the real cells. scCANDI's default hyperparameters can be found in Appendix C

5.2 UNDERSTANDING CONTINUOUS VS DISCRETE DIFFUSION

We analyze the individual contributions of continuous and masked diffusion to generation quality by ablating CANDI (Pynadath et al., 2026) into its generative processes of MDLM (Sahoo et al., 2024) and DiT (Peebles & Xie, 2023). We continue to utilize the same framework that scCANDI is in, and we will denote the MDLM and DiT variations as scMDLM and scDiT.

Results. scCANDI consistently scores as a middle ground between scDiT and scMDLM, see Table 1. scCANDI balances the MMD and WD performance, and we can see that scCANDI produces the closest distribution to the real data at steps 10, 20, and 100, see Figure 2. From Table 1, we see that the WD distance is very poor for scMDLM. We also notice in Figure 2 that the sparsity distribution is shifted to the left of the real distribution at every step. Meanwhile, scCanDI and scDiT both maintain their distributions near the real data and have strong WD scores. On the other hand, we can see that scMDLM is able to appropriately match the shape of the real data which aligns with the lowest MMD score in Table 1. In Figure 2, we notice scDiT has a extremely narrow distribution at higher steps. This aligns with the higher MMD scores from Table 1.

NFEs	MMD	WD
5		
scCANDI	0.547± 0.01	1.512±0.05
scDiT	0.470± 0.02	**1.498±0.04**
scMDLM	**0.220± 0.01**	4.865±0.06
10		
scCANDI	0.397± 0.01	**1.440 ± 0.07**
scDiT	0.500 ± 0.00	1.510 ± 0.04
scMDLM	**0.226 ± 0.02**	11.934 ± 0.02
20		
scCANDI	0.362± 0.01	**1.416± 0.05**
scDiT	0.586 ± 0.00	1.591 ± 0.04
scMDLM	**0.164**± 0.01	27.456 ± 0.03
100		
scCANDI	0.280± 0.01	1.866 ± 0.03
scDiT	0.539 ± 0.00	**1.547 ± 0.03**
scMDLM	**0.169 ±0.01**	32.703 ± 0.10

Table 1: The MMD and WD scores at various NFEs. The best results are bolded for each NFE step. scCANDI provides a strong balance between MMD and WD distance from the individual continuous and discrete methodologies.

Analysis. These results highlight the difference between the continuous and discrete processes on single-cell data. When the discrete process is able to find good discrete anchors, it can be unmatched in the distribution, shown my the MDLM's low MMD score. The continuous process is able to find an appropriate space to converge near the real data, shown by the consistent WD scores of scDiT. We can obtain various benefits from, scCANDI's ability to utilize both

the discrete and continuous information. Our method uses the discrete anchors to create a good representation of the distribution, evidenced by the stronger MMD scores compared to scDiT, and is able to place correct the discrete anchors with the continuous information, shown by the stronger WD scores compared to MDLM, causing the generated distribution to better match the real data, shown by the distributions in Figure 2.

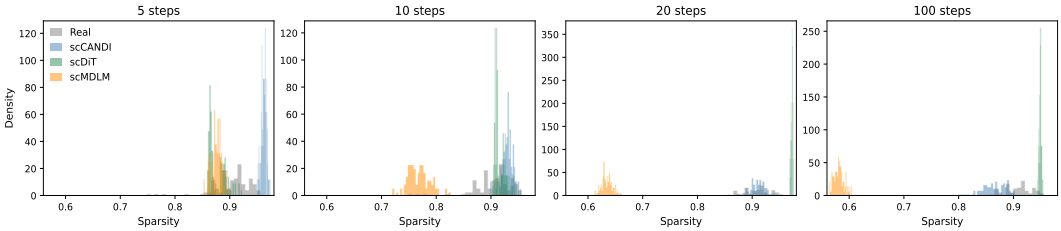


Figure 2: The sparsity, the inverse of mean gene expression level, for each method at various NFE steps. As the steps progress, left to right, we notice that the continuous information can be important to ensure that the discrete anchors better align with the real distribution.

5.3 Unconditional Generation

We evaluate the unconditional generation using the datasets and metrics in the experimental setup. The idea is to test the model's capability in reproducing cells similar to the underlying distribution of the original dataset without any kind of conditioning information like the cell type or perturbations (Bhattacharya et al., 2026).

Results. scCANDI demonstrates strong results over various datasets and metrics, see Table 2. In particular, scCANDI displays a slightly stronger WD compared to DCM over all the datasets. While the MMD has some mixed results depending on the dataset, overall, scCANDI has a competitive results to DCM. These results suggest that scCANDI is able to capture discrete diffusion properties while providing guidance in the single-cell structure with the continuous information.

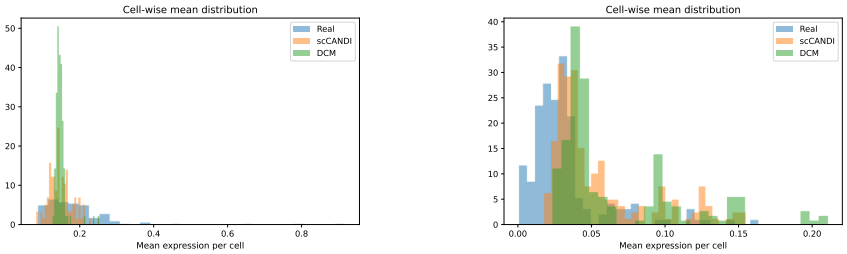
Dataset	Method	WD	MMD
Dentate Gyrus	scCANDI	**1.42±0.05**	0.36±0.01
	DCM	1.48±0.00	**0.022±0.01**
PBMC3K	scCANDI	**0.08±0.01**	**0.02±0.00**
	DCM	0.11±0.01	0.03±0.01
PBMC COVID	scCANDI	**0.07±0.01**	**0.02±0.00**
	DCM	0.13±0.01	0.05±0.01

Table 2: We compare the MMD and WD between the generated cells and the real cells. The best results for each dataset are bolded. scCANDI is comparable to DCM, a strong discrete diffusion method, while utilizing the same MMD for WD trade off.

Distribution Comparison. The results are further highlighted in Figure 3a, which plots the cell-wise mean expression level for generated and real cells on the Dentate Gyrus dataset. We can see that the DCM-generated cells are concentrated at a single sparsity value, while the real distribution is spread across a wider range of mean expression levels. scCANDI's generated cells are able to follow the real distribution more closely.

Analysis. These results suggests that the continuous score function helps prevent the over-concentration that arises can arise from discrete diffusion. We believe that the discrete diffusion is finding multiple discrete anchors that are shifted from the real distribution. In this case, we have overlapping causing the concentrated expression levels. The continuous score function allows for the discrete anchors to be placed in the direction of the real distribution. This causes scCANDI to generate cells that mirror the real distribution more closely. This still comes with a trade-off in MMD and WD, and aligns with our previous observation regarding the discrete diffusion for MMD and continuous to balance the WD.

Efficiency. It is important to note that scCANDI achieves these results at significantly fewer NFEs compared to DCM. DCM's standard implementation utilizes 100 NFE steps for each cell. Our method, scCANDI, only needed 10 steps for both PBMC datasets and 20 steps for Dentate Gyrus to generate each cell. This represents a 5-10× reduction in inference cost while maintaining comparable metric scores and improved generation distribution.



(a) (b)

Figure 3: **(a)** We compare the unconditionally generated cell's mean expression level for DCM and scCANDI on the Dentate Gyrus dataset. **(b)** We compare the conditionally generated cell's mean expression level for DCM and scCANDI on the PBMC COVID dataset. The conditioning was done with 29 cell types. We compare between 200 generated and real cells.

5.4 Conditional Generation

For conditional generation, we condition the model on biological attributes available in each dataset, either cell type or perturbation. During inference, scCANDI is queried with the target condition to generate expression profiles specific to that cell type or perturbation.

Results. The conditional generation setting similar to the unconditional setting, scCANDI shows improvement in the WD overall both datasets while the MMD performance is comparable DCM, see Table 3. This follows the trade-offs that scCANDI makes improving WD at the ccost of MMD. These results suggests that when giving specific cell types or perturbation profiles, scCANDI is able to generate a competitive realistic distribution compared to DCM. This is supported in Figure 3b.

Analysis. The cell-wise mean expression distribution for conditionally generated cells on the PBMC COVID dataset for scCANDI better reflects the real distribution compared to DCM, shown in Figure 3b. Aligned with previous results, this support the idea that the continuous information helps align the discrete anchors.

Dataset	Method	WD	MMD
PBMC3K	scCANDI	**0.34±0.05**	0.07±0.02
	DCM	0.39±0.14	**0.04±0.00**
PBMC COVID	scCANDI	**0.092±0.02**	**0.07±0.03**
	DCM	0.094±0.05	0.09±0.09

Table 3: The generated cells are conditioned on the 8 perturbations from the PBMC3k dataset, and the 29 cell types for PBMC COVID. The best results are bolded. scCANDI maintains comparable results to DCM under a conditional generation setting.

5.5 Importance of Inference Settings

In this section, we discuss the importance of finding the correct inference settings for scCANDI.

Results. From Figure 4, we notice that selecting the correct temperature and step size is very important. The WD and MMD performance can degrade quickly over high step counts and incorrect temperatures. They also provide an inverse trade off in that a higher temperature is generally better for MMD and worse for WD and flipped for a lower temperature. We also note that the MMD performance greatly varies at various temperatures for various NFE steps, while the WD has a very similar results at low temperatures.

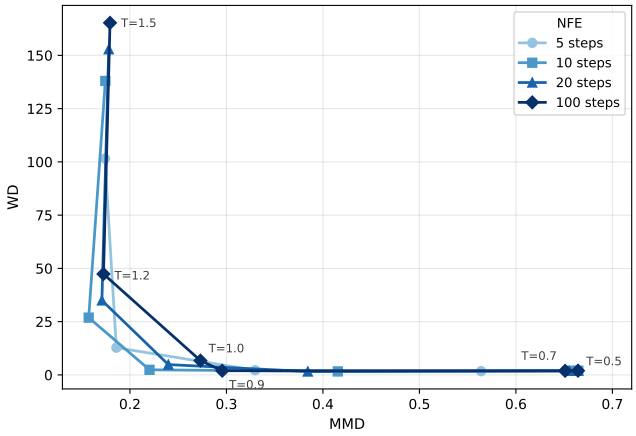


Figure 4: We compare the WD and MMD scores through a frontier plot. We utilize results on the Dentate Gyrus dataset from multiple temperatures and NFE sweeps. MMD is better at higher temperatures while WD is better at lower temperatures.

Analysis. This highlights how appropriate temperature tuning can also lead to compute savings, with slight tradeoffs of WD and MMD performance. We suspect that the tradeoffs between the WD and MMD at higher and lower temperatures is related to the influence of the continuous updates. As noted before, scMDLM has a stronger MMD and a weaker WD compared to scCANDI. The adjustment in the temperature affects the strength of the continuous updates. At higher temperatures, there is less influence thus creating this trade-off. Therefore, it may be worthwhile to do a temperature and NFE sweep specific settings. The sweep enables significant compute savings while maintaining strong results while also determining the how much influence the continuous updates should have.

6 Limitations

There are various limitations for scCANDI. An apparent one shown in the previous section is finding the appropriate parameters in terms of the temperature and the number of denoising steps. While scCANDI can utilize a much lower amount of NFE steps, finding the appropriate temperature to achieve that can be costly with larger datasets. Another limitation, is that the continuous refinement information does not result in stronger performance. As shown in our generation results, a carefully tuned DCM can match the real distribution without any continuous information. However, there is still the extra cost in the NFE as discussed in the results and in line with our expected performance.

7 Conclusion

In this work, we utilize a hybrid discrete-continuous framework that allows for a discrete diffusion framework to better tackle single-cell gene expression by utilizing continuous information. We take a deep dive on what continuous and discrete diffusion is doing individually to single-cell data through the lens of MMD scores, WD scores, and sparsity distribution in Section 5.2.We note that scCANDI is able to utilize the continuous updates to appropriately place the discrete anchors to match the real distribution We show that scCANDI performs on par or better than DCM at 5-10$\times$ lower NFE cost in Sections 5.3-5.4. Beyond fine-tuning and architecture improvements, scCANDI suggests that hybrid diffusion methods may have a great direction in other discrete generative tasks like molecular generation and protein generation.

AI use statement

In this work, we used generative AI tools for help implementing methods. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research

methodology or experiments, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, and interpret results are not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures, brainstorming, and improving readability. We have reviewed all AI-assisted work. LLM generated code was tested for correctness, figures and text were verified. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Manan Agarwal, Sheel Shah, Chanhyuk Lee, Jaehoon Yoo, Jerry Huang, Seunghoon Hong, Aditi Raghunathan, Jinwoo Kim, and Nicholas M. Boffi. Posterior refinement: Fast language generation via any-order flow maps, 2026. URL https://arxiv.org/abs/2606.24773.

Prabhu S Arunachalam, Florian Wimmers, Chris Ka Pun Mok, Ranawaka APM Perera, Madeleine Scott, Thomas Hagan, Natalia Sigal, Yupeng Feng, Laurel Bristow, Owen Tak-Yin Tsang, et al. Systems biological assessment of immunity to mild versus severe covid-19 infection in humans. *Science*, 369(6508):1210–1220, 2020.

Iskander Azangulov, Kianoosh Ashouritaklimi, Leo Zhang, Simon Vary, and Patrick Rebeschini. Masked language flow models. *arXiv preprint arXiv:2606.27617*, 2026.

Sanjukta Bhattacharya, Christian Gensbigler, Shaamil Karim, and Jon Lees. Discrete diffusion for single-cell gene expression modeling. In *ICLR 2026 Workshop on Machine Learning for Genomics Explorations*, 2026. URL https://openreview.net/forum?id=GPR1YXdE4U.

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, Maider Zabala, Ferenc A. Scheeren, Neethan A. Lobo, Dalong Qian, Feiqiao B. Yu, Frederick M. Dirbas, Michael F. Clarke, and Aaron M. Newman. Single-cell transcriptional diversity is a hallmark of developmental potential. *bioRxiv*, 2019. doi: 10.1101/649848. URL https://www.biorxiv.org/content/early/2019/05/30/649848.

Lukas Heumos, Anna C Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D Lücken, Daniel C Strobl, Juan Henao, Fabiola Curion, et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, 2023.

Hannah Hochgerner, Amit Zeisel, Peter Lönnerberg, and Sten Linnarsson. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell rna sequencing. *Nature neuroscience*, 21(2):290–299, 2018.

Pascal Jutras-Dubé, Patrick Pynadath, Jeremy Lu, Yuan Gao, and Ruqi Zhang. Sticky jump diffusions: A unifying view of masked, continuous, and hybrid diffusion. *arXiv preprint arXiv:2607.10951*, 2026.

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.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2024.

Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. scdiffusion: conditional generation of high-quality single-cell data using diffusion model, 2024. URL https://arxiv.org/abs/2401.03968.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables, 2017. URL https://arxiv.org/abs/1611.00712.

Matteo 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.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sMyXP8Tanm.

Giovanni Palla, Sudarshan Babu, Payam Dibaeinia, James D. Pearce, Donghui Li, Aly A. Khan, Theofanis Karaletsos, and Jakub M. Tomczak. Scalable single-cell gene expression generation with latent diffusion models, 2025. URL https://arxiv.org/abs/2511.02986.

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 *ICLR*, 2025. URL https://openreview.net/forum?id=3MnMGLctKb.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. Candi: Hybrid discrete-continuous diffusion models, 2026. URL https://arxiv.org/abs/2510.22510.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models, 2024. URL https://arxiv.org/abs/2406.07524.

Dario Shariatian, Alain Durmus, Umut Simsekli, and Stefano Peluchetti. Latent-augmented discrete diffusion models. *arXiv preprint arXiv:2510.18114*, 2025.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data, 2025. URL https://arxiv.org/abs/2406.04329.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Mingxuan Wang, Cheng Chen, Gaoyang Jiang, Zijia Ren, Chuangxin Zhao, Lu Shi, and Yanbiao Ma. Scdiva: Masked discrete diffusion for joint modeling of single-cell identity and expression. *arXiv preprint arXiv:2602.03477*, 2026.

Grace Hui Ting Yeo, Sachit D Saksena, and David K Gifford. Generative modeling of single-cell time series with prescient enables prediction of cell trajectories with interventions. *Nature communications*, 12(1):3222, 2021.

Tao Zeng and Hao Dai. Single-cell rna sequencing-based computational analysis to describe disease heterogeneity. *Frontiers in Genetics*, Volume 10 - 2019, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00629. URL https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2019.00629.

Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049, 2017.

Huangjie Zheng, Shansan Gong, Ruixiang Zhang, Tianrong Chen, Jiatao Gu, Mingyuan Zhou, Navdeep Jaitly, and Yizhe Zhang. Continuously augmented discrete diffusion model for categorical generative modeling, 2025. URL https://arxiv.org/abs/2510.01329.

Cai Zhou, Chenxiao Yang, Yi Hu, Chenyu Wang, Chubin Zhang, Muhan Zhang, Lester Mackey, Tommi Jaakkola, Stephen Bates, and Dinghuai Zhang. Coevolutionary continuous discrete diffusion: Make your diffusion language model a latent reasoner, 2025. URL https://arxiv.org/abs/2510.03206.

A CODE AND DATASETS

All datasets can be either found under CFGen or DCM's respective papers Bhattacharya et al. (2026); Palma et al. (2025).
The code will be included when ready.
Code implementations for DCM and CANDI are available in their respective papers Bhattacharya et al. (2026); Pynadath et al. (2026).

B ADDITIONAL FIGURES

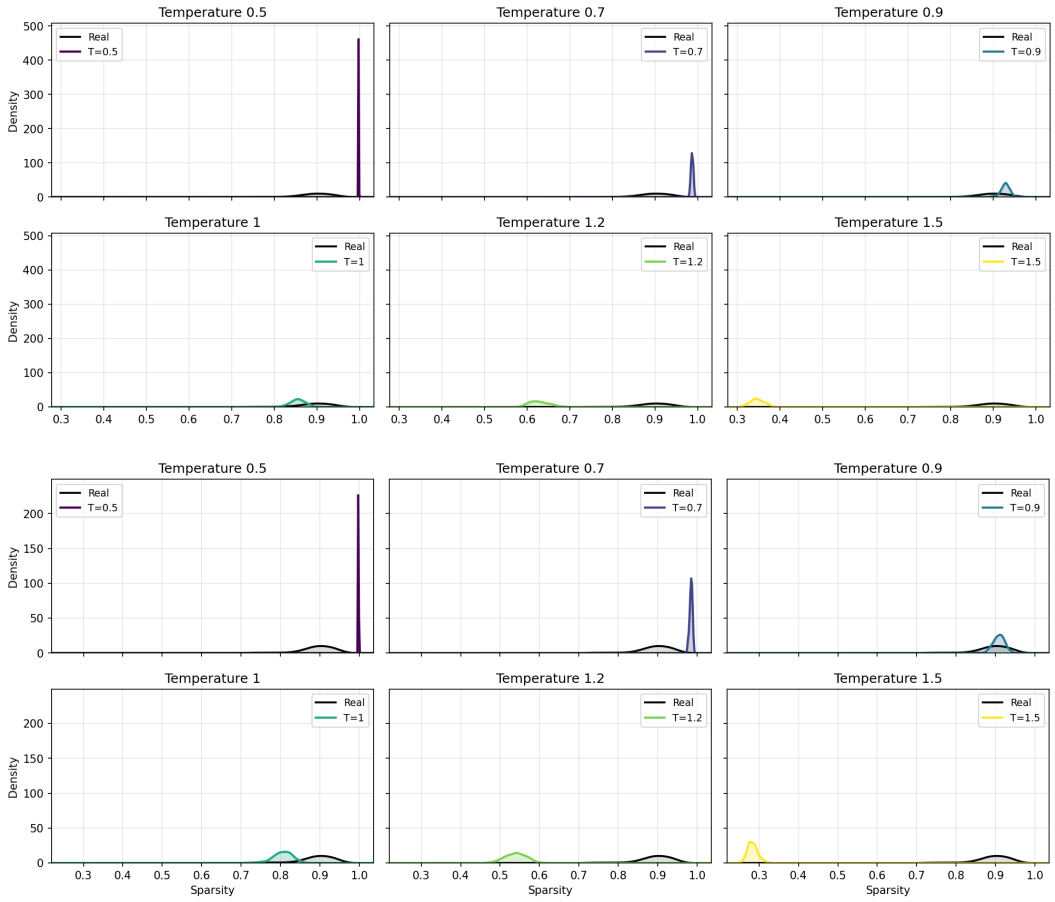


Figure 5: Sparsity plots between scCANDI and the real dataset. The top six graphs are for 10 steps while the bottom six are for 20 steps. These results highlight the importance of doing a temperature sweep as you can reduce the number of NFE steps by half or even more, while maintaining a decent generated distribution to the real.

C EXPERIMENTAL DETAILS

For all DCM experiments we utilized their standard setup in their config files. We only adjusted the datasets needed when running experiments.

The compute resources used was an A6000 48GB. We ran multiple training scripts for each dataset taking around 3-4 days total.

Table 4: CANDI Hyperparmeters

scCANDI Parameters	
Number of Blocks	4
Number of Heads	4
Embedding size	256

Remaining are the default for the tiny CANDI config file.