BALANCED LATENT SPACE OF DIFFUSION MODELS FOR COUNTERFACTUAL GENERATION

Baohua Yan, Qingyuan Liu, Zhaobin Mo, Kangrui Ruan & Xuan Di
Columbia University
500 West 120th Street, New York, NY 10025, USA
{by2348, ql2505, zm2302, kangrui.ruan, sharon.di}@columbia.edu

ABSTRACT

Counterfactual generation has demonstrated impressive performance in tasks such as image editing and synthesis, largely due to the development of diffusion models. However, existing diffusion-based counterfactual generation models suffer from instability due to a lack of understanding of the latent space. These models either retain too much of the original information or make excessive modifications, sacrificing crucial details, leading to inefficiency and inauthenticity. In this paper, we propose a framework that balances the latent space by incorporating signals that facilitate the transition to new counterfactuals while preserving factual information. We first identify the cause of this imbalance as the uncontrolled signal from the counterfactuals. Based on this understanding, we introduce a balancing method within the diffusion process. Our approach is evaluated on the colored MNIST dataset, a modified version of the standard MNIST dataset, with experimental results showing significant improvements over previous latent space methods.

1 INTRODUCTION

Counterfactual data generation has become an essential way to explore alternative modifications by generating causally coherent variations of input data. Previously, Pawlowski et al. (2020) proposed Deep-SCM that leverages Normalizing Flows (Tabak & Turner, 2013) and variational inference (Kingma & Welling, 2022) to infer exogenous noise and perform causal inference under the assumption of no unobserved confounders. Subsequent works further explore the application on Variational Autoencoders (VAEs) (Ribeiro et al., 2023; Kladny et al., 2024), Generative Adversarial Networks (GANs) (Kocaoglu et al., 2017; Dash et al., 2022), and real-world scenario (Wang et al., 2023; Yeganeh et al., 2024).

Recent work on diffusion models (DMs) has demonstrated impressive capabilities in counterfactual generation (Sanchez & Tsaftaris, 2022). Building on these developments, researchers have further explored the role of the latent space in DMs. Kwon et al. (2022) introduced a semantic latent space to address the semantic generative problems. Park et al. (2023) have analyzed the theoretical properties of the latent space in Riemannian geometry. Despite the remarkable success of diffusion models, there remains a lack of exploration into the latent space associated with the forward process, particularly in understanding its role in counterfactual generation.

To enhance the performance of counterfactual generation, we propose a novel latent space for DMs and an algorithm that leverages an auxiliary classifier and information knowledged by human. To the best of our knowledge, this is the first attempt to explicitly balance the preservation of factual features with the generation of counterfactual features. We design a counterfactual generative model and evaluate our method on the colored MNIST dataset, demonstrating improvements over the original latent space.

In this paper, our main contributions are: (i) We propose a latent space of DMs for counterfactual generation using auxiliary classifier; (ii) We propose an algorithm for mapping the image data to the balanced latent space using auxiliary classifier; (iii) We design a counterfactual generative model which can address OOD tasks.

2 PRELIMINARIES

2.1 DENOISING DIFFUSION PROBABILISTIC MODELS (DDPMS)

DDPMs (Ho et al. (2020)) are defined as a Markov chain that gradually adds Gaussian noise to images $x_0 \sim p_{data}(x)$. The latent variable x_t, with $t \in [0, T]$, can be expressed as a linear combination of x_0 and a noise z:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} z, \quad where \quad z \sim \mathcal{N}(\mathbf{0},\ \mathbf{I}) \tag{1}$$

Generative models aim to learn a denoising function ε_θ that predicts the noise component in Eq. 1. Once ε_θ has been trained, the sampling process is performed via a reverse Markov chain starting from $x_T \sim \mathcal{N}(\mathbf{0},\ \mathbf{I})$:

$$x_{t-1} = \frac{1}{\sqrt{1 - \beta_t}} \left[x_t + \beta_t \varepsilon_\theta(x_t, t) \right] + \sqrt{\beta_t} z, \quad t \in [0, T],\ z \sim \mathcal{N}(\mathbf{0},\ \mathbf{I}) \tag{2}$$

2.2 SAMPLING PROCESS WITH LATENT VARIABLES AND INTERVENTIONS

We assume that the denoising model ε_θ is pre-trained and fixed. Instead of denoising a white noise $x_T \sim \mathcal{N}(\mathbf{0},\ \mathbf{I})$, Denoising Diffusion Implicit Models (DDIMs, Song et al. (2021a)) formulate a forward process that allows a mapping from the image x_0 to a latent variable x_T:

$$x_{t+1} = \sqrt{\alpha_{t+1}} \left(\frac{x_t - \sqrt{1 - \alpha_t} \varepsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) + \sqrt{\alpha_{t+1}} \varepsilon_\theta(x_t, t), \quad t \in [0, T] \tag{3}$$

This process preserves certain features of the original image throughout the sampling process. To enable controllable intervention, Sanchez & Tsaftaris (2022) proposed a backward process based on Eq. 14 in Song et al. (2021b):

$$\begin{aligned} \varepsilon &:= \varepsilon_\theta(x_t, t) - s\sqrt{1 - \alpha_t} \nabla_{x_t} \log p_t(y \mid x_t) \\ x_{t-1} &= \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1 - \alpha_t} \varepsilon}{\sqrt{\alpha_t}} \right) + \sqrt{\alpha_{t-1}} \varepsilon, \quad t = T,\ T-1, \cdots,\ 1 \end{aligned} \tag{4}$$

where $\log p_t(y \mid x_t)$ is a conditional log probability that can be estimated by a separate model. In this case, the sampling process can be divided into two parts:

- **Forward process toward latent space:** Given an original image x_0, generate the latent variable x_T using Eq. 3.
- **Backward process with intervention:** Denoise the latent variable x_T and obtain x_0 by applying Eq. 4 iteratively.

2.3 ISSUES WITH THE LATENT SPACE OF DDIMS

Note that Eq. 3 implies a latent space in which each latent variable x_T retains certain features of x_0. However, the preserved features in x_T may be disproportionate, resulting in an unbalanced latent representation. For instance, in the MNIST dataset, the latent variable x_T closely resembles the original image x_0, as illustrated in Fig. 1.

When the original and target distributions differ significantly, such imbalance may cause the model to fail in generating accurate results.

3 METHODOLOGY

3.1 OVERVIEW OF THE BALANCED LATENT SPACE AND FRAMEWORK

We will introduce the definition and construction of our balanced latent space as shown in Fig. 1. The framework of our proposed model for counterfactual generation is illustrated in Fig. 2 in Appendix 6.2.

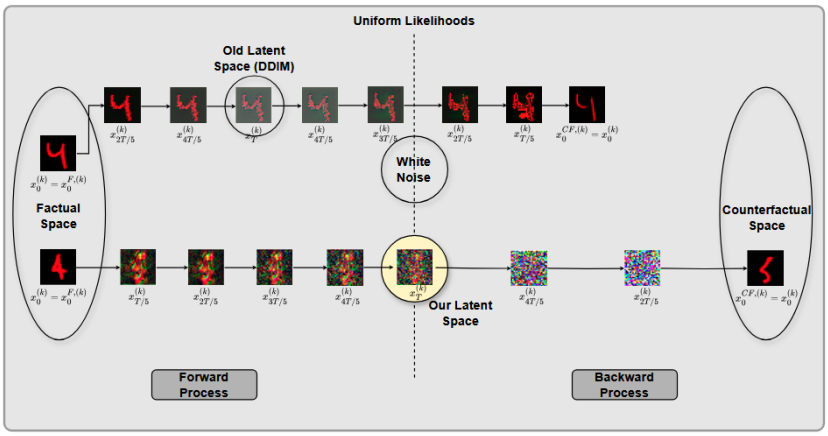

Figure 1: Overview of the balanced latent space and comparison with the old latent space.

3.2 Counterfactual Generation

3.2.1 Balanced Latent Space in Sampling Process

Suppose the set of data points X, together with a metric $d : X \times X \to \mathbb{R}$, forms a metric space (X, d). Let $(X_F, d|_{X_F \times X_F})$ and $(X_{CF}, d|_{X_{CF} \times X_{CF}})$ be the factual (original) and counterfactual (target) subspaces, separately. We define the balanced latent space as follows:

Definition 1 (Balanced latent space)**.** *A subspace $(X_T, d|_{X_T \times X_T}) \subset (X, d)$ is a balanced latent space if $\forall x_T \in X_T$, we have:*

$$d(x_T, x_0^F) = d(x_T, x_0^{CF}) \tag{5}$$

where $x_0^F \in X_F$ and $x_0^{CF} \in X_{CF}$ are the factual and counterfactual points corresponding to x_T.

Although the metric d can be defined as a simple Euclidean distance, we leverage the conditional probability $p_t(y \mid x)$ from Eq. 4 to define the balanced latent space with a specific metric:

Definition 2 (Balanced latent space with uniform likelihood)**.** *Consider a metric d defined by:*

$$d(x_1, x_2) := |p_t(y_{CF}|x_1) - p_t(y_{CF}|x_2)| + \rho(x_1, x_2) \tag{6}$$

where $y_{CF} \in \{0, 1\}$ is the label related to the counterfactual space and

$$\rho(x_1, x_2) = \begin{cases} 0, & x_1 = x_2 \\ 1, & x_1 \neq x_2 \end{cases} \tag{7}$$

Then $(X_T, d|_{X_T \times X_T}) \subset (X, d)$ in Def. 1 is a balanced latent space with uniform likelihood.

Theorem 1. *The function $d : X \times X \to \mathbb{R}$ defined by Eq. 6 is a metric.*

Theorem 2 (Uniform likelihood)**.** *$d(x_T, x_0^F) = d(x_T, x_0^{CF})$ if and only if $p_t(y_{CF} \mid x_T) = p_t(y_F \mid x_T)$.*

The corresponding proofs are provided in Appendix 6.1.

3.2.2 Forward Process Toward Balanced Latent Space

We can use auxiliary information that is easy to obtain to design an "auxiliary classifier" to estimate the conditional probability $p_t(y \mid x)$. Examples of auxiliary information include the color or shade of images that can be easily obtained from image metadata. After training the auxiliary classifier $p_\phi(y \mid x)$, we can derive a forward process toward the balanced latent space:

$$\Delta x_t = \varepsilon_\theta \left[x_t + \zeta_t \cdot \nabla_{x_t} p_\phi(y_{CF}|x_t), t\right] - \varepsilon_\theta \left(x_t, t\right) \tag{8}$$

$$x_{t+1} = x_t + \gamma_1 \Delta x_t + \gamma_2 z, \quad z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \; t = 0, \cdots, T-1 \tag{9}$$

where γ_1 and γ_2 are hyperparameters, $\nabla_{x_t} p_\phi$ is the gradient of the auxiliary classifier and ζ_t is a empirical function defined by:

$$\zeta_t = 2 \exp\left(-\frac{1}{4 \max\{10^{-6}, \Delta p\}}\right) \tag{10}$$

where

$$\Delta p = p_\phi(y_F \mid x_t) - p_\phi(y_{CF} \mid x_t) \tag{11}$$

ζ_t is an adaptive coefficient that rapidly decays to zero, as Δp approaches zero. This design ensures a large update step at the beginning of the forward process, while guaranteeing that the update gradually stops once x_T lies within the balanced latent space.

The complete sampling algorithm is provided in Appendix 6.2.

4 Experiments and Results

Dataset: The colored MNIST dataset is a modified version of the standard MNIST (Lecun et al., 1998) dataset, where each digit is assigned a specific color. In this dataset, the training set contains digits 0–4 in red and 5–9 in green, while the test set uses the opposite coloring.

Implementation: We apply our model to the colored MNIST dataset, where digits are represented in RGB color. Each digit class is assigned an auxiliary label corresponding to its shape. The diffusion model ε_θ is implemented using a UNet architecture, following the approach of Nichol & Dhariwal (2021). The auxiliary classifier p_ϕ is a partial UNet model, consisting only of the encoder part of ε_θ. Our goal is to generate red digit 5, which appears only in the test set and not in the training set.

Fig.3 illustrates how the likelihoods evolve during the forward process. By the end of the forward process ($t = 1000$), each sample becomes a latent variable with uniform likelihoods, indicating unbiased distances between the factual and counterfactual spaces. We observe that these latent variables incorporate both the original shape and the target shape (digit 5), consistent with the observed uniform likelihoods.

As shown in Fig.4, the diffusion model failed to generate digits 5 or preserve the original features from red digits when using the original latent space. In contrast, with our proposed latent space, most red digits are successfully transformed into digit 5 while retaining their original features.

Additional experimental results and details and be found in Appendix 6.4.

5 Conclusion

This paper proposes a framework that balances the latent variables in the diffusion process for more efficient and authentic counterfactual generation. In the proposed framework, the denoising process is guided by an auxiliary classifier, which induces a counterfactual generation signal. We further propose a balancing method for the diffusion process to ensure that the information retained in the latent variables of the original (factual) and target (counterfactual) states is balanced, enabling better counterfactual generation by facilitating the transfer between latent variables. This balancing method improves generation efficiency by avoiding two inefficiencies: retaining too much of the original latent information, resulting in minimal modification, or incorporating excessive latent signals from the target, leading to a loss of the original pattern. We demonstrate the effectiveness of the proposed method using the colored MNIST dataset, achieving more authentic and controlled counterfactual generation aligned with the counterfactual signal.

References

Saloni Dash, Vineeth N Balasubramanian, and Amit Sharma. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals, 2022. URL https://arxiv.org/abs/2009.08270.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Klaus-Rudolf Kladny, Julius von Kügelgen, Bernhard Schölkopf, and Michael Muehlebach. Deep backtracking counterfactuals for causally compliant explanations, 2024. URL https://arxiv.org/abs/2310.07665.

Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training, 2017. URL https://arxiv.org/abs/1709.02023.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. URL https://arxiv.org/abs/2102.09672.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36:24129–24142, 2023.

Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference, 2020. URL https://arxiv.org/abs/2006.06485.

Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High fidelity image counterfactuals with probabilistic causal models, 2023. URL https://arxiv.org/abs/2306.15764.

Pedro Sanchez and Sotirios A. Tsaftaris. Diffusion causal models for counterfactual estimation, 2022. URL https://arxiv.org/abs/2202.10166.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

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

E. G. Tabak and Cristina V. Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013. doi: https://doi.org/10.1002/cpa.21423. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21423.

Jingkang Wang, Sivabalan Manivasagam, Yun Chen, Ze Yang, Ioan Andrei Bârsan, Anqi Joyce Yang, Wei-Chiu Ma, and Raquel Urtasun. Cadsim: Robust and scalable in-the-wild 3d reconstruction for controllable sensor simulation, 2023. URL https://arxiv.org/abs/2311.01447.

Yousef Yeganeh, Ioannis Charisiadis, Marta Hasny, Martin Hartenberger, Björn Ommer, Nassir Navab, Azade Farshad, and Ehsan Adeli. Latent drifting in diffusion models for counterfactual medical image synthesis, 2024. URL https://arxiv.org/abs/2412.20651.

6 Appendix

6.1 Details of the Balanced Latent Space

Proof of Theorem 1. Recall that the function $d : X \times X \to \mathbb{R}$ is defined by:

$$d(x_1, x_2) := |p_t(y_{CF} \mid x_1) - p_t(y_{CF} \mid x_2)| + \rho(x_1, x_2) \tag{12}$$

1. If $x_1 = x_2$, then we have

$$d(x_1, x_2) = |p_t(y_{CF} \mid x_1) - p_t(y_{CF} \mid x_1)| + \rho(x_1, x_1) = 0 \tag{13}$$

If $d(x_1, x_2) = 0$, then we have

$$|p_t(y_{CF} \mid x_1) - p_t(y_{CF} \mid x_2)| = 0 \quad and \quad \rho(x_1, x_2) = 0 \tag{14}$$

From the definition of ρ, we know that $x_1 = x_2$. Therefore, $x_1 = x_2$ if and only if $d(x_1, x_2) = 0$.

2. (Symmetry) $\forall x_1, x_2 \in X$, we have

$$\begin{aligned} d(x_1, x_2) &= |p_t(y_{CF} \mid x_1) - p_t(y_{CF} \mid x_2)| + \rho(x_1, x_2) && (15)\\ &= |p_t(y_{CF} \mid x_2) - p_t(y_{CF} \mid x_1)| + \rho(x_2, x_1) && (16)\\ &= d(x_2, x_1) && (17) \end{aligned}$$

3. (Triangle inequality) Note that ρ is in fact a discrete metric and its triangle inequality naturally holds. $\forall x_1, x_2, x' \in X$, we have

$$\begin{aligned} d(x_1, x_2) &= |p_t(y_{CF} \mid x_1) - p_t(y_{CF} \mid x_2)| + \rho(x_1, x_2) && (18)\\ &= |p_t(y_{CF} \mid x_1) - p_t(y_{CF} \mid x') + p_t(y_{CF} \mid x') - p_t(y_{CF} \mid x_2)| + \rho(x_1, x_2) && (19)\\ &\leq |p_t(y_{CF} \mid x_1) - p_t(y_{CF} \mid x')| + |p_t(y_{CF} \mid x') - p_t(y_{CF} \mid x_2)| + \rho(x_1, x_2) && (20)\\ &\leq |p_t(y_{CF} \mid x_1) - p_t(y_{CF} \mid x')| + \rho(x_1, x') + |p_t(y_{CF} \mid x') - p_t(y_{CF} \mid x_2)| && (21)\\ &\quad + \rho(x', x_2) && (22)\\ &= d(x_1, x') + d(x', x_2) && (23) \end{aligned}$$

Hence, d is a metric. □

Proof of Theorem 2. We only consider the case where $x_T \neq x_0^F$ and $x_T \neq x_0^{CF}$. Suppose $d(x_T, x_0^F) = d(x_T, x_0^{CF})$, then

$$\begin{aligned} &|p_t(y_{CF} \mid x_T) - p_t(y_{CF} \mid x_0^F)| + 1 = |p_t(y_{CF} \mid x_T) - p_t(y_{CF} \mid x_0^{CF})| + 1 && (24)\\ \Rightarrow &|p_t(y_{CF} \mid x_T) - p_t(y_{CF} \mid x_0^F)| = |p_t(y_{CF} \mid x_T) - p_t(y_{CF} \mid x_0^{CF})| && (25)\\ \Rightarrow &|p_t(y_{CF} \mid x_T) - 0| = |p_t(y_{CF} \mid x_T) - 1| && (26) \end{aligned}$$

Since $0 \leq p_t(y_{CF} \mid x_T) \leq 1$, we have

$$p_t(y_{CF} \mid x_T) = \frac{1}{2} = p_t(y_F \mid x_T) \tag{27}$$

On the other hand, suppose $p_t(y_{CF} \mid x_T) = p_t(y_F \mid x_T) = \frac{1}{2}$, then

$$\begin{aligned} d(x_T, x_0^F) &= |p_t(y_{CF} \mid x_T) - p_t(y_{CF} \mid x_0^F)| + \rho(x_T, x_0^F) && (28)\\ &= |\frac{1}{2} - 0| + 1 && (29)\\ &= \frac{3}{2} && (30) \end{aligned}$$

and

$$\begin{aligned} d(x_T, x_0^{CF}) &= |p_t(y_{CF} \mid x_T) - p_t(y_{CF} \mid x_0^{CF})| + \rho(x_T, x_0^{CF}) && (31)\\ &= |\frac{1}{2} - 1| + 1 && (32)\\ &= \frac{3}{2} && (33)\\ &= d(x_T, x_0^F) && (34) \end{aligned}$$

Hence, $d(x_T, x_0^F) = d(x_T, x_0^{CF})$ if and only if $p_t(y_{CF} \mid x_T) = p_t(y_F \mid x_T)$. □

6.2 SAMPLING ALGORITHM

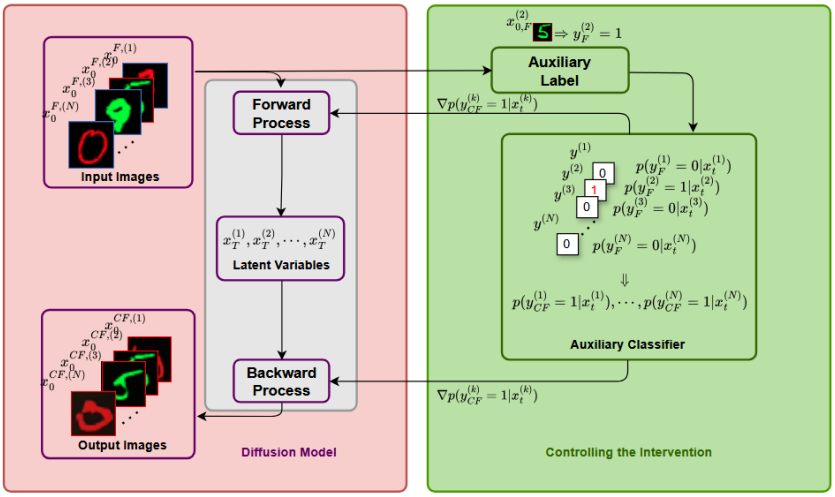

Figure 2: Overview of the proposed framework. Both the forward and backward processes are guided by the gradient of an auxiliary classifier. The gray block, which encompasses both processes, corresponds to our balanced latent space.

Algorithm 1 Sampling Process with Balanced Latent Space

Require: Input data x_0^F
Ensure: Output result x_0^{CF}

for $t = 0$ to T **do** ▷ Forward process toward balanced latent space
  $\Delta x_t = \varepsilon_\theta \left[x_t + \zeta_t \cdot \nabla_{x_t} p_\phi(y_{CF}|x_t), t\right] - \varepsilon_\theta\left(x_t, t\right)$
  $x_{t+1} = x_t + \gamma_1 \Delta x_t + \gamma_2 z, \quad z \sim (\mathbf{0}, \mathbf{I})$
end for

for $t = T$ to 0 **do** ▷ Generation under intervention
  $\varepsilon = \varepsilon_\theta(x_t, t) - s\sqrt{1-\alpha_t}\nabla_{x_t} \log p_\phi(y_{CF}|x_t)$
  $x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\varepsilon}{\sqrt{\alpha_t}}\right) + \sqrt{\alpha_{t-1}}\varepsilon$
end for
$x_0^{CF} = x_0$

6.3 EXPERIMENT DETAILS

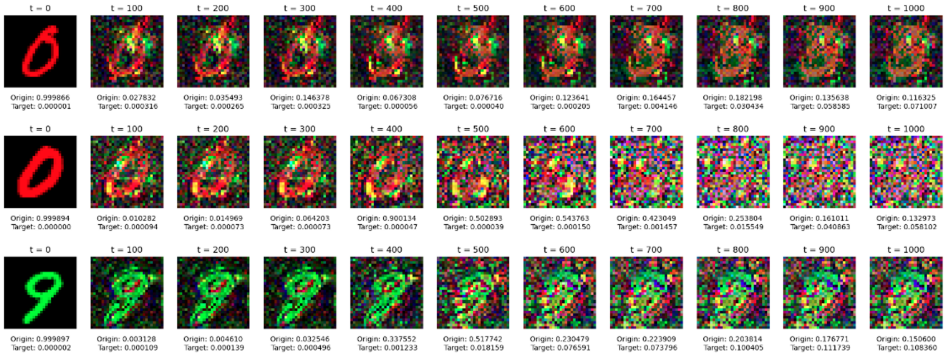

Figure 3: Sample likelihoods at different time steps during the forward process. We randomly select three different digits and compare their likelihoods of being classified as the original digit versus the target digit.

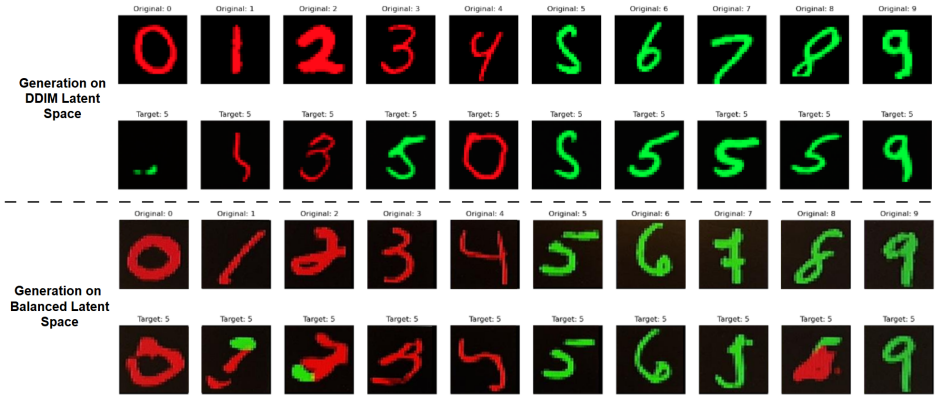

Figure 4: Generation results using the original latent space (DDIM) and the proposed balanced latent space on the colored MNIST dataset.

6.4 ADDITIONAL EXPERIMENTAL RESULTS AND DETAILS

Dataset: The masked MNIST dataset is a modified version of the standard MNIST (Lecun et al., 1998) dataset, where each digit is assigned a specific background color. In this dataset, the training set consists of digits 0–4 with a red background and 5–9 with a green background, while the test set has the opposite coloring.

Implementation: We apply our model to the masked MNIST dataset, where digits are assigned an auxiliary label corresponding to their background color. The diffusion model ε_θ is implemented using a UNet architecture, following the approach of Nichol & Dhariwal (2021). The auxiliary classifier p_ϕ is a partial UNet model that consists only of the encoder part of ε_θ. Our goal is to generate digits 0-4 with a green background, which appear only in the test set and not in the training set.

As shown in Fig.5, when given any digit with a red background, the diffusion model fails to change the background color using the original latent space. In contrast, with our proposed latent space, all red backgrounds are successfully transformed into green backgrounds while preserving their original features.

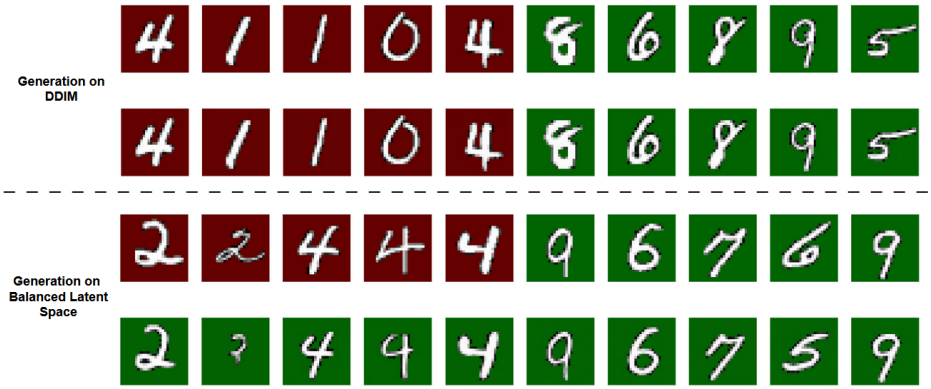

Figure 5: Generation results using the original latent space (DDIM) and the proposed balanced latent space on the masked MNIST dataset.