Interpretability of Diffusion Models

Sirui, Li
223040027
School of Data Science
The Chinese University of Hong Kong, Shenzhen
223040027@link.cuhk.edu.cn

Yangkuan, Li
223040028
School of Data Science
The Chinese University of Hong Kong, Shenzhen
223040028@link.cuhk.edu.cn

Linmeng, Li
223040026
School of Data Science
The Chinese University of Hong Kong, Shenzhen
223040026@link.cuhk.edu.cn

Abstract

Generative models are essential tools in machine learning, with Variational Autoencoders (VAEs) and diffusion models serving as primary instruments for synthesizing complex data distributions. Despite their strengths, current models face challenges in controllability and disentanglement, limiting their utility in refined application scenarios. Addressing these challenges, this study introduces a novel generative framwork that amalgamates the high-quality generative capabilities of diffusion models with the expressive and interpretable latent space of VAEs. This integration aims to leverage the structured nature of VAEs' latent space to direct the generative prowess of diffusion models, thereby achieving refined control and augmented sample diversity. To empirically validate the model's efficacy, we conducted extensive experiments focusing on image generation tasks, employing the CelebA-HQ Mask dataset to benchmark against standalone VAE and diffusion model baselines. Our findings indicate that the DiffusionVAE not only excels in generating images with superior fidelity, as evidenced by lower Frechet Inception Distance (FID) scores, but also demonstrates enhanced capability in controlling and disentangling generative factors. This research contributes a novel perspective to the generative model domain, offering theoretical and empirical insights into achieving controlled and interpretable generation processes and opening avenues for future explorations in highly controllable generative applications. The code for replicating the method is available at `https://github.com/Naukode/ML-course-Project`.

1 Introduction

With the rapid development of machine learning technology, generative models have become an important branch in the field of deep learning, especially demonstrating significant capabilities in the synthesis of images, audio, and text. Among them, diffusion models and VAEs serve as primary tools for generating complex data distributions, each showcasing unique advantages. Diffusion models are

particularly noted for their ability to produce high-quality images, while VAEs are widely applied in various tasks due to their powerful latent space expressiveness and higher interpretability. Nonetheless, current generative models still face challenges in terms of controllability and disentanglement, which limit their utility in more refined application scenarios.

Controllability refers to the ability of a model to manipulate specific features in the generation process in a predetermined manner, while disentanglement involves the model's ability to independently control and represent different generative factors of the data. High controllability and good disentanglement are crucial for enhancing the practicality and interpretability of generative models, especially when tasks require precise adjustments in generated content.

Based on reading a lot of papers, Pandey et al. (2022) gives us direction and ideas.To address these issues, this study introduces the DiffusionVAE model, a new model that attempts to integrate the advantages of VAEs and diffusion models, aiming to improve both the quality of generation and the model's controllability and disentanglement. By combining the latent space of VAEs with the generative process of diffusion models, DiffusionVAE aims to utilize the structured latent space of VAEs to control the output of diffusion models, thereby achieving finer control and greater sample diversity.

Furthermore, this research delves into controlling the generative process through operations in the latent space. Particularly in the DiffusionVAE model, we experimentally verify the impact of disentangling and controlling operations on the latent variables on the quality and diversity of the generated samples. We find that precise control over latent variables allows for adjustments to specific image features during the generation process, broadening the possibilities for applications of generative models.

Through these studies, we not only demonstrate the effectiveness of DiffusionVAE in image generation tasks but also provide new theoretical and empirical foundations for research into the controllability and disentanglement of generative models. These results foreshadow potential future directions for generative models, particularly in applications requiring highly controllable and interpretable generation processes.

2 Background

2.1 VAEs

Developed by Kingma and Welling (2013), Variational Autoencoders (VAEs) represent a significant advancement in generative models, employing a probabilistic framework to encode and reconstruct data efficiently. The core architecture of VAE comprises two primary components: an encoder that maps input data into a latent space distribution characterized typically by Gaussian parameters—means and variances—and a decoder that reconstructs the input data from sampled latent variables. VAEs aim to optimize the Evidence Lower Bound (ELBO), where the ELBO can be satisfied as

$$ELBO = E_{q_\phi(z|x)}[logp_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z)) \tag{1}$$

where $q_\phi(z|x)$ on the data's marginal likelihood, effectively balancing data likelihood maximization with latent distribution regularization towards a standard Gaussian. This optimization not only helps in accurate data reconstruction but also ensures that the latent space maintains meaningful and generalizable properties, making VAEs particularly useful for tasks requiring robust and interpretable generative models. The specific structure can be shown as Figure 1 below.

2.2 Diffusion Models

Diffusion models (Croitoru et al. (2023)) are a class of generative models that transform data into pure noise through a forward process and then reverse this transformation to regenerate the data from noise, effectively simulating a Markov chain where Gaussian noise is incrementally added to corrupt the data progressively. These models are particularly noted for producing high-quality, detailed outputs in image and audio generation. The forward process progressively shifts the data towards a standard Gaussian distribution, while the reverse process, which involves training a neural network to estimate and correct the noise at each step, aims to reconstruct the original data from

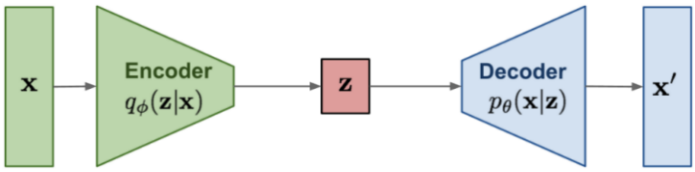

Figure 1: The Structure of VAEs

its noisy version, optimizing the model parameters to achieve accurate data recovery. The specific structure can be shown as Figure 2 below.

Figure 2: The Structure of Diffusion Models

3 Approach

3.1 Model Architecture and Training Objective

We present an innovative generative model architecture that combines the strengths of VAE and DDPM. Our model's architecture operates in two stages: the first focuses on optimizing the VAE, and the second optimizes the DDPM, integrating VAE's reconstructions into the diffusion process for enhanced interpretability and generative performance. The detailed structure is shown in the Figure 3 below.

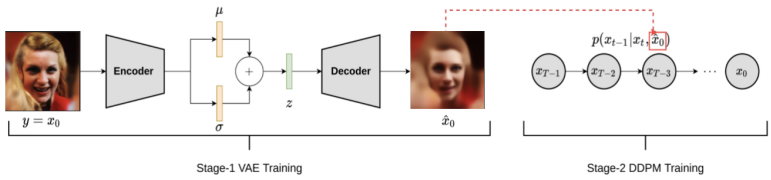

Figure 3: The Structure of our model

The joint distribution of our model is given by:

$$p(x_{0:T}, y, z) = p(z)p_\theta(y|z)p_\phi(x_{0:T}|y, z) \tag{2}$$

where x_0 is the original image, y is an auxiliary conditioning signal modeled using VAE, z represents the latent space associated with y, and $x_{0:T}$ is a sequence of T images learned by the diffusion model. Parameters θ and ϕ are associated with the VAE and DDPM components, respectively.

Due to the intractability of the joint posterior $p(x_{1:T}, z|y, x_0)$, we approximate it with a surrogate posterior:

$$q(x_{1:T}, z|y, x_0) = q_\psi(z|y, x_0)q(x_{1:T}|y, z, x_0) \tag{3}$$

The log-likelihood of the generative process is approximated by the Evidence Lower Bound (ELBO):

$$\log p(x_0, y) = \log \int p(x_{0:T}, y, z)\, dx_{1:T} dz \tag{4}$$

$$\geq \mathbb{E}_{q(x_{1:T}, z|x_0, y)} \left[\log \frac{p(x_{0:T}, y, z)}{q(x_{1:T}, z|x_0, y)} \right] \tag{5}$$

$$\geq \underbrace{\mathbb{E}_{q_\psi(z|y,x_0)}[p_\theta(y|z)] - D_{KL}(q_\psi(z|y, x_0)||p(z))}_{L_{VAE}} + \mathbb{E}_{z\sim q(z|y,x_0)} \left[\underbrace{\mathbb{E}_{q(x_{1:T}|y,z,x_0)} \left[\frac{p_\phi(x_{0:T}|y, z)}{q(x_{1:T}|y, z, x_0)} \right]}_{L_{DDPM}} \right] \tag{6}$$

To facilitate training, we assume:

- The conditioning signal y to be x_0 itself, ensuring a deterministic mapping.
- The second stage DDPM model is conditioned on the VAE reconstruction $\hat{x}_0$, a deterministic function of z.
- A two-stage training process: first optimizing L_{VAE} and then L_{DDPM}, with θ and ψ fixed during the second stage.

3.2 Parameterization

3.2.1 VAE Parameterization

Standard VAE is employed as discussed in the background section.

3.2.2 DDPM Parameterization

Forward Generative Process We integrate VAE reconstruction $\hat{x}_0$ into the forward process as a condition, modeled as:

$$q(x_t|x_{t-1}, \hat{x}_0) = N(\sqrt{1-\beta_t}x_{t-1} + (1 - \sqrt{1-\beta_t})\hat{x}_0, \beta_t\mathbf{I}) \tag{7}$$

$$q(x_t|x_0, \hat{x}_0) = N(\sqrt{1-\overline{\alpha}_t}x_0 + (1 - \sqrt{1-\overline{\alpha}_t})\hat{x}_0, (1-\overline{\alpha}_t)\mathbf{I}) \tag{8}$$

Reverse Generative Process We assume the transitions only depend on the VAE reconstruction $\hat{x}_0$:

$$p(x_{0:T}|z) \approx p(x_{0:T}|\hat{x}_0) \tag{9}$$

4 Experiment

4.1 Generator-Refiner Framework

Our proposed "generator-refiner" framework employs a two-stage training approach that significantly enhances the visual quality of generated images. In the first stage, a VAE constructs a basic visual outline. The second stage involves a DDPM, which refines this initial image to achieve superior detail and realism. This framework is particularly advantageous for high-quality applications such as facial generation and artistic creation. We trained our models using the CelebA-HQ Mask dataset (Karras et al. (2017)), and the results (illustrated in Figure 4 and detailed in Table 1) demonstrated that our model significantly outperforms the VAE baseline in terms of image fidelity, as quantified by Frechet Inception Distance (FID) scores (Heusel et al. (2017)), maintaining competitive benchmarks against standalone DDPM models.

Figure 4: Sampling Variations Across Different Models

Models	FID on 5k samples
VAE (baseline)	115.82
DDPM	43.87
Our Model	51.44

Table 1: Comparison of FID Scores Across Different Models

4.2 Enhanced Controllable Generation through Latent Interpolation

4.2.1 Model Interpolation Techniques

Interpolation (Vahdat et al. (2022)) within our model is implemented throough two distinct latent representations: the low-dimensional VAE latent code z_{vae} and the DDPM intermediate representations z_T.

- **Interpolation in VAE latent space** z_{vae}:

  By sampling two VAE latent codes $z_{vae}^{(1)}$ and $z_{vae}^{(2)}$ from a standard Gaussian distributionand interpolating between them, we generate an intermediate VAE latent code:

$$z_{vae} = \lambda z_{vae}^{(1)} + (1-\lambda) z_{vae}^{(2)} \tag{10}$$

  where λ is the interpolation factor ranging from 0 to 1. This method allows for the generation of Our model samples that, while maintaining overall image structure such as facial expressions and hairstyles, exhibit enhanced details compared to their VAE-generated counterparts.

Figure 5: Our model Samples Generated By Linearly Interpolating in the z_{vae} Latent Space

  Figure 5 displays samples generated by our VAE through interpolation between two sampled VAE codes, as outlined earlier. Correspondingly, Figure 4 (Top Row) presents samples refined by our model, obtained by interpolating within the z_{vae}space. These refined samples, derived from the initially blurry VAE outputs, notably retain the essential structure of the image, demonstrating the efficacy of our generator-refiner framework in preserving key image attributes while enhancing detail and clarity.

- **Interpolation in DDPM latent space** X_T:

  With a fixed z_{vae}, we interpolate between two sampled DDPM representations $X_T^{(1)}$ and $X_T^{(2)}$:

$$X_T = \lambda X_T^{(1)} + (1-\lambda) X_T^{(2)} \tag{11}$$

  This approach influences subtle features within the generated images, ensuring major structure across samples.

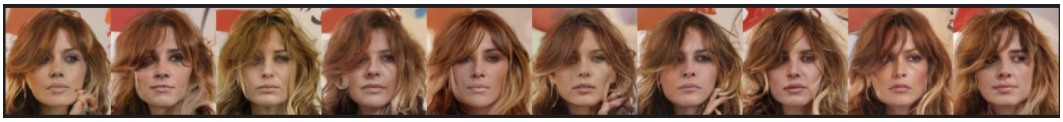

Figure 6: Our model Samples Generated By Linearly Interpolating In TheX_T Latent Space

Figure 6 illustrates the samples generated by our model with a fixed z_{vae} and interpolated x_T. It is evident that interpolating within the DDPM latent space results in variations in subtle features of the generated images. However, the overall structural integrity of the images remains consistent across different samples.

- **Consistency through Shared Stochasticity**:
  By sharing the stochasticity in the DDPM reverse process across all generated samples, we achieve consistent latent interpolations and enable deterministic sampling. This shared approach implies the same style being applied to all refined samples, leading to smoother transitions and uniformity in style across different interpolations, as depicted in Figure 7.

Figure 7: Samples Generated With Sharing The Stochasticity

Figure 7 demonstrates that sharing stochasticity in the DDPM sampling process across samples results in a uniform stylization for all refined images. This uniformity promotes smooth transitions between interpolations.

4.3 Improved β-VAE for Disentanglement

To enhance latent space disentanglement, we employ an advanced β-VAE (Burgess et al. (2018)) with a beta parameter greater than one, emphasizing KL divergence over reconstruction loss. This focus allows for significant disentanglement but could potentially degrade image quality. To mitigate this, we introduce a Dynamically Adjusted Parameter (C), initially prioritizing KL divergence to achieve disentanglement and subsequently adjusting to enhance reconstruction quality:

$$L(\theta, \phi; x, z, C) = E_{q_\phi(z|x)}[log_{p_\theta(x|z)}] - \gamma|D_{KL}q_\phi(z|x)||p(z) - C| \tag{12}$$

This dynamic adjustment ensures robust disentanglement while simultaneously maintaining high-quality image generation capabilities.

4.4 Experiment Details

To thoroughly evaluate the performance and robustness of our model framework, extensive experiments were conducted using carefully selected hyperparameters for both the β-VAE and DDPM stages. The following subsections detail the configuration settings and hyperparameters used, which have been optimized for the CelebA-HQ-Mask dataset to ensure that the results are both reliable and replicable.

4.4.1 Hyperparameters for improved β-VAE

The β-VAE stage of our model was rigorously trained with the following hyperparameters on Table 2:

- **Dataset:** High-resolution CelebA-HQ-Mask was chosen for its rich diversity in facial attributes, which is critical for assessing the performance of generative models in handling complex image details.
- **Resolution:** Each image was resized to 128x128 pixels to balance between computational efficiency and sufficient detail retention.
- **Batch Size:** Set at 128 to maximize hardware utilization without exceeding memory constraints.

- **Epochs:** The model was trained for 500 epochs to ensure thorough learning without overfitting.
- **Optimizer:** Adam optimizer was utilized for its adaptive learning rate capabilities, facilitating faster convergence.
- **Learning Rate:** Set at 1×10^{-4}, this rate was found optimal for steady progression in learning without causing instability in training dynamics.
- **Latent Code Size:** A size of 1024 was selected to capture a comprehensive amount of information about the data while maintaining manageable computational demands.
- α **Value:** Set at 100 to appropriately balance the trade-off between reconstruction fidelity and KL divergence.
- **Cmax and C_stop_iter:** These parameters were set at 25 and 100,000 iterations, respectively, to dynamically adjust the influence of KL divergence, promoting effective disentanglement over time.

Parameter	Value
Dataset	Celeb-A-HQ-Mask
Resolution	128x128 pixels
Batch Size	128
Epochs	500
Optimizer	Adam
Learning Rate	1×10^{-4}
Latent Code Size	1024
α value	100
Cmax	25
C_stop_iter	100,000

Table 2: Hyperparameters for the β-VAE Experiment

4.4.2 Hyperparameters for DDPM

The DDPM stage of our model was trained with the following hyperparameters on Table 3:

Parameter	Value
Dataset	CelebA-HQ-Mask
Resolution	128x128 pixels, 128 channels
Architecture Details	Scales of attention blocks: 16, Attention heads: 8, Residual blocks per scale: 2
Channel Multipliers	(1, 2, 2, 3, 4)
Total Parameters	95.2 million
Dropout	0.1
Noise Schedule	Linear from 1×10^{-4} to 0.02
Training Time Steps	100
EMA Decay Rate	0.9999
Batch Size	64
Optimizer	Adam
Learning Rate	2×10^{-5}
LR Annealing Steps	5000
Diffusion Loss Type	Noise prediction

Table 3: Hyperparameters and Configuration of the DDPM Experiment

- **Dataset and Resolution:** Consistent with the β-VAE, images from the CelebA-HQ-Mask dataset at 128x128 pixels were used.
- **Architecture Details:** The model features scales of attention blocks at 16, with 8 attention heads and 2 residual blocks per scale, ensuring detailed attention to features across various scales.

- **Channel Multipliers:** Configured as (1, 2, 2, 3, 4), these multipliers escalate the complexity and capacity of the model progressively through the network.
- **Total Parameters:** The model architecture comprises 95.2 million parameters, striking a balance between complexity and performance.
- **Dropout and Noise Schedule:** A dropout rate of 0.1 combined with a linear noise schedule from 1×10^{-4} to 0.02 was used to promote model robustness and effective noise management.
- **Training Time Steps, EMA Decay Rate, and Batch Size:** These were set at 100 steps, a decay rate of 0.9999, and a batch size of 64, respectively, to optimize the training process for quality and efficiency.
- **Optimizer and Learning Rate Adjustments:** The Adam optimizer with a learning rate of 2×10^{-5} and 5000 annealing steps ensured precise control over learning progression, minimizing potential for overfitting while accommodating for detailed feature capture.

4.4.3 Compute Resources

The experiments were conducted using high-performance computing resources to manage the substantial computational demands of training sophisticated generative models like β-VAE and DDPM. The specific hardware and execution times are detailed below, ensuring transparency and reproducibility of the results.

- **Hardware Configuration:** The training was performed on a robust setup involving high-end graphical processing units to effectively handle the training loads and memory requirements of the models.The specific data is shown in Table 4 below.

Resource Type	Specifications
Type of Compute Workers	Nvidia RTX 4090 GPU
Memory	24 GB

Table 4: Specifications of Compute Resources

- **Execution Time:** The execution time for each phase of the model training was carefully monitored to provide insights into the computational efficiency and practical deployment considerations of our framework. The data needed is shown below in Table 5.

Training Phase	Execution Time
Training for VAE	2 days
Training for DDPM	4 days

Table 5: Execution Time for Model Training Phases

The training of the β-VAE component was completed in 2 days, which involved extensive computational processing to optimize the latent space and ensure robust feature capture. Following this, the DDPM component required an additional 4 days of training, reflecting the intricate calculations and extensive sampling processes intrinsic to diffusion models. The longer training time for the DDPM phase highlights its complexity and the computational effort required to refine the initial outputs from the β-VAE stage to achieve high fidelity and visually appealing results.

5 Conclusion

In summary, this study presents a novel generative framework that synergistically combines Variational Autoencoders (VAEs) and diffusion models to address the challenges of controllability and disentanglement in generative models. Through extensive experimentation on the CelebA-HQ Mask dataset, our findings validate the model's superior capability to generate high-fidelity images while providing enhanced control over the generative process and improved disentanglement of generative

factors. Our model signifies a substantial step forward in the realm of generative models, offering a nuanced approach that leverages the structured latent space of VAEs to guide the generative prowess of diffusion models for controlled and diversified sample generation.

Limitations Despite the promising advancements introduced by our model, certain limitations warrant further investigation. The model's computational demand is significant, necessitating considerable resources for training and optimization, which could constrain its application in resource-limited environments. Additionally, while the model exhibits commendable performance in image generation tasks, its adaptability and efficacy across diverse data types and more complex generation tasks remain to be fully explored.

Future Work Looking ahead, there are multiple avenues for future research to build upon the foundations laid by this study. A primary area of interest involves exploring the potential of the framework of our model in the realm of Natural Language Processing analogies, akin to the queen-king = woman-man paradigm, for controllable image generation. This exploration could pave the way for sophisticated applications where nuanced adjustments to generated images are necessary, further bridging the gap between human cognitive processes and machine-generated content. Moreover, efforts to optimize the computational efficiency of the model could significantly broaden its applicability, making it more accessible for diverse applications. Additionally, extending the application of our model to other data types and generation tasks promises to unlock new capabilities and understandings in the field of generative models.

In conclusion, our model framework introduces a compelling approach to enhancing the quality, controllability, and interpretability of generative models. By addressing current limitations and pursuing outlined future directions, this research can contribute to the advancement of highly controllable and interpretable generative applications, fostering further innovation in the field of machine learning.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the section of Abstract and Introduction, we accurately show the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The information can be found in the conclusion part, such as significant computational cost.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer:[Yes]

Justification: In the Parameterization part, we assume the transitions only depend on the VAE reconstruction.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All the information can be found in Experiment part.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The accessible Github link can be found in Abstract.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The detailed information can be found in Experiment part.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The information can be found in Experiment part. Through the comparison between the VAE and DDPM, the results are significantly increasing.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The specific information can be found in the Experiment Details part.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research strictly adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper gives the positive societal impacts because it discussed a new model which can produce the picture with both great quality and reconstructions. Also, improved the model to β-VAE. Enhancing the interpretability.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, the creators or original owners of all the assets we use in our papers (such as code, data, models, etc.) have been properly recognized, and we have made explicit reference to the licenses and terms of use of these assets, ensuring that our use complies with these regulations.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The accessible link of code can be found in the abstract part.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

References

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. (2018). Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Pandey, K., Mukherjee, A., Rai, P., and Kumar, A. (2022). Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *arXiv preprint arXiv:2201.00308*.

Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K., et al. (2022). Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039.