

[← Go to ICML 2024 Conference homepage \(/group?id=ICML.cc/2024/Conference\)](#)

# Generating, Reconstructing, and Representing Discrete and Continuous Data: Generalized Diffusion with Learnable Encoding-Decoding



*Guangyi Liu (/profile?id=~Guangyi\_Liu1), Yu Wang (/profile?id=~Yu\_Wang24), Zeyu Feng (/profile?id=~Zeyu\_Feng2), Qiyu Wu (/profile?id=~Qiyu\_Wu2), Liping Tang (/profile?id=~Liping\_Tang2), Yuan Gao (/profile?id=~Yuan\_Gao11), Zhen Li (/profile?id=~Zhen\_Li6), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Julian McAuley (/profile?id=~Julian\_McAuley1), Eric P. Xing (/profile?id=~Eric\_Xing1), Zichao Yang (/profile?id=~Zichao\_Yang1), Zhiting Hu (/profile?id=~Zhiting\_Hu3)*



Published: 01 May 2024, Last Modified: 01 May 2024 ICML 2024 Conference, Senior Area Chairs, Area Chairs, Reviewers, Publication Chairs, Authors Revisions (/revisions?id=igRjCCAz2a) BibTeX CC BY 4.0  
(<https://creativecommons.org/licenses/by/4.0/>)

**Verify Author List:** I have double-checked the author list and understand that additions and removals will not be allowed after the submission deadline.

**Keywords:** generation, reconstruction, representation, image, text, protein

**TL;DR:** Generating, Reconstructing, and Representing Discrete and Continuous Data: Generalized Diffusion with Learnable Encoding-Decoding

## Abstract:

The vast applications of deep generative models are anchored in three core capabilities—generating new instances, reconstructing inputs, and learning compact representations—across various data types, such as discrete text/protein sequences and continuous images. Existing model families, like Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), autoregressive models, and diffusion models, generally excel in specific capabilities and data types but fall short in others. We introduce generalized diffusion with learnable encoder-decoder (DILED), that seamlessly integrates the core capabilities for broad applicability and enhanced performance. DILED generalizes the Gaussian noising-denoising in standard diffusion by introducing parameterized encoding-decoding. Crucially, DILED is compatible with the well-established diffusion model objective and training recipes, allowing effective learning of the encoder-decoder parameters jointly with diffusion. By choosing appropriate encoder/decoder (e.g., large language models), DILED naturally applies to different data types. Extensive experiments on text, proteins, and images demonstrate DILED's flexibility to handle diverse data and tasks and its strong improvement over various existing models.

**Primary Area:** Deep Learning (architectures, generative models, deep reinforcement learning, etc.)

**Position Paper Track:** No

**Paper Checklist Guidelines:** I certify that all co-authors of this work have read and commit to adhering to the Paper Checklist Guidelines, Call for Papers and Publication Ethics.

**Submission Number:** 144

Filter by rank type

Filter by author

Search keywords...

Sort: Newest First

☰

☰

☰

-

=

☰

🔗

👁

Everyone

Program Chairs

Submission144 Authors

Submission144...

34 / 37 replies shown

Submission144 Area...

Submission144...

Submission144...

Submission144...

Submission144...

Submission144...

✕

Add: **Withdrawal**

## Official Review of Submission144 by Reviewer LRww

Official Review Reviewer LRww 14 Mar 2024, 23:52 (modified: 21 Mar 2024, 05:11)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer LRww

Revisions (/revisions?id=bDnKShuyIm)

### Summary:

The paper proposes a reformulation of the diffusion models, called DILED, that jointly trains an encoder, a diffusion model in latent space, and a decoder. DILED can generate new samples, invert a sample to its representation, reconstruct, and manipulate representations in latent space. The paper applies DILED on image, text, and protein sequence data types.

### Strengths And Weaknesses:

Strengths:

1. The paper is articulate and detailed.
2. The relevant literature is rigorously discussed.
3. The extension of the DDPM loss to embrace the encoder and decoder losses is insightful.
4. The experiments encompass diverse tasks and datatypes.
5. The method is impactful on multiple applications.

Weakness:

1. The method is directly comparable to latent diffusion, which demands a comprehensive study of their differences in the experiment section.
2. The computational cost is not discussed.

### Questions:

1. What are the potential benefits of applying larger  $t$  in  $\ln 161$ ?
2. One of the advantages of latent diffusion is accelerated training. Did you run any experiments with pretrained encoder and decoder?

### Limitations:

The authors talk about the limitations and societal impact in the paper.

**Ethics Flag:** No

**Soundness:** 3: good

**Presentation:** 4: excellent

**Contribution:** 3: good

**Rating:** 7: Accept: Technically solid paper, with high impact on at least one sub-area, or moderate-to-high impact on more than one areas, with good-to-excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.

**Confidence:** 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

**Code Of Conduct:** Yes



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 01:27 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=7wKia9zty0)

**Rebuttal:**

## Rebuttals (Part I)

Thank you for your positive feedback and insightful comments. We greatly appreciate your recognition of the strengths of our paper, including its articulate and detailed presentation, rigorous discussion of relevant literature, insightful extension of DDPM, and diverse experiments! We provide responses to address each of your questions below.

### (W1) Comprehensive Study Comparing DiLED with Latent Diffusion Models

Thanks for the great comment!

- First, we've indeed compared with latent diffusion models (LDMs) on **images**, as shown in **Section 4.2 (Figure.3)**. Our approach achieves strong improvement over LDMs in image generation, reconstruction, and interpolation on all three datasets.

As discussed in both Line.76 (right column) and Line.240, LDMs learn a VAE (with a very small weight for the KL regularization) and diffusion model *separately*. That is, LDM's latent space is purely from VAE and shares the same limitations. For instance, LDM's latent vector tends to preserve the *spatial* information of images while lacking high-level *semantic* information (as also discussed in (Preechakul et al., 2022)). This limits the latent-space interpolation capability (as shown in Figure.3). In contrast, our DiLED with unified learning obtains more semantically-meaningful representation.

We have included more detailed discussion of DiLED vs LDMs in **Appendix D.1** (page 29).

- Second, we do not compare with LDMs on **text** because, to our best knowledge, there is not yet a successful application of LDMs on text data. The reason is again because LDMs rely on VAEs to learn a latent space, yet it's very difficult to train VAEs on text (Line.91). We've instead compared with more advanced models (**Table.1**), including *Optimus-DAE* and *LatentOps* (e.g., *LatentOps* is a combined VAE+ODE model). These models have shown to perform better than VAEs (and thus LDMs) on text. Our DiLED outperforms all these baselines across diverse metrics.

We'll make these clearer in the revised version.



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 01:28 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=EDh5VnW2rt)

**Rebuttal:**

## Rebuttals (Part II)

### (W2, Q2) Computational Cost, Pretrained Encoder/Decoder

DiLED is computationally efficient as with latent diffusion and VAEs, thanks to several factors:

- the use of pretrained encoder/decoder for initialization (Line.94)

- low-dimensional latent space
- simple diffusion network architectures (MLPs, as described in Lines.871, 980, 1092) on latent space

These factors accelerate convergence of the unified training (similar to latent diffusion). More specifically,

- In **text** experiments, our encoder and decoder are initialized from pretrained *BERT-small* and *GPT2-xl*. We train only a small portion of the pretrained models (Line.876) and freeze other parameters to accelerate the training. The latent dimension is 128.
- In **image** experiments, we initialize our model using pretrained *DiffAE* weights. The latent dimension is 512.
- In **protein** experiments, the autoencoder is initialized from pretrained *ReLSO*, with a latent dimension of 30.

Moreover, we measured the training cost of DiLED on text data. The results shown below validate that **training DiLED is as efficient as training VAEs and variants (e.g., DAAE, LatentOps)** with the same encoder-decoder architecture (Lines.267-271 and Lines.890-893). More specifically,

Model	Time / Epoch	Overall Training Cost
LatentOps	13.98 mins	6.98 hrs
Optimus-DAAE	14.49 mins	7.02 hrs
DiLED	19.35 mins	7.41 hrs

Due to the added diffusion process, the training time of DiLED for each epoch is  $\sim 1.3x$  that of LatentOps and Optimus-DAAE. However, thanks to the stable training process inherited from the well-established DDPM formulation (Line.145), our DiLED does not require the various training tricks necessary in VAEs and variants, such as beta-annealing [1], free bits [2], and cyclic annealing schedule [3]. This allows DiLED to converge in fewer steps, leading to similar overall training cost.

We'll include these and more results/discussion in the revised version.



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 01:29 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=LxWEGUQrxe)

Rebuttal:

## Rebuttals (Part III)

### (Q1) Potential Benefits of Larger $t$ :

When we have  $n > 1$ , i.e., multiple learnable diffusion steps, we obtain a series of hierarchical latent spaces  $\mathcal{X}_t$  where  $0 \leq t \leq n$ . As discussed in Line.234, when  $n$  is set to the total number of diffusion steps, the model in effect arrives at a *hierachical VAE* (similar to NVAE and HVAE).

As a result, the latent spaces with smaller  $t$  will be closer to the original data space, capturing more low-level information and preserving fine-grained details. The latent spaces with larger  $t$  will be more abstract, capturing higher-level features and global structures. Such a hierachical structure allows the model to progressively refine the latent representations with enhanced robustness as  $t$  increases, provides flexibility in representation at different abstract levels, and potentially improves the model's generalizability.

On the other hand, compared to traditional (*hierachical*) VAE (e.g., NVAE, HVAE), DiLED is compatible to the well-established DDPM training recipe, and thus allows for more effective stable training.

## Summary

We hope our responses have addressed your concerns and provided a clearer understanding of our work. We greatly appreciate your thorough review and valuable feedback, which have helped us improve the clarity and completeness of our paper.

If you have any further questions, please don't hesitate to let us know. We welcome the opportunity to discuss more details and ensure all your concerns are fully addressed.

## Reference

[1] Bowman, Samuel R., et al. "Generating sentences from a continuous space." arXiv preprint arXiv:1511.06349 (2015).

[2] Kingma, Durk P., et al. "Improved variational inference with inverse autoregressive flow." Advances in neural information processing systems 29 (2016).

[3] Fu, Hao, et al. "Cyclical annealing schedule: A simple approach to mitigating kl vanishing." arXiv preprint arXiv:1903.10145 (2019).



➔ *Replying to Rebuttal by Authors*

## Post rebuttal

Official Comment ✎ Reviewer LRww 📅 01 Apr 2024, 12:42

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

### Comment:

Thank you for your comments. I have reviewed the other comments and your rebuttals. I will maintain my original score.



➔ *Replying to Post rebuttal*

## Official Comment by Authors

Official Comment

✎ Authors (👁 Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

📅 01 Apr 2024, 23:28 👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

### Comment:

Thanks again for your insightful comments and valuable time!



## Official Review of Submission144 by Reviewer mAc2

Official Review ✎ Reviewer mAc2 📅 14 Mar 2024, 17:58 (modified: 21 Mar 2024, 05:11)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer mAc2

📄 Revisions (/revisions?id=F2j3YzyHkp)

### Summary:

The paper proposes an end-to-end training procedure for training an autoencoder model together with a diffusion process. The main goal of the paper is to achieve good performance on generation, reconstruction, and manipulation while being able to model effectively both continuous and discrete data representations. The model is tested on image, text, and protein sequences tasks, and compared to other state-of-the-art models.

### Strengths And Weaknesses:

While I think that the contribution is in principle important and promising, the presentation is confusing and dispersive. There is a lot of content and often the descriptions are only superficial. In my opinion, the paper should include fewer baselines and fewer experiments, but with a more detailed description and evaluation of the proposed method and selected baselines. To be more specific:

- The main drawback of the paper is the lack of a detailed comparison with the most closely related models, Latent Diffusion Models, Latent Score-based Generative Model (LSGM acronym is used throughout the paper but is only defined in the appendix) and Diffusion Autoencoders. A brief comparison is provided in Appendix D, but I think it should instead be a core part of the main text. I think the aforementioned methods should be clearly outlined and explained, and the differences with the proposed method should be highlighted.
- The training procedure should be more clearly described. When training the autoencoder and diffusion end-to-end with the procedure from DDPM, I would expect a very slow convergence if the time step  $t$  is sampled uniformly, due to unfrequent updates of the decoder, which from my understanding happens only when the first timesteps are sampled. Furthermore, the training of the autoencoder would slowly change the initial conditions of the forward and backward diffusion process, and I would like to see how that affects the training of the diffusion network.
- Obtaining training samples at time  $t$  now cannot be done in close form, as the first step of the diffusion is the encoder. Maybe it can still be done smartly, while still backpropagating the signal through the encoder, but this is not discussed in the paper.
- The code for the implementation and experiments is not provided, which makes it hard to verify the training algorithm and the evaluation pipeline.
- Figure 3 compares three different tasks at the same time, for models not necessarily designed for such tasks. I would find it more relevant to add numerical results for the different tasks in different tables, and perhaps only on a subset of the datasets and baselines. In that way, I could clearly understand which model is better for which tasks, and conclude which model is a better choice if I want a unique model to perform well on all the tasks.
- The numerical results reported in Table 6 seem misleading: the criteria for bold entries are not explained, and sometimes values on the DiLED row are highlighted even if they are not the lowest. Furthermore, sometimes the reported numbers don't match with the ones from the original paper. For example, the generation FID for Consistency Distillation on Bedroom is 5.22 for 2-step generation, but here the reported result is 7.01. The authors mention some differences in processing, but I would like to verify what is done in the code.

#### Questions:

- Point 3 in the introduction states that "The flexibility of the generalized diffusion formulation allows us to specify any desired encoder-decoder for modelling both discrete and continuous data." How does this differ from the other models using Autoencoder + diffusion?

#### Limitations:

The authors do not discuss the limitations of their approach. I think a section regarding which problems DiLED cannot handle well would be insightful for practitioners.

**Ethics Flag:** No

**Soundness:** 2: fair

**Presentation:** 1: poor

**Contribution:** 2: fair

**Rating:** 3: Reject: For instance, a paper with technical flaws, weak evaluation, inadequate reproducibility and incompletely addressed ethical considerations.

**Confidence:** 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

**Code Of Conduct:** Yes




### Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 01:31 (modified: 29 Mar 2024, 05:26)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

 Revisions (/revisions?id=9vqoqGgijD)

**Rebuttal:**

## Rebuttals (Part I)

Thank you for your valuable review and feedback. We sincerely appreciate your recognition of the importance and potential of our proposed end-to-end training procedure. We would like to address your concerns and provide some clarifications:

### (W1) Detailed Comparison with Related Work

Thanks for the comment. In **Section.1, Section.3.3, and Appendix.D**, we have discussed the differences between our method and related works from multiple perspectives (e.g., training processes, latent space patterns, pros/cons). In the experiments (**Section.4.2**), we have also empirically compared the performance and showed our advantages. We'll move relevant content from Appendix to the main text and add more discussion as suggested.



Here we reiterate some key differences/advantages of our approach:


- **Comparison with Latent Diffusion Models (LDMs)**[1]: LDMs use a two-stage training process, first training a VAE with strong reconstruction ability (using a small KL weight, e.g.,  $10^{-6}$ , Line.1549) but limited generation capability, followed by training a diffusion model in the latent space to enable generation. While LDMs perform well in generation, they lack a meaningful representation space (Line.1547), as evidenced by the interpolation results in **Figure 3** (green bar for LDM) and the  $\alpha = 0.4$  rows in **Figures.16 and 17** (Lines.1306 and 1347). The unified training of our DiLED resolves the difficulties.
- **Comparison with Latent Score-based Generative Model (LSGM)**[2]: LSGM can be viewed as a VAE with a Score-based Generative Model (SGM) as the prior. Unlike our approach, which derives the objective from a generalized diffusion model perspective without further assumptions, LSGM derives the objective from a VAE perspective, decomposing the KL term into two components and *approximating* each component with various additional techniques/tricks. Additionally, LSGM's training procedure (Line.1580) involves independent updates of VAE and SGM parameters, which could lead to unstable training (e.g., as discussed in the Github issue (<https://github.com/NVlabs/LSGM?tab=readme-ov-file#common-issues>) where the authors acknowledged). In contrast, our approach is fully compatible with the well-established DDPM training recipe (Line.145) and allows effective stable training.



## Rebuttal by Authors

Rebuttal

 Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

 29 Mar 2024, 01:37 (modified: 29 Mar 2024, 05:26)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

 Revisions (/revisions?id=zI4ExC7Djn)

**Rebuttal:**

## Rebuttals (Part II)

### (W1, continue from Part I)

- **Comparison with Diffusion Autoencoder (DiffAE)**[3]: DiffAE is an autoencoder with a diffusion model as the decoder, exhibiting good reconstruction and representation abilities. Similar to LDMs, DiffAE separately trains a diffusion model in the latent space after training the autoencoder. In contrast, our approach trains the diffusion model in a unified way with the encoder-decoder as learnable diffusion steps, resulting in improved latent representation and overall performance, as demonstrated in **Figure.4**.

Moreover, none of the LDMS, LSGM, and DiffAE has shown successful applications on **discrete text and protein** data, due to their limitations of unstable training and/or model formulations. For example, LDMS rely on VAEs to learn a latent space, yet it's very difficult to train VAEs on text (Line.91). We've instead compared with more advanced models (**Table.1**), including *Optimus-DAAE* and *LatentOps*, which have shown to perform better than VAEs (and thus LDMS) on text. Our DiLED outperforms all these baselines across diverse metrics.

We'll make the above clearer in the revised version.



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 01:42 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=f1Km2l7hi2)

Rebuttal:

## Rebuttals (Part III)

### (W2) Training Procedure, Convergence Speed, and Initial Condition of Diffusion Model

- **Training Procedure** We have outlined the complete training procedure in **Algorithm 1** on page.16. This algorithm describes the steps involved in a single training iteration, demonstrating how the encoder-decoder parameters and diffusion parameters are trained in a unified way.
- **Convergence Speed**

Thanks for the question. We used a simple way to enable frequent update of encoder/decoder and fast convergence, as described in **Eqs.30-33** on page.15.

More specifically, Eq.30 is the vanilla final objective, where the encoder/decoder would indeed be updated with low frequency, potentially leading to slow convergence. To address this, we have split the reconstruction loss term  $L_{rec}$  into  $T$  equal parts and distributed them among the other loss terms, as shown in Eq.32. This modification ensures that **the encoder/decoder is updated at every training step, regardless of the value of  $t$  that is sampled.**

As a result, the training convergence of our method is indeed comparable to that of a VAE (or DAAE) model which updates encoder/decoder frequently at each training step. Specifically, we measured the training cost of our DiLED on text data. The results shown below validate that **training DiLED is as efficient as training VAEs and variants (e.g., DAAE, LatentOps)** with the same encoder-decoder architecture (Lines.267-271 and Lines.890-893).

Model	Time / Epoch	Overall Training Cost
LatentOps	13.98 mins	6.98 hrs
Optimus-DAAE	14.49 mins	7.02 hrs
DiLED	19.35 mins	7.41 hrs



Due to the added diffusion process, the training time of DiLED for each epoch is  $\sim 1.3x$  that of LatentOps and Optimus-DAAE. However, thanks to the stable training process inherited from the well-established DDPM formulation (Line.145), our DiLED does not require the various training tricks necessary in VAEs and variants, such as beta-annealing [1], free bits [2], and cyclic annealing schedule [3]. This allows DiLED to converge in fewer steps, leading to **similar overall training cost.**




## Rebuttal by Authors




Rebuttal

 Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

 29 Mar 2024, 01:44 (modified: 29 Mar 2024, 05:26)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

 Revisions (/revisions?id=g89YZPS3qS)

Rebuttal:

## Rebuttals (Part IV)

### (W2, continue from Part III)

- **Initial Condition Changed** As one of the key advantages of our work, the proposed method uses a **unified training procedure** that updates the encoder/decoder parameters and diffusion parameters jointly with the single DDPM-like objective. This is a key distinction from previous approaches, such as LDMs that use a two-stage training process, or LSGM that alternates between training the autoencoder and the latent model. Just like training a common deep neural network where different neural layers are trained jointly with a single objective, training the DiLED parameters at different diffusion steps (including encoder/decoder) jointly allows all steps to co-evolve and co-adapt to each other to obtain a consistent latent space.

We will make the above points clearer in the main paper.

### (W3) Obtain Training Sample $\mathbf{x}_t$

As explained in **Section 3.1** (Lines.166-175) and illustrated in **Algorithm 1** (Lines.845-847), the only difference compared to the standard diffusion model lies in the first learnable step of the forward process. **Obtaining training samples at time  $t$  can be done in a straightforward manner, simply by applying encoding on the input  $\mathbf{x}_0$  and then deriving  $\mathbf{x}_t$  as usual.**

More specifically, to obtain a sample  $\mathbf{x}_t$  at an arbitrary timestep  $t$ , we first apply the encoding step to the input data  $\mathbf{x}_0$  to obtain the latent representation  $\mathbf{x}_1$ . We then follow the standard diffusion forward process, as described in the **Eq.1** (Line 127), to **obtain  $\mathbf{x}_t$  from  $\mathbf{x}_1$  in a closed form**. Thus, compared to the conventional diffusion process, the only additional step in our generalized diffusion process is the initial encoding step, which is straightforward and keeps it **easy to obtain samples  $\mathbf{x}_t$** .



### (W4) Implementation Code and Experiments


We have put our code at <https://anonymous.4open.science/r/DiLED> (<https://anonymous.4open.science/r/DiLED>). We will continue to clean up and release all code, data, and experiments upon acceptance.



## Rebuttal by Authors

Rebuttal

 Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

 29 Mar 2024, 01:49 (modified: 29 Mar 2024, 05:26)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

 Revisions (/revisions?id=9v8xksnHy7)

Rebuttal:

## Rebuttals (Part V)

### (W5) Numerical Results for Different Tasks

The detailed numerical results for different tasks in Figure.3 are provided in **Table.6** (Appendix C.2.2). The results for generation, reconstruction, and interpolation are listed in four columns, respectively, where each column has the comparisons of different models.

We'd like highlight that Figure.3 demonstrates the key advantage of our approach---a unique model performs well across all the tasks.

## (W6) Details in Table 6

**Criteria for bold entries:** We used two criteria for bold entries:

- The  $\{DDIM, DiffAE, DiLED\}$  models allow a configurable  $T$  ( $= 10, 20, \text{ or } 50$ ). We highlight the best results under different  $T$ , respectively. For example, for "Reconstruction rFID" on the "FFHQ 128" dataset, DiLED performs best under different  $T$  values, respectively. The three results by DiLED are thus highlighted.
- For other models that do not have a configurable  $T$ , including  $\{LDM, NVAE, StyleGAN-XL, CM-CD\}$ , we highlight their results if they surpass the best results of  $\{DDIM, DiffAE, DiLED\}$  at all  $T$ . Again, using "Reconstruction rFID" on the "FFHQ 128" dataset for example: LDM performs better than DiLED ( $T = 50$ ), thus its result is highlighted.

There is a typo by highlighting 9.58 (LDM's "reconstruction rFID" on on "CelebA64" dataset). We will fix the typo and rearrange the table to make the results clearer.



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 01:52 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=eZA0St7d36)

**Rebuttal:**

## Rebuttal (Part VI)

### (W6, continue from Part V)

**Results of Consistency Model (CM):** The original paper reports FID 5.22 while we obtained 7.01. Here we clarify the reasons:

1. As mentioned in Line.1016, we converted the 256x256 images by the pretrained CM to 128x128 for evaluating FID and comparing with other models.
2. When evaluating FID, we used the library here (<https://github.com/mseitzer/pytorch-fid>) for all the FID evaluations to ensure fair comparison. We found the evaluation script is slightly different from that used in the CM paper (link (<https://github.com/openai/guided-diffusion/tree/main/evaluations>)). Specifically, the reference mean/sigma of the Bedroom data used in the two scripts are slightly different, which could influence the resulting FID a bit. We give more details below. Note that even with the difference in FID, the **conclusions made in our paper is not impacted** (i.e., CM is good at generation, but not interpolation due to lack of a good representation space).
3. We'd like to note that, with the evaluation script we used, we were able to reproduce all results from the original DiffAE paper. This verifies that our evaluation is reliable.

Below, we provide more details from our investigation of the evaluation. In particular, we conduct a series of ablation experiments with CM's outputs in different evaluation settings, and show all results in the table below. Specifically:

- Using exact the same inference code and evaluation script from the CM's official repo (link ([https://github.com/openai/consistency\\_models](https://github.com/openai/consistency_models))), we obtained FID 5.82 (256x256 images), which is fairly close to the result 5.22 reported in the original CM paper.
- Next, on the same CM's output images, with the evaluation script used in our paper (and DiffAE), we obtain FID 6.56. This is slightly higher than the above FIDs. As mentioned above, by looking into the evaluation scripts, we found the reference mean/sigma are slightly different: the CM's evaluation script used *pre-*

*computed* reference mean/sigma, while the evaluation script `we/DiffAE` used computes the reference mean/sigma on the fly from the dataset.

- We then downscale the image resolution to 128x128. Using the same code as in (2) to calculate the FID between the down-scaled images and the down-scaled ground-truth images, we obtain FID 7.01, namely the results reported in our paper.



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 01:55 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=Ud2qPrmU3d)

**Rebuttal:**

## Rebuttal (Part VII)

(W6, continue from Part VI)

Evaluation Settings	CM Generation FID
256x256 with pre-computed statistics from link ( <a href="https://github.com/openai/guided-diffusion/tree/main/evaluations">https://github.com/openai/guided-diffusion/tree/main/evaluations</a> )	5.82
256x256 with reference statistics calculated on the fly	6.56
128x128 with reference statistics calculated on the fly	7.01

All our **evaluation code** is provided in this anonymous github repo link (<https://anonymous.4open.science/r/DiLED>). In this repo, `cm_sample.py` can be used for sampling 128x128 images, while `cm_sample_256.py` can be used for sampling 256x256 images. With the images saved to the folder `samples/consistency_model_twosteps/samples`, one can run the code

```
python fid.py "samples/consistency_model_twosteps/samples_lpips_256/*.png" "samples/bedroom/original_imgs_256/*.png"
```

to evaluate the FID, which would give the results 6.56; And the following command would yield the FID of images of resolution 128x128 (i.e. 7.01)

```
python fid.py "samples/consistency_model_twosteps/samples_lpips/*.png" "samples/bedroom/original_imgs/*.png"
```

Meanwhile, we follow link (<https://github.com/openai/guided-diffusion/tree/main/evaluations>), download the npz file ([https://openaipublic.blob.core.windows.net/diffusion/jul-2021/ref\\_batches/lsun/bedroom/VIRTUAL\\_lsun\\_bedroom256.npz](https://openaipublic.blob.core.windows.net/diffusion/jul-2021/ref_batches/lsun/bedroom/VIRTUAL_lsun_bedroom256.npz)) and convert all the images generated into another npz file using the script `convert_png_to_npz.py` :

```
python convert_png_to_npz.py samples/consistency_model_twosteps/samples_lpips_256 samples/consistency_model_twosteps/samples_lpips_256.npz
```

Then we can run the evaluation code using this script (<https://github.com/openai/guided-diffusion/blob/main/evaluations/evaluator.py>):

```
python evaluator.py VIRTUAL_lsun_bedroom256.npz samples/consistency_model_twosteps/samples_lpips_256.npz
```

which yields the FID 5.82.

## (Q1) Difference from Autoencoder + diffusion

We'd like to first clarify that Point 3 in the introduction is meant to highlight the advantage of our method vs *standard diffusion* and *GANs*, as stated in the second sentence in Point 3: "*DiLED thus overcomes the difficulty of standard diffusion and GANs on text and other discrete modalities.*"



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 01:57 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=Y3a6Vz4CkI)

**Rebuttal:**

## Rebuttal (Part VIII)

### (Q1, continue from Part VII)

Regarding autoencoder+diffusion, such as latent diffusion models (LDMs), we've highlighted our differences in multiple places, including **Point 1** in introduction, **Lines.240-256** in Section 3.3, experiments in **Section 4.2**, and discussion in **Appendix D**. To reiterate here briefly:

- Our DiLED learns a better latent representation space thanks to the new *unified* learning paradigm
- To our best knowledge, no/few prior studies have successfully applied existing autoencoder+diffusion models (e.g., LDMs, DiffAE) on discrete text data. This is because obtaining a good text (variational) autoencoder is very challenging (Line.91), making it infeasible to further add the diffusion component. Our unified training alleviates the challenges and achieves strong performance on text modeling. Note that previous text diffusion models (e.g., Li et al., 2022) do not use autoencoder components, which lack compact representation and get inferior performance than our method.

## Limitations

We'll add more discussion on the limitations of the approach. For example, our approach has not been verified on other data modalities such as videos, audios, and time series. It'll be interesting to see how the approach will perform on all the diverse forms of data, and even multi-modal data (such as text+images). In addition, we have only parameterized the first diffusion step with learnable encoding/decoding. A single latent vector representation could be limited to capture relevant information of data for different tasks. We'd like to investigate our approach with more encoding/decoding-based diffusion steps, to obtain hierarchical latent representations.

## Summary

We hope our responses have addressed your concerns and provided a clearer understanding of our work. If you have any further questions, please don't hesitate to let us know. We welcome the opportunity to discuss more details and ensure all your concerns are fully addressed.




## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 01:58 (modified: 29 Mar 2024, 05:26)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

 Revisions (/revisions?id=oe42Yn1tE8)

**Rebuttal:**

## Reference


[1]Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[2]Vahdat, Arash, Karsten Kreis, and Jan Kautz. "Score-based generative modeling in latent space." Advances in neural information processing systems 34 (2021): 11287-11302.

[3]Preechakul, Konpat, et al. "Diffusion autoencoders: Toward a meaningful and decodable representation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.



[4]Liu, Guangyi, et al. "Composable text controls in latent space with odes." arXiv preprint arXiv:2208.00638 (2022).



 *Replying to Rebuttal by Authors*

## Official Comment by Authors

Official Comment

 Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

 01 Apr 2024, 23:39  Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

**Comment:**

Dear *Reviewer mAc2*,

We wanted to reach out and express our sincere gratitude for the time and effort you have dedicated to reviewing our paper.

If there are any outstanding questions or concerns that you would like us to address, **please don't hesitate to let us know**. We are more than happy to provide further clarification or engage in a more detailed discussion to ensure that all of your points are addressed.



Thank you once again for your contribution to the peer-review process. We look forward to hearing from you.

Best regards,

The Authors



## Official Comment by Reviewer mAc2

Official Comment  Reviewer mAc2  03 Apr 2024, 08:28

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer mAc2

**Comment:**

I thank the authors for answering my questions and my concerns. While it is now more clear what the strengths of DiLED are, I still find the quality of the presentation insufficient for a top machine learning conference. Reading also the answers to the other reviewers, I see a lot of instances where the authors will make changes in the revised version, but there are so many that I think the paper would look substantially different from its current version.

Furthermore, I appreciate that the authors have shared the code for the paper, but is not in a final clean state yet, and it is impossible for me to go and verify the implementation within the short discussion period. Ideally, I think that the code should be already made available at the review stage, to give the reviewers the time to carefully check details if needed.

Overall, I am still of the opinion that the paper needs to improve in presentation and clarity, to convey the contributions more effectively.



## Follow-up response

Official Comment

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

04 Apr 2024, 00:12 (modified: 04 Apr 2024, 00:14)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=VumPe40TI6)

### Comment:

Thanks for the response!

- **"I still find the quality of the presentation insufficient for a top machine learning conference."**  
We'd appreciate if the reviewer could point out specific unclear aspects of the presentation after our initial response, so that we can clarify them.
- **"the paper would look substantially different from its current version"**  
We respectfully disagree. All main points, contributions, methods, experiments, results, and conclusions are indeed appreciated by all other reviewers and should be kept unchanged. We are happy to include more discussions about related works and computational efficiency as suggested by reviewers. This would enhance the paper but do not lead to "substantial difference".
- **Code**  
Again, we'd appreciate if the reviewer could point out specific unclear aspects of the code, so we can clarify them.



## Official Review of Submission144 by Reviewer 5fm8

Official Review Reviewer 5fm8 13 Mar 2024, 20:59 (modified: 04 Apr 2024, 00:21)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 5fm8

Revisions (/revisions?id=oq01JdD8fl)

### Summary:

Based on the VAE interpretation of diffusion models, the authors propose to use a parameterized VAE-like encoding-decoding instead of the vanilla Gaussian noising-denoising in the first layer of a diffusion model, resulting in the presented generalized diffusion with learnable encoder-decoder (DILED). The authors also mentioned that "DILED can be viewed as a VAE with a learned diffusion model prior." Extensive experiments on text, proteins, and images are conducted to demonstrate DILED's effectiveness.

### Strengths And Weaknesses:

Strengths:

- (1) The paper is well written and easy to follow.
- (2) The technique presented is sound and should be easy to use.
- (3) Experiments are diverse and abundant.

Weaknesses:

- (1) Important discussions with related work are missing.
- (2) Theoretical novelty is somewhat limited, as parameterized noising processes have been widely studied.

(3) Some experimental results are inconsistent. The code is not available for evaluation.

#### Questions:

(1) Parameterized noising processes have been investigated in [1]. So what's the main theoretical novelty of the proposed DILED?

(2) Some statements are not consistent with existing research on VAEs. For example, the two-stage VAE and many follow-up studies have shown that training the reconstruction and the prior separately improves generation performance over end-to-end joint training; why does the proposed DILED work differently?

(3) The beta-VAE shows that balancing the reconstruction term with the prior-related KL term is important for VAE. The hyperparameter  $w$  plays the same role for DILED. What's its influence on generation, reconstruction, and representation performance?

(4) Can DILED overcome the "posterior collapse" challenge on text sequences? How?

(5) The generation diversity results in Figure 1b and Table 1 are not consistent.

(6) It may not be fair to compare DILED with models with prompting, e.g., GPT-4, via generation distribution matching. It's challenging to ensure that the input prompts obey the training data distribution.

[1] Bartosh, Grigory, Dmitry Vetrov, and Christian A. Naesseth. "Neural Diffusion Models." arXiv preprint arXiv:2310.08337 (2023).

#### Limitations:

Yes.

**Ethics Flag:** No

**Soundness:** 3: good

**Presentation:** 4: excellent

**Contribution:** 3: good

**Rating:** 6: Weak Accept: Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.

**Confidence:** 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.

**Code Of Conduct:** Yes



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 02:01 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=SFf3ztO9Zw)

**Rebuttal:**

## Rebuttal (Part I)

Thank you for your positive feedback and constructive comments. We are pleased that you found our paper well-written and easy to follow, and that you recognize the soundness and ease of use of our proposed technique, DILED. We also appreciate your acknowledgment of the diversity and abundance of our experiments.

We would like to address the concerns and questions you mentioned and provide some clarifications:

### (W1) Missing Discussions with related works

We have discussed the related works in multiple places:

1. In **Section 3.3 (Connections with Other Generative Models)**, we have provided a high-level comparison of DILED with many related models, highlighting the connections and differences between our approach and various existing methods.

- Section 5 (Related Work)** presents additional discussion of related works, focusing on recent efforts to combine VAEs, GANs, and diffusion models. We have also mentioned the limitations of these approaches and how DiLED aims to overcome them.
- Appendix D** offers in-depth discussion of DiLED with two closely related methods: Latent Diffusion Models (LDMs, [2]) and Latent Score-based Generative Models (LSGM, [3]). We have analyzed the differences in terms of the autoencoder's functionality, training paradigm, and overall objectives.

We will make the discussions more clearly in the revised version.

## (W2, Q1) Difference with *Neural Diffusion Models*

The Neural Diffusion Models (NDM) [1] extends the forward process of standard diffusion models with learnable *time-dependent* Gaussian transformations, where  $\mathbf{z}_t$  from the forward process is sampled from  $\mathcal{N}(\text{tf}{z}; \alpha_t \theta(\text{tf}{x}, t), \sigma_t^2)$  for  $t=1, \dots, T$ . While NDM shares some similarities with our approach, there are several key differences that highlight the novelty and advantages of our DiLED:

- NDM redefines the entire forward process of standard diffusion by introducing *time-dependent* transformation  $F_\theta(\mathbf{x}, t)$ . Crucially, the time-dependent architecture makes NDM largely incompatible with **pretrained encoder/decoder** and have to train all parameters from scratch. In contrast, it's an important advantage of our DiLED to use pretrained encoder/decoder as initialization, which greatly improves performance and speeds up training convergence (Line.94).



➔ *Replying to Rebuttal by Authors*

## Rebuttal by Authors

Rebuttal

✍ Authors (👤 Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

📅 29 Mar 2024, 02:03 (modified: 29 Mar 2024, 05:26)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

📄 Revisions (/revisions?id=I5DqRLNgzp)

Rebuttal:

## Rebuttal (Part II)

### (W2, Q1, continue from Part I)

- In NDM, the forward process is formulated by a learnable time-dependent model, using the same architecture as the reverse process of the diffusion model which leads to doubling of the training parameters. Furthermore, NDM's training must utilize the **full vanilla objective** function (Eq.8 in [1]). In contrast, **DiLED uses a simpler objective inherited from DDPM**, which has been proven to be more efficient and leads to better sample quality. As a result, NDM requires 2.3x longer training than DDPM (Section.6 in [1]) yet achieves only marginal improvements over the baselines.
- Moreover, as above, NDM introduces extra training cost over standard diffusion. In contrast, DiLED shares the same advantage of latent diffusion models that **accelerates training** by mapping raw data into a low-dimensional latent space for diffusion modeling.
- Lastly, in terms of empirical performance, NDM shows only comparable or slightly better performance with vanilla DDPM (Table.2 in [1]). In contrast, DiLED outperforms diffusion and various stronger baselines in diverse domains/tasks with a large margin. DiLED shows applicability and strong performance on not only images but also text/proteins that have been difficult for diffusion modeling.

We'll add more discussion in the revised version.

## (Q2) Difference with Two-stage VAE

The two-stage VAE [4] demonstrated that training the reconstruction (first stage) and prior matching (second stage) separately can improve generation performance compared to joint end-to-end training of a standard VAE. The first stage focuses on reconstruction ability by setting the KL weight close to zero, while the second stage



uses another VAE to learn the data manifold in the first-stage VAE latent space to achieve good generation ability. Through this method, they can obtain a model with both good generation and reconstruction capabilities. The **two-stage approach is indeed similar to that of *latent diffusion models (LDMs)* and *LatentOps* which we've discussed and/or empirically compared extensively in our paper.**



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 02:04 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=s8V2cv8dID)

Rebuttal:

### Rebuttal (Part III)

#### (Q2, continue from Part II)

Specifically, a key shortcoming of the two-stage approaches (e.g., two-stage VAE, LDMs) is that the latent space is learned purely by the first-stage VAE. The latent space (due to small KL weight) is close to the latent space by a vanilla autoencoder, which is non-smooth with many "holes" [5]. The weak *interpolation* performance of LDMs demonstrates the limitation (LDM's Green Bar in **Figure.3**, Line.1306 in **Figure.16**, and Line.1347 in **Figure.17**).

In contrast, the unified learning of DiLED addresses the limitation. The experimental results in our paper show that DiLED's unified training avoids the generation-reconstruction tradeoff issues of the standard VAE, and leads to a more smooth and semantically meaningful latent space. Specifically, we compare DiLED with *LatentOps* (as a first-stage VAE) in **Table.1** on text, and with *LDMs* (as a two-stage VAE+Diffusion model) in **Figure.3**, **Figure.16**, and **Figure.17** on images.

We will add more discussion in the revised version.

#### (Q3) Influence of Weight $w$

Thanks for the question! In our experiments, we found the model is not very sensitive to  $w$  (compared to beta-VAE to its balancing  $\beta$ ). As mentioned in **Line.188**, we used  $w = 8$  for text and  $w = 1$  for both image and protein. **The choice of  $w$  is primarily influenced by the different loss functions used for the reconstruction term  $L_{Rec}$  (Eq.4) across various data types.** For example, cross-entropy is used for text, while mean squared error (MSE) is used for images. To ensure that the different loss terms have similar scales, we selected  $w$  values that roughly balanced their magnitudes. The  $w$  values are then fixed during training. As a comparison, beta-VAE on text typically requires careful scheduling/annealing of  $\beta$  values during training (Li et al., 2020), making the training more difficult and unstable.

We'll add the discussion in the revised version.



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 02:08 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=73kne68eAa)

Rebuttal:

### Rebuttals (Part IV)

## (Q4) Posterior Collapse

DiLED can alleviate the "posterior collapse" issue on text sequences thanks to the improved formulation and training:

1. As discussed in **Section 3.3**, DiLED can be viewed as a VAE with a jointly learned diffusion model prior (instead of a standard Gaussian prior in vanilla VAEs). More formally, the DiLED objective in **Eq.(7)** includes the  $L_{\text{align}}$  term that explicitly encourages alignment between the encoder's posterior distribution  $q_{\lambda}(\mathbf{x}_1|\mathbf{x}_0)$  and the diffusion model's "prior"  $p_{\theta}(\mathbf{x}_1|\mathbf{x}_2)$ . This learned prior is more flexible and helps avoid the unreasonable regularization posed by standard Gaussian prior that causes posterior collapse.
2. DiLED derives the training objective from the well-established DDPM framework, enabling stable and effective joint training of the encoder, decoder, and diffusion components. For example, as mentioned above, DiLED is robust to the balancing weight  $w$ . This is in contrast to the instability and sensitivity to hyperparameters in text-VAE training. To alleviate posterior collapse, training text-VAEs often requires tricks like beta-annealing [5], free bits [6], cyclic annealing schedule [7], and so on. This can be seen through the training loss curves ([https://anonymous.4open.science/r/DiLED/asset/loss\\_curve.png](https://anonymous.4open.science/r/DiLED/asset/loss_curve.png)) (LatentOps is a SOTA text-VAE).

We'll make the discussion clearer in the revised version.

## (Q5,W3) Inconsistency of Results, Implementation Code

**Inconsistency of results:** Thanks for the catch! There is a typo in Table.1: the correct diversity scores for GPT2 and GPT4 should be 0.65 and 0.87, respectively, which are consistent with Figure.1b. We will fix the typos in the revised version!

**Implementation code:** We have put our code at <https://anonymous.4open.science/r/DiLED> (<https://anonymous.4open.science/r/DiLED>). We wil continue to clean up and release all code, data, and experiments upon acceptance.



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 02:09 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=MatQgoLu7x)

**Rebuttal:**

## Rebuttals (Part V)

### (Q6) Comparison with GPT4

We agree that it's not a fair/rigorous comparison with GPT4 prompting. We included GPT4 just as a reference to provide a clue of how an LLM could perform. Our intention was to offer a broader context for understanding the effectiveness of DiLED.

On the other hand, this also indicates a limitation of LLMs for generating text from a desired distribution. That is, with only prompting, it's hard to control the output text distribution. The relatively lightweight DiLED (and smaller LMs like GPT2) allows quick finetuning for the task.

## Summary

We hope our responses have addressed your concerns and provided a clearer understanding of our work. We greatly appreciate your thorough review and valuable feedback, which have helped us improve the clarity and completeness of our paper.

If you have any further questions, please don't hesitate to let us know. We welcome the opportunity to discuss more details and ensure all your concerns are fully addressed.

## Reference

- [1]Bartosh, Grigory, Dmitry Vetrov, and Christian A. Naesseth. "Neural Diffusion Models." arXiv preprint arXiv:2310.08337 (2023).
- [2]Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
- [3]Vahdat, Arash, Karsten Kreis, and Jan Kautz. "Score-based generative modeling in latent space." Advances in neural information processing systems 34 (2021): 11287-11302.
- [4]Dai, Bin, and David Wipf. "Diagnosing and enhancing VAE models." arXiv preprint arXiv:1903.05789 (2019).
- [5] Bowman, Samuel R., et al. "Generating sentences from a continuous space." arXiv preprint arXiv:1511.06349 (2015).
- [6] Kingma, Durk P., et al. "Improved variational inference with inverse autoregressive flow." Advances in neural information processing systems 29 (2016).
- [7] Fu, Hao, et al. "Cyclical annealing schedule: A simple approach to mitigating kl vanishing." arXiv preprint arXiv:1903.10145 (2019).



➔ *Replying to Rebuttal by Authors*

### Official Comment by Authors

Official Comment

✍️ Authors (👁️ Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

📅 01 Apr 2024, 23:37 👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

#### Comment:

Dear *Reviewer 5fm8*,

We wanted to reach out and express our sincere gratitude for the time and effort you have dedicated to reviewing our paper.

If there are any outstanding questions or concerns that you would like us to address, **please don't hesitate to let us know**. We are more than happy to provide further clarification or engage in a more detailed discussion to ensure that all of your points are addressed.

Thank you once again for your contribution to the peer-review process. We look forward to hearing from you.

Best regards,

The Authors



➔ *Replying to Official Comment by Authors*

### Official Comment by Reviewer 5fm8

Official Comment ✍️ Reviewer 5fm8 📅 04 Apr 2024, 00:20

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

#### Comment:

Thank you for the detailed rebuttal, which has successfully addressed many of my concerns. However, I would strongly encourage the authors to provide theoretical discussions on comparisons with two-stage VAE, beta-VAE, and the treatment of posterior collapse. These are important but not analyzed in depth in the rebuttal. I have increased my score.



## Official Review of Submission144 by Reviewer Si5g

Official Review Reviewer Si5g 12 Mar 2024, 07:54 (modified: 01 Apr 2024, 12:06)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer Si5g

Revisions (/revisions?id=88BIYn2boM)

### Summary:

This paper introduces a novel framework called Generalized Diffusion with Learnable Encoding-Decoding (DILED), which enhances the capabilities of generative models across various data types such as text, images, and proteins. DILED extends traditional diffusion models by incorporating parameterized encoding-decoding mechanisms, allowing for the generation, reconstruction, and learning of compact representations with improved performance and flexibility. The approach is evaluated through extensive experiments demonstrating its effectiveness over existing models.

### Strengths And Weaknesses:

#### Strengths:

1. *Originality*: The introduction of learnable encoding-decoding to the diffusion process is a novel approach that broadens the applicability and effectiveness of generative models.
2. *Quality*: Extensive experiments across different data types and tasks demonstrate the high quality of the generated, reconstructed, and represented data.
3. *Clarity*: The paper is well-structured, with a clear exposition of the methodology and results.
4. *Significance*: DILED addresses key limitations of existing generative models, offering significant improvements in flexibility, performance, and the integration of generation, reconstruction, and representation capabilities.

#### Weaknesses:

1. *Generalization*: While the paper presents a broad range of applications, the generalizability of DILED to other domains or more complex datasets could be further explored.
2. *Complexity*: The complexity of the model, especially when integrating large pre-trained encoders or decoders, may present challenges in terms of computational resources and training time.
3. *Interpretability*: The paper could enhance its contribution by providing more insights into the interpretability of the learned representations, especially in comparison to those from other models.

#### Questions:

1. How does DILED perform in terms of computational efficiency and scalability, especially when using large pre-trained models as encoders or decoders?
2. Can the authors provide more details on the interpretability of the latent space learned by DILED, particularly in comparison with traditional diffusion models or VAEs?
3. What are the limitations in terms of data types and tasks where DILED might not perform as well, and how might these be addressed in future work?

#### Limitations:

The paper does an excellent job of addressing the limitations and potential societal impacts of their work. However, a more detailed discussion on the computational costs and the potential for bias when DILED is applied to sensitive or biased datasets could enhance the paper. Suggestions for mitigating these issues, such as efficiency improvements or bias detection mechanisms, would also be beneficial.

**Ethics Flag:** No

**Soundness:** 3: good

**Presentation:** 3: good

**Contribution:** 3: good

**Rating:** 7: Accept: Technically solid paper, with high impact on at least one sub-area, or moderate-to-high impact on more than one areas, with good-to-excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.

**Confidence:** 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

**Code Of Conduct:** Yes



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 02:13 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=BOgqaAi9mz)

Rebuttal:

### Rebuttal (Part I)

Thank you for your insightful feedback and positive comments. We greatly appreciate your recognition of the strengths of our paper, including (1) the originality of our approach in introducing learnable encoding-decoding to the diffusion process, (2) the high quality of our extensive experiments across various data types and tasks, (3) the clarity and structure of our manuscript, and (4) the significance of DiLED in addressing key limitations of existing generative models while offering improvements in flexibility, performance, and the integration of core capabilities! To address your questions and concerns, we'd like to give responses as below:

#### (Q1, W2) Computational Efficiency and Scalability

We'd like to highlight that it's a key advantage of DiLED to be compatible with pretrained encoders/decoders, as the **pretrained models substantially boost the model performance**. In contrast, for example, most previous text diffusion models are not compatible with pretrained models, leading to much inferior performance (e.g., GENIE in Table.1) and sometimes larger training cost due to training from scratch.

DiLED is computationally efficient thanks to several factors:

- the use of pretrained encoder/decoder for initialization that accelerates convergence (Line.94)
- low-dimensional latent space
- simple diffusion network architectures (MLPs, as described in Lines.871, 980, 1092) on latent space

These factors accelerate convergence of the unified training (similar to latent diffusion). More specifically,

- In **text** experiments, our encoder and decoder are initialized from pretrained *BERT-small* and *GPT2-xl*. We train only a small portion of the pretrained models (Line.876) and freeze other parameters to accelerate the training. The latent dimension is 128.
- In **image** experiments, we initialize our model using pretrained *DiffAE* weights. The latent dimension is 512.
- In **protein** experiments, the autoencoder is initialized from pretrained *ReLSO*, with a latent dimension of 30.



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 02:14 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=eijC8asJ8q)

Rebuttal:

### Rebuttal (Part II)

#### (Q1, W2, continue from Part I)

Moreover, we measured the training cost of DiLED on text data. The results shown below validate that **training DiLED is as efficient as training VAEs and variants (e.g., DAAE, LatentOps)** with the same encoder-decoder architecture (Lines.267-271 and Lines.890-893). More specifically,

Model	Time / Epoch	Overall Training Cost
LatentOps	13.98 mins	6.98 hrs
Optimus-DAAE	14.49 mins	7.02 hrs
DiLED	19.35 mins	7.41 hrs

Due to the added diffusion process, the training time of DiLED for each epoch is  $\sim 1.3x$  that of LatentOps and Optimus-DAAE. However, thanks to the stable training process inherited from the well-established DDPM formulation (Line.145), our DiLED does not require the various training tricks necessary in VAEs and variants, such as beta-annealing [1], free bits [2], and cyclic annealing schedule [3]. This allows DiLED to converge in fewer steps, leading to similar overall training cost.

We'll add more discussion in the revised version.

## (Q2, W3) Interpretability

We've included analysis of the structure and interpretability of learned latent spaces in the following experiments:

- **Protein Representation** (Section 4.3.1), such as the visualization of latent representations in Figure 6.
- **Latent Vector Arithmetic** on text (Section 4.1.3)
- **Image Manipulation** (Section 4.2.2)

More specifically, in the protein representation experiment, we train DiLED using the same autoencoder architecture as the baseline ReLSO [4]. After training, we **visualize the learned latent spaces** of both models, using different colors to represent the fitness values of the corresponding protein sequences, as shown in Figure 6.

The latent space visualization of DiLED (right) **exhibits a clear and interpretable structure**. Protein sequences with **similar fitness values are clustered together**, forming distinct groups in the latent space. Moreover, these **groups are arranged successively according to their fitness values**, with a **smooth transition from low-fitness to high-fitness regions**. This organization demonstrates that DiLED's latent space effectively captures the semantic properties of the proteins, such as their fitness levels, in a meaningful and interpretable way.



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 02:15 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=wcNpivQWEg)

Rebuttal:

## Rebuttal (Part III)

### (Q2, W3, continue from Part I)

In contrast, the latent space of ReLSO (left) appears more entangled and less interpretable. Different fitness values are overlapping and intermixed in the latent space, without a clear separation between different fitness levels. This lack of structure suggests that ReLSO's latent space struggles to capture the semantic properties of the proteins as effectively as DiLED.

The interpretability of DiLED's latent space is further supported by the results of the **Latent Vector Arithmetic** experiment on text data (Section 4.1.3) and the **Image Manipulation** experiment (Section 4.2.2). In the text experiment, DiLED's latent space **enables meaningful vector arithmetic operations to control the sentiment** of generated sentences. Similarly, in the image experiment, DiLED's latent space **allows for effective attribute**

**manipulation by modifying the latent representations.** These results demonstrate that DiLED learns a well-structured latent space that captures the semantic properties of the data, facilitating controlled generation and manipulation.

### (Q3, W1) Generalizability, Limitations on Data Types/Tasks

we have demonstrated the wide applicability and strong generalization capabilities of our proposed method across various domains, including text, image, and protein data. As discussed in Lines.49-85, none of the previous deep generative models shares a similar level of broad applicability and generalizability as DiLED. Our extensive experiments validates the improvement of DiLED over various baselines in handling diverse data types and tasks.

We'll add more discussion on the limitations of the approach. For example, our approach has not been verified on other data modalities such as videos, audios, and time series. It'll be interesting to see how the approach will perform on all the diverse forms of data, and even multi-modal data (such as text+images). In addition, we have only parameterized the first diffusion step with learnable encoding/decoding. A single latent vector representation could be limited to capture relevant information of data for different tasks. We'd like to investigate our approach with more encoding/decoding-based diffusion steps, to obtain hierachical latent representations.



## Rebuttal by Authors

Rebuttal

Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

29 Mar 2024, 02:17 (modified: 29 Mar 2024, 05:26)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=IAxbQKszH5)

Rebuttal:

## Rebuttal (Part IV)

### Summary

We hope our responses have addressed your concerns and provided a clearer understanding of our work. We greatly appreciate your thorough review and valuable feedback, which have helped us improve the clarity and completeness of our paper.

If you have any further questions, please don't hesitate to let us know. We welcome the opportunity to discuss more details and ensure all your concerns are fully addressed.

### Reference

[1] Bowman, Samuel R., et al. "Generating sentences from a continuous space." arXiv preprint arXiv:1511.06349 (2015).

[2] Kingma, Durk P., et al. "Improved variational inference with inverse autoregressive flow." Advances in neural information processing systems 29 (2016).

[3] Fu, Hao, et al. "Cyclical annealing schedule: A simple approach to mitigating kl vanishing." arXiv preprint arXiv:1903.10145 (2019).

[4] Castro, Egbert, et al. "Transformer-based protein generation with regularized latent space optimization." Nature Machine Intelligence 4.10 (2022): 840-851.



*Replying to Rebuttal by Authors*

## Official Comment by Reviewer Si5g

Official Comment  Reviewer Si5g  01 Apr 2024, 12:05

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors



**Comment:**


Thank you for your comments. I will increase my original score.



**Official Comment by Authors**

Official Comment

 Authors ( Qiyu Wu (/profile?id=~Qiyu\_Wu2), Zichao Yang (/profile?id=~Zichao\_Yang1), Shuguang Cui (/profile?id=~Shuguang\_Cui1), Eric Xing (/profile?id=~Eric\_Xing1), +8 more (/group/info?id=ICML.cc/2024/Conference/Submission144/Authors))

 01 Apr 2024, 23:30 (modified: 01 Apr 2024, 23:30)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

 Revisions (/revisions?id=1ODGApWj0a)

**Comment:**

Thank you for deciding to increase the score. We sincerely appreciate the time and effort you have invested in reviewing our work and providing valuable feedback!

[About OpenReview \(/about\)](/about)

[Hosting a Venue \(/group?id=OpenReview.net/Support\)](/group?id=OpenReview.net/Support)

[All Venues \(/venues\)](/venues)

[Sponsors \(/sponsors\)](/sponsors)

[Frequently Asked Questions](#)

[\(https://docs.openreview.net/getting-started/frequently-asked-questions\)](https://docs.openreview.net/getting-started/frequently-asked-questions)

[Contact \(/contact\)](/contact)

[Feedback](#)

[Terms of Use \(/legal/terms\)](/legal/terms)

[Privacy Policy \(/legal/privacy\)](/legal/privacy)

[OpenReview \(/about\)](/about) is a long-term project to advance science through improved peer review, with legal nonprofit status through [Code for Science & Society \(https://codeforscience.org/\)](https://codeforscience.org/). We gratefully acknowledge the support of the [OpenReview Sponsors \(/sponsors\)](/sponsors). © 2024 OpenReview