VAE-VDM: Representation Learning with Variational Diffusion Models

Egoitz Gonzalez, Diego Garcia Cerdas, Jacky Chu, Maks Kulicki

Motivation

- Diffusion-based models do not contain a module for capturing representations.
- *Abstreiter et al. (2022)* [1] proposed a method for representation learning using (conditional) score-based generative models.

Contribution

- We propose a **probabilistic** and **fully-generative** alternative using **Variational Diffusion Models (VDM)** [2].
- We explore its potential in terms of **representation learning** and **data generation.**

Limitations

- VAEs typically suffer from optimization challenges when using powerful decoders [3].
- We encountered *posterior collapse* during training.

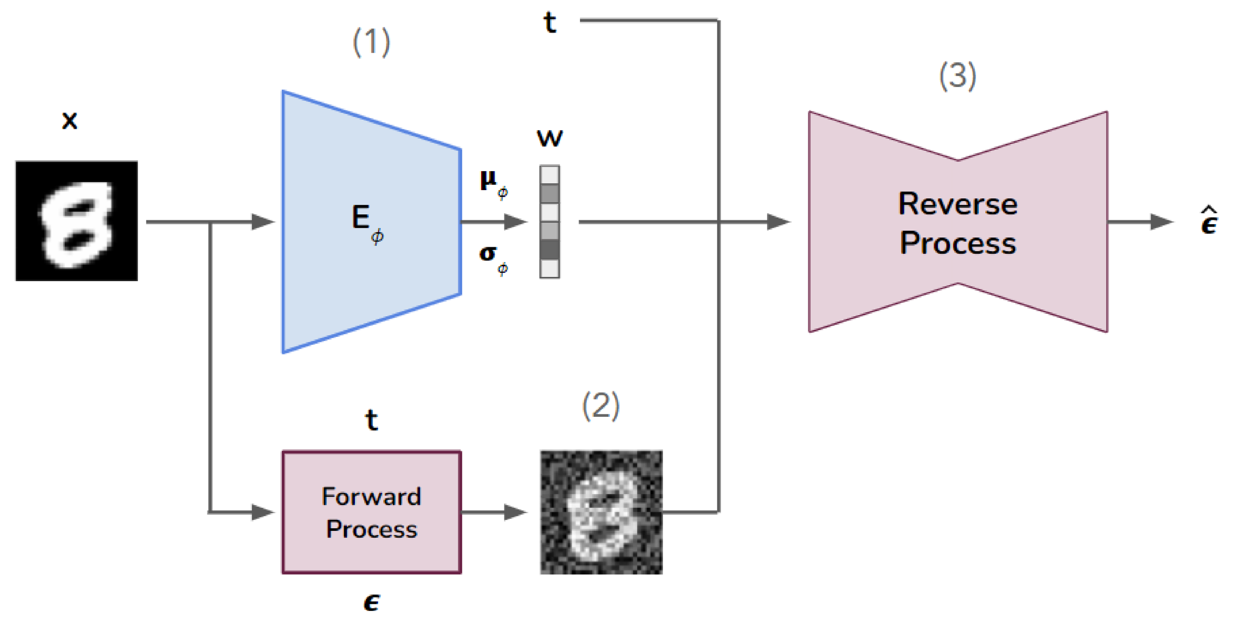
Approach

1. VAE using VDM as decoder

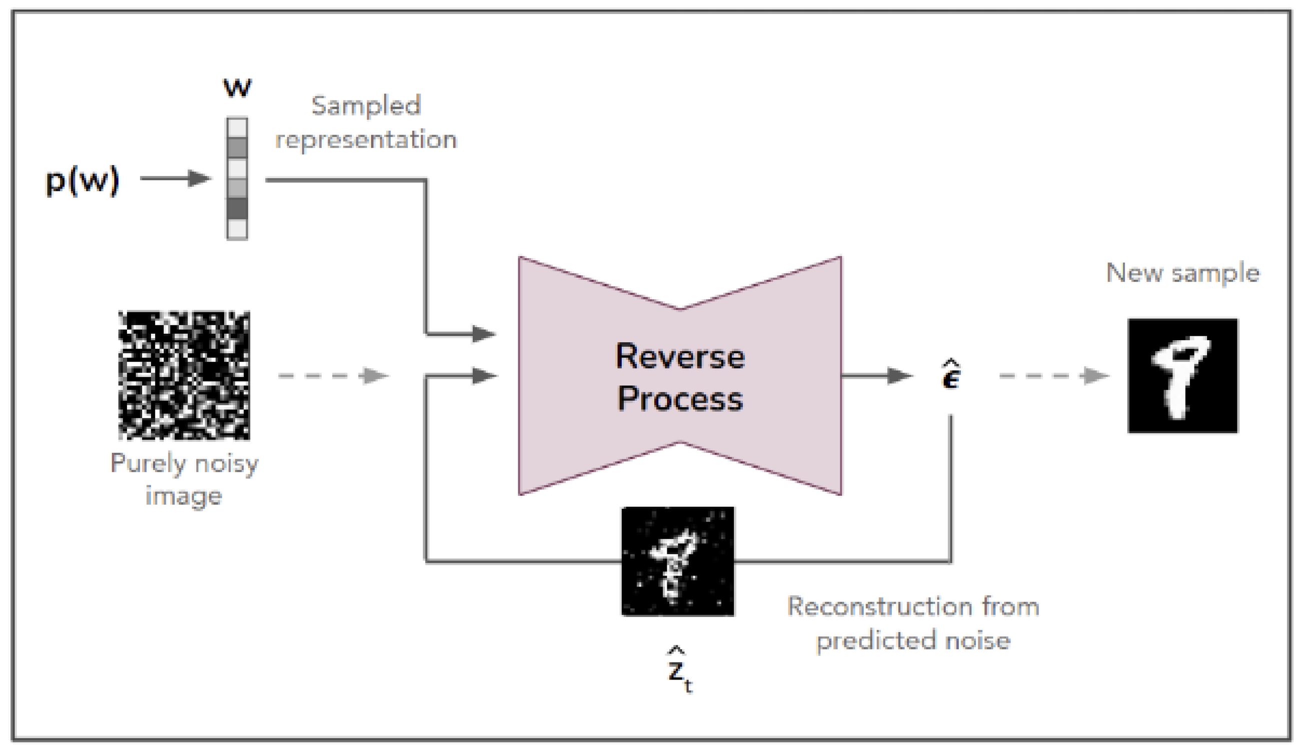
- **Derivation of Variational Lower Bound:**

$$\mathcal{L}_{VAE'} = \mathbb{E}_{\mathbf{w}\sim q_\phi(\mathbf{w}|\mathbf{x})}\mathcal{L}_{\text{VDM}} + D_{\text{KL}}(q_\phi(\mathbf{w}|\mathbf{x})||p(\mathbf{w})).$$

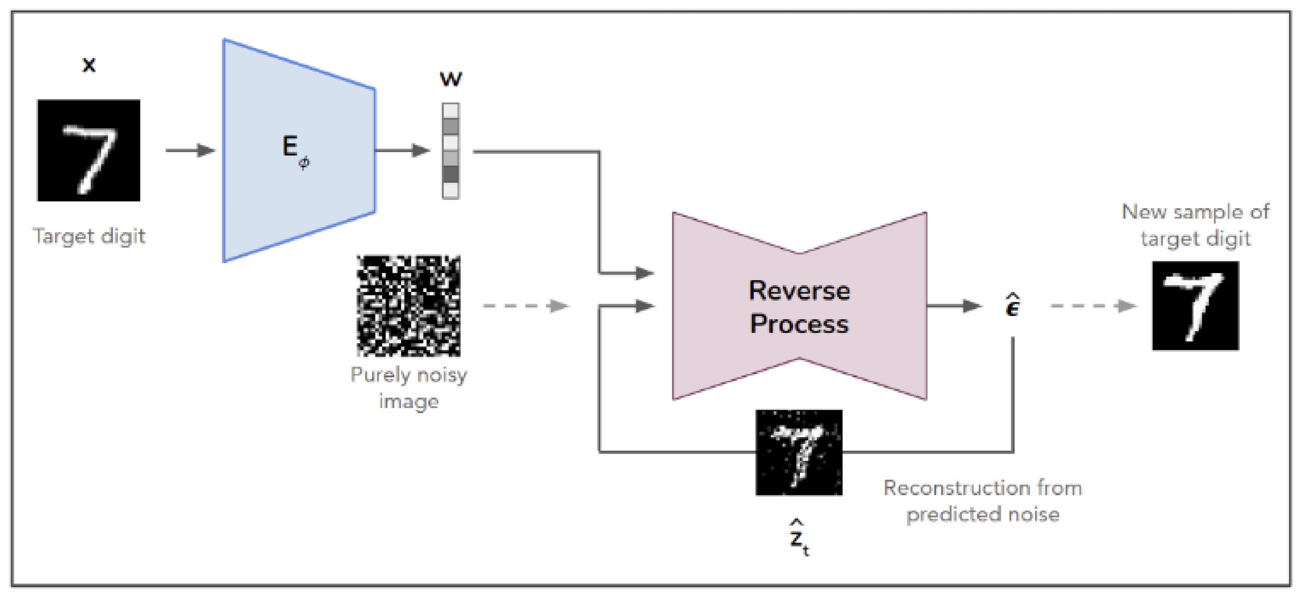
- **Architecture:**



Fully-Generative Setting



Guided Generation

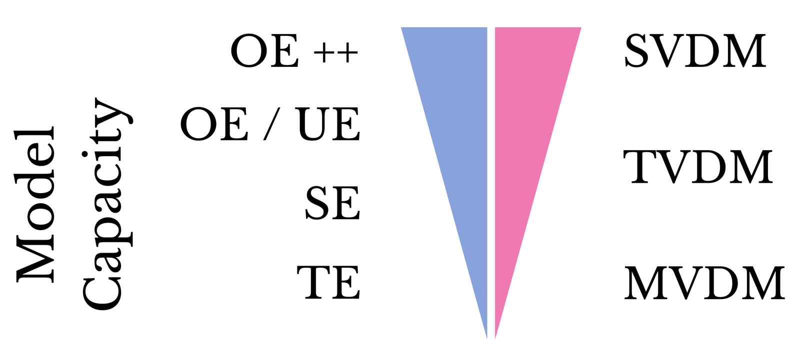


2. Research Questions

- Is the encoder learning a meaningful representation of the data?
- Is the representation useful for the diffusion-based model?

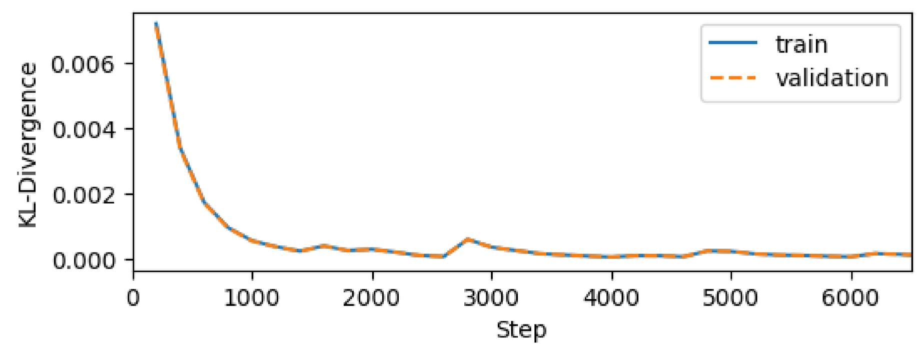
3. Experiments

- **Less powerful VDM**
  - More responsibility to the encoder
- **Smaller encoder**
  - Remove possible redundant parameters.
  - Analyze less meaningful encodings.
- **Unregularized training**
  - No collapse to the prior distribution.



Optimization Challenges

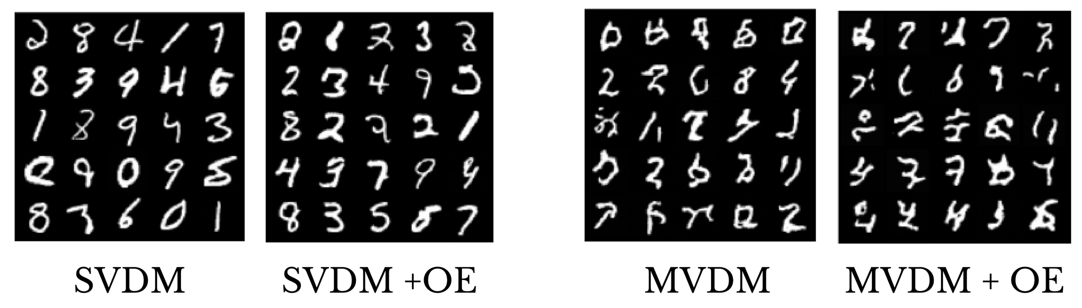
- **Posterior collapse** 😔
  - Encoder KL divergence falls to 0 for all experiments.



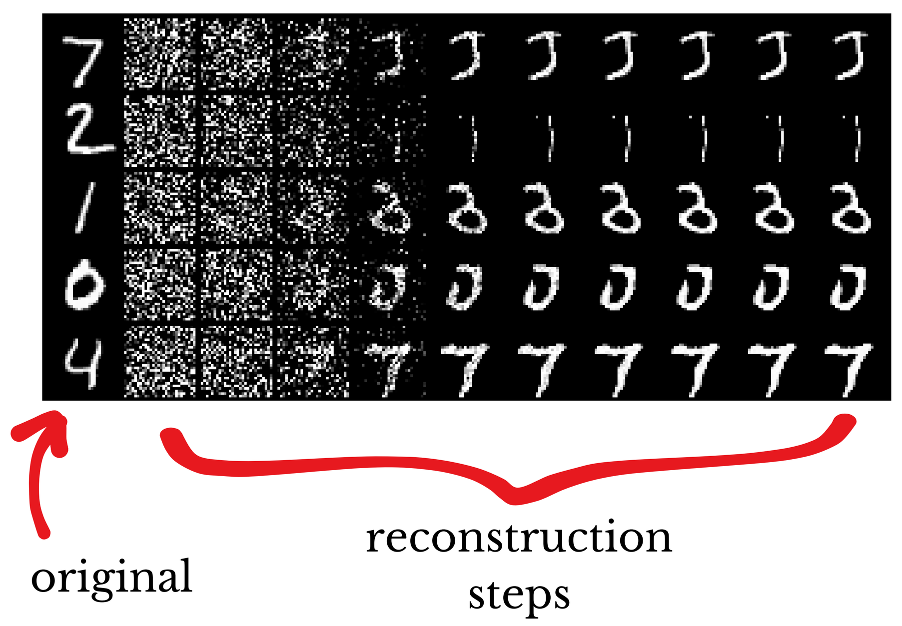
  - Decoder VDM is powerful enough to model data distribution and ignores **w**.

Qualitative Results

- **Sampling**
  - Smallest VDM (MVDM) cannot model the data distribution properly.
  - However, adding an encoder does not show qualitative improvement.

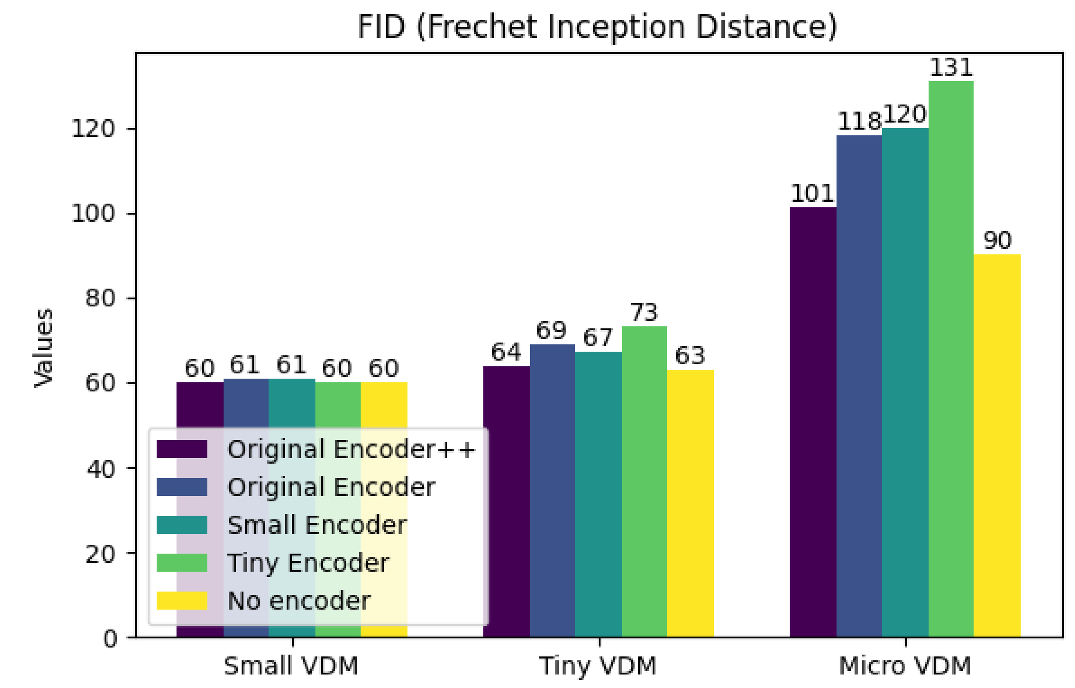


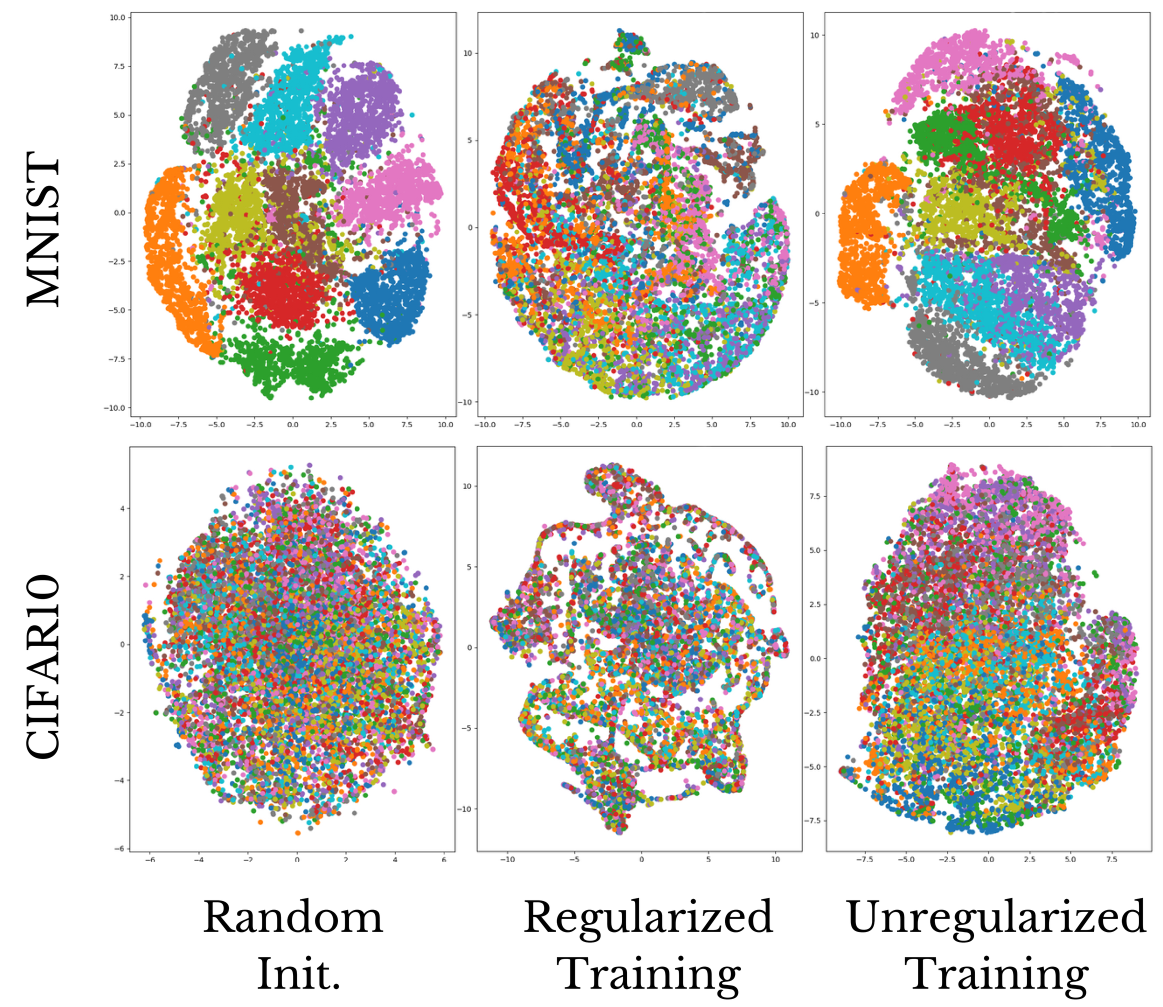
- **Latent space visualization (t-SNE)**
  - MNIST (very simple)
    - KL regularization breaks structure.
    - Unregularized retains random init. structure.
  - CIFAR10 (more complex)
    - Unregularized: slowly learns meaningful structure.
- **Reconstructions**
  - Analyze the effect of **w** during guided generation.
  - **w** ignored due to posterior collapse.



Quantitative Results

- No benefit to the diffusion-based model added by the encoder.
- FID and BPD lower or equal.





Discussion

- Posterior collapse limits encoder utility during reconstruction and sampling.
- Dataset simplicity might be the culprit: experiments on more diverse CIFAR10 show promising results on encoder's capability to learn structure.
- Future work: investigating techniques to mitigate posterior collapse and formalising the importance of the interplay between the decoder's power, the encoder's capability, and the complexity of the dataset in VAE-based representation learning.

References

[1] Abstreiter, K., Mittal S., Bauer S., Schölkopf B., Mehrjou A. (2021). Diffusion-Based Representation Learning.

[2] Kingma, D. P., Salimans, T., Poole B., Ho J. (2021). Variational Diffusion Models.

[3] Wang, Y., Blei, D., & Cunningham, P. (2021). Posterior Collapse and Latent Variable Non-identifiability.