# Few-Shot Diffusion Models Escape the Curse of Dimensionality

**Ruofeng Yang**[1], Bo jiang[1], Cheng Chen[2], Ruinan Jin[34], Baoxiang Wang[34], Shuai Li[1,*]

1. Shanghai Jiao Tong University  2. East China Normal University

3. The Chinese University of Hong Kong, Shenzhen  4. Vector Institute

# Motivations

- Diffusion models become an important paradigm in generation.
- The customized requirements → few-shot diffusion models
- Few-shot models uses only 5-10 images to fine-tune the pre-trained models.

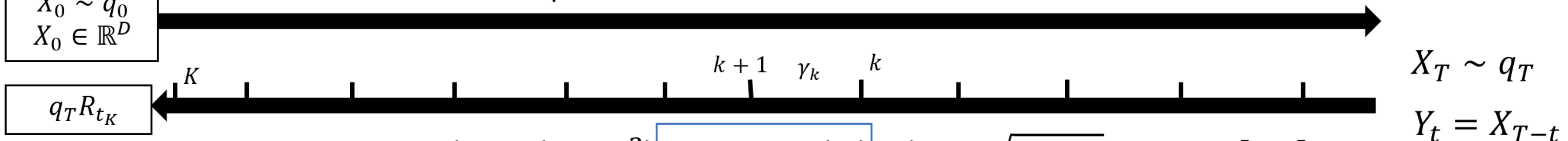
Why few-shot diffusion models achieve great performance with a limited target data?

# The Pretrained Phase of Diffusion Models

1. Large source data distribution: $\{X_{s,i}\}_{i=1}^{n_s} \sim q_0^s \in \mathbb{R}^D$

*Forward SDE: Data → Noise*

$$dX_t = -\beta_t X_t dt + \sqrt{2\beta_t} dB_t, \quad \beta_t \text{ is non}-\text{decreasing sequence}$$




$$dY_t = \beta_{T-t}(Y_t + (1+\eta^2)\boxed{\nabla \log q_{T-t}(Y_t)} dt) + \eta\sqrt{2\beta_{T-t}} dB_t, \eta \in [0,1]$$

*Score Function:* $s_\theta(t, Y_t)$

*Reverse SDE / ODE: Data ← Noise*

2. The pretrain score matching objective function

$$\min_{s \in S_{NN}} \hat{\mathcal{L}}_s(s) = \frac{1}{n_s(T-\delta)} \sum_{i=1}^{n_s} \int_\delta^T \mathbb{E}_{X_t|X_0 = X_{s,i}} \left[ \|\nabla \log q_t^s(X_t|X_0) - s(X_t, t)\|_2^2 \right]$$

# The Few-shot Diffusion Model

1. Limited target data distribution: $\{X_{ta,i}\}_{i=1}^{n_{ta}} \sim q_0^{ta} \in \mathbb{R}^D$

2. The few-shot objective function

$$\min_{s \in \mathcal{Q}_{NN}} \hat{\mathcal{L}}_{ta}(s) = \frac{1}{n_{ta}(T-\delta)} \sum_{i=1}^{n_{ta}} \int_{\delta}^{T} \mathbb{E}_{X_t|X_0=X_{ta,i}} \left[ \|\nabla \log q_t^{ta}(X_t|X_0) - s(X_t,t)\|_2^2 \right]$$

3. The function class $\mathcal{Q}_{NN}$ usually is a subset of $S_{NN}$ (e.g. cross-attention [1] or text-embedding layers [2])

# Current Results of Approximation Error

Many works foucs the requirement of data number $n_s$ to approximate the score

$$\frac{1}{T-\delta}\int_{\delta}^{T}\mathbb{E}_{q_t^s}\left[\|\nabla\log q_t^s(X_t)-\hat{s}(X_t,t)\|_2^2\right]dt$$

where $\hat{s}$ is the minimizer of pretrain objective function.

- Without strong assumption, [3] achieve the minmax error bound $n_s^{-1/D}$.
- With a linear subspace $X_s = A_s z, z \in \mathbb{R}^d$, [4] achieve $n_s^{-2/d}$ error bound.

Though $d \ll D$ [5], the results is heavliy influenced by $d$ $\rightarrow$

Trivially use current analysis achieve $n_{ta}^{-2/d}$ for few-shot models

# Assumption

**Assumption.** The source and target data distribution admit linear low dimensional subspace and share the latent space $X_s = A_s z$ and $X_{ta} = A_{ta} z, z \in \mathbb{R}^d$.

- The common image datasets admit low-dimensional [5].
- Diffusion models can adaptively find the manifold of data [6].
- The shared latent or representation is a standard assumption for few-shot learning.

# The Paradigm of Few-shot Models

- With the linear space assumption, $\nabla \log q_t$ is decomposed to (a) the latent score function $\nabla \log q_t^{LD}(\cdot)$ and (b) linear encoder, decoder $\boxed{A.}$ $\rightarrow$ Approximated by $V \in \mathbb{R}^{D \times d}$

$$\nabla \log q_t^s(X) = A_s \boxed{\nabla \log q_t^{LD}(A_s^\top X)} - \frac{1}{\sigma_t^2}(I_D - A_s A_s^\top)X$$

Shared the information of $\nabla \log q_t^{LD}(\cdot)$, approximated by NN $f_\theta$

- The parameters of $S_{NN}$ is $(V, \theta)$. Let $\left(\hat{V}_s, \hat{\theta}\right)$ be the minimizer of pretrain objective.

- The few-shot models freeze $\hat{\theta}$ and fine-tune $V$.

Let $\hat{V}_{ta}$ be the minimizer of the few-shot objective funciton.

# Main Results (Approximation Error)

Theorem 1 (Informal). With the above assumption, the approximation error of few-shot diffusion model is

$$\frac{1}{T-\delta}\int_{\delta}^{T} \mathbb{E}_{q_t^{ta}}\left[\left\|\nabla \log q_t^{ta}(X_t) - s_{\hat{V}_{ta},\hat{\theta}}(X_t,t)\right\|_2^2\right] dt \leq n_{ta}^{-\frac{1}{2}} + n_s^{-\frac{2}{d}}$$

## Discussion

- The dependence of $n_{ta}$ is independent $d$ $\rightarrow$

*The few-shot diffusion model Escape the Curse of Dimensionality*

- Intuition: The highly nonlinear latent score function is approximated by $n_s^{-2/d}$

The few-shot phase only pays the approxiamtion error of linear matrix $A_{ta}$

# The Real-world Requirement of $n_{ta}$

| Dataset | CIFAR-10 | CIFAR-100 | CelebA | MS-COCO | ImageNet |
| --- | --- | --- | --- | --- | --- |
| Dataset Size | $6 \times 10^4$ | $6 \times 10^4$ | $2 \times 10^5$ | $3.3 \times 10^5$ | $1.2 \times 10^6$ |
| Latent Dimension | 25 | 22 | 24 | 37 | 43 |
| The Requirement of $n_{ta}$ | 6 | 8 | 8 | 5 | 5 |

- Given $n_s$, we require $n_{ta}^{-1/2} = n_s^{-2/d}$ to achieve the same error bound.
- 5-10 target images is enough for few shot diffusion models.

# Main Results (Optimization)

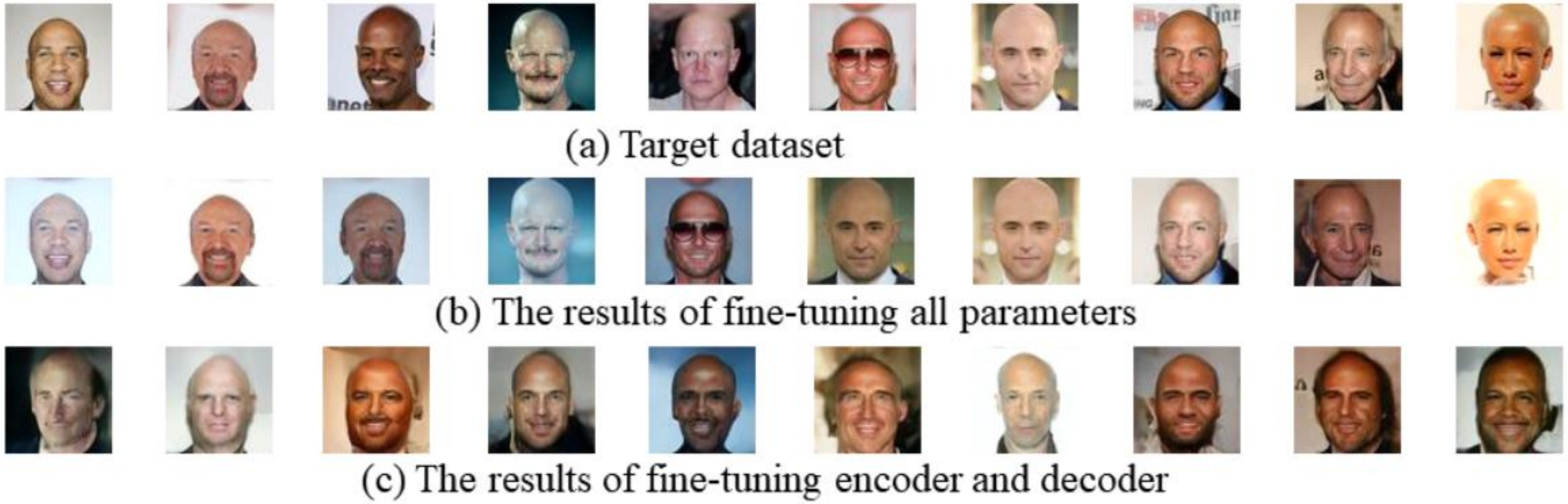
Intuition: Prior of pre-trained model → Simplied Optimization Problem

Theorem 2 (Informal). Assume Gaussian latent, the few-shot objective function has a closed-form minimizer, which (a) is equivalent to PCA and (b)has the following property

$$\left\|\hat{V}_{ta}\hat{V}_{ta}^{\top} - A_{ta}A_{ta}^{\top}\right\| \leq \tilde{O}\left(\frac{1}{\sqrt{n_s}} + \frac{1}{n_{ta}}\right)$$

- Though the Gaussian latent introduce a better $n_s$, the dependence of $n_{ta}$ is still better.

# Real-world Experiments (10 Target Dataset Images)



(a) Target dataset

(b) The results of fine-tuning all parameters

(c) The results of fine-tuning encoder and decoder

- 10 target images with bald feature
- Since fine-tuning all parameters lose the prior information, it also suffers from large $n_{ta}^{-2/d}$ error and memorization phenomenon.
- Our few-shot diffusion model generate novel images with target feature.

# Conclusion

- (Approximaiton) The few-shot diffusion models enjoys

$$n_{ta}^{-\frac{1}{2}} + n_s^{-\frac{2}{d}} \text{ bound instead of } n_{ta}^{-\frac{2}{d}}$$

- (Optimizaiton) The few-shot models simplfy the optimizaiton problem and enjoy closed-form minimizer (under the Gaussian latent).
- Furture work
  - (Approximation) The analysis for nonlinear manifold
  - (Optimization) Extend to general latent

Thanks!

Q&A

# References


- [1] Kumari, N., Zhang, B., Zhang, R., Shechtman, E., & Zhu, J. Y. (2023). Multi-concept customization of text-to-image diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1931-1941).
- [2] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618.
- [3] Oko, K., Akiyama, S., & Suzuki, T. (2023, July). Diffusion models are minimax optimal distribution estimators. In International Conference on Machine Learning (pp. 26517-26582). PMLR.
- [4] Chen, M., Huang, K., Zhao, T., & Wang, M. (2023, July). Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In International Conference on Machine Learning (pp. 4672-4712). PMLR.
- [5] Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., & Goldstein, T. (2021). The intrinsic dimension of images and its impact on learning. arXiv preprint arXiv:2104.08894.
- [6] Tang, R., & Yang, Y. (2024, April). Adaptivity of diffusion models to manifold structures. In International Conference on Artificial Intelligence and Statistics (pp. 1648-1656). PMLR.