A Sampling-Based Domain Generalization Study with Diffusion Generative Models

Ye Zhu[*1,2], **Yu Wu**[3], **Duo Xu**[4], **Zhiwei Deng**[5], **Yan Yan**[6], **Olga Russakovsky**[1]
[1]Department of Computer Science, Princeton Univeristy, USA
[2]LIX, École Polytechnique, IP Paris, France
[3]School of Computer Science, Wuhan University, China
[4]Canadian Institute for Theoretical Astrophysics (CITA), University of Toronto, Canada
[5]Google DeepMind, USA
[6]Department of Computer Science, University of Illinois Chicago, USA

Abstract

In this work, we investigate the domain generalization capabilities of diffusion models in the context of synthesizing images that are distinct from the training data. Instead of fine-tuning, we tackle this challenge from a sampling-based perspective using frozen, pre-trained diffusion models. Specifically, we demonstrate that arbitrary out-of-domain (OOD) images establish Gaussian priors in the latent spaces of a given model after inversion, and that these priors are separable from those of the original training domain. This OOD latent property allows us to synthesize new images of the target unseen domain by discovering qualified OOD latent encodings in the inverted noisy spaces, without altering the pre-trained models. Our cross-model and cross-domain experiments show that the proposed sampling-based method can expand the latent space and generate unseen images without impairing the generation quality of the original domain. We also showcase a practical application of our approach using astrophysical data, highlighting the potential of this generalization paradigm in data-sparse fields such as scientific exploration.

1 Introduction

Generalization ability, which enables a model to synthesize data from diverse domains, has long been a challenge for deep generative models. The current research trend focuses on leveraging larger models with more training data to facilitate improved generalization. The popularity of recent large-scale models, such as DALLE-2 [36], Imagen [18], and StableDiffusion [39], has demonstrated the impressive and promising representation capabilities of state-of-the-art (SOTA) diffusion generative models when trained on enormous image datasets. However, scaling up is not a panacea and does not fundamentally solve the generalization challenge. In other words, for data domains that remain sparse in these already giant datasets, such as astrophysical observation and simulation data, even SOTA models fail to synthesize data suitable for rigorous scientific research. In addition, scaling up requires extensive resources, severely limiting the number of research groups that are able to participate and contribute, and consequently hindering research progress. Given these concerns, our work focuses on studying generalization ability in a few-shot setup, where a pre-trained diffusion generative model and a small set of raw images different from its training domain are provided, with the ultimate objective of generating new data samples from the target OOD domain.

[*]Work mainly completed when YZ was a postdoc at Princeton University.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Frontiers in Probabilistic Inference: Sampling Meets Learning.

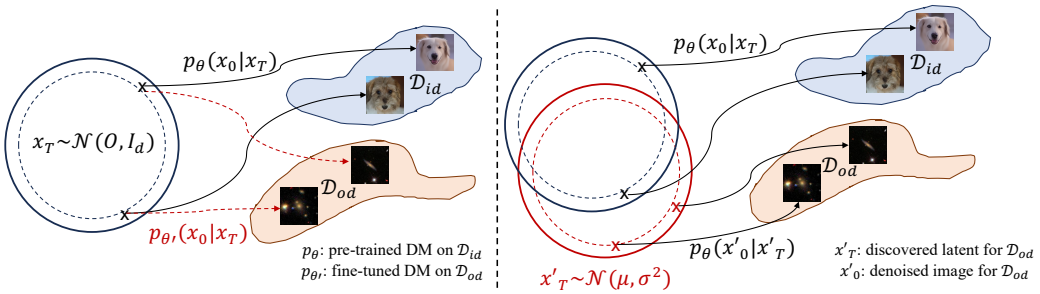


Figure 1: **Illustration of the trajectory-tuning based paradigm (left) and our proposed latent-sampling based paradigm (right) for OOD image synthesis with diffusion models.** Given a pre-trained DM p_θ on images from domain $\mathcal{D}_{id}$, most existing methods seek to finetune the generation trajectories $p_{\theta'}$ to synthesis data in a new domain $\mathcal{D}_{od}$. In contrast, we propose to discover unseen latent encodings to achieve the same goal via the frozen model p_θ by expanding the latent spaces.

To achieve this objective, one of the most straightforward and intuitive approaches is to fine-tune the pre-trained model using the target OOD images [26, 61, 28]. However, tuning-based methods have several drawbacks, particularly when there is a large gap between the training and target data domains. It is well established that tuning methods *do not generalize well* when the domain shift becomes too large [63, 64, 50]. In fact, our experiments show that fine-tuning pre-trained DMs using only image supervision with the vanilla variational lower bound loss [17] is *very difficult*; performance improves only when additional semantic guidance, such as the CLIP loss [35], is introduced. Moreover, modifying model parameters can degrade synthesis quality on the original training domain, and the computational cost of tuning fully depends on the pre-trained model, which can be substantial given the well-known expense of training diffusion models.

In this work, we propose a sampling-based alternative to fine-tuning for tackling this problem. Our core insight lies in reframing the generalization challenge from "learning a new mapping function" to "discovering new OOD latents," as illustrated in Fig.1. More specifically, we show that, after inversion, unseen OOD images exhibit several latent-space properties, including approximate Gaussianity and separability from the original training prior, as detailed in Sec.2. The former allows for the discovery of new OOD latent encodings using relatively simple sampling techniques, while the latter ensures that new OOD samples can be generated without interference from the original generation trajectories. We validate the effectiveness of this approach through a series of experiments.

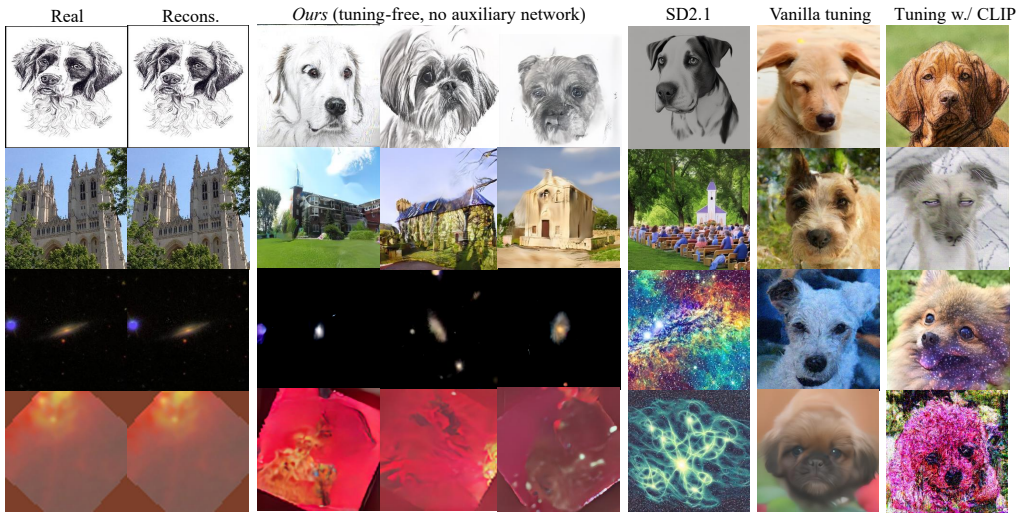
2 Problem Formulation and Method

2.1 Problem Formulation

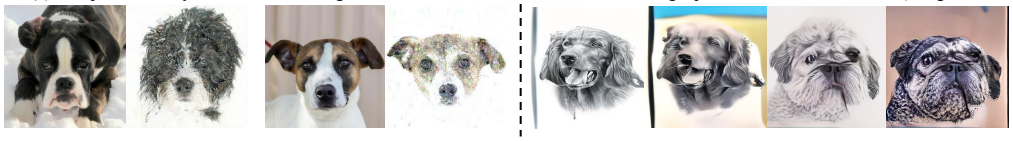
Given a diffusion denoising probablistic model (DDPM) p_θ trained on images from a domain $\mathcal{D}_{id}$, we aim to investigate the generalization properties of p_θ on other domain $\mathcal{D}_{od}$ using N data samples $\mathbf{x}_{od} \in \mathcal{D}_{od}$, and eventually generate new data samples $\mathbf{x}'_{od} \in \mathcal{D}_{od}$. The objective of DDPMs is similar to most previous generative models, which is to approximate an implicit data distribution $q(\mathbf{x}_0)$ with a learned model distribution $p_\theta(\mathbf{x}_0)$, as well as providing an easy-to-sample proxy (e.g., standard Gaussian). We further use p_s and p_i to represent the stochastic [17] and deterministic [42] generation processes, respectively. For the opposite direction, the pre-defined diffusion procedure is often denoted by $q(\mathbf{x}_{1:T}|q_0)$. Similar to existing literature, T denotes the total diffusion steps. We use $\mathcal{X}_t$ to represent the latent (noisy) spaces formed by $\mathbf{x}_t$ along denoising.

2.2 Latent Sampling to New Domain Generalization

Latent Representation of Unseen Images. We observe that a DDPM, trained even on a single-domain small dataset (*e.g.*, dog faces), already has sufficient representation ability to accurately reconstruct arbitrary unseen images (*e.g.*, human, church, and astrophysical data), as shown in Fig. 2. The reconstruction ability is subject to the deterministic inversion and denoising trajectories [42]. The findings above suggest that: with a good mapping approximator (i.e., pre-trained DDPM) and proper tool (i.e., deterministic trajectories with DDIMs), its intermediate latent spaces already have sufficient representation ability for arbitrary images, which opens up the possibility to leverage DDPMs for



(a) Comparisons of synthesized OOD images across data domains and methods, all using a pre-trained DDPM on AFHQ-Dog-256.



(b) In additional to unconditional synthesis, our proposed method can also achieve bi-directional domain transfers (e.g.. sketch - RGB).

Figure 2: **Examples of synthesized OOD images across data domains and methods.** (a) All the OOD samples are obtained via our proposed sampling method using a pre-trained DMs on AFHQ-Dog [7]. (b) The same sampling method can also be applied to achieve style transfer (e.g., between RGB and sketch images).

synthesizing images from new domains *without tuning* the model parameters. Details about the deterministic inversion and reconstruction test can be found in the Appendix C.

Separability of Latent Gaussians. We then present two key properties of the latent OOD distributions after inversion, which are critical for developing an effective sampling-based method for domain generalization. First, the inverted OOD latent encoding exhibits (approximate) Gaussians in the latent spaces. The Gaussianity is established with both theoretical and empirical groundings, with details about these proofs included in the appendix. Second, the OOD priors are separable from the original Gaussian of the pre-trained image domains after the same inversion. We also show statistical support for this second property and its significance in the Appendix C.

Latent Sampling from OOD Priors. With the above properties, we then propose to generate new target images by first sampling from the OOD priors, and then denoise the newly sampled noisy latents via the deterministic DDIM denoising trajectories. Under the approximate Gaussian condition, the sampling is rather straightforward using the interpolation between arbitrary inverted OOD latents. Similarly, due to space constraints, we present additional details in the Appendix D.

3 Experiments and Analysis

3.1 Experimental Setup

Model Zoos and Datasets. We adopt four pre-trained DDPMs on different single domain datasets as our base models for experiments: improved DDPM [32] trained on AFHQ-Dog [7], and DDPM [17] trained on CelebA-HQ [23], LSUN-Church [59], and LSUN-Bedroom [59]. Each model generates images in the original resolution of 256×256, resulting in a total dimensionality of the latent spaces $d = 256 \times 256 \times 3 = 196,608$.

In addition to the above commonly used natural image datasets, we further experiment with two astrophysical datasets to cover a wide range of domain differences and to showcase the application

Table 1: **General quality evaluation in cross model and domain setup.** *T* stands for *Tuning-based*, and *TF* denotes *Tuning-Free*. We report the FID scores (↓) for natural image domains and the Mean Opinion Scores (MOS) (↑) from subjective evaluations with astrophysicists. Note that most baseline methods perform the *image-to-image translation* and use *additional CLIP loss* to tune the model, and thus largely facilitate the task by bypassing the actual sampling stage and with extra strong semantic guidance. We therefore call for special attention when comparing the scores for a comprehensive and objective assessment, qualitative examples in Fig. 2.

Methods	Dog	CelebA	Church	Bedroom	Galaxy	Radiation
Vanilla tuning	213.6±4.8	229.7±4.3	192.5±3.7	191.1±4.0	-	-
CLIP tuning	**73.6±2.9**	**63.6±3.0**	66.3±2.8	68.1±2.8	-	-
Ours	78.2±2.8	64.5±2.7	**64.8±2.7**	**62.9±2.6**	**2.88±0.93**	**1.52±0.80**

scenarios with scientific data. Specifically, we adopt the GalaxyZoo [52] and the radiation simulation data [56], the latter has been investigated using DMs for prediction purposes. Details about those astrophysical datasets, their scientific interpretations, and evaluations are included in the Appendix E for interested readers, which differs from the usual interpretation of natural images.

Comparisons. Our main experiments focus on the comparison with the trajectory tuning-based approaches using the given DMs and OOD samples. Specially, we compare with vanilla tuning, where we finetune the pre-trained model with classic VLB loss [17] on unseen images, as well as several different SOTA image-to-image translation methods via diffusion models as baselines, as they also output images in domains that are different to the trained ones. However, it is worth noting that those baseline methods (*i.e.*, DiffusionClip [26], and Asyrp [28]) are *not really generating* unseen images, but rather *editing an given original ID image* to a target unseen domain. Moreover, those are learning-based methods trained on each unseen domain with extra CLIP loss [35], while our method operates on a single mutual latent space from the pre-trained base diffusion models *without* additional external supervision, echoing our conceptual design of "space expansion".

Implementations and Learning Budget. We use $N = 1000$ images for OOD domains, and the univariant Gaussian estimation for inverted OOD latent encodings. We use 2 RTX 3090 GPUs for all experiments including baselines. For baseline methods that perform image translation and editing [26, 28], we use their respective officially released implementations, each tuning process takes approximately 30 mins, with an initial default learning rate of 2e-6. To ensure a fair comparison, we also stop the vanilla tuning after around 30 minutes, resulting in around 20 epochs of finetuning, with an even larger learning rate of 1e-5. For our proposed method, the inversion takes the same time as those baseline methods, but the core proposed sampling methods take *negligible time*.

Main Results. As the general quality evaluation, we calculate the FID scores [16] for natural images and report the Mean Opinion Scores (MOS) for astrophysical data. The FID scores are averaged over four DDPMs pre-trained on different image domains, and the Mean Opinion Scores (MOS) with a scale between 1-5 are collected from subjective evaluations performed by astrophysicists with respect to the ground truth observation and simulation data in a *non cherry-picky manner*. As shown in Fig. 2 and Tab. 1, *vanilla tuning* with only image supervisions can *hardly* alter the original generation trajectories and synthesize desired images, always synthesizing in-domain images after comparable tuning time with other tuning baselines. As for methods that finetune the model with additional CLIP loss [35], such as DiffusionCLIP [26] and Asyrp [28], they relatively perform better for domains closer to their trained domains as expected. Our proposed method shows an opposite trend by achieving better performance in data domains with bigger differences, as it is easier to avoid mode interference with larger domain gaps in the latent spaces.

4 Conclusion

To sum up, we study the generalization abilities of diffusion models in the few-shot scenario. From the analytical point of view, we explore the generalization properties of diffusion models on unseen OOD domains. From the methodological perspective, our analytical study allows us to propose a sampling-based method for synthesizing images from new domains without tuning the pre-trained generative trajectories. In addition to experiments on natural images, we also showcase the superiority of our method in data-sparse cases with large domain gaps, such as in astrophysics.

Acknowledgments and Disclosure of Funding

This research was primarily conducted while YZ was a postdoctoral researcher at Princeton University. YZ also acknowledges travel funding from the French National Research Agency (ANR) via the "GraspGNNs" JCJC grant (ANR-24-CE23-3888), coordinated by Johannes F. Lutzeyer from École Polytechnique.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, 2020.

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *ICCV*, 2021.

[4] Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in t2i models by identifying semantic directions. *arXiv preprint arXiv:2403.17064*, 2024.

[5] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006.

[6] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of data science*. Cambridge University Press, 2020.

[7] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020.

[8] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE TNNLS*, 2018.

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

[10] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4): 1–13, 2022.

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 2016.

[12] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *CVPR*, 2019.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[14] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022.

[15] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 1998.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeuIPS*, 2017.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS Workshop*, 2022.

[20] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*. Pmlr, 2018.

[21] Minghui Hu, Yujie Wang, Tat-Jen Cham, Jianfei Yang, and PN Suganthan. Global context with discrete diffusion in vector quantised modelling for image generation. *arXiv preprint arXiv:2112.01799*, 2021.

[22] Minyoung Huh, Richard Zhang, Jun-Yan Zhu, Sylvain Paris, and Aaron Hertzmann. Transforming and projecting images into class-conditional generative networks. In *ECCV*. Springer, 2020.

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[24] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022.

[25] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, page 102846, 2023.

[26] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022.

[27] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*, 2020.

[28] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *ICLR*, 2023.

[29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017.

[30] Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. *arXiv preprint arXiv:2103.16091*, 2021.

[31] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*. PMLR, 2013.

[32] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*. PMLR, 2021.

[33] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019.

[34] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021.

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[37] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *NeurIPS*, 2017.

[38] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2021.

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[40] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020.

[41] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*. PMLR, 2015.

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021.

[43] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 2019.

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020.

[45] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *ICML*, 2023.

[46] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 2021.

[47] Jianzhong Wang. *Geometric structure of high-dimensional data and dimensionality reduction*. Springer, 2012.

[48] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[49] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *CVPR*, 2022.

[50] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *NeurIPs*, 2019.

[51] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Lu Yuan, Gang Hua, and Nenghai Yu. E2style: Improve the efficiency and effectiveness of stylegan inversion. *IEEE TIP*, 2022.

[52] Kyle W Willett, Chris J Lintott, Steven P Bamford, Karen L Masters, Brooke D Simmons, Kevin RV Casteels, Edward M Edmondson, Lucy F Fortson, Sugata Kaviraj, William C Keel, et al. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 2013.

[53] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, pages 1623–1639. PMLR, 2024.

[54] KE Wu, KK Yang, R van den Berg, JY Zou, AX Lu, and AP Amini. Protein structure generation via folding diffusion. *arXiv preprint arXiv:2209.15611*, 2022.

[55] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE TPAMI*, 2022.

[56] Duo Xu, Stella SR Offner, Robert Gutermuth, Michael Y Grudić, Dávid Guszejnov, and Philip F Hopkins. Predicting the radiation field of molecular clouds using denoising diffusion probabilistic models. *The Astrophysical Journal*, 958(1):97, 2023.

[57] Duo Xu, Jonathan C Tan, Chia-Jung Hsu, and Ye Zhu. Denoising diffusion probabilistic models to predict the density of molecular clouds. *The Astrophysical Journal*, 950(2):146, 2023.

[58] Yongqi Yang, Ruoyu Wang, Zhihao Qian, Ye Zhu, and Yu Wu. Diffusion in diffusion: Cyclic one-way diffusion for text-vision-conditioned generation. *ICLR*, 2024.

[59] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[60] Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *ICML*. PMLR, 2021.

[61] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *NeurIPS22*, 2022.

[62] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. *NeurIPS*, 2020.

[63] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, 2020.

[64] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

[65] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *ECCV*. Springer, 2020.

[66] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *ECCV*. Springer, 2016.

[67] Ye Zhu, Yu Wu, Zhiwei Deng, Olga Russakovsky, and Yan Yan. Boundary guided learning-free semantic control with diffusion models. In *NeurIPS*, 2023.

[68] Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. Discrete contrastive diffusion for cross-modal music and image generation. In *ICLR*, 2023.

A Technical Appendices and Supplementary Material

We structure the appendices as follows: We provide a detailed discussion on the related work in Appendix B. In Appendix C, we present additional analysis of the latent OOD properties, including the background of deterministic diffusion models, the reconstruction test, the Gaussianity and the separability from the original ID latent prior.

B Related Work

B.1 Generalization in Generative Models

Domain Generalization [48] that aims to generalize models' ability to extended data distributions has been an important research topic in broad machine learning area [11, 62, 64, 31, 29], with various computer vision applications such as recognition [33, 37], detection [60] and segmentation [20, 12]. In the vision generative field, it becomes an even more challenging task with the extra demand to sample from the generalized distributions. One popular recent trend in the computer vision community is scaling up the model and dataset sizes as the most intuitive and obvious solutions [36, 18, 39]. Another scenario to study the domain generalization of generative models is within the few-shot scenario, where we only have a limited amount of data compared to the training set. In this case, fine-tuning the given model on the limited images [26] is the most straightforward way to go.

Our work falls into the second category: provided with a pre-trained model and a small set of unseen images different from the model's training domain, we seek to better understand the generalization abilities of DDPMs.

B.2 Diffusion Models and Deterministic Variants

Diffusion Models (DMs) [41, 17, 43] are the state-of-the-art generative models for data synthesis in images [36, 39, 32, 14, 9, 21], videos [19], and audio [27, 68, 30]. There are currently two mainstream fundamental formulations of diffusion models, i.e., the denoising diffusion probabilistic models (DDPMs) [17] and score-based models [44]. One common perspective to understand both formulations is to consider the data generation as solving stochastic differential equations (SDEs), which characterize a stochastic process. Based on vanilla models, both branch develops their own deterministic variants, i.e., denoising diffusion implicit models (DDIMs) [42] and consistency models [45], with their core idea to follow the marginal distributions in denoising. Compared to initial DDPMs and Score-based DMs with ancestral sampling, the deterministic variants are solving ODEs instead of SDEs and largely accelerate the generation speed with fewer steps.

We leverage the deterministic variant (DDIMs [42]) as the tool to achieve bidirectional transition between latent noisy space and data space in this work.

B.3 Latent Space of Deep Generative Models

Comprehensive studies of latent space of generative models [23, 1, 10] help to better understand the model and also benefit downstream tasks such as data editing and manipulation [66, 40, 28, 65]. A large portion of work has been exploring this problem within the context of GAN inversion [55], where the typical methods can be mainly divided into either learning-based [66, 38, 51, 3] or optimization-based categories [1, 2, 22, 8]. More recently, with the growing popularity of diffusion models, researchers have also focused on the latent space understanding of DMs for better synthesis qualities or semantic control [39, 67, 58].

Our work also contributes to a better understanding of latent spaces, and aims to introduce a new synthesis paradigm to explore the intrinsic potential of DMs.

B.4 Diffusion Models in Science

While DMs have been extensively applied in data generation and editing within the multimodal context [39, 19, 68, 58, 67], recent works have extended their application domains to scientific explorations, such as astrophysics [57, 56], medical imaging [25, 53], and biology [54]. Compared to conventional computer vision applications, scientific tasks usually exhibit several distinct features.

For instance, data acquisition and annotation are generally more expensive due to their scientific nature, resulting in a relatively smaller amount of available data for experiments. Additionally, the evaluation of these works adheres to established conventions within their respective contexts, which are usually different from image synthesis evaluation based on perceptual quality.

Our work also experiments with several astrophysical datasets to showcase the potential of applying our proposed paradigm and method to such specific domains with limited data.

C Additional Analysis of Latent OOD Properties

C.1 Background of Deterministic Diffusion

Our analytical studies and methodology designs are built upon a specific variant of diffusion formulations, i.e., the deterministic diffusion process. While the original diffusion denoising probabilistic models (DDPMs) involve a stochastic process for data generation via denoising (*i.e.*, the same latent encoding will output different denoised images every time after the same generative chain), there is a variant of diffusion model that allows us to perform the denoising process in a deterministic way, known as the Denoising Diffusion Implicit Models (DDIMs) [42]. DDIMs were initially proposed for the purpose of speeding up the denoising process, however, later research works extend DDIMs from faster sampling application to other usages including the inversion technique to convert a raw image to its arbitrary latent space in a deterministic and tractable way. As briefly stated in our main paper, the core theoretical difference between DDIMs and DDPMs lies within the nature of forward process, which modifies a Markovian process to a non-Markovian one.

The key idea in the context of non-Markovian forward is to consider a family of $\mathcal{Q}$ of inference distributions, indexed by a real vector $\sigma \in \mathbb{R}^T_{\geq 0}$:

$$q_\sigma(\mathbf{x}_{1:T}|\mathbf{x}_0) := q_\sigma(\mathbf{x}_T|\mathbf{x}_0)\prod_{t=2}^{T} q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0), \tag{1}$$

where $q_\sigma(\mathbf{x}_T|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_T}\mathbf{x}_0, (1-\alpha_T)\mathbf{I})$ and for all $t > 1$,

$$q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}\mathbf{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\frac{\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\mathbf{I}). \tag{2}$$

The choice of mean function from Eqn. 2 ensures that $q_\sigma(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I})$ for all t, so that it defines a joint inference distribution that matches the "marginals" as desired. The non-Markovian forward process can be derived from Bayes' rule:

$$q_\sigma(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) = \frac{q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)q_\sigma(\mathbf{x}_t|\mathbf{x}_0)}{q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_0)}. \tag{3}$$

In fact, in the original paper, the authors also explicitly stated that: " The forward process from Eqn. 3 is also Gaussian (although we do not use this fact for the remainder of this paper)". [2] While this Gaussian property was not emphasized and leveraged in the DDIMs paper, we find it useful in our context to explore the representation and generalization ability of pre-trained DDPMs.

In particular, the hyper-parameters for Gaussian scheduling α and β in the context of DDIMs are slightly different from the original formulation in DDPMs [17]. Denote the original sequences from DDPMs as α'_t, then the α_t in this work follows the definition of DDIMs to be $\alpha_t = \prod_{t=1}^{T}\alpha'_t$.

In addition to DDIMs, we note that the score-based formulation has also recently marked a deterministic variant, namely the Consistency Models [45]. The core idea of the consistency model is, to some extent, similar to DDIMs, which allows the vanilla score-based stochastic diffusion models to achieve "one-step" denoising, by following the marginal distributions.

As mentioned in our main paper, the deterministic diffusion is mainly used as a tool in this work for our proposed tuning-free paradigm.

[2] This paper refer to the DDIM paper [42].

Table 2: **Reconstruction results for arbitrary images via deterministic diffusion.** We use an iDDPM [32] trained on AFHQ-Dog and 1K testing OOD images to compute the MAE (mean absolute error) reconstruction metric. Note DDIMs [42] was initially proposed to accelerate DDPMs sampling, but have not been studied in this OOD reconstruction setting.

Method	Recons. Domain	MAE (↓)
pSp [38]	CelebA (ID)	0.079
e4e [46]	CelebA (ID)	0.092
ReStyle [3]	CelebA (ID)	0.089
HFGI [49]	CelebA (ID)	0.062
DDIMs [42]	Dog (ID)	0.073 ± 6e-4
	CelebA (OOD)	0.073 ± 8e-4
	Church (OOD)	0.074 ± 8e-4
	Bedroom (OOD)	0.072 ± 7e-4
	Galaxy (OOD)	0.067 ± 1e-3
	Radiation (OOD)	0.077 ± 9e-4

C.2 Latent Representation Ability via Reconstruction

In our work, we evaluate the latent representation capability through reconstruction experiments and report the corresponding quantitative results in Tab. 2. These experiments serve as a sanity check under the assumption that, given plausible noisy latent encodings, the model should at least be able to faithfully reconstruct arbitrary target OOD images.

C.3 Parameter-Independent Properties: Gaussian Priors

We seek a theoretically grounded explanation to the generalization properties of pre-trained DDPMs after the inversion. The takeaway message is: *In theory*, the inverted latent encodings also establish Gaussian priors as presented in Lemma C.1.[3]

Lemma C.1 *For $q_\sigma(\mathbf{x}_{1:T}|\mathbf{x}_0)$ defined in Eqn. 1 and $q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ defined in Eqn. 2, we have:*

$$q_\sigma(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I}). \tag{4}$$

As also mentioned in [42], one can derive Lemma C.1 by assuming for any $t \leq T$, $q_\sigma(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I})$ holds, if:

$$q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}\mathbf{x}_0, (1-\alpha_{t-1})\mathbf{I}), \tag{5}$$

and then prove the statement with an induction argument for t from T to 1, since the base case $(t = T)$ already holds by definition. We provide the detailed proof below.

Proof:
Assume for any $t \leq T$, $q_\sigma(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I})$ holds, if:

$$q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}\mathbf{x}_0, (1-\alpha_{t-1})\mathbf{I}), \tag{6}$$

then we can prove that the statement with an induction argument for t from T to 1, since the base case $(t = T)$ already holds.

First, we have that

$$q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_0) := \int_{\mathbf{x}_t} q_\sigma(\mathbf{x}_t|\mathbf{x}_0) q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) d\mathbf{x}_t, \tag{7}$$

$$q_\sigma(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I}), \tag{8}$$

$$q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}\mathbf{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\mathbf{I}). \tag{9}$$

[3] However, in practice, due to the fact that pre-trained DMs themselves are function approximators, the samples after inversion do not establish perfect Gaussians but rather approximations.

Table 3: **Geometric properties of inverted ID and OOD latent encodings.** The results are computed based on 1K sample pairs. We report the mean and std for each geometric measurement to ensure the statistical significance. The base model is trained on AFHQ-Dog [7] in 256x256.

$\mathcal{D}$	Dog (ID)	Human (O)	Bedroom (O)	Church (O)	Astro. Galaxy (O)	Astro. Turbulence (O)
Pair-Angle	60.0±0	60.0±0	60.0±0.1	60.0±0.1	60.0±0.1	60.0 ± 0
Angle-Origin	89.7±0.01	89.7±0	89.8±0.01	89.7±0.01	89.1±0.01	87.6 ±0.03
Pair-Distance	607.4±0.01	611.4±0.05	611.2±0.07	609.9±0.02	612.37±0.05	609.13 ± 0.1
Center-Distance	-	33.3	26.0	33.2	54.7	60.8
Clf. Acc.	-	0.97	0.99	0.99	1.0	1.0

According to [5] *2.3.3 Bayes' theorem for Gaussian variables*, we know that $q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_0)$ is also Gaussian, denoted as $\mathcal{N}(\mu_{t-1}, \Sigma_{t-1})$ where:

$$\mu_{t-1} = \sqrt{\alpha_{t-1}}\mathbf{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\frac{\sqrt{\alpha_t}\mathbf{x}_0 - \sqrt{\alpha_t}\mathbf{x}_0}{\sqrt{1-\alpha_t}} = \sqrt{\alpha_{t-1}}\mathbf{x}_0, \tag{10}$$

$$\Sigma_{t-1} = \sigma_t^2\mathbf{I} + \frac{1-\alpha_{t-1}-\sigma_t^2}{1-\alpha_t}(1-\alpha_t)\mathbf{I} = (1-\alpha_{t-1})\mathbf{I}. \tag{11}$$

Therefore, $q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}\mathbf{x}_0, (1-\alpha_{t-1})\mathbf{I})$, which allows to apply the induction argument.

Q.E.D

C.4 Data-Dependent Properties: Mode Interference and Separability

In the literature of GANs-based generative models [13], "mode collapse" is a common issue that describes the training failure when generated images tend to be very similar given randomly sampled starting encodings from the Gaussian prior. Within the context of diffusion models in our work, we explicitly reveal a phenomenon analog to the "mode collapse" in GANs, which we refer to as *"mode interference"*, as qualitatively illustrated in Fig. 3 (a).

Intuitively, *"mode interference"* describes the case when the denoised images fall into the model's original training domain $\mathcal{D}_{id}$ instead of the target unseen domain $\mathcal{D}_{od}$ due to the prior interference in the latent spaces. Specifically, when we sample directly from the standard Gaussian to obtain a latent encoding $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, then the denoised image will surely fall into the original training domain $\mathbf{x}_0 \in \mathcal{D}_{id}$ with $\mathbf{x}_0 \sim p_\theta(\mathbf{x}_0)$, which is the vanilla generation process of a trained DDPM. However, it contradicts our task objective to synthesize images $\mathbf{x}'_0 \in \mathcal{D}_{od}$. As illustrated in Fig. 1, since we are denoising the latent encoding via deterministic trajectories p_i, the remaining critical technical challenge to generate $\mathbf{x}'_0$ is to find additional qualified latent encoding $\mathbf{x}'_T$ *free from the interference* of the ID Gaussian mode in the sampling stage.

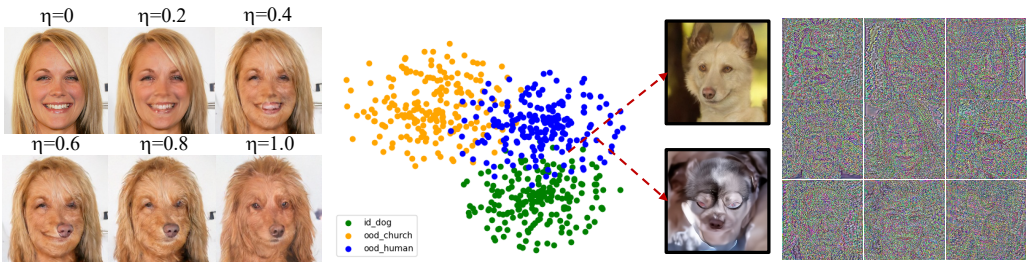
Notably, a key precondition to achieving the effective OOD latent sampling is that the established OOD prior mode *should be separable* from the ID Gaussian prior mode (i.e., a standard Gaussian). Otherwise, the denoised image would fall into the training domain as in Fig. 3 (b). The separability is further supported and validated by our empirical verification below in Sec. C.5.

C.5 Analytical Experiments

We show empirical verification from multiple perspectives to support our parameter-independent and data-dependent properties described above.

Geometrical Properties of Gaussians. We leverage the geometrical measurements established of the high-dimensional studies in mathematics [6], as additional empirical support for the Gaussian priors in Sec. C.3. Specifically, we compute several geometric metrics, including the pair-wise angles (angles formed by three arbitrary samples), sample-to-origin angles (angles formed by two arbitrary samples and the origin), pair-wise distance (euclidean distance between two arbitrary samples) and distance between OOD and ID Gaussian centers, and list the results in Tab. 3.

Characteristics above are typical geometric properties possessed by isotropic high-dimensional Gaussians [6]. Notably, three randomly sampled points from a high-dimensional Gaussian are almost surely form an equilateral triangle and are almost surely nearly orthogonal, which corresponds to the constant 60° pair-wise angle and 90° sample-to-origin angle in the first and second rows of Tab. 3, respectively.



(a) Qualitative illustration. (b) T-SNE of inverted latent encodings and failure cases. (c) Inverted OOD latents.

Figure 3: **Various visualizations of "mode interference".** Given an example setting of synthesizing human faces from DDPMs trained on dogs. *(a):* An interfered image of human faces gradually becomes similar to its original trained domain as the denoising trajectory shifts from deterministic (η = 0) to stochastic (η = 1). *(b):* Failure cases happen when sampled latent OOD encodings are captured by the model's original probabilistic concentration mass. *(c):* Inverted OOD latent encodings preserve slight perceptible low-level visual features and are not *perfect* Gaussians but rather approximations.

Mode Separability. As revealed by our analysis in Sec. C.4, the separability between ID and OOD Gaussian modes is critical for synthesizing target unseen domain images without modifying the model parameters and for avoiding the "mode interference". We further provide validation from the statistical and learning-based classifier perspectives to support the separability claim.

Statistical Validation. The separability of high-dimensional Gaussians follows Lemma C.2 [6], which states that spherical Gaussians can be relaxingly separated by $\Omega(d^{\frac{1}{4}})$, or even $\Omega(1)$ with more sophisticated algorithms. In other words, for a DDPM trained on 256× 256 images with dimensionality $d = 3 \times 256 \times 256$, ID and OOD modes can be well separated and avoid interference given a distance larger than $d^{\frac{1}{4}} \approx 21$, which is further validated by the empirical distance between centers, listed in the forth row of Tab. 3.

Lemma C.2 *Mixtures of spherical Gaussians in d dimensions can be separated provided their centers are separated by more than $d^{\frac{1}{4}}$ distance (i.e., a separation of $\Omega(d^{\frac{1}{4}})$). and even by $\Omega(1)$ separation with more sophisticated algorithms.*

Proof:

According to existing established understanding (Lemma 2.8 from [6]), for a d-dimensional spherical Gaussian of variance 1, all but $\frac{4}{c^2}e^{-\frac{c^2}{4}}$ fraction of its mass is within the annulus $\sqrt{d-1} - c \leq r \leq \sqrt{d-1} + c$ for any $c > 0$, as illustrated in Fig. 4.

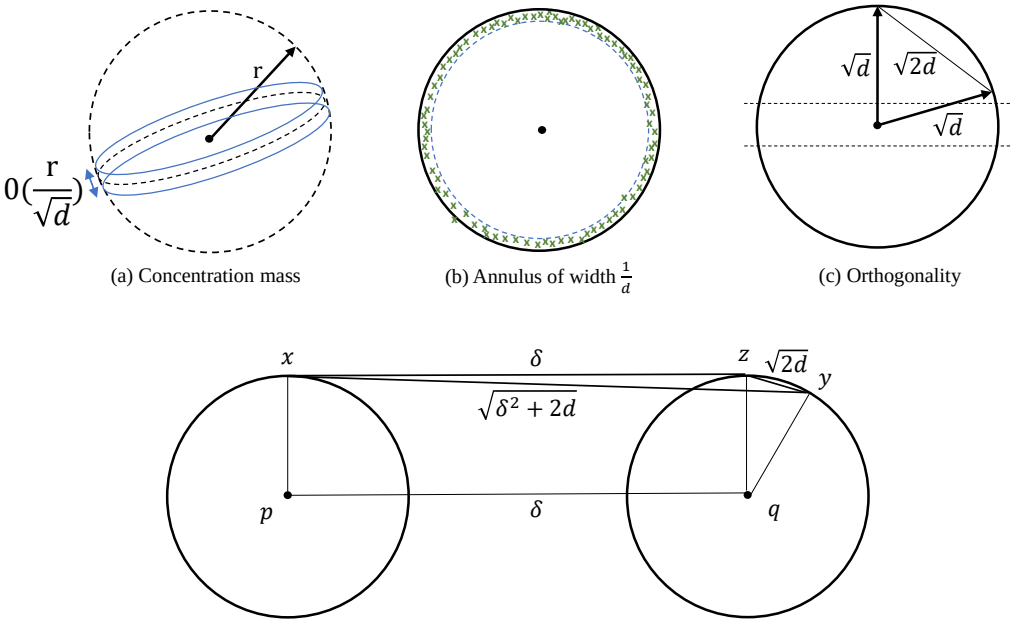
Given two spherical unit variance Gaussians, we have most of the probability mass of each Gaussian lies on an annulus of width $O(1)$ at radius $\sqrt{d-1}$. Also, $e^{-|x|^2/2}$ factors into $\prod_i e^{-x_i^2/2}$ and almost all of the mass is within the slab $\{\mathbf{x}| - c \leq x_1 \leq c\}$, for $c \in O(1)$.

Now consider picking arbitrary samples and their separability. After picking the first sample $\mathbf{x}$, we can rotate the coordination system to make the first axis point towards $\mathbf{x}$. Next, independently pick a second point $\mathbf{y}$ also from the first Gaussian. The fact that almost all of the mass of the Gaussian is within the slab $\{\mathbf{x}| - c \leq x_1 \leq c, c \in O(1)\}$ at the equator says that $\mathbf{y}$'s component along $\mathbf{x}$'s direction is $O(1)$ with high probability, which indicates $\mathbf{y}$ should be nearly perpendicular to $\mathbf{x}$, and thus we have $|\mathbf{x} - \mathbf{y}| \approx \sqrt{|\mathbf{x}|^2 + |\mathbf{y}|^2}$.

More precisely, we note $\mathbf{x}$ is at the North Pole after the coordination rotation with $\mathbf{x} = (\sqrt{(d)} \pm O(1), 0, ...)$. At the same time, $\mathbf{y}$ is almost on the equator, we can further rotate the coordinate system so that the component of $\mathbf{y}$ that is perpendicular to the axis of the North Pole is in the second coordinate, with $\mathbf{y} = (O(1), \sqrt{(d)} \pm O(1), ...)$. Thus we have:

$$(\mathbf{x} - \mathbf{y})^2 = d \pm O(\sqrt{d}) + d \pm O(\sqrt{d}) = 2d \pm O(\sqrt{d}), \tag{12}$$

and $|\mathbf{x} - \mathbf{y}| = \sqrt{(2d)} \pm O(1)$.



(a) Concentration mass (b) Annulus of width $\frac{1}{d}$ (c) Orthogonality

(d) Separability of two Gaussians

Figure 4: **Illustration of various geometric properties of high-dimensional Gaussians.** (a) and (b) show the probability concentration mass is mainly centered around a thin annulus around the equator. (c) illustrates the geometric observation on the orthogonality of sample pairs. (d) illustrates the idea of separating two Gaussian distributions in high-dimensional spaces.

Given two spherical unit variance Gaussians with centers $\mathbf{p}$ and $\mathbf{q}$ separated by a distance δ, the distance between a randomly chosen point $\mathbf{x}$ from the first Gaussian and a randomly chosen point $\mathbf{y}$ from the second is close to $\sqrt{\delta^2 + 2d}$, since $\mathbf{x}-\mathbf{p}$, $\mathbf{p}-\mathbf{q}$, and $\mathbf{q}-\mathbf{y}$ are nearly mutually perpendicular, with:

$$|\mathbf{x} - \mathbf{y}|^2 \approx \delta^2 + |\mathbf{z} - \mathbf{q}|^2 + |\mathbf{q} - \mathbf{y}|^2 = \delta^2 + 2d \pm O(\sqrt{d}). \tag{13}$$

To ensure that the distance between two points picked from the same Gaussian are closer to each other than two points picked from different Gaussians requires that the upper limit of the distance between a pair of points from the same Gaussian is at most the lower limit of distance between points from different Gaussians. This requires the following criterion to be satisfied:

$$\sqrt{2d} + O(1) \leq \sqrt{2d + \delta^2} - O(1), \tag{14}$$

which holds when $\delta \in \Omega(d^{1/4})$.

Thus, mixtures of spherical Gaussians can be separated provided their centers are separated by more than $d^{1/4}$.

Q.E.D

Classifier Validation. Another empirical perspective to validate the separability between modes in the latent spaces is using the classifiers as in existing literature [40, 67]. Specifically, a linear classifier such as SVMs [15] can be fitted to test the separability between ID and OOD encodings in the latent spaces. In our analytical experiments, we fit SVMs on 1K inverted ID and OOD samples following the 7:3 training-testing ratio, and report the test accuracy in Tab. 3. As additional clarification, the classification results are obtained with the test on the latent space $\mathcal{X}_T$. Our rationale behind the choice of T corresponds to the recent findings of DMs [67, 58], which indicates that $\mathcal{X}_T$, as the departure latent space, has the largest probabilistic support for the trained domain. In other words, if the latent ID and OOD modes can be separated in $\mathcal{X}_T$, they can be separated more easily in other $\mathcal{X}_t$, for $t = \{T-1, ..., t, ...1\}$.

C.6 Geometric Properties

We consistently observe three geometric properties for the inverted OOD latent encodings. We provide a more detailed discussion on what each property implies in this sub-section.

The three geometric properties as below:

Observation 1: *For any OOD sample pairs $\mathbf{x}_{inv,i}^{out}$ and $\mathbf{x}_{inv,j}^{out}$ from the sample set, the Euclidean distance between these two points is approximately a constant d_o.*

Observation 2: *For any three OOD samples $\mathbf{x}_{inv,i}^{out}$, $\mathbf{x}_{inv,j}^{out}$ and $\mathbf{x}_{inv,k}^{out}$ from the sample set, the angle formed between $\mathbf{x}_{inv,k}^{out}\vec{\mathbf{x}}_{inv,i}^{out}$ and $\mathbf{x}_{inv,k}^{out}\vec{\mathbf{x}}_{inv,j}^{out}$ is always around 60°.*

Observation 3: *For any OOD sample pairs $\mathbf{x}_{inv,i}^{out}$ and $\mathbf{x}_{inv,j}^{out}$ from the sample set, let O denote the origin in the high-dimensional space, the angle formed between $O\vec{\mathbf{x}}_{inv,i}^{out}$ and $O\vec{\mathbf{x}}_{inv,j}^{out}$ is always around 90°.*

For the first observation, when the sample pairs keep approximately the same distance, the direct implication is that those samples are likely to be drawn from some convex region in the high-dimensional space [47]. One typical example is the spherical structure, where every data points exhibit an equal distance from the center.

The second geometric property suggests that the unknown samples could lie on a regular lattice near a low-dimensional manifold or sub-manifold, where the local geometry of the manifold is approximately Euclidean. However, a less evident implication is that for samples drawn from a high-dimensional Gaussian, this property also holds, as detailed in the next section C.7, and illustrated in Fig. 4(c).

The third geometry property implies that the sample points might be isotropic in nature, who are rotationally symmetric around any point in the space. Therefore, any two points drawn from the distribution are equally likely to lie along any direction in the space. This property is also observed for a high-dimensional Gaussian [6], whose covariance matrix is proportional to the identity matrix.

We acknowledge that to deduce a distribution in high-dimensional space solely based on its geometric properties is very challenging, and there may exist other complex distributions that exhibit similar properties we have observed. However, combined with our theoretical analysis and empirical observations, the OOD Gaussian assumption seems to hold well. Explicitly, we find the above geometric properties do not hold for images $\mathbf{x}_0$ from the data space. For instance, the angle of samples to the origin is approximately 75° rather than 90°.

C.7 High-Dimensional Gaussian

Gaussian in high-dimensional space establishes various characteristic behaviors that are not obvious and evident in low-dimensionality. A better understanding of those unique geometric and probabilistic behaviors is critical to investigate the latent spaces of DDMs, since all the intermediate latent spaces along the denoising chain are Gaussian as demonstrated and proved in our previous sections.

We present below several properties of high-dimensional Gaussian from [6], note those are known and established properties, we therefore omit the detailed proofs in this supplement, and ask readers to refer to the original book if interested.

Property D.1. *The volume of a high-dimensional sphere is essentially all contained in a thin slice at the equator and is simultaneously contained in a narrow annulus at the surface, with essentially no interior volume. Similarly, the surface area is essentially all at the equator.*

This property above is illustrated in Fig. 4(a)(b), where the sampled ID encodings are presented in a narrow annulus.

Lemma D.2. *For any $c > 0$, the fraction of the volume of the hemisphere above the plane $x_1 = \frac{c}{\sqrt{d-1}}$ is less than $\frac{2}{c}e^{-\frac{c^2}{2}}$.*

Lemma D.3. *For a d-dimensional spherical Gaussian of variance 1, all but $\frac{4}{c^2}e^{-c^2/4}$ fraction of its mass is within the annulus $\sqrt{d-1} - c \leq r \leq \sqrt{d-1} + c$ for any $c > 0$.*

Algorithm 1 Latent sampling-centric approach for domain generalization

Input: Arbitrary pre-trained DDPM p_θ in $\mathcal{D}_{id}$, N images $\mathbf{x}_{od} \in \mathcal{D}_{od}$.
Output: images of the unseen target domain $\mathbf{x}'_{od} \in \mathcal{D}_{od}$
// Get the inverted latent encodings $\mathbf{x}_{od,T}$
Define $\{\tau_s\}_{s=1}^{S_{inv}}$ s.t. $\tau_1 = 0$, $\tau_{S_{inv}} = T$
for $i = 1, 2, ..., N$ **do**
  for $s = 1, 2, ..., T-1$ **do**
    $\epsilon \leftarrow p(\mathbf{x}^i_{od,\tau_s}, \tau_s)$
    $\mathbf{x}^i_{od,\tau_{s+1}} = \sqrt{\alpha_{\tau_s}}\mathbf{x}^i_{od,\tau_s} + \sqrt{1-\alpha_{\tau_s}}\epsilon$
  end for
end for
// Get new ood encodings $\mathbf{x}'_{od,T}$ and denoise via DDIMs
$\mathbf{x}'_{od,T} \leftarrow \text{Interp}(\mathbf{x}^i_{od,T}, \mathbf{x}^j_{od,T})$, for $i \neq j,\ (i,j) \in \{1, 2, ..., N\}$
$\mathbf{x}'_{od} \leftarrow p_\theta(\mathbf{x}'_{od,T}, T)$

The lemmas above imply that the volume range of the concentration mass above the equator is in the order of $O(\frac{r}{\sqrt{d}})$, also within an annulus of constant width and radius $\sqrt{d-1}$. In fact, the probability mass of the Gaussian as a function of r is $g(r) = r^{d-1}e^{-r^2/2}$. Intuitively, this states the fact that the samples from a high-dimensional Gaussian distribution are mainly located within a manifold, which matches our second geometric observation.

Lemma D.4. *The maximum likelihood spherical Gaussian for a set of samples is the one over center equal to the sample mean and standard deviation equal to the standard deviation of the sample.*

The above lemma is used as the theoretical justification for the proposed empirical search method in [67]. We also adopt the search method using the Gaussian radius for identifying the operational latent space along the denoising chain to perform the OOD sampling.

Property D.5. *Two randomly chosen points in high dimension are almost surely nearly orthogonal.*

The above property corresponds to the *Observation 3*, where two inverted OOD samples consistently form a 90° angle at the origin.

D More Details about the Latent Sampling Methods

Intuitively, unlike sampling from a pure Gaussian, sampling from an approximate-Gaussian in non-trivial. Alternatively, the general property of Gaussian distributions ensures that any convex combination of two samples lies within the same distribution. This motivates our choice of a simple yet effective approach to identify additional qualified $\mathbf{x}'_{ood,T} \in \mathcal{X}_{ood,T}$ using latent interpolation, a technique widely adopted in generative modeling for geometric and spatial analysis [42, 40, 67, 4, 34]. Specifically, as shown in Algo. 1, new latent samples $\mathbf{x}'_{od,T}$ are obtained by interpolating between two known latent encodings $\mathbf{x}^i_{od,T}$ and $\mathbf{x}^j_{od,T}$, both inverted from raw data.

E More Details for Generative Experiments

E.1 Background and Evaluation about the Astrophysical Data

Galaxy Data. The images from the GalaxyZoo dataset [52] are observation data of galaxies that belong to one of six categories - elliptical, clockwise spiral, anticlockwise spiral, edge-on, star/don't know, or merger. The original data format of those galaxy images are also RGB images, thus "somewhat" similar to natural images, but they contain important morphological information to study the galaxies in astronomy.

The evaluation of the synthesized galaxy data is based on the expertise of astrophysicists if they could reliably classify the generated images into one of the known categories.

Radiation Data. For the radiation data from [56], the original format is physical quantity instead of RGB images, which correspond to the dust emission.

Dust is a significant component of the interstellar medium in our galaxy, composed of elements such as oxygen, carbon, iron, silicon, and magnesium. Most interstellar dust particles range in size from a few molecules to 0.1 mm (100 μm), similar to micrometeoroids. The interaction of dust particles with electromagnetic radiation depends on factors like their cross-section, the wavelength of the radiation, and the nature of the grain, including its refractive index and size. The radiation process for an individual grain is defined by its emissivity, which is influenced by the grain's efficiency factor and includes processes such as extinction, scattering, absorption, and polarization.

In RGB images of dust emission, different colors represent emissions at three wavelengths: blue for 4.5 μm, green for 24 μm, and red for 250 μm. The blue color typically indicates short-wavelength dust emission from point sources, such as young stars or young stellar objects. The green color represents mid-wavelength dust emission from warm and hot dust. The red color signifies long-wavelength dust emission from cold dust.

Warm/hot dust emission (green) is usually found around stars, which appear as blue-colored dots. Since warm dust often mixes with cold dust on the outer edges of bubble structures, the resulting color is often yellowish. Cold dust extends farther from the stars, giving the background or areas outside star clusters a red appearance. In the case of massive star clusters, stellar feedback, such as radiation and stellar winds, can blow away the surrounding gas and dust, creating black or blank areas. Typically, RGB images show more extensive red emission with some orange/yellow emission, displaying filamentary and bubble structures, along with blue and/or white dotted point source emissions.

The above background is considered as part of the underlying evaluation criteria when performing subjective evaluation on the quality of generated radiation data.

Qualitative evaluation from astrophysicists. As the evaluation of astrophysical data requires deep domain expertise, we collaborated with astrophysicists to subjectively assess the quality of the generated data using Mean Opinion Scores (MOS). Specifically, we provided 50 non-cherry-picked generated samples from our unseen domain generalization experiments alongside 50 raw data samples from true physical simulations. For each generated sample across the two astrophysical datasets, a score ranging from 1 to 5 was assigned relative to the raw data, where 5 indicates the highest quality and 1 is the lowest.

Overall, the final MOS ratings for our generated galaxy data and radiation data are 2.88 ± 0.93 and 1.52 ± 0.80, respectively. In comparison, the same number of samples generated using the CLIP fine-tuning method received lower MOS ratings of 1.84 ± 0.97 and 1.35 ± 0.74 for the galaxy and radiation datasets, respectively. These results highlight the relative superiority of our proposed method in generating higher-quality samples, particularly for domains with complex structures and significant domain gaps.

E.2 More Experimental Results

We provide extended discussions in this section for the readers who are interested in more subtitle experimental details.

E.2.1 Discussion on the Latent Step t, Stochasticity and Mode Interference

While we empirically find that $t \approx 800$ is a reasonable range for the choice of t, we note there exists an entangled mechanism for the trade-off between the sampling difficulty and the mode interference issue.

For the diffusion step t, recent studies [67, 58] suggest that t characterizes the formation of image information at different stages of the denoising process. Intuitively, the early stage of the denoising process (e.g., $t > 800$) represents a rather chaotic process, the mixing step t_m [67] signifies a critical stage where the image semantic information starts to form, and the later stage where t is close to 0 demonstrates a stage during which more fine-grained pixel-level information are introduced to the final generated data. From the distribution point of view, the influence of t can be interpreted as the convergence of distributions, where $t = T$ is a standard Gaussian by definition, thus the ID and OOD modes are more difficult to separate. However, as the denoising process gets closer to the real image

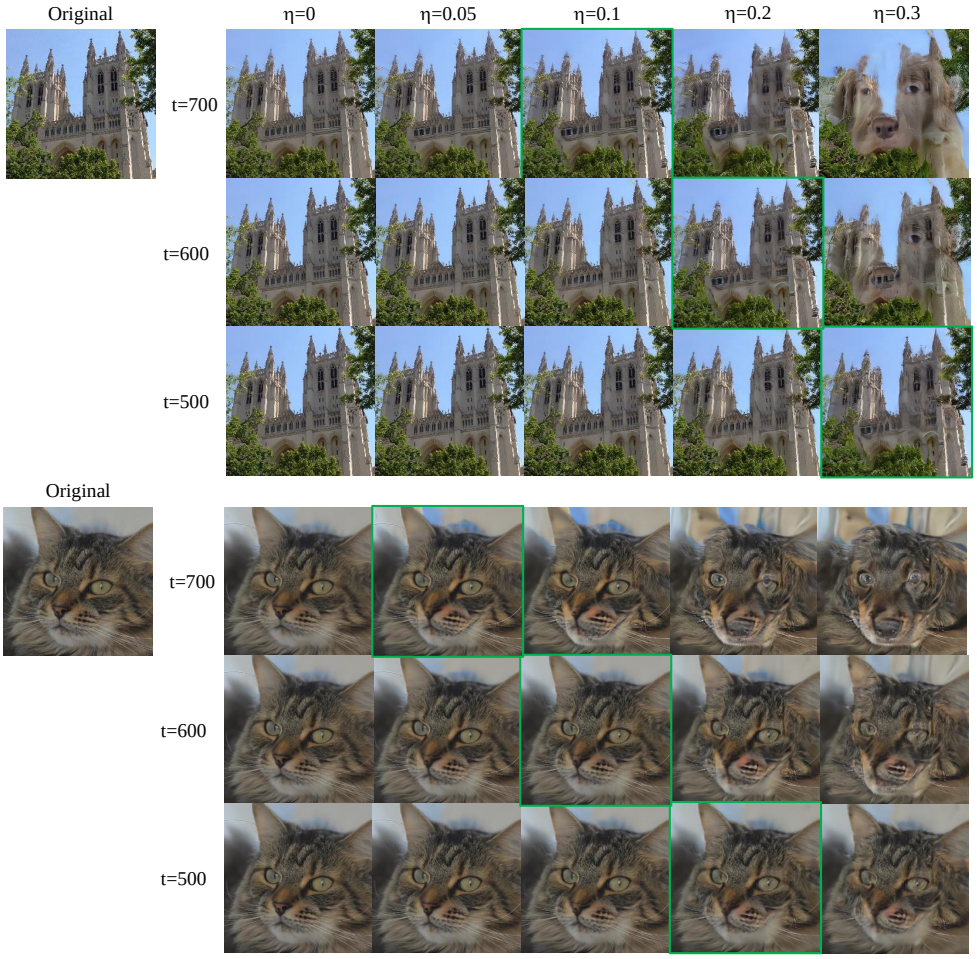


Figure 5: **Illustration of unseen trajectory bandwidth at different diffusion steps.** We show qualitative examples using the iDDPM [32] trained on AFHQ-Dog-256 as the base model, the examples of church and cat are both unseen domain images. The image in green boxes indicates the bandwidth we have empirically selected to preserve the reconstruction quality. Compared to the trained image domain (*i.e.*, *dogs*), *cats* have a smaller domain gap than *churches*. Different from the conventional understanding that a smaller domain gap is beneficial for better and easier generalization from a trained model, we observe a larger domain gap signifies a larger bandwidth, making it easier to perform the OOD sampling and synthesis.

space at $t = 0$, the sampling difficulty increases as the implicit distribution moves away from the standard Gaussian.

Meanwhile, the diffusion step t is not the only factor that impacts the trade-off between sampling difficulty and mode interference. While scarcely discussed in the main paper, we note the stochasticity of the denoising trajectory also plays a similar role as the diffusion step in this work. The stochasticity of the denoising trajectory in DMs has been proven to be generally beneficial in improving the synthesis quality [24, 26, 28, 67]. In this work, while we choose the $\eta = 0$ for the main paper, a tolerance for a certain range of stochasticity allows us to follow a "relatively deterministic" denoising process $p_{\eta=k}$, with $k \neq 0$, instead of the completely deterministic p_i. We hereby refer to it as "bandwidth of the unseen trajectories," denoted as $\mathcal{B}_{\eta,t}$, which can be used to quantify the "mode interference". Another interpretation is to analog the trajectory bandwidth $\mathcal{B}_{\eta,t}$ to the actual subspace volume occupied by the OOD latent samples. Fig. 5 shows more qualitative results for the bandwidth search in the reconstruction task and reveals its connection to the diffusion step t. Overall, the

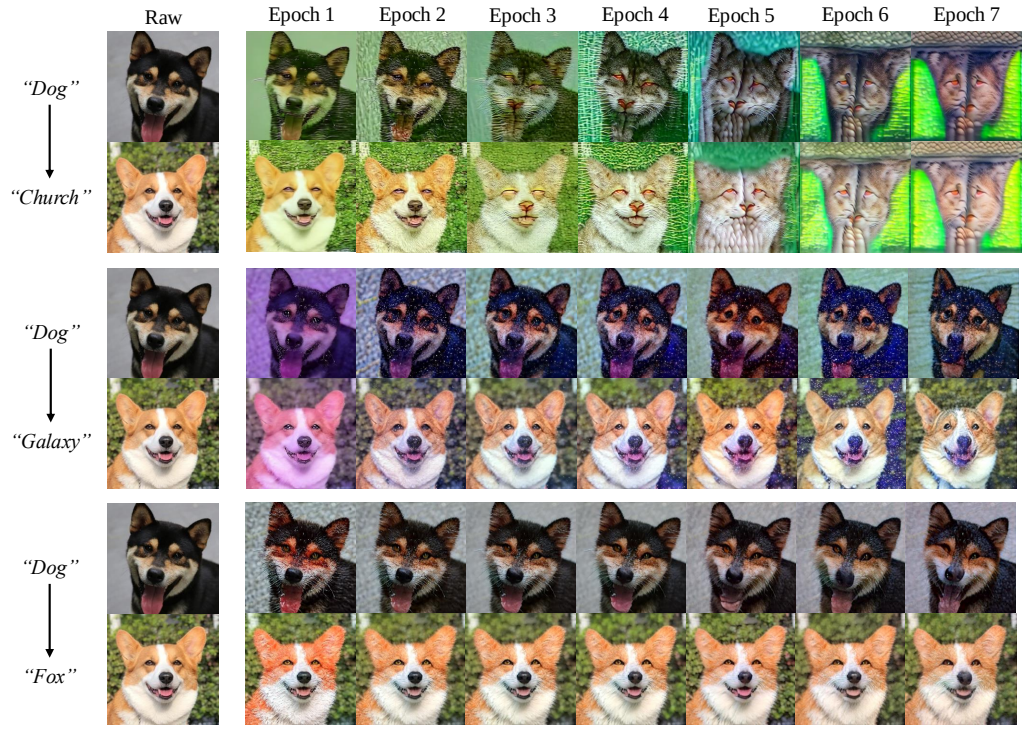


Figure 6: **Fine-tuning methods often fail to transfer the original trained domain to the target unseen domain with large distributional shifts.** We qualitatively show how a given ID sample (e.g., a dog RGB image) changes as the tuning epoch increases, using extra CLIP semantic guidance.

bandwidth is a hyper-parameter that relates to the base model and the unseen domains, and the diffusion step t, while the bandwidth gets larger at the latent spaces closer to the raw image domains, sampling from OOD unseen distributions also gets more difficult.

E.2.2 Discussion on Model Designs

Among four base DDPMs we have tested, there are two architecture variants namely the improved DDPM [32] and vanilla DDPM [17]. The difference between the two variants lies within the scheduler design for the Gaussian perturbation kernels: improved DDPM uses a cosine scheduler while vanilla DDPM adopts a linear one. Our experiments suggest that iDDPM in general synthesizes images with better quality in terms of FID scores, which aligns with previous studies [32, 67]. One implication from the above observation is that the domain generalization abilities studied in this context is inherited from the performance of model's original performance.

E.2.3 More Qualitative Results

We show more qualitative samples from the CLIP-tuned methods in Fig. 6 and note that the tuning-based methods often fail to generalize to new image domains with large distributional shofts.

In Fig. 7, we first show the correlation between the diversity among generated samples with respect to the number of raw samples from the new target domains used in our proposed method. Specifically, we measure the diversity using the LPIPS scores, and note two takeaway information: First, the diversity increases when more samples are available, and our empirical findings suggest that $N = 800 - 1000$ is a reasonable choice. Next, the generated data from our proposed sampling method exhibits reasonable diversity among samples within the same target domain. Finally, it is important to note that the LPIPS score does not directly reflect the quality of the generated images. For example, in the case of galaxy images under large domain shift, the LPIPS score may appear lower despite the good perceptual quality of the outputs, since such images often share similar dark backgrounds and structural patterns by nature.

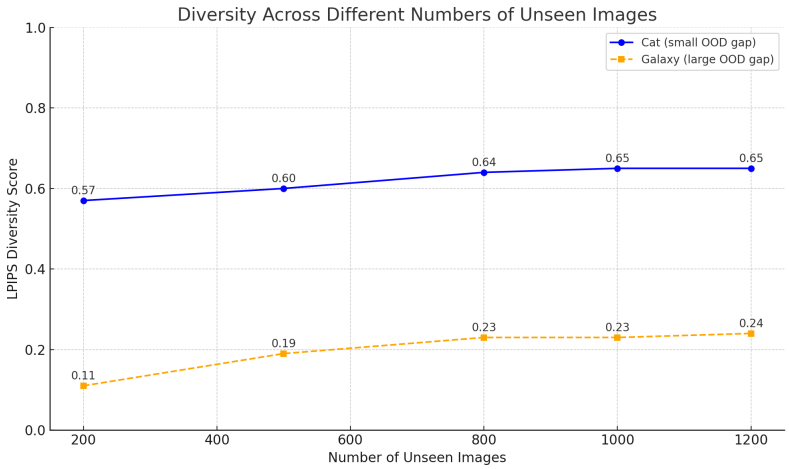


Figure 7: **Effect of the number of new samples from the target domains on generation diversity.** Results are shown for two OOD domains: Cat (small distribution gap) and Galaxy (large distribution gap) via diffusion model pre-trained on dogs.

Figure 8: **Visualization of the interpolation between two raw samples.** Uncurated samples are shown for two OOD domains: Cat (small distribution gap) and Galaxy (large distribution gap) via diffusion model pre-trained on dogs.

We also qualitatively visualize the generated data between the interpolation between two raw samples in Fig. 8. Overall, the results demonstrate noticeable visual variation across the full interpolation path, with higher similarity observed between samples that are spatially closer in the latent space. In practice, we typically select the samples from the middle of the interpolation, as it offers a better trade-off between quality and diversity.

Finally, Fig. 9 includes more uncurated generation results from our proposed sampling-centric method from frozen pre-trained unconditional diffusion models.

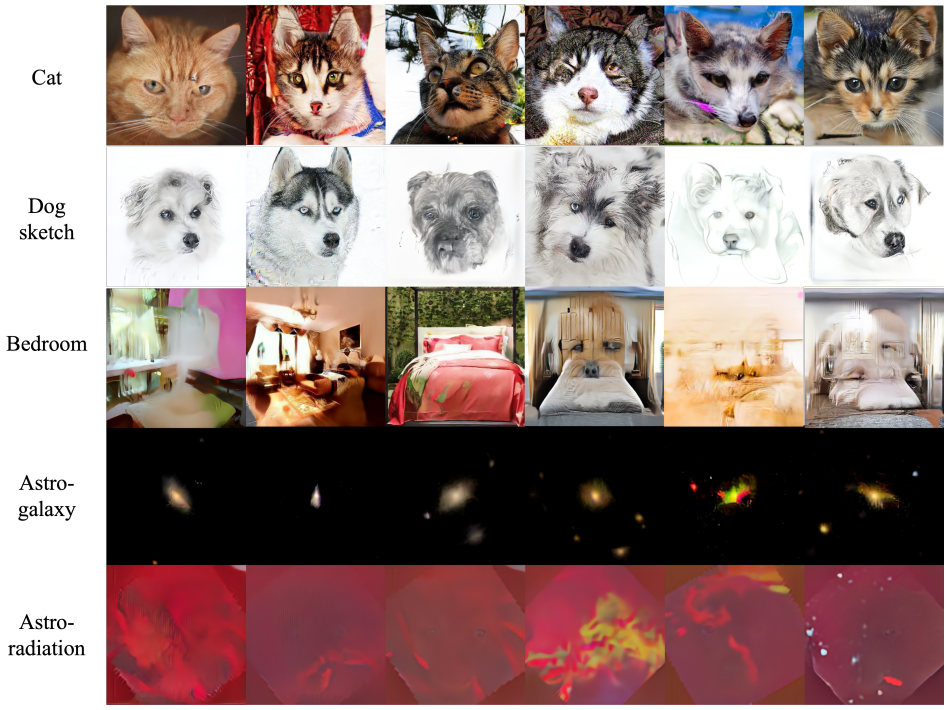


Figure 9: **More uncurated samples for various OOD domains generated via our proposed method from frozen diffusion models.**