Conditional Sampling from Frozen Generative Models: From Explicit Rules to Example-Based Guidance

Niccolò Dalmasso, Vamsi K. Potluru, Manuela Veloso
J.P. Morgan AI Research
New York
{niccolo.dalmasso, vamsi.k.potluru, manuela.veloso}@jpmchase.com

Abstract

Deep generative models have emerged as both scalable and high-fidelity solutions for generating high-quality synthetic data, effectively capturing the bulk of the training data distribution. However, these models often struggle to adequately generate samples that are rare, underrepresented or that satisfy user-defined conditions or constraints, which are valuable in fields such as finance and healthcare. Retraining generative modes from scratch or using expensive sampling-based methods to capture these targeted outcomes can be computationally prohibitive. To address this challenge, we propose a general framework that enables targeted generation of user-defined conditions from pretrained deep generative models without extensive retraining. Specifically, we address two practical scenarios. In scenarios where explicit rules can evaluate whether generated samples satisfy desired conditions, we propose to use contrastive learning to learn a latent space prior to guide generation towards rule-satisfying outcomes. In settings where only examples of the desired outcomes are provided, we adapt methodologies from the simulation-based inference literature to condition the generation process. Experiments demonstrate that our approach reliably produces condition-satisfying samples, significantly outperforming existing techniques on tabular data in terms of generation quality.

1 Introduction

Generating synthetic data that satisfy user-defined constraints or corresponding to rare events is a critical problem in generative modeling, especially with class imbalance or low-incidence scenarios. In finance, rare events such as fraudulent transactions or market crashes carry disproportionate importance despite their low frequency [56, 34], while in healthcare, the scarcity of data on rare diseases motivates the use of generative models to create synthetic data for research purposes [29]. Additionally, with the recent development of large language models (LLMs), fine-tuning of pretrained models and in-context learning approaches [8, 66] require user-defined data to cover specific tasks.

However, standard generative modeling approaches predominantly capture high-probability regions of the training data distribution, typically underrepresenting distribution tails or minority classes [9]. Training deep generative models on inherently imbalanced datasets often results in models biased toward common events, inadequately representing or missing lower-incidents events altogether [11, 7, 44]. This limitation motivates the adoption of post-hoc conditioning methods designed to bias synthetic data generation towards critical outcomes without requiring expensive modifications or retraining of the original generative architectures.

Many modern deep generative models synthesize data by first sampling from a latent space, typically Gaussian, and then transforming these samples into realistic data points. Models such as variational autoencoders (VAE, [42]), generative adversarial networks (GAN, [31]), normalizing flows [54] and

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance.

diffusion models [38] are usually trained unconditionally, meaning they aim to replicate the statistical distribution of the entire training dataset without specific constraints. In this work, we address two practical scenarios for such deep generative models by exploring the structure in their latent space.

In the first scenario, we assume an on-the-fly evaluator that can quantify whether generated samples satisfy desired conditions or rules. Examples include specific patterns in tabular datasets or temporal patterns in time series, such as volatility levels or trends. Our approach requires the evaluator to be able to quantitatively assess a condition but we do not restrict to differentiable conditions, thus providing versatility across diverse applications and domains. Given a pretrained generative model G, we introduce **C**ontrastive **L**atent **A**mplification via **I**nterpretable **M**apping (CLAIM), which uses contrastive learning to learn a mapping from a low-dimensional interpretable prior to embedding regions that preferentially generate condition-satisfying outputs, optimizing only the auxiliary parameters while keeping G frozen. Our work stems directly from post-hoc conditioning methods such a latent constraints [23], where regions in the latent space are explored using an actor-critic discriminative model and identify that samples from the same region of the latent space share similar structure, as well as more recent approaches on latent space explorations [3, 71, 1, 2], providing a lightweight interpretable alternative which does not require external labeled data or semantic direction in the latent space. Additionally, while other successful methods have achieved conditional generation by modifying specific generative models such as diffusion models [32, 14] or normalizing flows [25, 4], our approach is applicable to any generative model which generates samples by decoding them from a latent space. Finally, we note that the same goal could also be achieved by applying model corrections during the generation phase, as shown in robotics [46, 47].

In the second scenario, we assume direct condition evaluation may be computationally expensive, but practitioners possess examples of condition-satisfying samples that can provide guidance. We propose **S**imulation-based **P**osterior **I**nference for **R**elevant **E**xamples (SPIRE), which casts the conditional generation as a simulation based inference (SBI, [16]) problem, treating samples from a latent embedding as parameters and the frozen pretrained generative model G as a simulator. This is in line with what noted by [30], who show that one can use a transformer-based diffusion model to approximate any function model and conditionally sample from it. As in our case a frozen deep generative model is available, we directly apply SBI methods to infer a posterior distribution over latent embeddings responsible for generating similar condition-satisfying events.

2 CLAIM: Contrastive Latent Amplification via Interpretable Mapping

Our first scenario addresses cases where practitioners have access to an on-the-fly evaluator capable of assessing whether generated samples satisfy a given set of criteria or rules. This evaluator does not need to be differentiable and can incorporate complex domain-specific logic. Our proposed approach CLAIM learns a compact mapping g_ϕ from a low-dimensional space $\mathcal{H} \subseteq \mathbb{R}^k$ to the embedding of the pretrained generation $\mathcal{Z} \subseteq \mathbb{R}^d$, where ($k \ll d$). A low-dimensional space $\mathcal{H}$ with $k = 2, 3$ offers an interpretable mechanism to explore how different regions correspond to various types of condition-satisfying samples. We learn the mapping g_ϕ using a lightweight multi-layer perceptron (MLP), with LeakyReLu activations [68].

The training proceeds through three phases:

Phase I: Latent Space Exploration. We begin by exploring the pretrained generator's latent space to identify regions associated with the condition-satisfying samples, sampling a set of m d-dimensional space vectors $\{\mathbf{z}_i\}_{i=1}^M \sim \mathcal{N}(0, I_d)$, generating corresponding data samples $\mathbf{x}_i = G(\mathbf{z}_i)$, and evaluating each sample using a (potentially multivariate) condition function $c(\mathbf{x}_i)$[1]. This yields two sets of samples, $\mathcal{Z}^+ = \{\mathbf{z}_i : c(G(\mathbf{z}_i)) = 1\}$, the set of condition-satisfying samples, and $\mathcal{Z}^- = \{\mathbf{z}_i : c(G(\mathbf{z}_i)) = 0\}$, a set of samples that do not satisfy the conditions. While it might be likely that $|\mathcal{Z}^+| \ll |\mathcal{Z}^-|$, according to the given rules and constraints, the positive set provides crucial anchor points for the mapping training.

Phase II: Distribution Alignment. We pretrain the MLP for the mapping g_ϕ by learning the mapping between two Gaussians of dimensions k and d respectively, using the following loss:

[1]While this approach corresponds to Monte Carlo sampling [36], one could use more efficient techniques such as Bayesian optimization [60] or Parzen tree estimators [67].

$$\mathcal{L}_{\text{pretrain}} = \mathbb{E}_{\mathbf{h}\sim\mathcal{N}(0,I_k)}\left[\|\mathbf{z}_{\text{target}} - g_\phi(\mathbf{h})\|_2^2\right] + \lambda_{\text{cov}}\sum_{j=1}^{d}(\text{Var}[g_\phi(\mathbf{h})_j] - 1)^2 \tag{1}$$

where $\mathbf{z}_{\text{target}} \sim \mathcal{N}(0, I_d)$ represents samples from the original latent distribution, and the second term regularizes the covariance of the mapping g_ϕ to avoid dimension collapse.

Phase III: Contrastive Learning. Finally, we train the mapping g_ϕ via constrastive learning, pushing generated latent samples towards condition-satisfying samples $\mathbf{z}^+ \in \mathcal{Z}^+$ and away from negative samples $\mathbf{z}^- \in \mathcal{Z}^-$. More specifically, for each mapped sample $\mathbf{z} = g_\phi(\mathbf{h})$, we optimize:

$$\mathcal{L}_{\text{contrastive}}(\mathbf{z}) = -\log\left(\frac{\sum_{\mathbf{z}^+\in\mathcal{Z}^+}\exp(\text{cs}(\mathbf{z},\mathbf{z}^+)/\tau)}{\sum_{\mathbf{z}^+\in\mathcal{Z}^+}\exp(\text{cs}(\mathbf{z},\mathbf{z}^+)/\tau) + \beta\sum_{\mathbf{z}^-\in\mathcal{Z}^-}\exp(\text{cs}(\mathbf{z},\mathbf{z}^-)/\tau)}\right), \tag{2}$$

where $\text{cs}(\mathbf{z}_1, \mathbf{z}_2) = \mathbf{z}_1^T\mathbf{z}_2/\|\mathbf{z}_1\|\|\mathbf{z}_2\|$ is the cosine similarity. The loss function (2) is a version of the N-pair loss function [61], to encourage the generation of samples that exhibit high cosine similarity with condition-satisfying synthetic data, where the temperature parameter τ controls the sharpness of the similarity, and β emphasizes the importance of avoiding non-condition-satisfying samples.

Additionally, in order to prevent mode collapse and avoid the network focusing on a single region of the latent embedding, we incorporate an maximum mean discrepancy (MMD, [33]) regularization term: $\mathcal{L}_{\text{reg}} = \text{MMD}(\{g_\phi(\mathbf{h}^{(i)})\}_{i=1}^N, \mathcal{N}(0, I_d))$. By choosing a Gaussian distribution $\mathcal{N}(0, I_d)$ to regularize our mapped samples $\mathbf{h}$, we ensure coverage of the embedding space. Finally, we note that in case of conditionally sampling based on a different set of constraints, only Phase III would need to be run, as the samples in Phase I can be evaluated according to the new set of constraints.

3 SPIRE: Simulation-based Posterior Inference for Relevant Examples

The second scenario addresses situations where evaluating sample conditions is computationally expensive or time-consuming, but modelers or practitioners might have a curated set of examples representing the conditions of interest. This setting is common in domains like medical diagnosis, where obtaining expert annotations is costly, or financial risk modeling, where historical rare events are key in the modeling process.

We propose SPIRE, which reframes the generation of condition-satisfying samples as a simulation based inference task. We start with a collection of observed condition-satisfying samples $\{\mathbf{x}_{\text{obs}}^{(l)}\}_{l=1}^{L_{\text{obs}}}$, and our objective becomes inferring the posterior distribution of $p(\mathbf{z}|\mathbf{x}_{\text{obs}})$, to capture the regions of the embedding space $\mathcal{Z}$ which can generate similar events. In this context, the forward simulator is the frozen pretrained generator: $\text{simulator}(\mathbf{z}) = \mathbf{f}(G(\mathbf{z} + \epsilon))$, to which we add a small regularization noise term ($\sigma \approx 10^{-4}$). This noise term enforces the stochasticity of the simulator, under the assumptions that samples within the same regions of the embedding generate data points $\mathbf{x}$ that are close to each other. For our approach, we use neural posterior score estimation (NPSE, [28]), with a large uniform prior of the embedding space $p(\mathbf{z}) = \mathcal{U}([-a, a]^d)$, setting $a = 5$.

Note that by recasting condition-satisfying sampling as a SBI problem we not only inherit the wealth of posterior inference models developed in the SBI literature, but we gain in efficiency as our approach works even with *only a single* condition-satisfying sample, i.e., when $L = 1$.

4 Experiments

We evaluate both our proposed approaches CLAIM and SPIRE on three tabular UCI datasets, choosing conditions with varying degrees of occurrence in the training data: (1) Adult dataset [6], with conditions being individuals younger than 25 and earning more than USD50,000 (0.3% occurrence), (2) Wine Quality dataset [15], with condition being high-quality wine with rating 8+ (1.5% occurrence) and (3) Abalone dataset [49], with condition being the animal being older than 10 years (12.1% occurrence). We train three tabular models, TVAE (tabular variation autoencoder [69]), CTGAN (conditional tabular generative adversarial model [69]) and TabDDPM (tabular diffusion model [43]). We compare claim against two baselines, Latent Constraints [23], which learns value functions to identify condition-satisfying latent regions in a variational autoencoder, and NCP (Neural Conditional Priors [1]), which provides a better exploration of the latent space of a variational autoencoder and represents a baseline for the true occurrence of the condition-satisfying samples in the embedding space. Note that for SPIRE, we utilize a single sample $\mathbf{x}_{\text{obs}}$.

Tables 1 and 2 report results by evaluating the *condition sampling ratio* of the generated data, the *improvement ratio* over the occurrence in the training data, the *Wasserstein distance* from condition-satisfying samples in the test set versus the generated ones, the *Vendi score* [26] to evaluate diversity of the generated samples and the *average Pearson correlation* of the generated samples with the condition-satisfying samples in the test set. Values reported include mean and average over 5 runs. CLAIM provides a consistent alternative to the baselines, by generating condition-satisfying samples with a better distributional quality and diversity than Latent Constraint. On the challenging Adult dataset with only 0.3% natural occurrence, CLAIM with TVAE achieves 100% condition satisfaction with 296x improvement, while Latent Constraints fails to generate due to mode collapse in the actor/critic architecture. SPIRE demonstrates remarkable effectiveness despite using only a single condition-satisfying sample. TabDDPM consistently show lower performance with both methods, likely due to to a potential overfitting to the training data or to the nature of their decoding process.

Table 1: Comparison of conditional generative methods with an on-the-fly evaluator available.

Dataset	Model	Method	% Rare (↑)	Improvement (↑) Ratio	Wasserstein (↓) Distance	Vendi (↑) Score	Avg Pearson (↑) Corr
Wine Red (1.5%)		NCP	0.28 ± 0.10	0.18 ± 0.06	2.42 ± 0.11	6.41 ± 0.11	0.90 ± 0.02
		Latent Constraint	**100.00 ± 0.00**	**64.00 ± 0.00**	179.74 ± 0.00	3.41 ± 0.02	0.94 ± 0.02
	TVAE	Ours	65.24 ± 0.71	41.75 ± 0.45	**2.42 ± 0.11**	3.15 ± 0.02	0.93 ± 0.01
	CTGAN	Ours	67.40 ± 1.04	43.14 ± 0.67	5.04 ± 0.02	3.05 ± 0.02	0.93 ± 0.02
	TabDDPM	Ours	24.81 ± 0.73	15.90 ± 0.46	3.01 ± 0.23	**6.75 ± 0.02**	**0.95 ± 0.01**
Abalone (12.1%)		NCP	11.50 ± 0.87	0.95 ± 0.07	0.12 ± 0.02	4.50 ± 0.04	0.25 ± 0.01
		Latent Constraint	**100.00 ± 0.00**	**8.28 ± 0.00**	8.49 ± 0.02	3.74 ± 0.03	0.25 ± 0.00
	TVAE	Ours	**100.00 ± 0.00**	**8.28 ± 0.00**	1.26 ± 0.01	3.97 ± 0.03	0.25 ± 0.01
	CTGAN	Ours	**100.00 ± 0.00**	**8.28 ± 0.00**	1.73 ± 0.02	3.82 ± 0.01	0.24 ± 0.01
	TabDDPM	Ours	54.56 ± 1.98	4.52 ± 0.16	**0.32 ± 0.02**	**5.13 ± 0.04**	**0.26 ± 0.01**
Adult (0.3%)		NCP	0.44 ± 0.15	1.30 ± 0.44	417.06 ± 109.86	7.19 ± 0.04	0.32 ± 0.01
		Latent Constraint	0.00 ± 0.00	0.00 ± 0.00	-	-	-
	TVAE	Ours	**100.00 ± 0.00**	**296.02 ± 0.00**	1066.39 ± 27.59	3.71 ± 0.06	0.32 ± 0.01
	CTGAN	Ours	13.68 ± 0.76	40.50 ± 2.25	311.58 ± 0.03	3.68 ± 0.04	0.34 ± 0.01
	TabDDPM	Ours	21.82 ± 0.69	64.5 ± 1.57	**226.25 ± 31.82**	**6.59 ± 0.11**	**0.37 ± 0.02**

Table 2: Conditional generation with our approach SPIRE, given one condition-satisfying example.

Dataset Name	Model Type	% Rare (↑)	Improvement (↑) Ratio	Vendi Score (↑) Rare	Wasserstein (↓) Distance
Wine Red (1.5%)	TVAE	**47.32 ± 2.02**	**30.28 ± 1.29**	3.61 ± 0.06	3.28 ± 0.09
	CTGAN	43.64 ± 1.05	27.93 ± 0.67	3.62 ± 0.05	4.87 ± 0.11
	TabDDPM	25.24 ± 0.53	5.91 ± 0.34	**4.46 ± 0.24**	**2.78 ± 0.27**
Abalone (12.1%)	TVAE	**92.44 ± 0.30**	**7.65 ± 0.02**	**5.14 ± 0.05**	**0.38 ± 0.01**
	CTGAN	92.06 ± 0.72	7.62 ± 0.06	4.67 ± 0.05	0.50 ± 0.01
	TabDDPM	42.58 ± 1.89	3.52 ± 0.16	3.61 ± 0.12	0.48 ± 0.05
Adult (0.3%)	TVAE	**52.02 ± 1.53**	**153.99 ± 4.53**	2.97 ± 0.02	2483.17 ± 40.57
	CTGAN	11.88 ± 0.49	26.29 ± 1.44	**3.34 ± 0.11**	**675.76 ± 106.20**
	TabDDPM	9.80 ± 0.20	5.33 ± 0.59	2.89 ± 0.12	976.65 ± 700.78

5 Conclusions and Future Work

We introduce two complementary approaches for conditional sampling from frozen generative models: CLAIM for scenarios with on-the-fly constraints evaluation and SPIRE for example-based guidance. Our experimental results demonstrate that both methods achieve effective conditional sampling without expensive retraining, with CLAIM generating samples close in quality to the original condition-satisfying data and SPIRE providing robust performance even when provided a single condition-satisfying example. Future research directions include extending our work to additional data modalities, as well as targeting language generation by large language models (LLMs).

Disclaimer

This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan") and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

[1] J. Aneja, A. Schwing, J. Kautz, and A. Vahdat. A contrastive learning approach for training variational autoencoder priors. *Advances in neural information processing systems*, 34:480–493, 2021.

[2] G. Arvanitidis, B. Georgiev, and B. Schölkopf. A prior-based approximate latent riemannian metric. *arXiv preprint arXiv:2103.05290*, 2021.

[3] G. Arvanitidis, L. K. Hansen, and S. Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

[4] S. Asghar, Q.-X. Pei, G. Volpe, and R. Ni. Efficient rare event sampling with unsupervised normalizing flows. *Nature Machine Intelligence*, 6(11):1370–1381, 2024.

[5] A. Bansal, H.-M. Chu, A. Schwarzschild, S. Sengupta, M. Goldblum, J. Geiping, and T. Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 843–852, 2023.

[6] B. Becker and R. Kohavi. Adult. *UCI Machine Learning Repository*, 10:C5XW20, 1996.

[7] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7327–7347, 2021.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[9] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[11] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. P. Chen. A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, 57(6):137, 2024.

[12] W.-C. Cheng, T.-H. Mai, and H.-T. Lin. From smote to mixup for deep imbalanced classification. In *International Conference on Technologies and Applications of Artificial Intelligence*, pages 75–96. Springer, 2023.

[13] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

[14] A. Coletta, S. Gopalakrishnan, D. Borrajo, and S. Vyetrenko. On the constrained time-series generation problem. *Advances in Neural Information Processing Systems*, 36:61048–61059, 2023.

[15] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553, 2009.

[16] K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[17] D. Dablain, B. Krawczyk, and N. V. Chawla. Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE transactions on neural networks and learning systems*, 34(9):6390–6404, 2022.

[18] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[19] A. Doucet, W. Grathwohl, A. G. Matthews, and H. Strathmann. Score-based diffusion meets annealed importance sampling. *Advances in Neural Information Processing Systems*, 35:21482–21494, 2022.

[20] Y. Du, C. Durkan, R. Strudel, J. B. Tenenbaum, S. Dieleman, R. Fergus, J. Sohl-Dickstein, A. Doucet, and W. S. Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pages 8489–8510. PMLR, 2023.

[21] Y. Du, S. Li, and I. Mordatch. Compositional visual generation with energy based models. *Advances in Neural Information Processing Systems*, 33:6637–6647, 2020.

[22] Y. Du and I. Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019.

[23] J. H. Engel, M. D. Hoffman, and A. Roberts. Latent constraints: Learning to generate conditionally from unconditional generative models. In *ICLR (Poster)*, 2018.

[24] V. A. Fajardo, D. Findlay, C. Jaiswal, X. Yin, R. Houmanfar, H. Xie, J. Liang, X. She, and D. B. Emerson. On oversampling imbalanced data with deep conditional generative models. *Expert Systems with Applications*, 169:114463, 2021.

[25] S. Falkner, A. Coretti, S. Romano, P. Geissler, and C. Dellago. Conditioning normalizing flows for rare event sampling. *arXiv preprint arXiv:2207.14530*, 2022.

[26] D. Friedman and A. B. Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

[27] M. Gabrié, G. M. Rotskoff, and E. Vanden-Eijnden. Adaptive monte carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119(10):e2109420119, 2022.

[28] T. Geffner, G. Papamakarios, and A. Mnih. Compositional score modeling for simulation-based inference. In *International Conference on Machine Learning*, pages 11098–11116. PMLR, 2023.

[29] L. Gibson, M. Hoerger, and D. Kroese. A flow-based generative model for rare-event simulation. *arXiv preprint arXiv:2305.07863*, 2023.

[30] M. Gloeckler, M. Deistler, C. Weilbach, F. Wood, and J. H. Macke. All-in-one simulation-based inference. *arXiv preprint arXiv:2404.09636*, 2024.

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[32] A. Graikos, S. Yellapragada, and D. Samaras. Conditional generation from unconditional diffusion models using denoiser representations. *arXiv preprint arXiv:2306.01900*, 2023.

[33] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

[34] J. Gu, X. Zhang, and G. Wang. Beyond the norm: A survey of synthetic data generation for rare events. *arXiv preprint arXiv:2506.06380*, 2025.

[35] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020.

[36] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[37] Y. He, N. Murata, C.-H. Lai, Y. Takida, T. Uesaka, D. Kim, W.-H. Liao, Y. Mitsufuji, J. Z. Kolter, R. Salakhutdinov, et al. Manifold preserving guided diffusion. *arXiv preprint arXiv:2311.16424*, 2023.

[38] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[39] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[40] S. Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *Advances in Neural Information Processing Systems*, 37:66743–66772, 2024.

[41] B. Kim and J. C. Ye. Denoising mcmc for accelerating diffusion-based generative models. *arXiv preprint arXiv:2209.14593*, 2022.

[42] D. P. Kingma, M. Welling, et al. Auto-encoding variational bayes. 2013.

[43] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International conference on machine learning*, pages 17564–17579. PMLR, 2023.

[44] B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in artificial intelligence*, 5(4):221–232, 2016.

[45] M. Kwon, J. Jeong, and Y. Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

[46] Ç. Meriçli, M. Veloso, and H. L. Akın. Task refinement for autonomous robots using complementary corrective human feedback. *International Journal of Advanced Robotic Systems*, 8(2):16, 2011.

[47] C. Meriçli, M. Veloso, and H. L. Akın. Multi-resolution corrective demonstration for efficient task execution and refinement. *International Journal of Social Robotics*, 4(4):423–435, 2012.

[48] T. Müller, B. McWilliams, F. Rousselle, M. Gross, and J. Novák. Neural importance sampling. *ACM Transactions on Graphics (ToG)*, 38(5):1–19, 2019.

[49] W. Nash, T. Sellers, S. Talbot, A. Cawthorn, and W. Ford. Abalone. UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C55C7W.

[50] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4467–4477, 2017.

[51] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in neural information processing systems*, 29, 2016.

[52] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[53] W. Nie, A. Vahdat, and A. Anandkumar. Controllable and compositional generation with latent-space energy-based models. *Advances in Neural Information Processing Systems*, 34:13497–13510, 2021.

[54] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[55] Y.-H. Park, M. Kwon, J. Choi, J. Jo, and Y. Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36:24129–24142, 2023.

[56] V. K. Potluru, D. Borrajo, A. Coletta, N. Dalmasso, Y. El-Laham, E. Fons, M. Ghassemi, S. Gopalakrishnan, V. Gosai, E. Kreačić, et al. Synthetic data applications in finance. *arXiv preprint arXiv:2401.00081*, 2023.

[57] X. Ren, T. Yang, Y. Wang, and W. Zeng. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. *arXiv preprint arXiv:2102.10543*, 2021.

[58] Y. Shen, C. Yang, X. Tang, and B. Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2004–2018, 2020.

[59] M. H. Shirvan, M. H. Moattar, and M. Hosseinzadeh. Deep generative approaches for oversampling in imbalanced data classification problems: A comprehensive review and comparative analysis. *Applied Soft Computing*, 170:112677, 2025.

[60] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

[61] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

[62] S. Spanbauer, C. Freer, and V. Mansinghka. Deep involutive generative models for neural mcmc. *arXiv preprint arXiv:2006.15167*, 2020.

[63] C. Vashist, S. Peng, and K. Li. Rejection sampling imle: Designing priors for better few-shot image synthesis. In *European Conference on Computer Vision*, pages 441–456. Springer, 2024.

[64] A. Verine, M. S. Pydi, B. Negrevergne, and Y. Chevaleyre. Optimal budgeted rejection sampling for generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 3367–3375. PMLR, 2024.

[65] A. Voynov and A. Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning*, pages 9786–9796. PMLR, 2020.

[66] Z. Wang, A. W. Yu, O. Firat, and Y. Cao. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*, 2021.

[67] S. Watanabe. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint arXiv:2304.11127*, 2023.

[68] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[69] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*, 2019.

[70] Y. Yang, X. Liu, D. Wang, Q. Sui, C. Yang, H. Li, Y. Li, and T. Luan. A ce-gan based approach to address data imbalance in network intrusion detection systems. *Scientific Reports*, 15(1):7916, 2025.

[71] O. K. Yüksel, E. Simsar, E. G. Er, and P. Yanardag. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14263–14272, 2021.

[72] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

A Literature Review

In this section we elaborate on the details on how our proposed methods relate to the broader literature of post-hoc conditioning approaches for frozen pretrained models.

Diffusion Model Guidance Methods. A category of post-hoc conditioning approaches has emerged for diffusion models [38], for which guidance techniques can be incorporated without architectural changes. Among methods that require training-time modifications, [39] train a diffusion model both conditionally and unconditionally, randomly dropping conditioning information during training, and by interpolating between the two at inference time. [20] show that by predicting the scalar energy directly rather than the diffusion score, one can enable compositional operatoins, while [72] improve conditioning of frozen text-to-image models by creating trainable copies of encoder layers. Other approaches include the training of a separate classifiers. [18] train a separate classifier on noisy images across all diffusion timesteps, and include the the gradient information in the diffusion model score, with [5] extending classifier guidance to any differentiable loss function. However, [52] note that classifier-free guidance achieves better human evaluation. Finally, among the training-free approaches, [13] approximate posterior guidance without training by computing likelihood in inverse problem settings, [40] use attention energy modulation for condition-free guidance by blurring attention queries, and [37] refine the guided diffusion by projecting onto the data manifold using gradient projection. Our proposed methods CLAIM and SPIRE do not require any training-time adjustment, and operate in latent space rather than relying on gradient-based guidance. However, one would expect bespoke techniques for diffusion models to achieve better results in conditional generation especially for high-dimensional generation task, like image or text-to-image generation.

Latent Space Manipulation and Exploration. Exploring latent spaces has first been a prominent research direction for generative adversarial networks. [35] provides an unsupervised discovery of interpretable directions in pretrained GAN models by applying PCA on intermediate layer activations of pretrained GANs, with layer-wise perturbations along principal components enabling semantic editing without attribute classifiers. A similar approach is also presented by [65], where a reconstructor network is trained to identify interpretable directions through orthogonality constraints and Jacobian penalties, and by [58], which use linear SVM in latent space to find semantic boundaries. The introduction of a secondary classification model over any generative model latent space was pioneered by [51, 50], who show that a plug & play approach with a prior over the latent space can dramatically improve generation quality. Finally, energy-based models have also been show to have desirable properties in conditional generations [22, 21, 53]. Our proposed approach CLAIM is inspired by the success of disentanglement of latent spaces via contrastive learning in generative models [71, 57]. We also note that further work has explored the native semantic properties of latent space in diffusion models [45], as well as the geometric structure of diffusion models latent spaces [55].

Sampling Methods. Although simple rejection sampling might be computationally inefficient to generate conditionally at scale, recent work has developed rejection sampling approaches in settings with limited sampling budgets or data availability [64, 63]. Additionally, methods that incorporate traditional annealed importance sampling [19] or Monte Carlo Markov Chain (MCMC) in diffusion models [41] could also be modified conditionally for the generation of condition-satisfying samples. Our proposed approach SPIRE would also benefit from methods that improve the fit and the sampling from the posterior distribution over the latent space, such as [27, 62, 48]. Finally, the generation of data points is also key for supervised learning in presence of imbalanced datasets. Oversampling approaches like SMOTE [10] and more recent deep learning variants [17, 12, 59, 24, 70] aim to increase the number of rare points in the training set by creating synthetic rare events. Our proposed approaches CLAIM and SPIRE could directly be used to generate data points from the minority class to augment imbalanced datasets for downstream supervised learning performance improvement.