Quantifying Feature Emergence in the Reverse Processes of Diffusion Probabilistic Models

Asher Stout*, Andrew Lensen†, Marcus Frean‡
School of Engineering and Computer Science, Te Herenga Waka Victoria University of Wellington
Wellington, New Zealand
Email: *fufupignz@gmail.com, †andrew.lensen@vuw.ac.nz, ‡marcus@ecs.vuw.ac.nz

***Abstract*—Diffusion Probabilistic Models (DPMs) are a nascent class of Score-based generative model that have routinely demonstrated high sample quality in a variety of applications. Unlike other generative models DPMs do not rely on a well-behaved latent space during sampling, instead utilizing a parameterized Markov chain called the *reverse process* to iteratively generate samples from pure noise. Although passive observations have been made about the dynamics of the reverse process no formal study has analysed the critical transition from noise to high-level features. Motivated by this dearth of research we propose three methods that capture the rate of feature emergence in the reverse process. We name these methods *Stepwise Entropy*, *Label Variance* and *Sample Consistency*. From experiments across several popular DPMs and image datasets we report previously unobserved behaviours of the reverse process.**

***Index Terms*—Diffusion processes, deep learning, generative models, Markov processes**

I. Introduction

Deep Generative Models (DGMs) have long captivated the public and academic imagination and have proved a motivating force in Machine Learning research. DGMs leverage deep neural networks to learn a transformative process that enables the synthesis of novel samples qualitatively identical to an original domain. DGMs have become highly specialized across an ever-broadening variety of generative tasks and data domains including image translation [1], inpainting [2], and superresolution [3] tasks.

Historically two models have dominated most generative applications, Variational Autoencoders [4] and Generative Adversarial Networks [5]. VAEs and GANs endeavour to learn a lower-dimensional representation of an input data distribution x by encoding it in a *latent space* z with a predefined dimensionality. Synthesis is then performed by generating a random sample in z and transforming it back into the x domain via a DNN. This approach has the added benefit of requiring the model to learn an efficient latent representation of abstract features present in x, providing critical insight into a model's learning process and allowing for direct influence over the features of synthesized data through manual modifications in the z-space of samples.

In recent years these latent-space driven DGMs have been outclassed by a nascent Score-based approach known as Diffusion Probabilistic Models (DPMs) [6]. The approach taken by DPMs to synthesis is antithetical to the dimensionality reduction principle that drives the learning of the latent space in VAEs and GANs. DPMs operate in the same dimensions as the original data domain and do not explicitly encode any information into the initial states of samples. Instead, DPMs iteratively recover a synthesized signal from pure noise by removing small noise perturbations at each step of a parameterized Markov Chain called the *reverse* or *denoising process*. DPMs have rapidly achieved state-of-the-art performances in a plethora of generative tasks and consistently outperform VAEs and GANs in terms of fidelity in both the image and audio domains [7, 8].

While diffusion models have witnessed a massive influx of academic interest the trends that emerge during the reverse process remain poorly explored. Unlike the aforementioned DGMs that generate the initial sample in a meaningful z-space, the starting state of samples in diffusion models is pure noise that encodes no prior information. Early denoising steps convert this noise into high-level features that are gradually refined into low-level details over the remainder of the process; we provide a visualization of this behaviour in Figure 1. The transitional phase that sees the creation of high-level features from noise is crucial to the performance of DPMs for tasks in the generative domain yet a formal exploration of this timeframe has proved elusive.

We intend to further our understanding of this littoral zone by undertaking an analysis of the timeframe and rates of feature emergence in the reverse process. Our definition of feature emergence is inclusive to the appearances of both high- and low-level features. To assist us we propose three novel approaches to quantify the level of feature emergence at different timesteps of the reverse process. We verify our methods on three seminal DPM implementations, DDPM [6], DDIM [9] and Improved DDPM [10].

In summary, our contributions are:

1. We review previously observed patterns of feature emergence in the reverse processes of a range of diffusion models
2. We propose three approaches tailored for diffusion models that measure rates of feature emergence, which we call *stepwise entropy*, *label variance* and *sample consistency* [9]
3. We provide results from experiments on CIFAR-10 [11], CelebA [12] and LSUN-bedroom [13]

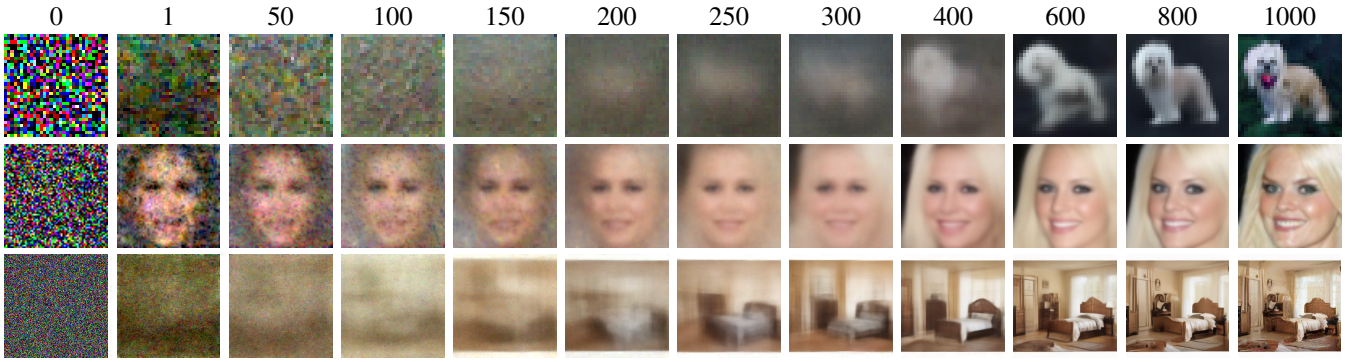

Fig. 1. Latents of the reverse process of [6] at different stages of denoising on the CIFAR-10 (top), CelebA (middle), and LSUN-bedroom (bottom) datasets. Note the qualitative trends in sample features; by $t = 400$ most high-level features have been synthesized and the remaining denoising steps provide detailed refinements to the content. The delayed emergence of features in CIFAR-10 can be attributed to the variety of image labels in the dataset. The lack of visually concrete qualities in $t < 400$ does not necessarily discredit the presence of latent information. This is the region we seek to explore using quantitative methods to track the rate of feature emergence.

II. Background

Here we provide a general overview of diffusion models beginning with the original formulation in [6] and continuing into two extensions, Denoising Diffusion Implicit Models [9] and Improved DDPMs [10].

A. Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DPMs) [6] consist of two parameterized Markov chains known as the *forward* or *diffusion process* and *reverse process*. Under the forward process a distribution $x_0 \sim q(x_0)$ is incrementally perturbed with Gaussian noise with variance set according to a linear schedule β_t over T steps:

$$\begin{aligned} q(x_{1:T}|x_0) &= \prod_{t=1}^{T} q(x_t|x_{t-1}), \text{ where} \\ q(x_t|x_{t-1}) &= \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \end{aligned} \tag{1}$$

A critical parameter choice lies in the selection of T, as at $t = T$ the distribution $q(x_t)$ should be able to be approximated by the Gaussian noise signal; a common value to achieve is $T = 1000$. For both the forward and reverse processes the diffused steps $x_{1:T}$ are known as the *latents* of the model. In line with the encoder network of a VAE the forward process is only leveraged during training and is discarded at sampling time.

The reverse process aims to recover x_0 from the final latent x_T over $\mathcal{T}$ denoising steps (where $T = \mathcal{T}$ in [6]):

$$\begin{aligned} p_\theta(x_{0:T}) &= p_\theta(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \text{ where} \\ p_\theta(x_{t-1}|x_t) &= \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \end{aligned} \tag{2}$$

which is approximated via a parameterization θ, represented in [6] as a DNN following a U-Net [14] architecture that takes an additional input in the form of a sinusoidal t embedding [15]. In [6] sample quality improves when the learned Σ_θ term in (2) is substituted with a scheduled variance $\sigma_t^2 I$ where $\sigma_t^2 = \beta_t$. This change results in the network having to learn exclusively the $\mu_\theta(x_t, t)$ term at each denoising step.

[6] defines a parameterization of μ_θ based on estimating the perturbation at each t in (1), $\epsilon_\theta(x_t, t)$. This noise prediction form was found to perform better than both predicting μ_t directly as well as a parameterization based on x_0. Under noise prediction θ is trained by minimizing a reweighted version of the original variational lower bound (VLB) with a closed-form definition of (1):

$$\mathcal{L}(\theta) := \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|\right] \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, I)$, $\bar{a}_t = \prod_{s=0}^{t} \alpha_s$, and $a_t = 1 - \beta_t$. A major departure of (3) from the original VLB term lies in its approach to weighting loss terms. Losses corresponding to low t contribute less to learning than high t owing to the reduction of the noise signal as $T \to 0$. The model therefore prioritizes denoising early into the reverse process when the values of $\sqrt{1-\bar{\alpha}_t}\epsilon$ are large.

By providing synthetic values of $x_T \sim q(x_T)$ the reverse process becomes a generative one. Sampling is performed by removing the estimated noise from the sample at each t beginning with the final latent x_T:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z \tag{4}$$

where $z \sim \mathcal{N}(0, I)$ if $t > 1$, else 0. Figure 1 visualizes the generative process for DDPM.

B. Denoising Diffusion Implicit Models

A drawback of DDPM is its restriction to Markovian forward processes. This requires the reverse process model each latent sequentially during sampling, making diffusion models significantly more computationally inefficient to generate from compared to other high-performance DGMs. Denoising Diffusion Implicit Models [9] overcome this limitation by generalizing the forward process, showing that non-Markovian forward processes can be substituted at sampling time without requiring model retraining. To achieve this the authors define

a sequence of inference distributions indexed by a vector $\sigma \in \mathbb{R}^T_{\geq 0}$ and rewrite (1) in non-Markovian form,

$$q_\sigma(x_t|x_{t-1}, x_0) = \frac{q_\sigma(x_{t-1}|x_t, x_0)q_\sigma(x_t|x_0)}{q_\sigma(x_{t-1}|x_0)} \tag{5}$$

Furthermore, while noise prediction is still leveraged for training the model the reverse process is redefined in terms of x_0, $f^t_\theta(x_t) = (x_t - \sqrt{1-\alpha_t}\epsilon^t_\theta(x_t))/\sqrt{\alpha_t}$:

$$p^t_\theta(x_{t-1}|x_t) = \begin{cases} \mathcal{N}(f^1_\theta(x_1), \sigma^2_1 I) & \text{if } t = 1 \\ q_\sigma(x_{t-1}|x_t, f^t_\theta(x_t)) & \text{otherwise} \end{cases} \tag{6}$$

where training is attained by optimizing a surrogate in the form of the original VLB from DDPM. When the parameters of ϵ_θ are not shared across t the optimal solutions of DDIM's VLB and (3) are shown to be identical [9]. A pretrained DDPM model therefore becomes a candidate solution for modelling many non-Markovian forward processes parameterized by σ.

Sampling follows a modified form of (4) while conforming to (6),

$$\begin{aligned} x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{\alpha_{t-1}}\epsilon^t_\theta(x_t)}{\sqrt{\alpha_t}}\right) \\ + \sqrt{1 - \alpha_{t-1} - \sigma^2_t}\epsilon^t_\theta(x_t) + \sigma_t z \end{aligned} \tag{7}$$

where the parameter σ_t controls the choice of generative process, with certain values being equivalent to Markovian diffusion [9]. A case of interest is when $\sigma_t = 0$ for all t; with this (5) becomes deterministic and the z term is nullified in (7), giving rise to an *implicit probabilistic model* where the forward process is known and no stochasticism occurs at sampling time.

An important observation of DDIM is the decoupling of $\mathcal{T}$ from T. (3) does not depend on a specific forward process provided a fixed $q_\sigma(x_t|x_0)$, which enables the reverse process to consider timesteps from a subset of T where $\mathcal{T} < T$. With minor adjustments to (7) sampling can be performed in $\mathcal{T}$ denoising steps, potentially providing large improvements to the sampling speed. Experiment results validate this approach, with samples comparable to DDPM generated in as few as 20 denoising steps [9].

C. Improved DDPMs

While diffusion models are capable of synthesizing high-quality samples, in practice they have not attained log-likelihoods competitive with alternative DGMs. Better log-likelihoods are attributed to greater mode coverage and may yield further improvements to sample quality. Here lies the motivation for the proposals made by Improved DDPM [10], which extend DDPM in several simultaneous capacities parallel to DDIM [9]. The alterations encompass three prominent changes: learned Σ_θ, cosine β_t and importance sampling during training.

Under the original DDPM formulation Σ_θ is set as a time-dependent constant $\sigma^2_t I$ equivalent to β_t. While this approximation is understandable in the sample quality context a learned Σ_θ has much greater potential for improved log-likelihoods, however predicting the variance directly is difficult [6]. To provide stability Improved DDPM parameterizes Σ_θ as an interpolation between β_t and $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ in the log domain,

$$\Sigma_\theta(x_t, t) = \exp(v \log \beta_t + (1 - v) \log \tilde{\beta}_t) \tag{8}$$

where v is the model output with a single component per dimension. As no learning signal for Σ_θ is propagated in (3) a hybrid objective function is defined leveraging the original VLB:

$$\mathcal{L}_{hybrid} = (3) + \lambda \mathcal{L}_{vlb} \tag{9}$$

with λ down-weighting $\mathcal{L}_{vlb}$ to prevent overwhelming (3) and a stop-gradient on $\mathcal{L}_{vlb}$ preventing double-feedback to the μ_θ term.

Setting β_t as a linear schedule was found to negatively impact sampling quality at lower image resolutions. Uneven noise perturbations in the forward process cause $\bar{\alpha}_t$ to converge to ≈ 0 after 80% of the diffusion, resulting in later noising steps perturbing an already-diffused input and providing very little effective learning in the reverse process. Instead, a cosine noise schedule defined in terms of $\bar{\alpha}_t$ is leveraged to ensure noising occurs evenly across all timesteps,

$$\bar{\alpha}_t = \frac{f(t)}{f(0)},\ f(t) = \cos\left(\frac{t/T + s}{1 + s} \cdot \frac{\pi}{2}\right)^2 \tag{10}$$

s represents a small offset to prevent insignificant values of β_t when $t \to 0$ and was found to improve ϵ predictions early in the forward process.

An unexpected side-effect of optimizing the $\mathcal{L}_{vlb}$ term in (9) was the severity of gradient noise. In [6] t is sampled uniformly during each training step; owing to the difference in magnitudes of the terms in (8) this potentially injects unnecessary noise. To combat this importance sampling is utilized for t selection, where a history of 10 recent ts is maintained for each step's loss term. Using this technique gradients were smoother for the $\mathcal{L}_{vlb}$ term and the highest log-likelihoods of the model were achieved.

While other important modifications are proposed by the authors in [10] we direct the reader to review these at their leisure as we did not utilize the proposed faster sampling scheme during out experiments, opting instead to sample using the original T steps.

III. Feature Emergence in the Reverse Process

Early improvements to the quality of samples are a well-established pattern in the reverse processes of diffusion models. Qualitative trends of CIFAR-10 sample fidelity presented in [6] are consistent with our provided examples in Figure 1 where most high-level features are established by the middle of the reverse process. A similar analysis of Fréchet Inception Distances confirms this, with scores seeing only limited improvement until the 350^{th} timestep where a rapid improvement occurs until $t = 600$ that precedes more tempered refinements, finishing with FID$= 3.17$ at $t = 1000$. In another experiment

independent samples conditioned on the same latent x_t at different stages of denoising were found to share similar high-level features even before the emergence of recognizable qualities in x_t. [6] concludes the intermediate latents encode meaningful features of the sample prior to the point of visual recognition around $t \approx 350$.

The authors of DDIM exploit the determinism of (7) and explore the *sample consistency* of the reverse process [9]. Since $\sigma_t = 0$ the final sample x_T only depends on the initial latent x_0, which is found to be an informative encoding of x_T's features. At different selections of $\mathcal{T}$ the high-level features of DDIM samples remain consistent, with extra denoising steps serving to improve the quality of emerged features. This behaviour is broadly similar to the patterns observed in [6] when taking the reverse process' altered form into consideration, albeit more restrictive.

While the qualitative attributes of the reverse process are widely reported quantitative analyses have proven less commonplace. This is not wholly unsurprising as the challenge of defining stable yet adaptable methods presents several definitive issues. FID and Inception Score have witnessed previous use in quantifying gains in the reverse process, however we do not find either a convincing measure of feature emergence since contrasting reverse process latents with $q(x_0)$ is irrelevant for our analysis. Achieving consistent metrics also proves difficult owing to the diversity of sampling methods for diffusion models and the variance introduced by datasets. In the remainder of this section we propose three methods that reliably quantify the prevalence of features in reverse process latents. We begin by introducing *stepwise entropy* that leverages information theory to track trends in information across latents. Next we exploit a pretrained classifier to compute the *label variance* of a single sample. Finally, we extend the works of [6] and [9] and provide a formal definition of *sample consistency* by introducing a partially deterministic schedule for z in (4).

A. Stepwise Entropy

Information theory provides a natural starting point to measure the prevalence of features in the reverse process' latents, and there is precedence in previous works to leverage information metrics for diffusion models. [6] analyzes the rate-distortion of the reverse process, finding that the majority of bits are allocated to invisible distortions in early steps. [16] introduces an auxiliary cross-entropy loss term to (3) which improves predictions of their x_0-parameterized model at each timestep. Lastly [17] introduces two entropy-driven methods to guide the sampling and training of conditional diffusion models.

We define stepwise entropy as the conditional entropy between our final sample $p(x_0)$ and the latents:

$$H(p(x_0)|p(x_t)) = H(p(x_0), p(x_t)) - H(p(x_t)) \qquad (11)$$

We choose this for several reasons. First, the conditional entropy can be calculated directly as the joint distributions are known for each x_t. We find this method to be computationally cheaper than other comparative measures like Kullback-Liebler Divergence and Wasserstein Distance while maintaining high informativeness. In tracing the stepwise entropy we aim to highlight where $p(x_t)$ begins to inform $p(x_0)$.

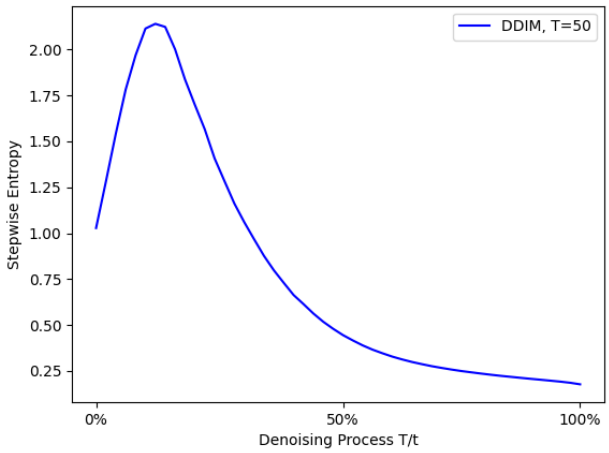

Fig. 2. Stepwise Entropy on CIFAR-10

IV. Experiments

To verify our proposed method we perform experiments on three seminal diffusion models, DDPM [6], DDIM [9], and Improved DDPM [10] on the CIFAR-10 dataset. We select these models due to their influence on later work and the significant differences in sampling behaviour. The CIFAR-10 dataset provides a robust testing ground for our methods owing to its content diversity and small image size (32x32 px).

We do not re-train our chosen diffusion models as all offer pretrained CIFAR-10 examples, allowing us to directly evaluate on the same networks where qualitative remarks are already reported. We leverage these models in all of our experiments. For DDIM we perform two rounds of sampling, one at $\mathcal{T} = 50$ and another at $\mathcal{T} = 100$ to reflect changes in sampling behaviour as $\mathcal{T} \to 0$, while for DDPM we sample with 1000 denoising steps and 4000 for IDDPM per the recommendations of the authors. For stepwise entropy, to obtain the probability distributions of samples we convert all images to greyscale and sort pixels into 256 bins based on their intensity value.

V. Results

A. Stepwise Entropy

We provide our results on the CIFAR-10 dataset for our DDIM50 and DDIM100 models in Figure 2, while in Figure 3 we display trends of the joint probability over the reverse process between $p(x_0)$ (horizontal) and $p(x_t)$ (vertical).

The mapping of stepwise entropy presents a bizarre pattern for our DDIM results, namely a rapid increase in entropy early in the reverse process followed by a gradual decline over

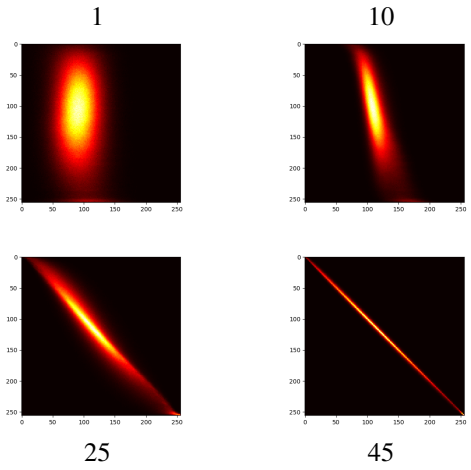

Fig. 3. Joint Probabilities of Pixel Intensities at different timesteps

the remaining denoising steps. The initial increase in Entropy can be attributed to a behaviour of DDIM, a convergence towards a "mean" grey color value. We visualize this behaviour for DDPM in 1, where the saturation of the sample rapidly diminishes prior to the emergence of features. Fascinatingly, this implies latents around $\mathcal{T} = 10$ for DDIM are less informative than the original noise embedding, however this is likely a side-effect of using entropy and requires further investigations to verify. The joint probabilities confirm the general trends of our stepwise entropy analysis, with the distribution gradually converging to the diagonal over time; this corresponds with limited, local changes in pixel value in latter steps, which we again witness in DDPM in Figure 1

VI. Conclusion

In this work we propose an information-based metric called stepwise entropy to measure the prevalence of features embedded in the latents of the reverse process. We evaluate our approach across several diffusion models and datasets, finding our measure to be a stable indicator that confirms previously-held qualitative beliefs about sample features. There are several avenues future work could extend our findings. Class-conditional diffusion models [7, 18] may yield different patterns in the reverse process depending on the label when evaluated with our method. Moreover, we only evaluate stepwise entropy on discrete diffusion models configured for the image domain. Adapting our method for use in continuous diffusion models [19] and for those in the audio domain [20] would provide invaluable insight into a broader range of reverse processes.

VII. Addendum

Progress on this report was severely hampered by multiple medical events over the August-October period of 2022. A two week extension was granted to balance lost time; unfortunately this additional period proved insufficient to achieve the aims set out. Much of the additional period was spent adapting the implementations of diffusion models for the implementation and testing of our methods, the significant effort required to gather which cannot be included in this report. Any further academic extensions to this project are constrained by personal financial obligations. Though incomplete, the submission of this work for marking cannot be delayed any further.

References

[1] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," 2021.

[2] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. V. Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," 2022.

[3] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," 2021.

[4] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020.

[7] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," 2021.

[8] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," 2020.

[9] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020.

[10] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," 2021.

[11] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[12] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[13] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[16] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. van den Berg, "Structured denoising diffusion models in discrete state-spaces," 2021.

[17] S. Li, G. Zheng, H. Wang, T. Yao, Y. Chen, S. Ding, and X. Li, "Entropy-driven sampling and training scheme for conditional diffusion generation," 2022.

[18] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," 2021.

[19] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2020.
[20] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," 2020.