Critique and Analysis of Latent Diffusion Models for High-Resolution Image Synthesis

Bright Liu and Victor Cai

***Abstract*—Latent Diffusion Models (LDMs) advance generative modeling by shifting the computationally intensive diffusion process from pixel space to compressed latent space, enabling efficient training and high-quality output. This paper explores the theoretical foundations, architecture, and innovations of LDMs, including pre-trained autoencoders, reweighted objectives, and time-conditional UNet architectures. We highlight applications in text-to-image synthesis, super-resolution, and video generation, as well as recent advances in model distillation and temporal alignment. By integrating principles from information theory, we demonstrate the scalability and versatility of LDMs, emphasizing their transformative role in generative AI and future research potential.**

***Index Terms*—diffusion models, latent diffusion models, autoencoders, denoising, image synthesis**

I. Introduction

THE development of image synthesis frameworks over the past two decades have found tremendous applications, including in the medical field for denoising, segmentation, anomaly detection, and image reconstruction, as well as in computer vision for image recognition, inpainting, and view synthesis. Many flavors of generative models have contributed to this growth, including autoregressive transformers, generative adversarial networks (GANs), and autoencoder-based methods. In particular, we conduct a case study centered on the latent diffusion model (LDM), introduced by Rombach in 2022 [1].

GANs [2] as generative models were often considered the state-of-the-art before the rise of diffusion models [3]. GANs train a generative model and a simultaneous discriminative model, where the generative model training seeks to maximize the discriminative model's probability of error. Approaches to image synthesis with GANs tend to give good perceptual quality. However, they can suffer from mode collapse [4], where the generative model settles on a few dominant modes without learning the full distribution, thus reducing diversity of generation. GANs become unstable as they scale to complex and multi-modal distributions [5]. Other approaches include variational autoencoders (VAEs) [6], which employs an encoder that encodes the input into a low-dimensional latent space and a decoder to recover the original input. VAEs are more efficient at synthesis, but do not give as good quality as GANs [1].

In recent years, diffusion probabilistic models (DMs) have reigned supreme in sample detail and diversity of image synthesis. With origins in nonequilibrium thermodynamics [7], they achieve impressive synthesis through a series of sequential denoising autoencoders. DMs first use a forward diffusion stage to perturb the input data with Gaussian noise over several steps, then do a reverse diffusion stage, where the model learns to reverse the diffusion by removing Gaussian noise to recover the input data. Finally, new images are generated by sampling from the generated distribution by passing it noise [8]. These models have found use in waveform generation [9], audio and music synthesis [10] [11], and are taking off in image synthesis [1]. One of the most influential works in image synthesis with diffusion models is the latent diffusion model (LDM) [1], which [4] claims LDMs such as [1] introduce a new level in the area of generative modeling.

First, we provide background on the landscape of generative models and motivate latent diffusion models (Section II). Then, we dive into the details of latent diffusion models (Section III). Finally, we look at recent advances beyond the original LDM (Section IV).

II. Generative Models, Diffusion Models, and Notation

Generative models aim to learn the underlying distribution of a dataset and generate new samples from this distribution. These models have been widely applied in various domains, including image synthesis, text-to-image generation, video synthesis, and more. Three prominent categories of generative models have emerged: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models. Each class of models offers unique advantages and trade-offs, and we outline their characteristics to provide a broader context for understanding Latent Diffusion Models (LDMs).

A. GANs and VAEs

Generative Adversarial Networks (GANs) are perhaps the most well-known generative models. Introduced by Goodfellow et al. [2] in 2014, GANs involve a two-player game between a generator and a discriminator. The generator produces its generated samples, while the discriminator attempts to distinguish between generated samples and original data. The training objective is a minimax game:

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))],$$

where G is the generator, D is the discriminator, p_{data} is the real data distribution, and p_z is the prior distribution over latent variables.

All authors are with Harvard University.

While GANs are capable of generating high-quality samples, they suffer from instability during training and issues like mode collapse. Training instability comes from needing to delicately balance the generator and the discriminator so that one does not dominate the other. Mode collapse is where the model fails to capture the diversity of the data distribution, since often the generator gravitates towards a few dominant modes within the distribution that are enough to get past the discriminator. Various methods have been proposed to resolve these issues, such as Wasserstein loss in the discriminator, which removes vanishing gradients by replacing the binary discriminator with a "critic" with a differentiable gradient that can be trained to optimality [12].

Variational Autoencoders (VAEs) [13], on the other hand, take a probabilistic approach to modeling the data distribution. VAEs encode input data x into a latent space z using a probabilistic encoder and then reconstruct x using a decoder. The training objective combines reconstruction loss and a regularization term that enforces the latent distribution to approximate a prior distribution (e.g., Gaussian):

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{\text{KL}}(q_\phi(z|x)\|p(z)),$$

where q_ϕ and p_θ are the encoder and decoder distributions, respectively, and D_{KL} is the Kullback-Leibler divergence. The first term can be interpreted as a reconstruction loss between the decoding and encoding distributions. The second term encourages the VAE to use q_ϕ close to the standard normal $p(z)$. VAEs are more stable to train compared to GANs but often produce lower-quality samples.

B. Diffusion Models: An Overview

Diffusion models are a class of generative models inspired by nonequilibrium thermodynamics. Unlike GANs and VAEs, diffusion models rely on a sequential denoising process. A diffusion model comprises two stages:

1) *Forward Diffusion Process:* This process gradually adds noise to the data x_0 over T timesteps, resulting in a highly noisy data representation x_T:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I),$$

where α_t controls the noise variance at each timestep.

2) *Reverse Diffusion Process:* The generative model learns to reverse this process by progressively denoising x_T to reconstruct the original data x_0. This is achieved by training a neural network to predict the added noise ϵ at each step:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{x,\epsilon,t}[\|\epsilon - \epsilon_\theta(x_t, t)\|^2].$$

Diffusion models have demonstrated exceptional performance in image synthesis tasks, rivaling and often surpassing GANs and VAEs [3]. However, they are typically computationally intensive due to the large number of sequential steps required during both training and generation, often taking hundreds of GPU days to train the most powerful models [3].

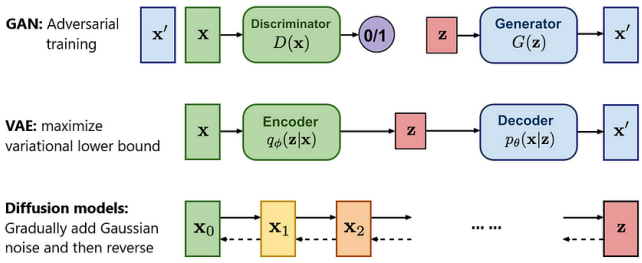

Fig. 1: General GAN, VAE, and DM models visualized as block diagrams.

C. Forms of Diffusion Models

Croitoru [4] provides a detailed categorization of diffusion models into three major forms, each of which builds upon the foundational principles described above:

1) Denoising Diffusion Probabilistic Models (DDPMs): Introduced by Ho et al. (2020) [14], DDPMs are one of the earliest and most influential implementations of diffusion models. They define the forward diffusion process as a series of Gaussian perturbations and train a neural network to approximate the reverse process. The objective minimizes the variational lower bound of the negative log-likelihood:

$$\mathcal{L}_{\text{vlb}} = \mathbb{E}[-\log p_\theta(x_0|x_1)] + \sum_{t=2}^{T} \mathbb{E}[D_{\text{KL}}(q(x_{t-1}|x_t, x_0)\|p_\theta(x_{t-1}|x_t))].$$

DDPMs achieve high-quality synthesis but require hundreds or thousands of timesteps during inference.

2) Score-Based Generative Models (SGMs): Score-based models [15] reformulate the reverse diffusion process using the score function $\nabla_x \log p_t(x)$. These models learn the score function via score matching and use Langevin dynamics to iteratively sample from the learned distribution. SGMs provide a mathematically elegant framework and are particularly effective for high-dimensional data, though it can be sensitive to parameters such as numerical stability and the step size for the Langevin dynamics.

3) Stochastic Differential Equation (SDE)-Based Models: SDE-based models [16] generalize DDPMs and SGMs by representing the forward diffusion process as a continuous-time stochastic differential equation:

$$dx = f(x,t)dt + g(t)dW_t,$$

where W_t is a Wiener process, and $f(x,t)$ and $g(t)$ control the drift and diffusion coefficients. This formulation allows for a unified perspective on diffusion models and opens up new possibilities for adaptive noise schedules and efficient sampling, though it is difficult to optimize over a number of sampler hyperparameters.

D. Motivating Latent Diffusion Models (LDMs)

While diffusion models offer unmatched performance in generating diverse and high-fidelity samples, their computational cost remains a significant bottleneck. For instance, DDPMs and SDE-based models require hundreds of iterations

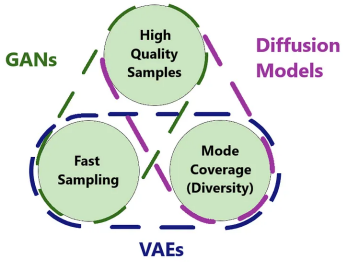

Fig. 2: High-level view of the tradeoffs between GAN, VAE, and DM-based approaches on mode coverage, sample quality, and computational speed. LDMs seek to speed up DMs to cover all three.

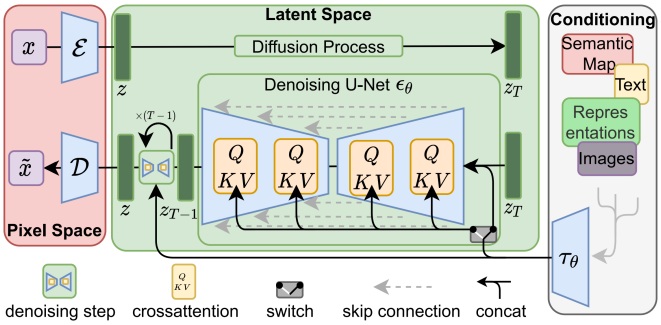

Fig. 3: LDM schematic from Rombach [1]

for both training and sampling, making them impractical for many real-world applications. Various fast sampling strategies [17] can be implemented to speed up training, but this does not address the dimensionality problem in that training still occurs on the large pixel space.

LDMs address this limitation by utilizing a latent space representation, where the diffusion process operates on a compressed representation of the data. This innovation significantly reduces computational overhead while retaining the quality and diversity of the generated samples. The next section dives into the specifics of LDMs and their contributions to the field.

III. Latent Diffusion Models

Latent Diffusion Models (LDMs) [1] represent a breakthrough in generative modeling, addressing the computational inefficiencies inherent in pixel-space diffusion models while maintaining high-quality generation. The success of diffusion models arises from their likelihood-based framework, which avoids mode collapse—a common issue in GANs—thereby enabling the production of diverse outputs. However, DMs' success comes at a steep computational cost, often requiring hundreds of GPU days for training and thousands of iterative steps for sampling.

The primary innovation of LDMs lies in shifting the diffusion process from the high-dimensional pixel space x to a compressed, lower-dimensional latent space z. This transformation significantly reduces computational demands while preserving critical semantic information, making LDMs highly scalable for real-world applications. This section delves into the key mechanisms, efficiency strategies, and applications of LDMs.

A. Motivation and Problem Statement

Traditional diffusion models operate directly in pixel space, where the dimensionality is exceedingly high for tasks like high-resolution image synthesis. This results in high computational cost and scaling issues to larger images.

LDMs address these limitations by pretraining autoencoders to map images into a continuous latent space. This latent representation, which captures the most informative semantic features, serves as the input for the diffusion process, enabling efficient and high-fidelity synthesis. It is important to note that the latent space is not just a spacial downsampling of the original pixel space, but has been processed by a pre-trained autoencoder to the specified dimensional downsampling factor.

B. Diffusion Process in Latent Space

Diffusion models aim to approximate the data distribution $p(x)$ by gradually denoising a noisy variable $x_T \sim \mathcal{N}(0, I)$ over T timesteps, with each denoising step corresponding to a probabilistic denoising autoencoder. The reverse diffusion process is learned by training a U-Net ϵ_θ over steps to predict the noise in a given noisy input.

For LDMs, the diffusion process is defined in the latent space z instead of the pixel space x. This process is governed by a fixed Markov chain, where each step gradually adds noise to z, and the model learns the reverse process of denoising:

$$z_t = \sqrt{\alpha_t} z_{t-1} + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

where α_t controls the variance schedule across timesteps t. Given a noisy latent z_t, the model ϵ_θ is trained to predict the added noise ϵ through a simplified objective:

$$\mathcal{L}_{DM} = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,I),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|^2\right].$$

C. Reweighted Objective for Efficiency

To improve computational efficiency, LDMs utilize a reweighted variant of the diffusion model objective, focusing the learning on denoising at intermediate stages where the signal-to-noise ratio is more balanced. The objective in latent space is

$$\mathcal{L}_{LDM} := \mathbb{E}_{\varepsilon(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|^2\right],$$

where ε is the encoder, z_t represents the latent variable at timestep t, and ϵ_θ is implemented as a time-conditional UNet architecture, leveraging 2D convolutional layers that exploit the spatial structure of z.

D. Experimental Results and Conditioning

Rombach [1] demonstrates state-of-the-art or competitive performance across a range of generative tasks, including unconditional image synthesis, text-to-image synthesis, super-resolution, and inpainting. In particular, conditioning on text

is done through cross-attention layers in the decoder on the UNet backbone.

The latent space also facilitates efficient conditioning on additional modalities, such as semantic maps or text prompted inputs, further expanding the applicability of LDMs.

E. Information-Theoretic Perspective on Latent Diffusion Models

LDMs can be understood through the lens of information theory, where the encoding and decoding processes resemble source coding and channel decoding, respectively. The autoencoder learns a compressed latent representation z that minimizes the rate-distortion trade-off:

$$R(D) = \inf_{q(z|x)} \mathbb{E}\left[\|x - D(E(x))\|^2\right].$$

Here, the latent diffusion process optimizes the trade-off between efficient compression (low rate) and reconstruction fidelity (low distortion). Specifically, [1] found experimentally that a compression factor of 4 to 8 in the dimensions of the latent space compared to the pixel space is ideal for quality measured by FID and Inception Score. This perspective highlights the theoretical foundations of LDMs and their ability to balance efficiency and quality.

F. Key Advantages and Limitations

The shift to latent space provides several advantages. The dimensionality reduction provided by the latent space significantly lowers the computational cost of training and sampling. The latent representation supports diverse conditioning mechanisms, making LDMs adaptable to various generative tasks. LDMs also scale more efficiently to high-resolution tasks, enabling applications in domains such as medical imaging, gaming, and virtual reality.

The open-sourced pretrained autoenocders help to democratize LDMs for those with less computational resources. However, the reliance on a pretrained autoencoder introduces potential limitations, such as sensitivity to the quality of the learned latent space and additional pretraining requirements. Addressing these challenges remains an active area of research.

In summary, LDMs represent a paradigm shift in diffusion-based generative modeling, combining efficiency, scalability, and high-quality synthesis. The next section will explore how these models extend to specific applications and recent advancements in the field.

IV. Recent Progress

The rapid development of generative modeling techniques has catalyzed significant advancements in latent diffusion models (LDMs). Building on the foundational principles of LDMs, recent research has focused on improving efficiency, enhancing temporal consistency for video synthesis, and extending the application scope to new modalities. This section highlights two major areas of progress: model distillation and video synthesis.

A. Model Distillation for Efficiency

One of the key limitations of diffusion models, including LDMs, is their reliance on computationally expensive iterative processes during both training and inference. Recent work by Meng et al. [18] addresses this challenge through a two-stage distillation framework designed to reduce the number of reverse diffusion steps while maintaining high-quality outputs.

The distillation framework simplifies the reverse diffusion process as follows:

1) *Single-Step Approximation.* A student model is trained to match the output of a teacher model for each diffusion step. The objective minimizes the discrepancy between the teacher's noise prediction and the student's approximation:
$$\min_{\theta} \mathbb{E}_{z_t,t} \left\|\epsilon_{\text{teacher}}(z_t, t) - \epsilon_{\text{student}}(z_t, t; \theta)\right\|^2.$$
2) *Step Reduction.* The student model is progressively distilled to reduce the number of reverse diffusion steps. This involves learning a compressed set of inference steps while preserving the overall denoising trajectory:
$$z_{t-1} = g_\phi(z_t, t), \quad \phi \text{ reduces steps from } T \to T/N.$$

The distillation framework significantly reduces inference time, achieving a reduction of up to $10\times$ in computational cost. This improvement is critical for real-time applications such as interactive image synthesis and autonomous systems, enabling faster generation of high-resolution images for creative tools and reducing latency in systems requiring rapid environmental understanding or simulation.

B. Video Synthesis with Latent Diffusion Models

Extending LDMs to video synthesis presents unique challenges, such as ensuring temporal consistency across frames. Blattmann et al. [19] introduce a novel method for high-resolution video generation by aligning latent representations over time. The proposed method employs temporal alignment mechanisms to ensure that the latent variables for successive frames remain consistentby conditioning across frames in the latent space and by adapting the loss function to include temporal smoothness to reduce flickering.

The integration of temporal alignment techniques enables LDMs to generate high-quality videos for a range of applications, including autonomous driving simulations and test platforms; cinematic rendering and virtual reality; and personalized content creation based on user preferences. These advances establish LDMs as a powerful framework for generating temporally consistent video content, addressing key challenges in dynamic generative modeling.

V. Conclusion

Latent Diffusion Models (LDMs) represent a significant milestone in the evolution of generative modeling. By addressing the computational inefficiencies inherent in diffusion models through a latent space approach, LDMs have achieved a remarkable balance between scalability and fidelity. This paper has explored the fundamental principles of LDMs, their connections to prior frameworks, and the transformative potential they hold across various applications.

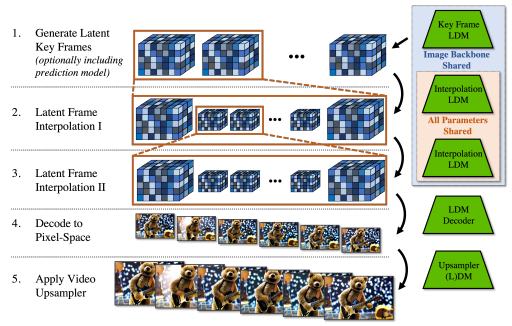

Fig. 4: LDM Video Generation

A. Key Insights and Contributions

This review of LDMs and recent progress has revealed several important insights:

- **Efficiency Without Compromise:** By shifting the computationally intensive diffusion process from pixel space to latent space, LDMs reduce resource demands without sacrificing the quality or diversity of generated outputs. This efficiency unlocks practical deployment in real-world scenarios, including interactive systems and edge devices.
- **Strong Foundations in Information Theory:** The LDM framework elegantly ties together concepts from information theory, such as compression, rate-distortion trade-offs, and efficient representation learning. These principles underscore the robustness of LDMs and their theoretical underpinnings.
- **Versatility Across Modalities:** LDMs have demonstrated versatility in a wide range of applications, including unconditional image synthesis, text-to-image generation, super-resolution, and inpainting. Recent advancements extend these capabilities to video synthesis and multi-modal generation, broadening their scope.
- **Advances in Optimization and Scalability:** Recent innovations such as model distillation have enhanced the practicality of LDMs by reducing inference time and computational costs. These methods ensure that LDMs remain at the forefront of generative modeling research.

B. Implications for Future Research

The field of generative modeling remains vibrant and full of unanswered questions. LDMs provide a foundation upon which new research can build. Promising directions include:

1) **Dynamic and Adaptive Diffusion Models:** Investigating the incorporation of dynamic latent spaces that adapt to input complexity could further enhance the efficiency and robustness of LDMs.
2) **Integration with Multi-Modal Systems:** Expanding LDMs to handle complex multi-modal tasks, such as generating video synchronized with audio or creating 3D assets based on text descriptions, is a natural progression.
3) **Applications in Emerging Domains:** Beyond traditional applications in image and video synthesis, LDMs could be adapted for scientific computing, industrial design, and educational tools, opening new avenues for interdisciplinary collaboration.
4) **Theory-Driven Optimization:** Leveraging insights from information theory and probabilistic modeling to develop more efficient training objectives and robust architectures will further advance the field.

C. Final Reflections

The journey from traditional generative models to latent diffusion has exemplified the power of combining theoretical rigor with practical innovation. LDMs highlight a paradigm shift in generative modeling, proving that computational efficiency and output quality need not be mutually exclusive. Their impact spans creative industries, scientific research, and technological innovation, providing tools that are as versatile as they are powerful.

As the field continues to evolve, LDMs stand as a testament to the potential of deep generative models. They not only offer a window into the future of AI-driven creativity but also lay the groundwork for solving some of the most pressing challenges in scalable and high-fidelity generation. In embracing these opportunities, the community is poised to explore new frontiers in artificial intelligence, one latent space at a time.

Acknowledgment

B. Liu's contribution focused on the latent diffusion model paper [1] and the details of the underlying approach, as well as recent developments beyond LDMs. V. Cai's contribution focused on contextualizing the works that influenced [1] and giving credit where credit is due.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, *High-resolution image synthesis with latent diffusion models*, 2022. arXiv: 2112.10752 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2112.10752.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, *Generative adversarial networks*, 2014. arXiv: 1406.2661 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1406.2661.

[3] P. Dhariwal and A. Nichol, *Diffusion models beat gans on image synthesis*, 2021. arXiv: 2105.05233 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2105.05233.

[4] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 850–10 869, Sep. 2023, ISSN: 1939-3539. DOI: 10.1109/tpami.2023.3261988. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2023.3261988.

[5] A. Brock, J. Donahue, and K. Simonyan, *Large scale gan training for high fidelity natural image synthesis*, 2019. arXiv: 1809.11096 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1809.11096.

[6] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, 2022. arXiv: 1312.6114 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1312.6114.

[7] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, *Deep unsupervised learning using nonequilibrium thermodynamics*, 2015. arXiv: 1503.03585 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1503.03585.

[8] Z. Chang, G. A. Koulieris, and H. P. H. Shum, *On the design fundamentals of diffusion models: A survey*, 2023. arXiv: 2306.04542 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2306.04542.

[9] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, *Wavegrad: Estimating gradients for waveform generation*, 2020. arXiv: 2009.00713 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2009.00713.

[10] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, *Diffwave: A versatile diffusion model for audio synthesis*, 2021. arXiv: 2009.09761 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2009.09761.

[11] G. Mittal, J. Engel, C. Hawthorne, and I. Simon, *Symbolic music generation with diffusion models*, 2021. arXiv: 2103.16091 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2103.16091.

[12] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, Jun. 2017, pp. 214–223. [Online]. Available: https://proceedings.mlr.press/v70/arjovsky17a.html.

[13] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proceedings of the 28th international conference on international conference on machine learning*, 2011, pp. 833–840.

[14] J. Ho, A. Jain, and P. Abbeel, *Denoising diffusion probabilistic models*, 2020. arXiv: 2006.11239 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2006.11239.

[15] Y. Song and S. Ermon, *Generative modeling by estimating gradients of the data distribution*, 2020. arXiv: 1907.05600 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1907.05600.

[16] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, *Score-based generative modeling through stochastic differential equations*, 2021. arXiv: 2011.13456 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2011.13456.

[17] Z. Kong and W. Ping, *On fast sampling of diffusion probabilistic models*, 2021. arXiv: 2106.00132 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2106.00132.

[18] C. Meng, R. Rombach, R. Gao, *et al.*, "On distillation of guided diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 14 297–14 306.

[19] A. Blattmann, R. Rombach, H. Ling, *et al.*, *Align your latents: High-resolution video synthesis with latent diffusion models*, Presented at CVPR 2023, 2023. DOI: 10.48550/arXiv.2304.08818. arXiv: 2304.08818 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2304.08818.