International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct- 2026, PP 01-6

Diffusion Models For Image Generation: A Review Of Denoising Probabilistic Models, Latent Diffusion, Guidance And Controllable Synthesis

Sayyed Rida Fatima Sharif[1] Prof. S. R. Borhade[2] Prof. M. C. Kolse[3] Prof. S. K. Said[4]
Department of Artificial Intelligence and Data Science
Jaihind College of Engineering
Pune, Maharashtra, India

Abstract—Diffusion models are generative models that create high-quality, realistic images by learning to reverse a gradual noising process. This paper reviews the working principles, architectures and applications of diffusion models for image generation. It explains the forward process, which adds Gaussian noise to training images over many steps, and the reverse process, in which a neural network, usually a U-Net, learns to predict and remove noise. The review traces the field from its thermodynamic origins to denoising diffusion probabilistic models (DDPM), the score-based stochastic differential equation view, and faster samplers such as DDIM. It examines classifier and classifier-free guidance, latent diffusion models that operate in a compressed latent space for efficiency, text-to-image systems conditioned on language embeddings, and spatial control methods that add edges, poses or depth as conditions. Applications in digital art, image editing, restoration, medical imaging and animation are discussed. A five-stage architecture is presented, along with evaluation metrics, research gaps, objectives, a comparison with GANs and VAEs and challenges such as slow sampling, computational cost, bias, copyright and misuse for deepfakes. The review concludes that diffusion models offer an excellent balance of quality and diversity, with efficiency and responsible use as the main open problems.

Index Terms—Diffusion models, DDPM, latent diffusion, score-based models, classifier-free guidance, text-to-image, U-Net, generative AI

I. Introduction

Until recently, generative adversarial networks (GANs) [1] produced the most realistic synthetic images, with models such as StyleGAN [2] generating convincing faces. GANs, however, are difficult to train and may cover only part of the data distribution. Diffusion models offer an alternative with stable training and high diversity.

Sohl-Dickstein *et al.* [3] introduced the idea of learning to reverse a gradual diffusion process inspired by non-equilibrium thermodynamics. Ho *et al.* [4] made the approach practical with denoising diffusion probabilistic models, and Dhariwal and Nichol [5] showed that improved diffusion models surpass GANs in image quality on ImageNet. Diffusion models now power popular text-to-image systems.

This paper reviews the principles, key variants and applications of diffusion models and presents a reference generation pipeline.

II. Theoretical Background

2.1 Forward and Reverse Processes

In the forward process, Gaussian noise is added to an image x_0 over T steps until it becomes almost pure noise. A noisy sample at step t can be written directly as

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct- 2026, PP 01-6

$$x_t = \sqrt{\bar{\alpha}_t} \cdot x_0 + \sqrt{(1 - \bar{\alpha}_t)} \cdot \varepsilon, \quad \varepsilon \sim N(0, I) \quad (1)$$

where $\bar{\alpha}_t$ follows a noise schedule. The reverse process uses a network ε_θ to predict the noise and gradually denoise from x_t to x_0.

2.2 Family of Diffusion Models

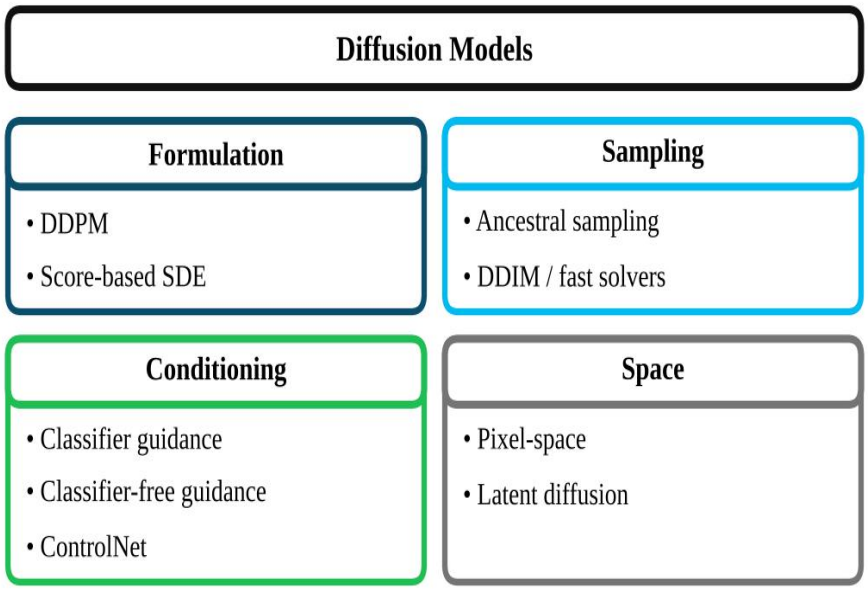


Fig. 1 organizes diffusion models by formulation, sampling, conditioning and representation space.

Fig. 1. Taxonomy of diffusion models for image generation.

III. Survey of Existing Approaches

Foundations. Sohl-Dickstein *et al.* [3] proposed diffusion probabilistic models. Ho *et al.* [4] trained a U-Net [6] to predict noise with a simple loss and achieved high-quality samples. Song *et al.* [7] unified diffusion and score-based models through stochastic differential equations.

Faster sampling. Song *et al.* [8] introduced DDIM, a deterministic sampler that generates images in far fewer steps.

Guidance. Dhariwal and Nichol [5] used classifier guidance to trade diversity for fidelity, and Ho and Salimans [9] proposed classifier-free guidance, which mixes conditional and unconditional predictions without a separate classifier.

Latent and text-to-image models. Rombach *et al.* [10] performed diffusion in the latent space of an autoencoder, greatly reducing compute, and conditioned generation on text through cross-attention. Ramesh *et al.* [11] generated images from CLIP [12] embeddings, and Saharia *et al.* [13] showed that large language-model text encoders improve photorealism and text alignment.

Control. Zhang *et al.* [14] proposed ControlNet, which adds spatial conditions such as edges, poses or depth maps to pretrained diffusion models. Table I summarizes these works.

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct- 2026, PP 01-6

Ref.	Authors	Year	Technique Used	Merits	Demerits
[3]	Sohl-Dickstein et al.	2015	Diffusion probabilistic models	Introduced reverse diffusion	Low sample quality
[4]	Ho et al.	2020	DDPM with U-Net	High-quality samples	Slow sampling
[7]	Song et al.	2021	Score-based SDE	Unified framework	Complex solvers
[8]	Song et al.	2021	DDIM sampler	Far fewer steps	Some quality loss
[5]	Dhariwal and Nichol	2021	Improved diffusion + guidance	Beat GANs on ImageNet	High compute
[10]	Rombach et al.	2022	Latent diffusion	Efficient text-to-image	Autoencoder artefacts
[13]	Saharia et al.	2022	LLM text encoder	Better text alignment	Not openly released
[14]	Zhang et al.	2023	ControlNet	Spatial control	Extra training per condition

TABLE I Summary of Literature Survey

IV. Comparative Study

Table II compares diffusion models with GANs and VAEs. Diffusion models offer the best quality and diversity with stable training, at the cost of slower sampling.

Model	Quality	Diversity	Sampling Speed
VAE	Medium	High	Fast
GAN	High	Medium	Fast
Pixel diffusion	Very high	High	Slow
Latent diffusion	Very high	High	Moderate

TABLE II Comparison of Generative Model Families

V. Research Gaps

- Sampling is slower than single-pass generators such as GANs.
- Training large models requires very high compute and data.
- Outputs may reflect societal biases in training data.
- Copyright and consent for training images are unresolved.
- Realistic outputs can be misused for deepfakes and misinformation.

VI. Aims and Objectives

- 1. To explain the forward and reverse diffusion processes.
- 2. To review major diffusion model variants and samplers.
- 3. To examine guidance, latent diffusion and controllable generation.

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct- 2026, PP 01-6

- 4. To present a reference text-to-image diffusion pipeline.
- 5. To discuss evaluation, challenges and responsible use.

VII. Proposed Architecture

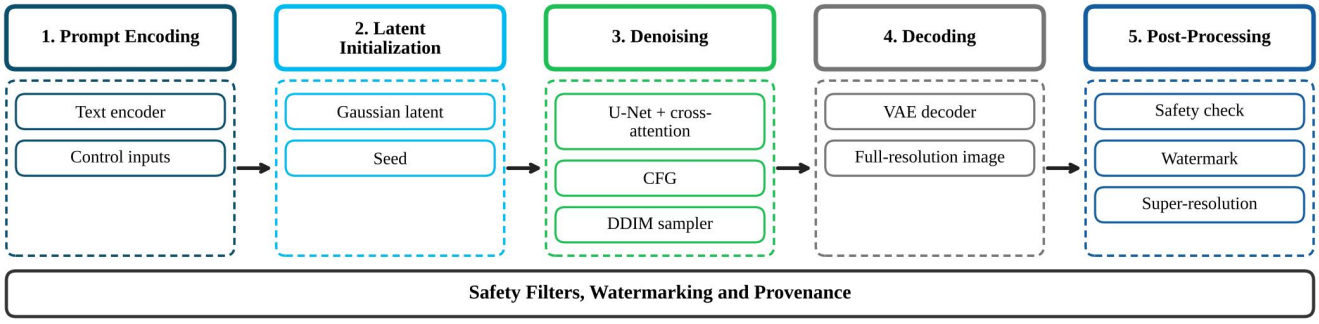


Fig. 2 presents a reference latent text-to-image diffusion pipeline.

7.1 Prompt Encoding Stage

A text encoder converts the user's prompt into embeddings; optional control inputs such as sketches or poses are encoded separately.

7.2 Latent Initialization Stage

A random Gaussian latent of reduced resolution is sampled.

7.3 Denoising Stage

A U-Net with cross-attention predicts noise at each step; classifier-free guidance strengthens prompt adherence; a sampler such as DDIM updates the latent.

7.4 Decoding Stage

The autoencoder decoder converts the final latent into a full-resolution image.

7.5 Safety and Post-Processing Stage

Safety filters check outputs, watermarks or provenance metadata are added and optional super-resolution is applied.

VIII. Process Flow and Evaluation

Fig. 3 describes the generation process. Training minimizes the noise-prediction loss

$$L = E\,\|\varepsilon - \varepsilon_\theta(x_t, t, c)\|^2 \quad (2)$$

where c is the condition. With classifier-free guidance, the predicted noise is

$$\hat{\varepsilon} = \varepsilon_\theta(x_t, \varnothing) + s \cdot (\varepsilon_\theta(x_t, c) - \varepsilon_\theta(x_t, \varnothing)) \quad (3)$$

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct- 2026, PP 01-6

where s is the guidance scale. Image quality is commonly measured with the Fréchet Inception Distance [15], and text alignment with CLIP scores.

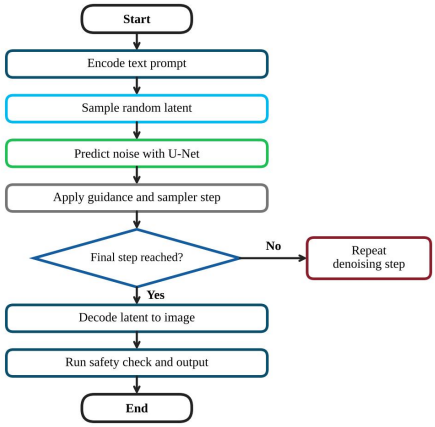


Fig. 3. Text-to-image generation with a latent diffusion model.

IX. Limitations and Future Scope

Table III summarizes challenges. Future directions include few-step and distilled samplers, video and 3D diffusion, on-device generation and reliable watermarking of generated content.

Challenge	Mitigation
Slow sampling	DDIM, distillation, fewer steps
High compute	Latent diffusion; efficient U-Nets
Bias	Curated data; bias audits
Copyright	Licensed data; opt-out mechanisms
Deepfake misuse	Safety filters; watermarking

TABLE III Challenges and Mitigation

X. Conclusion

Diffusion models generate images by learning to remove noise step by step, and have become the leading approach to image synthesis. The reviewed literature traces their development from DDPM to latent diffusion, guidance and controllable generation. Faster sampling, lower compute requirements and responsible deployment are the main directions for future work.

10.1 Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

References

1. I. Goodfellow et al., “Generative adversarial nets,” in Proc. NeurIPS, vol. 27, 2014.

2. T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. IEEE/CVF CVPR, 2019, pp. 4401–4410.
3. J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in Proc. 32nd ICML, 2015, pp. 2256–2265.
4. J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in Proc. NeurIPS, vol. 33, 2020, pp. 6840–6851.
5. P. Dhariwal and A. Nichol, “Diffusion models beat GANs on image synthesis,” in Proc. NeurIPS, 2021, pp. 8780–8794.
6. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. MICCAI, 2015, pp. 234–241.
7. Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” in Proc. ICLR, 2021.
8. J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in Proc. ICLR, 2021.
9. J. Ho and T. Salimans, “Classifier-free diffusion guidance,” arXiv:2207.12598, 2022.
10. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in Proc. IEEE/CVF CVPR, 2022, pp. 10684–10695.
11. A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with CLIP latents,” arXiv:2204.06125, 2022.
12. A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. ICML, 2021, pp. 8748–8763.
13. C. Saharia et al., “Photorealistic text-to-image diffusion models with deep language understanding,” in Proc. NeurIPS, 2022.
14. L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in Proc. IEEE/CVF ICCV, 2023, pp. 3836–3847.
15. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” in Proc. NeurIPS, 2017.