IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 18, April 2026
Impact Factor: 8.2

Diffusion Models in Image Generation: A Comprehensive Survey

Pratiksha Sawant[1], Vinay Shinde[2], Shreyash Sutar[3], Sanket Tambekar[4], Piyush Kondar[5]
Assistant Professor, Department of Artificial Intelligence and Data Science[1]
Student, Department of Artificial Intelligence and Data Science[2-5]
AISSMS Institute of Information Technology, Pune, India

Abstract: *Diffusion models are the primary approach for generative work in computer vision, used for creating new images from scratch, tweaking existing ones, or performing complex edits. This survey traces the evolution of AI image generation from early methods to the latest diffusion models. Beginning with the basics of Denoising Diffusion Probabilistic Models (DDPM)—how noise is gradually added in the forward pass and how the model learns to remove it step by step in the reverse process—this paper then covers Latent Diffusion Models (LDMs), which advance the process by operating in a compressed latent space using Variational Autoencoders (VAEs). This approach encodes images into lower-dimensional representations before adding noise and decoding them after denoising, dramatically reducing computational cost while preserving perceptual quality. The survey offers a complete look at diffusion models tracing their path from core fundamentals to cutting-edge advances and real-world tools for image generation..*

Keywords: Diffusion Models, DDPM, Latent Diffusion Models, VAE, U-Net, Image Generation, Generative AI, Flow Matching, Stable Diffusion

I. INTRODUCTION

Generative modeling has evolved considerably over the past few years. Among the various approaches, diffusion models have emerged as the primary engine behind modern visual content creation, largely surpassing older methods such as GANs and VAEs. They are now the default choice for generating sharp, realistic synthetic images—appearing in applications ranging from photorealistic portraits and vivid textures to coherent 3D scenes, both in research laboratories and commercial products.

The core idea is conceptually simple: random Gaussian noise is added to an image, and a model is trained to progressively remove that noise. This gradual denoising approach has proven highly flexible, handling text-to-image generation, interactive editing, super-resolution, and even video synthesis.

Diffusion models offer a key advantage in training stability. Unlike GANs—which suffer from mode collapse and unstable adversarial dynamics—diffusion models train smoothly and cover the full data distribution more reliably. Compared to VAEs, they deliver noticeably sharper results while retaining a solid mathematical foundation. The step-by-step construction also makes it far easier to guide the generation process using conditions or extra guidance signals.

Recent breakthroughs have made these models much more practical. Distribution Matching Distillation (DMD) compresses 50–100 denoising steps into just one or two passes, achieving up to 30× faster inference with minimal quality loss. Latent diffusion models shift computation into a compact learned space instead of full-resolution pixels, slashing hardware requirements and enabling high-quality generation on consumer GPUs.

Beyond images, diffusion models drive video synthesis (e.g., via cascaded latents) and audio generation, with extensions to protein design and graph generation. Consistency models further accelerate sampling to 2-8 steps, blending distillation and progressive distillation. In 2026, hybrid approaches combine them with flow matching for faster, invertible generation.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-34353

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 18, April 2026 Impact Factor: 8.2

Distribution Matching Distillation (DMD) trains one-step generators to mimic multi-step diffusion teachers by matching output distributions via KL divergence, cutting inference from 50+ steps to 1-4 with near-identical FID scores. Improved DMD variants enhance gradient estimation for even stabler training, achieving 30× speedups on benchmarks like CIFAR-10. These enable real-time applications on edge devices without retraining from scratch.

Literature Review

This section establishes the conceptual foundation and key research background for diffusion models and generative modeling in general. It provides an overview of what diffusion models are and how they fit into broader generative frameworks. It then covers major research directions—including methods for handling heterogeneous data distributions, new architectural ideas, and techniques to accelerate sampling. Finally, it addresses emerging trends focused on making diffusion models more reliable and practical for real-world deployment.Recent research tackles multimodal and irregular distributions by integrating diffusion processes with mixture models or graph structures, enabling robust generation across domains like single-cell RNA sequencing and knowledge graphs. For instance, scVIC combines VAEs with Gaussian mixtures to disentangle biological variance from batch effects in scRNA-seq data, outperforming baselines in clustering accuracy. Block Diffusion employs semi-autoregressive block unmasking for code and text, balancing throughput gains with quality via KV cache reuse

II. MATHEMATICAL FOUNDATIONS

A. Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs) generate new samples through a two-stage process. First, a forward diffusion process gradually adds Gaussian noise to the original data until it becomes pure noise. Then, a learned reverse process undoes this corruption step by step, turning noise back into realistic data.

B. Forward Diffusion Process

The forward process corrupts a clean data sample by adding Gaussian noise step by step over T timesteps (typically T = 1000). At each timestep t, a small amount of noise is added according to a fixed variance schedule $\beta_1, \beta_2, ..., \beta T$. The conditional distribution at each step is:

$$q(x_t \mid x_{t-1}) = N(x_t; \sqrt{(1-\beta_t)}\, x_{t-1}, \beta_t I) \quad (1)$$

A key property is that a noisy sample at any timestep t can be obtained directly in closed form without simulating all preceding steps. Defining $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_t)$, the marginal distribution is:

$$q(x_t \mid x_0) = N(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)I) \quad (2)$$

In practice, a linear variance schedule is used and the data is scaled at each step to prevent signal collapse as noise accumulates.

C. Reverse Diffusion Process

The reverse process starts from pure Gaussian noise $xT \sim N(0, I)$ and gradually denoises to recover a clean sample. Since the exact reverse posterior is intractable, a neural network (U-Net) parameterised by θ is trained to approximate it:

$$p\theta(x_{t-1} \mid x_t) = N(x_{t-1}; \mu\theta(x_t, t), \Sigma\theta(x_t, t)) \quad (3)$$

By repeatedly sampling from this learned reverse process starting from random noise, the model progressively constructs the final image.

D. Training Objective

The training criterion is derived from a variational lower bound (ELBO) on the log-likelihood of the data. After mathematical simplification, this reduces to a simple noise prediction task—the network learns to estimate the noise ε injected at each timestep, given the noisy observation x_t and the timestep index t:

$$L = E[||\varepsilon - \varepsilon\theta(x_t, t)||^2] \quad (4)$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 18, April 2026
Impact Factor: 8.2

This simplified objective frames diffusion model training as a structured denoising regression problem, which is stable to optimise and scales well with model size.

III. ARCHITECTURE COMPONENTS

A. U-Net Backbone

The predominant neural architecture in diffusion models is based on a U-Net incorporating ResNet-style residual blocks. The U-Net encodes noisy input images through a hierarchical downsampling path that captures multi-scale features, then decodes through a symmetric upsampling path to produce the denoised output. Lateral skip connections between encoder and decoder stages at matching spatial resolutions preserve fine-grained spatial detail throughout the reconstruction process.

Temporal conditioning is introduced via sinusoidal timestep embeddings, which encode the current denoising step and are incorporated into the network through adaptive normalisation mechanisms or additive/multiplicative modulation layers, enabling the model to adapt its behaviour appropriately at each stage of the reverse process.

B. Attention Mechanisms

To facilitate coherent synthesis at a global level, modern diffusion architectures augment the U-Net with self-attention modules placed at multiple resolution stages. These components allow each spatial location to attend to information from all other positions, enabling long-range dependency modelling. For conditioning on external signals such as text, cross-attention layers are introduced, routing information from embedding representations into the image generation process and enabling versatile conditional generation.

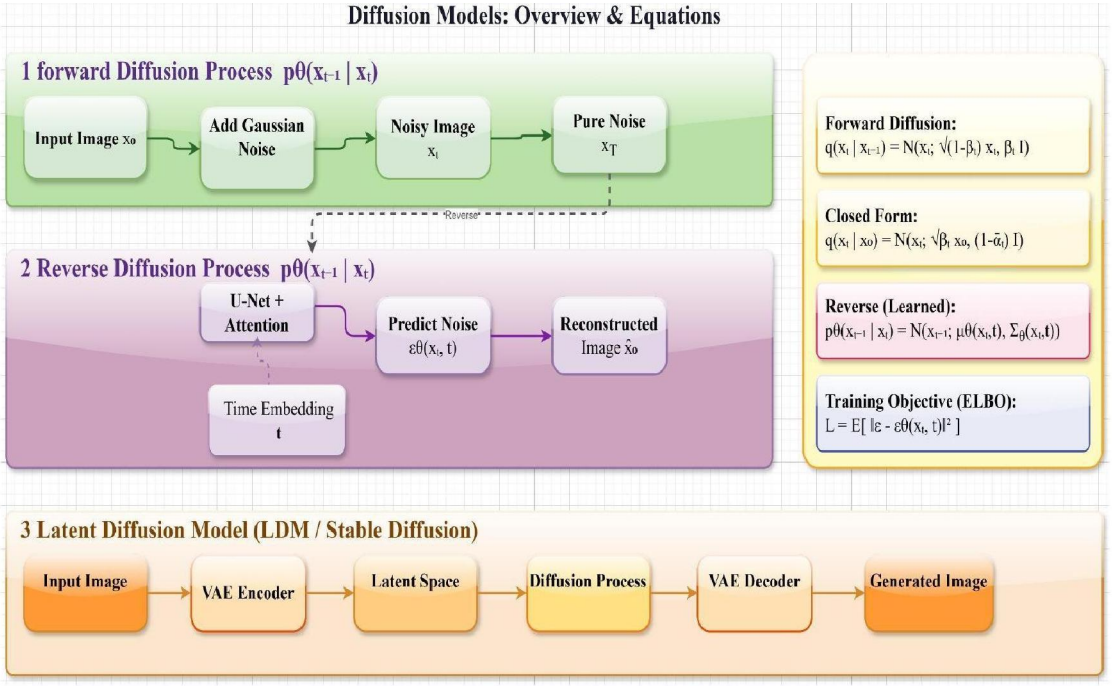

Fig. 1. Diffusion Models: Overview & Equations

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 18, April 2026
Impact Factor: 8.2

C. Conditional Generation

Conditional diffusion models steer the generation process using various types of additional inputs. For text-to-image tasks, the prompt is passed through a pre-trained language model such as CLIP or T5 to create a rich embedding, which is then injected into the denoising network via cross-attention. Tools like Stable Diffusion, DALL·E, and Midjourney all use this architecture.

Beyond text, other conditioning signals include class labels, semantic layouts, edge maps, depth maps, or reference images. Classifier-free guidance has gained particular traction as a technique that encourages the model to pay much closer attention to the conditioning signal, resulting in outputs that better match the input description.

IV. LATENT DIFFUSION MODELS

A. Motivation and Architecture

Although pixel-space diffusion models produce impressive results, they carry a heavy computational cost because they must process every pixel at full resolution. Latent Diffusion Models (LDMs) solve this by moving the entire diffusion process into a compressed latent space that still captures the essential semantic information. The LDM approach consists of three components: (1) a VAE encoder that compresses high-resolution images into lower-dimensional latent representations—typically shrinking spatial size by a factor of 8; (2) the diffusion model, which performs forward noising and reverse denoising entirely within this compact latent space; and (3) a VAE decoder that reconstructs a full high-quality image from the denoised latent.

B. VAE Design Considerations

The quality of the VAE encoder-decoder pair plays a critical role in final image quality. The VAE must balance accurate reconstruction against creating a clean, well-organised latent space for the diffusion model. A known challenge is semantic entanglement in standard VAE latent spaces, which can interfere with diffusion model learning. While VAE parameters have traditionally been frozen during diffusion training, newer approaches that jointly train both components in an end-to-end fashion have shown promising results, automatically improving the latent space structure across various VAE designs.

C. Computational Benefits

By working in a compressed latent space, LDMs handle 64× fewer spatial elements per denoising step (8×8 spatial compression). Combined with the typically lower channel count of latent representations compared to full RGB images, this makes high-resolution image generation practical on consumer GPUs—something that would be extremely slow or infeasible with pixel-space diffusion.

V. RECENT ADVANCES

A. Flow Matching

Flow Matching (FM) has emerged as a strong alternative to traditional diffusion models, offering faster inference while delivering comparable or superior results. Instead of a discrete-time Markov chain, Flow Matching defines a smooth, continuous-time Ordinary Differential Equation (ODE) that creates a direct path from a simple base distribution (Gaussian noise) to the target data distribution. This results in a straighter, more efficient sampling trajectory compared to the curved paths of standard DDPM.

The Diff2Flow framework enables converting existing pre-trained diffusion models into the Flow Matching setup by realigning timesteps, adjusting the interpolation schedule, and converting model outputs into a compatible velocity field. Diff2Flow outperforms both standard Flow Matching trained from scratch and conventional diffusion fine-tuning approaches, especially under limited parameter budgets.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 18, April 2026
Impact Factor: 8.2

B. Accelerated Sampling

A major research focus has been reducing the number of denoising steps required at inference. The Distribution Matching Distillation (DMD) method achieves up to 30× faster generation by training a compact student model to match the full output distribution of powerful multi-step teachers such as DALL·E 3 and Stable Diffusion. This collapses up to 100 denoising steps into a single network pass while maintaining or even improving output quality.

C. Ultra-High-Resolution Synthesis

Recent advances have enabled diffusion-based generation at true 4K resolution (3840×2160 pixels). The Diffusion-4K framework addresses the challenges of this scale through two key innovations: (1) a wavelet-based fine-tuning strategy that supports direct training on 4K photorealistic images across various latent diffusion architectures; and (2) architectural modifications that make the high memory and computation demands of 4K generation feasible without exploding resource requirements.

D. Cascaded Generation

Rather than generating large images all at once, cascaded diffusion pipelines break the job into stages. The process typically begins by synthesising a small low-resolution draft (e.g., 64×64 pixels) from the text prompt, and then progressively upsamples through intermediate stages to the final target resolution (e.g., 1024×1024). By resolving overall layout and composition at low resolution, the system saves substantial computation and reserves high-resolution processing for fine texture and detail rendering

VI. PRACTICAL IMPLEMENTATIONS

A. Stable Diffusion

Stable Diffusion, released in 2022, quickly became one of the most prominent open-source generative AI systems. By applying latent diffusion—operating in the compressed latent space of a pre-trained VAE—it made high-resolution text-to-image generation feasible on standard consumer hardware. The decision to release model weights under a permissive open licence triggered a large ecosystem of fine-tuned models, tooling, and downstream research.

B. DALL-E 3

DALL-E 3, developed by OpenAI, represents the state of the art in commercial text-to-image systems. It combines advanced diffusion techniques with strong language understanding from large language models, resulting in consistently high benchmark performance. It is particularly effective at handling complex compositional prompts involving many objects, with superior spatial coherence and fewer artefacts compared to most alternatives.

C. Comparative Analysis

While DALL·E 3 often leads in photorealism and prompt adherence, Stable Diffusion remains competitive in image quality and offers full open-source accessibility and customisability. Stable Cascade employs a cascaded latent diffusion approach, matching Stable Diffusion XL in overall performance while reducing major errors across several benchmarks. Selecting the appropriate model depends on output quality requirements, available hardware, customisation needs, and budget.

TABLE I: COMPARISON OF MAJOR DIFFUSION-BASED IMAGE GENERATION SYSTEMS

Model	Architecture	Strengths	Availability
Stable Diffusion	Latent Diffusion	Image fidelity, detail	Open-source
DALL-E 3	Advanced Diffusion	Text adherence, quality	Commercial API
Midjourney	Proprietary	Artistic quality	Commercial

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 18, April 2026
Impact Factor: 8.2

Model	Architecture	Strengths	Availability
Stable Cascade	Cascaded LDM	Quality, flexibility	Open-source

VII. CHALLENGES AND FUTURE DIRECTIONS

A. Current Limitations

Despite impressive progress, diffusion models still face several important challenges. Inference speed remains a real obstacle: even with distillation techniques, diffusion models are noticeably slower than single-pass generators, and real-time high-resolution generation remains an open engineering problem.

Fine-grained control is another difficult area. While high-level conditioning through text is well established, precisely editing specific image regions or adjusting individual semantic attributes without heavy prompt engineering remains challenging. Evaluation methodology has also not kept pace with model capabilities, as there is still no widely accepted comprehensive metric set for jointly measuring image quality, diversity, and prompt fidelity.

Training large foundation models requires massive curated datasets and enormous computing resources, effectively limiting serious development to well-funded organisations. Additionally, diffusion models still struggle with complex compositional reasoning—accurately counting objects, placing them in correct relative positions, and maintaining logical scene consistency.

While excelling in sample quality, diffusion models can still under-represent rare data modes due to their reliance on score matching, which prioritizes high-density regions. This leads to subtle biases in tail distributions, such as underrepresented skin tones in portraits or atypical viewpoints in 3D scenes. Mitigation strategies like adaptive sampling or mode-seeking losses help, but fully equitable coverage across diverse real-world datasets remains elusive.

B. Emerging Research Directions

Several promising directions are shaping the near future of diffusion models. On the inference side, researchers are developing smarter sampling schedules and learnable samplers that drastically reduce step counts while preserving output quality. Cross-modal generation is another active area, aiming to build unified models capable of handling images, video, 3D content, and audio using a single diffusion backbone.

Improved conditioning techniques are needed to give users more precise and intuitive control over generated outputs. For practical deployment, pruning, quantisation, and knowledge distillation will be essential for making large models efficient enough to run on standard hardware. Theoretical work on generalisation bounds and developing robust content safety and alignment mechanisms will remain important as these systems are deployed at scale.

C. Long-term Vision

Looking further ahead, diffusion models have the potential to become core infrastructure for next-generation AI creative tools. Tighter integration with large language models is bringing us closer to interfaces that can translate detailed natural language descriptions directly into complex visual outputs. Extending these capabilities to video and 3D generation will transform workflows in animation and virtual reality. As mathematical understanding deepens and sampling speeds increase, diffusion models are poised to transition from specialised research tools into standard everyday instruments for creative and professional work.

D. Ethical and Safety Concerns

Unintentional memorization from web-scale training data risks generating near-duplicates of copyrighted works, prompting defenses like DALL-E's CLIP filtering. Bias amplification occurs through uneven conditioning response, where prompts for "CEO" skew male despite balanced training corpora. Safety alignments via human preference optimization (e.g., RLHF on red-teamed outputs) reduce harms, but adversarial prompt engineering can still elicit unsafe content, necessitating dynamic runtime safeguards.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 18, April 2026
Impact Factor: 8.2

VIII. CONCLUSION

Over the past few years, diffusion models have fundamentally transformed generative modelling for images, setting new standards for output quality, training stability, and architectural flexibility. From the original DDPM framework to efficient latent diffusion approaches, these models have demonstrated the power of iterative denoising as a generative paradigm.

Breakthroughs in flow matching and accelerated sampling have significantly advanced the field—enabling higher resolutions without prohibitive compute costs. The successful deployment of Midjourney, DALL-E 3, and Stable Diffusion in commercial workflows confirms that diffusion models have moved well beyond the research stage. As the community continues to address remaining challenges in inference speed, fine-grained control, and semantic understanding, diffusion models are well positioned to define the next era of visual AI.

Extensions like Video Diffusion Models apply cascaded 3D U-Nets to generate temporally coherent clips from text, achieving realistic motion in tools such as Sora by factorizing space-time into spatial then temporal diffusion passes. AudioLDM tokenizes waveforms into semantic tokens for music and speech synthesis, bridging visual conditioning with sound via shared CLIP embeddings. In 3D, DreamFusion optimizes Neural Radiance Fields (NeRFs) via score distillation sampling, rendering text-prompted objects with consistent geometry from novel viewpoints.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of the research community whose work on denoising diffusion probabilistic models, latent diffusion, and accelerated sampling forms the foundation of this survey.

REFERENCES

[1] Geometry and Computing Group, UCL. (2024). Diffusion Models for Image and Video Generation: From Foundations to Emerging Directions. SIGGRAPH 2025 Course.
[2] Yin, T., et al. (2024). Distribution Matching Distillation for accelerating diffusion models. MIT News.
[3] AI Research Blog. (2026). DDPM: Denoising Diffusion Probabilistic Models.
[4] DigitalOcean. (2025). 8 Stable Diffusion Alternatives for Image Generation in 2025.
[5] National Institutes of Health. (2024). Evaluating Text-to-Image Generated Photorealistic Images. PMC.
[6] Nichol, A., & Dhariwal, P. (2021). Improved Denoising Diffusion Probabilistic Models. ICML.
[7] Zhang, J., et al. (2025). Diffusion-4K: Ultra-High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2025.
[8] IJCRT. (2024). Stable Diffusion, Dall-E and Dream by WOMBO: Comparative Analysis.
[9] LearnOpenCV. (2024). In-Depth Guide to Denoising Diffusion Probabilistic Models DDPM.
[10] BentoML. (2023). The Best Open-Source Image Generation Models in 2026.
[11] Wikipedia. (2022). Diffusion model.
[12] Li, Y., et al. (2025). REPA-E: Unlocking VAE for End-to-End Tuning with Latent Diffusion Models. arXiv:2504.10483.
[13] Schusterbauer, D., et al. (2025). Diff2Flow: Training Flow Matching Models via Diffusion Model Alignment. CVPR 2025.
[14] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. NeurIPS.
[15] Song, Y., & Ermon, S. (2019). Generative Modeling by Estimating Gradients of the Data Distribution. NeurIPS.
[16] Song, Y., et al. (2021). Score-Based Generative Modeling through Stochastic Differential Equations. ICLR.
[17] Rombach, R., et al. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. CVPR.
[18] Dhariwal, P., & Nichol, A. (2021). Diffusion Models Beat GANs on Image Synthesis. NeurIPS.
[19] Kingma, D. P., et al. (2021). Variational Diffusion Models. NeurIPS.
[20] Karras, T., et al. (2022). Elucidating the Design Space of Diffusion-Based Generative Models. NeurIPS