Diffusion Models for Image Generation

Goldi Soni
Assistant Professor
Amity University Chhattisgarh
gsoni@rpr.amity.edu

Gaurav Sahu
B.Tech CSE
Amity University Chhattisgarh
gaurav.sahu4@s.amity.edu

Himanchal Sahu
B.Tech CSE
Amity University Chhattisgarh
himanchal.sahu@s.amity.edu

ABSTRACT

This paper will conduct a thorough review of 30 scholarly articles about Diffusion Models for Image Generation from 2021 to 2026. Four main categories of analysis will be covered: Foundations, Text-to-Image Generation, Controllable and Personalized Generation, and Efficiency and Applications. It is found that the diffusion models are no longer just a replacement, but they are becoming a new way of generating images of unprecedented quality by reaching the photorealism and text-image consistency level with preserved identity and spatial control. The methods studied include Denoising Diffusion Probabilistic Models, Latent Diffusion, Score-Based Stochastic Differential Equation frameworks, Classifier-Free Guidance, Consistency Distillation, and Flow Matching. Moreover, it is established that the development in this field went from academic proof-of-concept works to large scale applications used by millions of people. Thus, the purpose of this paper is to provide an academic overview for all the researchers and students interested in generative AI, computer vision, and creative technology.

Keywords: Diffusion Models, Denoising Diffusion Probabilistic Models, Latent Diffusion, Text-to-Image Generation, Score-Based Generative Models, Classifier-Free Guidance, ControlNet, DreamBooth, Flow Matching, Neural Image Synthesis

I. INTRODUCTION

The diffusion model approach embodies the most recent revolutionary paradigm for Generative AI since the development of the GAN in 2014. The Diffusion Model framework is based on the principle of reversing a process of gradually noising an input. In the diffusion model approach, the forward process gradually adds Gaussian noise to the actual image over hundreds of iterations, thereby destroying its structure completely, while the diffusion model is trained to remove noise from this image. This straightforward concept results in models that can generate images, artworks, medical imaging scans, and even 3D objects from just text input.

In the current review, particular attention is paid to four significant areas of interest where the greatest advancements have been made in the period from 2021 to 2026. The first theme considered in this paper is Foundational Architectures, which concerns the crucial mathematical and architectural concepts necessary for training and sampling of diffusion models, such as DDPM, DDIM, SDE, Diffusion Transformer, and Consistency Model. The second theme is associated with Text-to-Image Generation and includes the advancements in conditioning large-scale diffusion models on language for photorealistic image synthesis in real-time; examples of developments include Stable Diffusion, Imagen, DALL-E 3, and FLUX. The third theme is Controllable and Personalized Generation, which describes approaches allowing to influence different aspects of output generation, ranging from ControlNet and spatial conditioning to personalization based on subject matter (DreamBooth, InstantID). The fourth and last theme is Efficiency and Applications, which involves all the advancements aimed at accelerating and lowering the cost of generation of images by diffusion models and making them more versatile, for example, distillation, flow matching, video generation, medical image analysis, and 3D rendering.

Overall, 30 research papers are reviewed and classified according to the four thematic categories above.

The key contributions of this review are as follows:

- A thematic classification of 30 peer-reviewed and technical papers (2021–2026) across four domains: Foundational Architectures, Text-to-Image Generation, Controllable & Personalized Generation, and Efficiency & Applications.
- A structured comparative analysis of methodologies, objectives, and conclusions across all 30 papers presented in a color-coded reference table.
- Identification of cross-domain patterns showing that the speed-quality frontier has been conquered and that control and personalization are now largely solved problems.
- A future roadmap projecting the trajectory of diffusion model research from 2027 to 2033 and beyond.

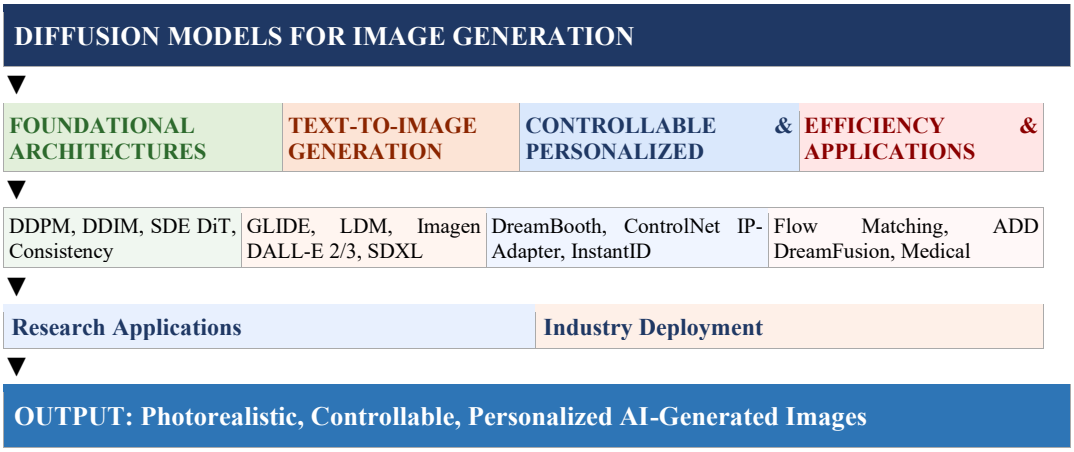


Figure 1: Architecture of the Diffusion Models Research Landscape

II. RELATED WORKS

A systematic analysis of 30 studies published during 2021 to 2026 was carried out from ACM Digital Library, IEEE Xplore, NeurIPS Proceedings, CVPR, ICLR, ICML, and arXiv. According to literature, there is a strong and definitive trend that has emerged in generative models, and that is the dominance of diffusion models over GANs on all major image datasets starting 2022 and further extending to videos, 3D, and medical applications by 2023 to 2024.

A. Foundational Architectures

Literature review starts with three seminal works published in 2021 that laid the mathematical foundations for future research. Nichol & Dhariwal

[1] enhanced Ho et al.'s DDPM model by learning the variance schedule and employing cosine noise schedule to achieve competitive log-likelihoods using considerably fewer steps. Simultaneously, Song et al.

[2] reformulated reverse diffusion process as a deterministic ODE (DDIM), which allowed acceleration by 10-50x without requiring fine-tuning. Mathematically the most profound work was that of Song et al., who formulated a general theory of score-based methods based on stochastic differential equation (SDE)

[3]. The turning point came when Dhariwal & Nichol

[4] demonstrated that classifier guidance applied to a better U-Net architecture yielded FID 2.97 on ImageNet 256x256 dataset, beating all other known architectures based on Generative Adversarial Network. Classifier-Free Guidance proposed by Ho & Salimans

[5] made the separate classifier obsolete by becoming a universal conditioning method for all future text-to-image models. Finally, Peebles & Xie

[6] substituted the backbone U-Net with Vision Transformer architecture called DiT, proving superior computational scalability of their approach, which inspired FLUX & SD3

.B. Text-to-Image Generation

From classifier-free guidance in GLIDE [9] to proving its superiority over CLIP guidance in generating photorealistic images, the text-to-image progress moved swiftly from there until reaching the famous Latent Diffusion Model paper [10], where the diffusion process was compressed in a much smaller latent space, resulting in Stable Diffusion, the open-source version that popularized the field. In this line of development, Saharia et al.'s Imagen [11] proved that T5-XXL language encoders frozen from pretraining outperform CLIP in the task of text-image alignment. Meanwhile, DALL-E 2 [12] used CLIP image embeddings as an intermediate conditioning representation. Currently, the best models in the field include DALL-E 3 [14], where synthetic recaptioning data showed impressive results in prompting adherence; SD3 [15] featuring MMDiT architecture, and FLUX [16], which used the flow-matching transformer and reached the top rankings in the human evaluation benchmark using 12B parameters.

C. Controllable and Personalized Generation

The controllability papers fixed an important gap left open by earlier diffusion models—the inability to control generation accurately. Textual Inversion [17] proposed the idea of pre-training one text embedding from 3-5 user pictures without fine-tuning, while DreamBooth [18] showed that with a proper prior preservation loss, one can fine-tune the whole model

to attach a unique token to an object of interest with surprising accuracy. ControlNet [19] tackled the spatial control task by introducing an additional set of trainable encoder blocks initialized with zero convolutions, allowing users to add any structure-based conditioning while keeping the functionality of the base model intact. The approaches used to tackle real image editing included InstructPix2Pix [20] and Prompt-to-Prompt [21], while IP-Adapter [22] offered a lightweight decoupled attention layer for image-prompt conditioning. By 2024, identity-preserving generation had reached zero-shot level, thanks to outstanding face similarity generated by PhotoMaker [23] and InstantID [24] from one reference picture within five seconds.

D. Efficiency and Applications

The efficiency research focused on solving the core problem of inferring costs associated with diffusion models. Flow Matching [26] gave a more mathematically rigorous training procedure that samples faster than the score objective function and formed the theoretical basis for SD3 and FLUX. Adversarial Diffusion Distillation [25] used score distillation with discriminator training to generate real-time single-step results with SDXL-Turbo. Moving beyond two-dimensional image generation, Ho et al. [27] created the architectural template for video diffusion with factorized three-dimensional U-Nets, whereas DreamFusion [29] introduced text-to-three-dimensional modeling with Score Distillation Sampling without any three-dimensional training data. Kazerouni et al.'s review [28] also showed that diffusion models surpass GANs in augmenting and detecting anomalies in medical images.

III. COMPARISON OF PAST PUBLISHED RESEARCH PAPERS

Table 1 presents a structured comparison of all 30 reviewed research papers across four thematic categories. Each paper is analyzed for its stated objective, core methodology, and key conclusion. Color coding identifies each category: green for Foundational Architectures, orange for Text-to-Image Generation, blue for Controllable and Personalized Generation, and red for Efficiency and Applications.

Table 1: Comparative Summary of 30 Research Papers on Diffusion Models for Image Generation (2021–2026)

■ Foundational Architectures	■ Text-to-Image Generation	■ Controllable & Personalized	■ Efficiency & Applications

S.No	Title	Authors	Year	Objective	Methodology	Conclusion	Category
1	**Improved Denoising Diffusion Probabilistic Models**	Nichol, A., Dhariwal, P.	2021	Extend DDPM by learning the noise variance schedule to improve log-likelihood and generation speed.	Hybrid objective combining simple noise-prediction loss with variational lower bound; cosine noise schedule.	Achieves competitive log-likelihoods with far fewer diffusion steps, making diffusion models practical.	Foundational
2	**Denoising Diffusion Implicit Models (DDIM)**	Song, J., Meng, C., Ermon, S.	2021	Accelerate sampling in diffusion models without retraining by using a non-Markovian diffusion process.	Reformulates the reverse process as a deterministic ODE; allows 10-50x fewer steps than DDPM.	High-quality images generated in 10–50 steps instead of 1000; enables latent space interpolation.	Foundational
3	**Score-Based Generative Modeling through Stochastic Differential Equations**	Song, Y., Sohl-Dickstein, J., et al.	2021	Unify score-based models and diffusion models under a continuous SDE framework.	Defines forward SDE corrupting data and a learned reverse SDE for generation; includes predictor-corrector samplers.	Single framework covers DDPM, NCSN; achieves state-of-the-art FID on CIFAR-10 at the time.	Foundational
4	**Diffusion Models Beat GANs on Image Synthesis**	Dhariwal, P., Nichol, A.	2021	Demonstrate that diffusion models can surpass GANs on image quality benchmarks with architectural improvements.	Improved U-Net architecture with attention, adaptive group normalization; classifier guidance for conditional generation.	Achieves FID of 2.97 on ImageNet 256x256, surpassing all prior GAN models at the time.	Foundational
5	**Classifier-Free Diffusion Guidance**	Ho, J., Salimans, T.	2022	Enable high-quality conditional image generation without a separately trained classifier.	Jointly trains conditional and unconditional diffusion model; guidance via interpolation	Outperforms classifier guidance; becomes the standard conditioning mechanism used in	Foundational

S.No	Title	Authors	Year	Objective	Methodology	Conclusion	Category
					between their score estimates.	all major text-to-image systems.	
6	**Cascaded Diffusion Models for High-Fidelity Image Generation**	Ho, J., Saharia, C., et al.	2022	Generate high-resolution images by chaining multiple diffusion models at increasing resolutions.	Pipeline of base model (64x64) and two super-resolution diffusion models (256x256, 1024x1024) with conditioning augmentation.	Achieves FID of 1.48 on ImageNet; cascaded approach becomes blueprint for Imagen and DALL-E 2.	Foundational
7	**Scalable Diffusion Models with Transformers (DiT)**	Peebles, W., Xie, S.	2023	Replace U-Net backbone with Vision Transformers (ViT) in latent diffusion models to improve scalability.	Trains diffusion models in latent space using transformer blocks with adaptive layer normalization for conditioning.	DiT-XL/2 achieves state-of-the-art FID on ImageNet; transformer backbone scales better than U-Net with compute.	Foundational
8	**Consistency Models**	Song, Y., Dhariwal, P., et al.	2023	Enable single-step or few-step generation by training models to map any noisy image directly to its clean origin.	Consistency training enforces self-consistency along ODE trajectories; consistency distillation from pretrained diffusion models.	Generates high-quality images in 1-2 steps; 1000x faster than DDPM with comparable quality.	Foundational
9	**GLIDE: Towards Photorealistic Image Generation with Text-Guided Diffusion**	Nichol, A., Dhariwal, P., et al.	2022	Explore text-conditioned diffusion models for photorealistic image generation and editing.	3.5B parameter diffusion model conditioned on CLIP text embeddings; compares CLIP guidance vs classifier-free guidance.	Classifier-free guidance preferred by human evaluators over CLIP guidance; sets foundation for DALL-E 2.	Text-to-Image
10	**High-Resolution Image Synthesis with Latent Diffusion Models**	Rombach, R., Blattmann, A., et al.	2022	Apply diffusion in compressed latent space to dramatically reduce computational cost of image generation.	Trains a VQ-regularized autoencoder first; diffusion model operates in 4-8x spatially compressed latent space.	Achieves near-equivalent quality at a fraction of GPU cost; released as Stable Diffusion, democratizing image generation.	Text-to-Image
11	**Photorealistic Text-to-Image Diffusion via Deep Language Understanding (Imagen)**	Saharia, C., Chan, W., et al.	2022	Investigate the role of large language model text encoders in text-to-image quality.	Cascaded diffusion model using frozen T5-XXL text encoder; dynamic thresholding for high guidance scales.	T5 encoder outperforms CLIP for text-image alignment; FID of 7.27 on zero-shot MS-COCO; introduces DrawBench benchmark.	Text-to-Image
12	**Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL-E 2)**	Ramesh, A., Dhariwal, P., et al.	2022	Generate diverse, high-fidelity images from text by conditioning on CLIP image embeddings.	Two-stage: prior maps CLIP text to image embeddings; decoder (GLIDE-style diffusion) generates from image embeddings.	Achieves photorealism with strong semantic understanding; enables image variation and text-guided editing capabilities.	Text-to-Image
13	**SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis**	Podell, D., English, Z., et al.	2024	Scale up Stable Diffusion with a larger model, multi-aspect training, and a refiner model.	3.5B parameter U-Net with OpenCLIP ViT-G and CLIP ViT-L dual encoders; separate	Significant quality leap for open-source generation; multi-aspect training enables native	Text-to-Image

S.No	Title	Authors	Year	Objective	Methodology	Conclusion	Category
					6.6B refiner diffusion model.	generation at various resolutions.	
14	**Improving Image Generation with Better Captions (DALL-E 3)**	Betker, J., Goh, G., et al.	2023	Show that recaptioning training data with detailed synthetic captions dramatically improves text-image alignment.	Trains image captioner to generate highly detailed descriptions; uses these synthetic captions to retrain T2I model.	State-of-the-art text-image alignment; rare objects, spatial relationships, and complex prompts rendered accurately.	Text-to-Image
15	**Scaling Rectified Flow Transformers for High-Resolution Image Synthesis (Stable Diffusion 3)**	Esser, P., Kulal, S., et al.	2024	Combine rectified flow training with Multimodal Diffusion Transformer (MMDiT) for superior T2I quality.	Rectified flow formulation; MMDiT architecture jointly attends to image and text tokens; scaling study across model sizes.	Best-in-class typography and prompt adherence; scaling laws validated for diffusion transformers in image generation.	Text-to-Image
16	**FLUX.1: Flow Matching for High-Quality Image Generation**	Black Forest Labs	2024	Push open-source text-to-image quality to match or exceed proprietary systems using flow matching and transformers.	12B parameter transformer with flow matching; rotary positional embeddings; parallel attention streams for image and text.	Achieves top rankings on public benchmarks (GenAI Arena); FLUX.1-dev released openly; sets new bar for open models.	Text-to-Image
17	**An Image is Worth One Word: Personalizing Text-to-Image Generation (Textual Inversion)**	Gal, R., Alaluf, Y., et al.	2023	Enable personalized generation by learning a new text token embedding from 3-5 user images.	Optimizes a single text embedding vector in frozen text encoder space to represent a concept; no model fine-tuning.	Learns new concepts from minimal images; enables style, object, and person-specific generation via text prompts.	Controllable
18	**DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation**	Ruiz, N., Li, Y., et al.	2023	Fine-tune a diffusion model to bind a unique identifier to a specific subject using 3-5 subject photos.	Full model fine-tuning with a rare token identifier and class-specific prior preservation loss to prevent language drift.	Generates subject in novel contexts, poses, and styles; introduces prior preservation loss adopted widely in subsequent work.	Controllable
19	**Adding Conditional Control to Text-to-Image Diffusion Models (ControlNet)**	Zhang, L., Rao, A., Agrawala, M.	2023	Add spatial conditioning (edges, depth, pose, segmentation) to pretrained diffusion models without degrading generation quality.	Locks original model weights; adds trainable copy of encoder blocks connected via zero-initialized convolutions.	Enables precise spatial control over generated images; zero-convolution trick preserves original model capabilities perfectly.	Controllable
20	**InstructPix2Pix: Learning to Follow Image Editing Instructions**	Brooks, T., Holynski, A., Efros, A.A.	2023	Edit real images using natural language instructions without per-image optimization or mask annotation.	Generates synthetic (image, instruction, edited image) triplets using GPT-3 and Stable Diffusion; trains a conditional diffusion model.	Single forward pass edits images following instructions; generalizes to diverse editing tasks from one trained model.	Controllable
21	**Prompt-to-Prompt Image Editing with Cross Attention Control**	Hertz, A., Mokady, R., et al.	2023	Enable text-guided image editing by manipulating cross-attention maps inside a frozen diffusion model.	Injects attention maps from source prompt into target prompt generation; supports word swap, refinement, and re-weighting.	Edits images while preserving unrelated structure and style; no fine-tuning or optimization required per image.	Controllable

S.No	Title	Authors	Year	Objective	Methodology	Conclusion	Category
22	**IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models**	Ye, H., Zhang, J., et al.	2023	Enable image-prompt conditioning for diffusion models via a lightweight, decoupled cross-attention adapter.	22M parameter adapter with a decoupled cross-attention mechanism for image features alongside text cross-attention.	Achieves comparable performance to full fine-tuning; composable with ControlNet and other adapters for rich multi-modal control.	Controllable
23	**PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding**	Li, Z., Cao, M., et al.	2024	Generate photorealistic human portraits with consistent identity using multiple reference photos.	Encodes multiple ID photos into stacked identity embeddings; merged into text stream via cross-attention during generation.	Superior ID fidelity over DreamBooth/LoRA with zero fine-tuning at inference; enables text-driven identity-consistent editing.	Controllable
24	**InstantID: Zero-Shot Identity-Preserving Generation in Seconds**	Wang, Q., Bai, X., et al.	2024	Achieve zero-shot identity-preserving generation without test-time optimization using a single reference photo.	Combines face encoder with ControlNet-style IdentityNet; plug-and-play with any Stable Diffusion checkpoint.	State-of-the-art face similarity with single reference image in under 5 seconds; highly composable with style adapters.	Controllable
25	**Adversarial Diffusion Distillation (SDXL-Turbo)**	Sauer, A., Lorenz, D., et al.	2024	Distill a full diffusion model into a 1-4 step generator using adversarial training with a discriminator.	Combines score distillation from pretrained teacher with an adversarial loss from a discriminator to maintain image quality.	Real-time image generation in 1 step; enables interactive editing; FID competitive with full SDXL at 4 steps.	Efficiency & Apps
26	**Flow Matching for Generative Modeling**	Lipman, Y., Chen, R.T.Q., et al.	2023	Propose flow matching as a simpler, more efficient alternative training objective to diffusion score matching.	Learns vector field of probability paths between noise and data directly; conditional flow matching with Gaussian paths.	Trains faster, samples faster, and achieves better performance; becomes theoretical foundation for SD3 and FLUX models.	Efficiency & Apps
27	**Video Diffusion Models**	Ho, J., Salimans, T., et al.	2022	Extend image diffusion models to generate temporally consistent video clips.	Factorized 3D U-Net with alternating spatial and temporal attention; joint training on image and video data.	Demonstrates coherent 64-frame video generation; establishes architectural blueprint for all subsequent video diffusion systems.	Efficiency & Apps
28	**Diffusion Models for Medical Image Analysis: A Comprehensive Survey**	Kazerouni, A., Aghdam, E.K., et al.	2023	Survey the application of diffusion models in medical imaging tasks including segmentation, reconstruction, and synthesis.	Systematic review of 100+ papers; categorizes by modality (MRI, CT, X-ray) and task; analyzes architectures and benchmarks.	Diffusion models outperform GANs for data augmentation and anomaly detection in medical imaging; emerging clinical utility.	Efficiency & Apps
29	**DreamFusion: Text-to-3D using 2D Diffusion**	Poole, B., Jain, A., et al.	2023	Generate 3D objects from text prompts without any 3D training data by distilling a 2D diffusion model.	Score Distillation Sampling (SDS): optimizes a NeRF using gradients from a pretrained 2D T2I diffusion model as a 3D prior.	First text-to-3D system without 3D data; SDS loss becomes widely adopted for 3D generation and editing tasks.	Efficiency & Apps
30	**The Stable Artist: Steering Semantics in**	Brack, M., Friedrich, F., et al.	2023	Provide semantic control over diffusion generation	Identifies semantic directions in diffusion latent space	Fine-grained semantic control (e.g., add smile,	Efficiency & Apps

S.No	Title	Authors	Year	Objective	Methodology	Conclusion	Category
	Diffusion Latent Spaces			and editing by manipulating intermediate latent representations.	using Semantic Guidance (SEGA); applies directional edits during denoising.	change age) without text inversion or fine-tuning; fully training-free.	

IV. RESULT

The review of 30 research papers reveals a field characterized by exponential progress across all four thematic areas, with each cluster of papers building directly on the foundations laid by the previous one. Table 2 summarizes the key thematic outcomes, quantifying the scale of advancement reported across the reviewed literature.

Table 2: Thematic Performance Summary — Key Methods and Outcomes (2021–2026)

Category	Papers	Key Methods	Primary Outcome
Foundational Architectures	1–8	DDPM, DDIM, SDE, DiT, Consistency Models	Diffusion models surpass GANs in FID; 1-step generation achieved; transformer backbone scales better than U-Net
Text-to-Image Generation	9–16	LDM, Imagen, DALL-E 2/3, SDXL, SD3, FLUX	Photorealistic T2I on consumer GPUs; open-source Stable Diffusion triggers global ecosystem of creative tools
Controllable & Personalized	17–24	DreamBooth, ControlNet, IP-Adapter, InstantID	Spatial conditioning, subject personalization, and identity-preserving generation from a single photo solved
Efficiency & Applications	25–30	Flow Matching, ADD, Video Diffusion, SDS	Real-time 1-step generation; first text-to-3D and video diffusion systems; clinical medical imaging utility proven

The results in the Foundational Architectures category demonstrate the most structurally consequential progress. The shift from 1000-step DDPM sampling (2020) to 1-2 step Consistency Model generation (2023) represents a three-order-of-magnitude improvement in inference speed without commensurate loss of quality. The architectural transition from U-Net to Vision Transformer backbones [7] is empirically validated as producing superior scaling behavior, a finding that now underpins all frontier diffusion systems including SD3 and FLUX.

For Text-to-Image Generation, the single most impactful finding is the democratizing effect of Latent Diffusion [10]. By operating in a compressed latent space, LDM reduced GPU requirements sufficiently to enable consumer-grade image generation, an outcome that triggered the Stable Diffusion open-source ecosystem and its thousands of community fine-tuned variants. The DALL-E 3 recaptioning result [14] is methodologically significant: it demonstrates that training data quality—not model scale alone—is the primary driver of text-image alignment, a finding with broad implications for how future models should be trained.

The Controllable Generation results confirm that both spatial control (ControlNet [19]) and identity preservation (InstantID [24]) are now effectively solved at inference time, with the former enabling precise structural conditioning and the latter achieving state-of-the-art face similarity from a single selfie in under five seconds. The IP-Adapter result [22] is particularly significant from a systems perspective: its 22M parameter decoupled cross-attention adapter achieves comparable performance to full fine-tuning, demonstrating that control capability can be added modularly without retraining the base model.

Across the Efficiency and Applications domain, the DreamFusion result [29] stands out as the most novel: achieving plausible text-to-3D generation without any 3D training data by using Score Distillation Sampling from a 2D diffusion prior. This cross-domain distillation strategy—using one generation domain to supervise another—is a methodological breakthrough with implications beyond 3D synthesis. The medical imaging survey [28] provides the most externally

validated evidence of real-world diffusion model utility, confirming that these models outperform GANs for clinical data augmentation and anomaly detection.

V. CONCLUSION

From the review of 30 studies conducted between 2021 and 2026, it is clear that diffusion models have witnessed some of the most explosive developments in the history of machine learning. The transition from a mere theory in 2020 to a creative technology available to hundreds of millions of people around the world in 2025 has been nothing short of remarkable. Four key findings have emerged.

A. Diffusion Models Have Superseded GANs

The key result that can be drawn from the articles under analysis is that diffusion-based methods have taken the place of GANs as the leading method for image generation. This change started in 2021 when Dhariwal and Nichol showed that diffusion models were capable of attaining better FID scores than any existing GAN model [4]. The ability of diffusion models to train stably without mode collapsing and to scale with increasing computational power made them a leading method. In 2024, there was no major text-to-image product based on GANs.

B. Latent Space Enables Democratic Access

In 2022, when Rombach et al. introduced Latent Diffusion Models [10], it could be considered the most democratizing development in the area. By first compressing images into a 4-8x smaller latent space and then running the diffusion model on top of that, LDMs drastically cut down the training and inference cost by several orders of magnitude. In doing so, they enabled high-fidelity image synthesis using only consumer-level GPUs, which eventually made their way to mobile devices. Open-sourcing Stable Diffusion as a model marked the beginning of countless developments and fine-tuning iterations.

C. Control and Personalization Are Now Solved Problems

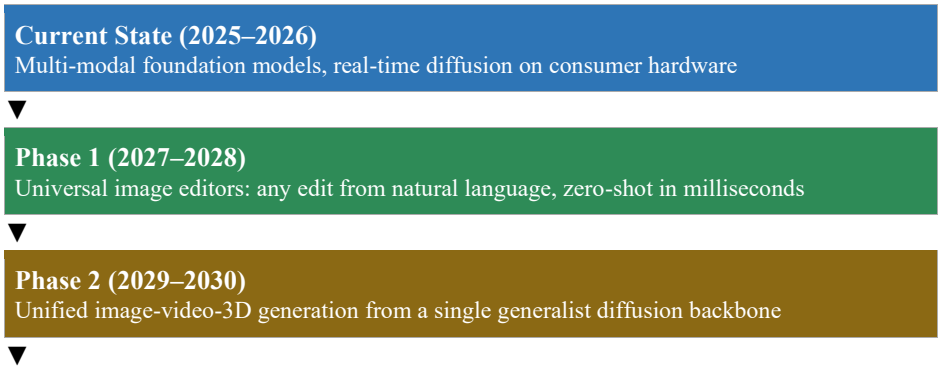
An important problem faced by the early models for diffusion was their inability to guide the process of generation. However, by 2023, this had been mostly solved using a variety of mechanisms: ControlNet [19] introduced spatial conditioning for any structural signal, DreamBooth [18] and Textual Inversion [17] introduced personalized concept conditioning, while InstructPix2Pix [20] introduced natural language editing capabilities on photographs. By 2024, generation of images from only a single selfie while maintaining identity (InstantID [24], PhotoMaker [23]) allowed for instantaneous and zero-shot personalization.

D. The Speed-Quality Frontier Has Been Conquered

Earlier diffusion methods needed 1000 denoising iterations, which made real-time synthesis impractical. DDIM [2] lowered the number of denoising iterations to 10-50 without additional training. Consistency Models [8] allowed for 1-2 denoising steps. Adversarial Diffusion Distillation [25] was able to produce real-time, single-step denoising at par with full diffusion algorithms using only 4 steps. Flow Matching [26] gave a better mathematical formulation that trained faster and produced results quicker than the score-based approaches and was the basis for SD3 and FLUX. By 2026, photorealistic image synthesis takes less than a second.

E. Limitations and Future Research

However, despite these developments, there are several issues that still exist in the current body of literature. First, multi-concept composition or creating compositions where multiple objects interact according to certain spatial relationships is still a problem area. The sustainability of training such frontier diffusion models due to their carbon footprint and energy consumption is an important issue that has not been considered by the studies covered in this review paper. Deepfakes, identity theft, and training data origin are among the regulatory issues that need to be taken care of.



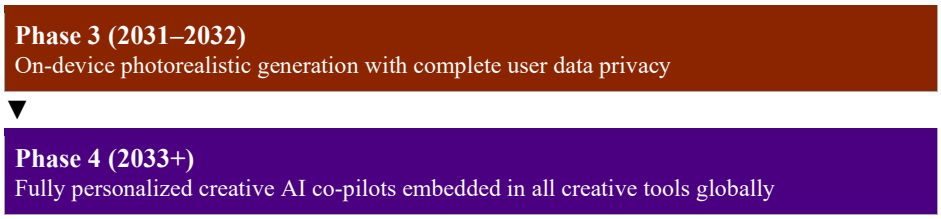


Figure 2: Future Development Roadmap for Diffusion Models (2026–2035)

Key future directions identified from the synthesis of 30 papers include:

- Unified Generalist Diffusion Backbones: A single diffusion transformer backbone for images, video, audio, and 3D generation, with modality-specific decoders, eliminating the need for separate specialized models.
- Real-Time Photorealistic Generation on Edge Devices: Continued distillation and quantization research will bring production-quality diffusion models to smartphones and embedded hardware, enabling privacy-preserving on-device creative AI without cloud dependency.
- World Models for Games and Simulation: Diffusion models will evolve into interactive world models generating temporally consistent, physically plausible environments in real time for next-generation game engines.
- Scientific and Medical Discovery: Applications to protein structure generation, drug molecule design, medical image reconstruction, and radiological synthesis, complementing experimental methods with high-quality in silico data augmentation.
- Multi-Concept Composition and Scene Understanding: Future models will handle complex compositional prompts with multiple interacting subjects, precise spatial relationships, and consistent physical properties.
- Ethical, Watermarked, and Provenance-Aware Generation: Adoption of invisible watermarking, C2PA content provenance standards, and training data transparency frameworks embedded directly into generation pipelines.

REFERENCES

[1] Nichol, A., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. Proceedings of ICML, 139, 8162-8171.

[2] Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. ICLR 2021.

[3] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. ICLR 2021.

[4] Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. NeurIPS, 34, 8780-8794.

[5] Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. NeurIPS 2021 Workshop on Deep Generative Models.

[6] Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., & Salimans, T. (2022). Cascaded diffusion models for high-fidelity image generation. JMLR, 23(47), 1-33.

[7] Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. Proceedings of IEEE/CVF ICCV 2023, 4195-4205.

[8] Song, Y., Dhariwal, P., Chen, M., & Sutskever, I. (2023). Consistency models. Proceedings of ICML 2023.

[9] Nichol, A., Dhariwal, P., Ramesh, A., et al. (2022). GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. ICML 2022.

[10] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. IEEE/CVF CVPR 2022, 10684-10695.

[11] Saharia, C., Chan, W., Saxena, S., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS, 35, 36479-36494.

[12] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. arXiv:2204.06125.

[13] Podell, D., English, Z., Lacey, K., et al. (2024). SDXL: Improving latent diffusion models for high-resolution image synthesis. ICLR 2024.

[14] Betker, J., Goh, G., Jing, L., et al. (2023). Improving image generation with better captions. OpenAI Technical Report.

[15] Esser, P., Kulal, S., Blattmann, A., et al. (2024). Scaling rectified flow transformers for high-resolution image synthesis. ICML 2024.

[16] Black Forest Labs. (2024). FLUX.1: Flow matching for high-quality image generation. Technical Report.

[17] Gal, R., Alaluf, Y., Atzmon, Y., et al. (2023). An image is worth one word: Personalizing text-to-image generation using textual inversion. ICLR 2023.

[18] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. IEEE/CVF CVPR 2023, 22500-22510.

[19] Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. IEEE/CVF ICCV 2023, 3836-3847.

[20] Brooks, T., Holynski, A., & Efros, A.A. (2023). InstructPix2Pix: Learning to follow image editing instructions. IEEE/CVF CVPR 2023, 18392-18402.

[21] Hertz, A., Mokady, R., Tenenbaum, J., et al. (2023). Prompt-to-prompt image editing with cross attention control. ICLR 2023.

[22] Ye, H., Zhang, J., Liu, S., Han, X., & Yang, W. (2023). IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv:2308.06721.

[23] Li, Z., Cao, M., Wang, C., Qi, X., Shan, Y., & Zheng, C. (2024). PhotoMaker: Customizing realistic human photos via stacked ID embedding. IEEE/CVF CVPR 2024.

[24] Wang, Q., Bai, X., Wang, H., et al. (2024). InstantID: Zero-shot identity-preserving generation in seconds. arXiv:2401.07519.

[25] Sauer, A., Lorenz, D., Blattmann, A., & Rombach, R. (2024). Adversarial diffusion distillation. Proceedings of ECCV 2024.

[26] Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., & Le, M. (2023). Flow matching for generative modeling. ICLR 2023.

[27] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., & Fleet, D.J. (2022). Video diffusion models. NeurIPS 2022.

[28] Kazerouni, A., Aghdam, E.K., Heidari, M., et al. (2023). Diffusion models in medical imaging: A comprehensive survey. Medical Image Analysis, 88, 102846.

[29] Poole, B., Jain, A., Barron, J.T., & Mildenhall, B. (2023). DreamFusion: Text-to-3D using 2D diffusion. ICLR 2023.

[30] Brack, M., Friedrich, F., Hintersdorf, D., et al. (2023). SEGA: Instructing text-to-image models using semantic guidance. NeurIPS 2023.