

AI Diffusion Models for Image Generation: A Review

Shaikh Iqra Arif, Prof. Bhosale Sachin Bajirao, Prof. Bangar Abhishek Popat

Dept of Computer Engineering Jaihind College of Engineering pune, india

Abstract- Diffusion models have become an important approach for generating images from text descriptions, random noise, and other conditions. They work by learning how to gradually remove noise from data until a meaningful image is formed. This review studies five publications from Springer and IEEE related to diffusion-based image generation: a broad survey of image-generation applications, a survey of lightweight diffusion models, the original Latent Diffusion Models paper, a survey of diffusion models in computer vision, and a review of text-to-image generative models. The selected works are compared in terms of their focus, main contribution, practical value, and limitations. Together, they show that image-generation research has progressed in three connected directions: improving visual quality and controllability, reducing computation, and better measuring how closely generated images follow user prompts. The review also identifies open issues involving evaluation, computational cost, dataset bias, privacy, copyright, and responsible use. This is a literature-based review; it does not report new experiments or claim that the five papers were tested under identical conditions.

Keywords— Diffusion Models, Image Generation, Text-to-Image Synthesis, Latent Diffusion, Lightweight Models, Computer Vision, Generative

I. INTRODUCTION

Artificial intelligence can now produce visual content from short written instructions. This task, commonly called text-to-image generation, is useful in design, education, advertising, entertainment, and image editing. A major family of methods behind this progress is diffusion models. Instead of producing the final image in one direct step, a diffusion model learns to reverse a gradual process in which noise is added to training images. During generation, it starts from noise and refines the result through repeated denoising steps [1], [4].

The quality of a generated image is only one part of the problem. A useful system should also follow the text prompt, preserve important details, run in a reasonable amount of time, and behave responsibly. The survey by Chen et al. discusses image-generation tasks such as style transfer, image completion, editing, and super-resolution, while also considering privacy and copyright concerns [1]. Song et al. focus on making diffusion models lighter because their repeated computations can make training and generation expensive [2].

Latent Diffusion Models (LDMs) offer one way to reduce this cost by performing the diffusion process in a compressed representation instead of directly in the full image-pixel space [3]. Croitoru et al. provide a broader view of diffusion models used in computer vision [4], and Alhabeeb and AlShargabi examine text-to-image methods, datasets, evaluation metrics, and remaining challenges [5].

The aim of this review is to connect the main ideas in these five publications, compare their contributions and limitations, and identify practical research gaps. The paper does not introduce a new diffusion model or perform a new benchmark. Its contribution is a clear synthesis of the selected literature for understanding the current direction of AI-based image generation.

II. REVIEW METHODOLOGY

This paper follows a narrative review approach. Five publications were selected because they cover complementary parts of the topic and are available through recognised publisher platforms. Two articles are published by Springer in Artificial Intelligence Review [1], [2]. The other three are an IEEE/CVF conference paper [3], an IEEE Transactions survey [4], and an IEEE Access review [5]. The selection covers general image-generation applications, model

efficiency, latent-space generation, computer-vision foundations, and text-to-image evaluation.

For each publication, the review records its main objective, approach, contribution, and limitations or open questions. The findings are then grouped into common themes rather than treated as a direct performance ranking. This distinction matters because the selected works do not all study the same task: some are broad surveys, while [3] presents a specific model design. They also draw on different datasets, metrics, and experimental settings. Therefore, their numerical results cannot be compared as if they came from one controlled experiment.

The review is limited to the five selected publications and their publisher descriptions and article content. It is not presented as a systematic review, and no claim is made that every relevant paper in the field has been included.

III. LITERATURE REVIEW

1. Comprehensive Survey of Diffusion Models in Image Generation

Chen et al. provide a broad survey of diffusion models and their uses in image generation [1]. The article explains the foundations of diffusion models and reviews applications including style transfer, image completion, image editing, and super-resolution. It also discusses the wider effects of generated images, including data privacy, security, malicious use, originality, and copyright. This makes the paper useful for understanding both the technical scope of diffusion models and the responsibilities that come with deploying them.

A strength of this work is its wide coverage: it connects the model's underlying process with real image-generation tasks and social implications. However, a broad survey cannot provide the same depth for every individual method, and the field continues to change quickly. Its discussion is therefore best used as a map of the research area rather than as a single experimental comparison of all current systems.

2. Lightweight Diffusion Models

Song et al. focus on a key limitation of diffusion models: the computation and storage needed during training and inference [2]. Their survey organises lightweight methods into categories such as knowledge distillation, quantisation, pruning, fine-tuning, signal-domain transformation, algorithm optimisation, and hybrid strategies. These approaches aim to reduce resource use while retaining useful image quality.

This work is important because a high-quality model is less practical if it needs excessive computing power or takes too long to generate each image. The survey helps explain that efficiency is not achieved through one universal trick; different methods reduce different sources of cost and may createtrade-offs. A limitation is that the best choice depends on the target model, hardware, image task, and acceptable quality loss. More consistent reporting of runtime, memory use, energy use, and visual quality would help users compare efficiency methods fairly.

3. High-Resolution Image Synthesis with Latent Diffusion Models

Rombach et al. introduce Latent Diffusion Models, a model approach that carries out diffusion in the latent representation of a pretrained autoencoder rather than directly on every image pixel [3]. The image is first encoded into a compressed representation, the diffusion model denoises that representation, and a decoder converts the final result back into an image. The paper also uses cross-attention to support conditioning inputs such as text and other spatial information.

The main contribution is a practical balance between image detail and computational demand. Working in latent space reduces the cost of operating on high-resolution images while maintaining strong synthesis quality. This approach influenced later text-to-image systems, including latent diffusion-based image generators. Its limitations include dependence on the quality of the autoencoder and training data; compression may also lose some fine details. In addition, generation remains iterative, so

efficiency and prompt accuracy still depend on the full system design.

4. Diffusion Models in Computer Vision

Croitoru et al. review diffusion models applied to computer vision and explain three broad frameworks: denoising diffusion probabilistic models, noiseconditioned score-based models, and stochastic differential equation approaches [4]. The survey also relates diffusion models to other generative approaches, including GANs, variational autoencoders, autoregressive models, energy-based models, and normalising flows. It categorises the ways diffusion models are used in vision and discusses their current limitations and future directions.

The value of this publication is its conceptual overview. It helps readers understand that diffusion is not one single architecture but a family of related approaches with different mathematical views and practical designs. The survey also highlights a well-known drawback: generating an image through many denoising steps can be slow. Because the paper covers computer vision broadly, it is less focused on the detailed evaluation of one particular text-to-image product or deployment setting.

5. Text-to-Image Synthesis with Generative Models

Alhabeeb and Al-Shargabi review text-to-image synthesis methods, including the development of GAN-based and diffusion-based systems [5]. Their article discusses commonly used datasets, evaluation metrics, previous tools and systems, and open challenges. This is relevant because text-to-image generation is not judged only by whether an image looks realistic. The image should also match the user's description and represent the requested objects and relationships.

The paper's strength is its attention to evaluation. It discusses measures such as Fréchet Inception Distance (FID), Inception Score, R-precision, CLIP Score, and human evaluation. These measures capture different aspects of quality, so one score alone cannot fully describe a system. The article also points to continuing issues in prompt alignment,

data quality, bias, computational efficiency, and fair comparison. Since the field develops rapidly, datasets and evaluation practices need to be revisited as new models appear.

6. Synthesis of the Reviewed Literature

Across the five publications, three connected priorities appear repeatedly. First, image quality and controllability remain central: a model should produce detailed images that follow the prompt. Second, efficiency matters because repeated denoising and large neural networks can require substantial time and computing resources [2], [3], [4]. Third, evaluation and responsible use matter: prompt alignment, diversity, bias, privacy, and copyright are not fully captured by a single image-quality score [1], [5].

The papers therefore complement each other rather than compete in a direct ranking. The broad surveys explain the field and its applications [1], [4]; the lightweight-model survey focuses on reducing resource requirements [2]; the LDM paper demonstrates a concrete architectural direction [3]; and the text-to-image review highlights datasets and evaluation [5].

IV. COMPARATIVE ANALYSIS

Table I summarises the main focus of each selected publication. The comparison is qualitative because the papers use different methods and scopes; the table should not be read as a numerical ranking.

Ref. / Publication	Main focus	Key contribution	Limitation / open issue
[1] Chen et al., 2023 (Springer)	Image generation applications and social impact	Connects diffusion foundations with editing, completion, style transfer, super-resolution, privacy and copyright.	Broad coverage means individual methods and benchmarks are not all examined in equal depth.
[2] Song et al., 2024 (Springer)	Reducing computation and storage cost	Organises lightweight methods such as distillation, quantisation, pruning and algorithm optimisation.	Efficiency gains depend on the model, task, hardware and acceptable quality trade-off.
[3] Roessbach et al., 2022 (IEEE CVF)	Latent-space high-resolution synthesis	Runs diffusion in a compressed latent space and uses cross-attention for conditioning.	Quality depends on the autoencoder and data; compression may lose details.
[4] Croitoru et al., 2023 (IEEE TPAMI)	Diffusion frameworks in computer vision	Explains DDPM, score-based and SDE views, and compares diffusion with other generative families.	Broad vision scope; iterative sampling can still be slow.
[5] Alhabeeb & Al-Shargabi, 2024 (IEEE Access)	Text-to-image methods, datasets and metrics	Reviews generative methods and evaluation metrics, including text-image alignment measures.	Metrics capture different qualities; prompt understanding and fair evaluation remain difficult.

V. GENERAL WORKING OF A DIFFUSION-BASED IMAGE GENERATOR

A simplified text-to-image diffusion system can be described as a sequence of steps. First, the user enters a text prompt. A text encoder converts the prompt into a representation that can guide image generation. The image-generation process begins from random noise. At each denoising step, the model predicts how the current representation should be updated. The process repeats until a meaningful image is formed; if the model works in latent space, a decoder converts the final latent representation into the output image [3], [5].

Text Prompt → Text Encoder → Random Noise → Iterative Denoising → Decoder → Generated Image

Applications

Diffusion models are studied and used in digital art, concept design, advertising, educational illustrations, game assets, and visual media. In image editing, they can support inpainting, where a selected region is filled, and super-resolution, where an image is reconstructed at a higher resolution. Style transfer can help adapt the appearance of an image, while text-to-image generation allows a user to explore a visual idea before producing a final design [1], [3], [5].

The usefulness of these applications depends on the task. A generated image may look realistic but still contain incorrect objects, text, counts, or spatial relationships. In areas such as scientific or medical imaging, the output should be checked by a qualified person and should not be treated as reliable evidence merely because it appears convincing.

Challenges and Research Gaps

- **Computational cost:** Diffusion generation often requires repeated model evaluations. Lightweight methods and latent-space generation can reduce cost, but there is still a need to report runtime, memory, energy use, and quality together [2]–[4].
- **Prompt alignment:** A model may produce a visually attractive image while missing an object,

miscounting items, or misunderstanding the spatial relation described in a prompt. Evaluation should therefore measure both visual quality and semantic alignment [5].

- **Inconsistent evaluation:** The selected publications discuss different tasks, datasets, and metrics. A fair comparison needs common prompts, image sizes, hardware settings, and evaluation procedures. Automated metrics should be combined with careful human assessment where appropriate [4], [5].
- **Dataset bias and privacy:** Models learn from large image collections, which may contain stereotypes, personal information, or unbalanced representation. Data curation, privacy-aware training, and testing across diverse groups are important research directions [1], [5].
- **Copyright, originality and misuse:** Generated images can raise questions about training-data permissions, attribution, impersonation, and misleading visual content. Technical safeguards, clearer disclosure, and responsible-use policies should develop alongside model capability [1].
- **Reproducibility and deployment:** Model quality can vary with prompts, random seeds, sampling steps, guidance settings, and hardware. Future studies should describe these settings clearly and include failure cases rather than reporting only the strongest outputs.
- **Fine-detail preservation:** Compression in latent-space systems makes generation more efficient, but very small features or exact structures can be lost. Research should continue to examine how to balance compression, speed, and faithful detail [3].

VIII. CONCLUSION

This review examined five recognised publications on AI diffusion models for image generation. Chen et al. provide a broad view of image-generation applications and social concerns [1]. Song et al. organise approaches for reducing the resource demands of diffusion models [2]. Rombach et al. show how latent-space diffusion can support high-resolution synthesis with lower computational requirements than pixel-space diffusion [3]. Croitoru

et al. explain the main diffusion frameworks used in computer vision [4], while Alhabeeb and Al-Shargabi review text-to-image methods, datasets, metrics, and open challenges [5].

Taken together, the literature shows that progress is not only about generating more realistic images. Models also need to follow prompts accurately, run efficiently, be evaluated fairly, and reduce risks linked to bias, privacy, copyright, and misuse. Future work should report performance and resource use more consistently, improve fine-detail and prompt alignment, and make responsible use part of model development. Since this paper is based on published literature and includes no new experiments, its conclusions reflect the selected works rather than a new experimental ranking.

REFERENCES

1. H.Chen, Q. Xiang, J. Hu, M. Ye, C. Yu, H. Cheng, et al., "Comprehensive exploration of diffusion models in image generation: A survey," *Artificial Intelligence Review*, vol. 58, article 99, 2025. doi: 10.1007/s10462-025-11110-3. Available: <https://doi.org/10.1007/s10462-025-11110-3>
2. W. Song, W. Ma, M. Zhang, Y. Zhang, et al., "Lightweight diffusion models: A survey," *Artificial Intelligence Review*, vol. 57, article 161, 2024. doi: 10.1007/s10462-024-10800-8. Available: <https://doi.org/10.1007/s10462-024-10800-8>
3. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVFConference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042. Available: <https://doi.org/10.1109/CVPR52688.2022.01042>
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. 10850–10869, 2023. doi: 10.1109/TPAMI.2023.3261988. Available: <https://doi.org/10.1109/TPAMI.2023.3261988>
5. S. K. Alhabeeb and A. A. Al-Shargabi, "Text-to-image synthesis with generative models: Methods, datasets, performance metrics, challenges, and future direction," *IEEE Access*, vol. 12, pp. 24412–24427, 2024. doi: 10.1109/ACCESS.20