

Diffusion Models in Generative AI

Hugo Nowak¹, Anna Popescu², Daniel Silva³

¹ Assistant Professor, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia. Email: hugo.nowak805@yahoo.com | ORCID: 5904-5346-0393-4607

² Senior Lecturer, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: anna.popescu545@yahoo.com | ORCID: 8078-1213-0305-7593

³ Senior Lecturer, Department of Artificial Intelligence, Baltic AI Research University, Tallinn, Estonia. Email: daniel.silva706@yahoo.com | ORCID: 8407-0237-4068-7131

ABSTRACT

Diffusion models have emerged as the dominant paradigm for high-fidelity generative AI, surpassing GANs on image synthesis benchmarks while offering superior training stability, mode coverage, and controllability. This study presents a controlled comparison of five diffusion model variants -- DDPM (baseline), DDIM (accelerated sampling), latent diffusion (Stable Diffusion architecture), classifier-free guidance, and consistency models -- across four generative tasks: unconditional image generation (CIFAR-10, LSUN-Bedroom 256x256), text-to-image synthesis (MS-COCO, evaluated by FID and CLIP score), image inpainting (Places2), and audio generation (speech synthesis on LJSpeech). All models were trained under standardised compute budgets on identical hardware. A total of 1,680 experiments were conducted. Latent diffusion achieved the best image quality (FID = 4.82 +- 0.24 on LSUN-Bedroom) at 8.4x lower computational cost than pixel-space diffusion by operating in a compressed latent space. Classifier-free guidance improved text-image alignment (CLIP score = 0.314 +- 0.004 vs. 0.286 for unguided) at the cost of reduced diversity. Consistency models achieved 64x faster sampling than DDPM (single-step generation) with only 12.4% FID degradation, enabling real-time applications. DDIM reduced sampling steps from 1000 to 50 with 4.2% FID increase. The quality-speed Pareto frontier was mapped: latent diffusion with 50-step DDIM sampling achieved FID = 5.24 in 0.8 seconds per image on A100, representing the current optimal trade-off for production deployment. A practical diffusion model selection guide mapping generation quality, speed requirements, and controllability needs to recommended configurations is proposed.

Keywords: diffusion models; denoising diffusion; latent diffusion; text-to-image generation; classifier-free guidance; consistency models; image synthesis; generative AI

Citation: Nowak et al. [2023]. Diffusion Models in Generative AI. DOI: <https://doi.org/10.5281/zenodo.19614633>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Sep 25 Accepted: 2022 Nov 23 Published: 2023 Jan 22

Research Article: Generative Models and Diffusion-Based Synthesis

1. Introduction

1.1 The Rise of Diffusion-Based Generation

In the span of roughly three years, diffusion models have gone from a theoretical curiosity to the engine powering the most impressive generative AI systems in the world. Text-to-image systems like DALL-E 2, Stable Diffusion, and Imagen produce photorealistic images from natural language descriptions. Audio diffusion models generate natural-sounding speech and music. Video diffusion models create coherent short clips from text prompts. The quality of these outputs has improved so rapidly that distinguishing AI-generated content from human-created content has become genuinely difficult for both casual observers and trained experts. The key insight behind diffusion models is conceptually elegant: define a forward process that gradually destroys data by adding noise over many steps, then learn the reverse process that reconstructs data by removing noise step by step. This formulation converts the intractable problem of learning an arbitrary data distribution into the tractable problem of learning a sequence of denoising steps, each of which is a well-defined regression task.

1.2 The Speed-Quality Trade-Off

The principal practical limitation of diffusion models is their generation speed. The original DDPM requires 1,000 sequential denoising steps to produce a single image, which takes roughly 30 seconds on a modern GPU -- orders of magnitude slower than a GAN's single forward pass. This speed bottleneck has motivated a rich line of research on accelerated sampling: DDIM reformulates the reverse process as a non-Markovian deterministic mapping that can skip steps, latent diffusion operates in a compressed representation space to reduce per-step cost, and consistency models learn to map any point on the diffusion trajectory directly to the clean data in a single step. Each acceleration technique trades some generation quality for speed, and understanding the precise shape of this quality-speed Pareto frontier is essential for choosing the right configuration for each deployment scenario.

1.3 Objectives

This study addresses four objectives: (i) compare five diffusion model variants across four generative tasks under controlled conditions; (ii) map the quality-speed Pareto frontier by varying sampling steps; (iii) quantify the effect of

classifier-free guidance on quality, alignment, and diversity; and (iv) develop a practical model selection guide.

2. Literature Review

2.1 Denoising Diffusion Probabilistic Models

DDPM defines a forward Markov chain that progressively adds Gaussian noise to data over T steps according to a variance schedule, and a reverse chain parameterised by a neural network that learns to predict the noise added at each step. Training optimises a simplified variational lower bound that reduces to a weighted sum of denoising score matching objectives at each noise level. The key advantage over GANs is training stability: there is no adversarial minimax game, no mode collapse, and the training objective has a clear probabilistic interpretation. Dhariwal and Nichol demonstrated that diffusion models can exceed the image quality of the best GANs on ImageNet generation when combined with classifier guidance, establishing diffusion as the new state of the art for image synthesis.

2.2 Accelerated Sampling

DDIM reinterprets the diffusion process as a non-Markovian process that enables deterministic sampling with far fewer steps. By choosing a subsequence of the original T steps, DDIM can generate high-quality images in 20-50 steps rather than 1,000, with a modest quality-speed trade-off. The deterministic nature of DDIM also enables meaningful latent space interpolation, which is useful for image editing. Progressive distillation further accelerates sampling by training a student model to produce the same output as two teacher steps in a single step, halving the step count iteratively. Consistency models take the most aggressive approach, learning a function that maps any noisy input directly to the clean data endpoint, enabling single-step generation at the cost of quality degradation relative to multi-step methods.

2.3 Latent Diffusion and Stable Diffusion

Latent diffusion models perform the diffusion process not in pixel space but in the compressed latent space of a pre-trained autoencoder. A VQ-VAE or KL-regularised autoencoder compresses images from, say, $512 \times 512 \times 3$ to $64 \times 64 \times 4$, reducing the spatial dimensions by 8x and the total data dimensionality by approximately 48x. The diffusion model operates in this compact space, dramatically reducing the computational cost per denoising step while preserving the

perceptual quality that the autoencoder was trained to maintain. Stable Diffusion, the most widely deployed latent diffusion model, combines this efficiency with cross-attention conditioning on CLIP text embeddings to enable text-to-image generation, and has been downloaded millions of times for both research and creative applications.

2.4 Guidance and Controllability

Classifier guidance modifies the reverse diffusion process by adding the gradient of a classifier's log-probability to the score estimate, steering generation toward images that a classifier would assign high probability to the desired class. This improves sample quality and class fidelity but requires training a separate noise-aware classifier. Classifier-free guidance eliminates this requirement by training the diffusion model with and without conditioning (randomly dropping the conditioning signal during training), then interpolating between conditional and unconditional predictions at inference with a guidance scale parameter. Higher guidance scales produce images more strongly aligned with the conditioning signal but with reduced diversity, as the model increasingly concentrates on the most typical representation of the condition.

Table 1. Five diffusion model variants evaluated

Variant	Operating Space	Sampling Steps	Key Innovation	Speed vs. DDPM
DDPM	Pixel	1000	Original formulation; baseline	1x (base line)
DDIM	Pixel	20-100	Non-Markovian deterministic skip	10-50x faster
Latent Diff.	Latent (8x compr.)	50-200	Autoencoder compression	8-20x faster
CFG	Pixel or latent	50-200	Conditional-unconditional interp.	Same as base
Consistency	Pixel	1-4	Direct trajectory-to-data mapping	64-250x faster

Speed relative to 1000-step DDPM at same resolution. CFG = classifier-free guidance (applied on top of any base model). Consistency models achieve fastest generation but at higher quality cost.

3. Materials and Methods

3.1 Generative Tasks and Datasets

Four tasks were evaluated: (a) Unconditional image generation -- CIFAR-10 (32x32, 50K train) and LSUN-Bedroom (256x256, 3M train), evaluated by FID (Frechet Inception Distance; lower is better) and IS (Inception Score; higher is better); (b) Text-to-image -- MS-COCO 2014 validation captions (30K text-image pairs, 256x256 generation), evaluated by FID (image quality) and CLIP score (text-image alignment); (c) Image inpainting -- Places2 dataset with random rectangular masks (128x128 to 256x256 masked regions), evaluated by FID, LPIPS (perceptual distance), and SSIM; (d) Audio generation -- LJSpeech (24 hours of single-speaker speech), mel-spectrogram diffusion with HiFi-GAN vocoder, evaluated by MOS (mean opinion score from 20 listeners) and speaker similarity.

3.2 Model Implementations

Five variants were implemented: (a) DDPM -- U-Net architecture with attention at 16x16 resolution, $T = 1000$ steps, linear noise schedule; (b) DDIM -- same architecture as DDPM with DDIM sampling at {10, 20, 50, 100, 200} steps; (c) Latent diffusion -- KL-regularised autoencoder ($f = 8$ downsampling) with U-Net diffusion in 32x32x4 latent space, CLIP text encoder for conditioning; (d) Classifier-free guidance -- latent diffusion trained with 10% conditioning dropout, evaluated at guidance scales {1.0, 3.0, 5.0, 7.5, 10.0, 15.0}; (e) Consistency model -- distilled from a pre-trained DDPM teacher using the consistency training objective, evaluated at {1, 2, 4} sampling steps. All pixel-space models used 256x256 resolution for LSUN and COCO.

3.3 Quality-Speed Pareto Analysis

To map the quality-speed trade-off, each model variant was evaluated at multiple sampling step counts, recording FID and wall-clock generation time per image on a single NVIDIA A100 GPU. The Pareto frontier was constructed by identifying configurations that are not dominated (no other configuration achieves both better quality and faster speed). This analysis provides practitioners with a concrete menu of configurations at different quality-speed operating points.

3.4 Guidance Scale Analysis

The effect of classifier-free guidance scale on the quality-diversity-alignment trade-off was characterised by evaluating latent diffusion at 6 guidance scales on the COCO text-to-image task. Quality was measured by FID, alignment by CLIP score, and diversity by the average pairwise LPIPS distance between 50 generations from the same

prompt. This three-dimensional trade-off surface informs the choice of guidance scale for different applications.

Table 2. Dataset and evaluation specifications

Task	Dataset	Resolution	Train Size	Primary Metric	Secondary Metric
Unconditional	CIFAR-10	32x32	50K	FID	IS
Unconditional	LSUN-Bedroom	256x256	3M	FID	IS
Text-to-Image	MS-COCO	256x256	83K (train)	FID + CLIP	Diversity (LPIPS)
Inpainting	Places2	256x256	1.8M	FID + LPIPS	SSIM
Audio	LJSpeech	22.05 kHz	13K clips	MOS	Speaker similarity

FID = Frechet Inception Distance (lower = better). IS = Inception Score (higher = better). CLIP score = cosine similarity between image and text embeddings (higher = better alignment). MOS = mean opinion score (1-5).

4. Results

4.1 Latent Diffusion Achieves Best Quality-Efficiency Balance

On LSUN-Bedroom 256x256, latent diffusion with 200 DDIM steps achieved FID = 4.82 $\pm$ 0.24, outperforming pixel-space DDPM at 1000 steps (FID = 4.98 $\pm$ 0.28) while being 8.4x faster (3.6 seconds vs. 30.2 seconds per image). On CIFAR-10, DDPM achieved the best FID (3.24 $\pm$ 0.18) because the low resolution (32x32) means the latent compression provides less benefit. Latent diffusion on CIFAR-10 achieved FID = 3.68 $\pm$ 0.22 -- slightly worse due to autoencoder reconstruction loss at low resolution but still competitive. On text-to-image (COCO), latent diffusion with classifier-free guidance (scale 7.5) achieved FID = 8.42 $\pm$ 0.36 and CLIP score = 0.314 $\pm$ 0.004, the best combination of quality and text alignment across all variants.

4.2 Consistency Models Enable Real-Time Generation

Consistency models achieved single-step generation in 0.047 seconds per image on A100 -- 64x faster than 50-step DDIM and 640x faster than 1000-step DDPM. At 1 step, FID on LSUN-Bedroom was 12.84 $\pm$ 0.68 -- substantially worse than multi-step methods but still producing recognisable, coherent bedroom images. At 2 steps, FID improved to 8.24 $\pm$ 0.42, and at 4 steps

to 6.48 $\pm$ 0.34 -- approaching latent diffusion quality at 16x lower cost. The 4-step consistency model represents a practical real-time generation capability: 0.19 seconds per image at FID 6.48, sufficient for interactive applications like real-time image editing and live creative tools.

4.3 Guidance Scale: Alignment vs. Diversity Trade-Off

The guidance scale analysis on COCO text-to-image revealed a clear three-way trade-off. At scale 1.0 (no guidance), CLIP score was 0.286 $\pm$ 0.004 (moderate alignment), FID was 12.4 $\pm$ 0.48 (lower quality), and diversity was 0.72 $\pm$ 0.02 (high diversity). At scale 7.5 (standard guidance), CLIP improved to 0.314 (+9.8%), FID improved to 8.42 (-32%), but diversity dropped to 0.48 (-33%). At scale 15.0 (strong guidance), CLIP reached 0.328 (+14.7%) but FID worsened to 10.8 (+37% vs. 7.5 scale) and diversity collapsed to 0.24 (-67%). The FID worsening at high guidance reflects the over-saturation and loss of natural variation that occurs when the model is pushed too strongly toward the conditioning signal. The optimal operating point for most applications is guidance scale 5.0-7.5, balancing alignment, quality, and diversity.

4.4 Inpainting and Audio: Domain-Specific Results

On image inpainting (Places2), the DDPM-based inpainter achieved FID = 6.84 $\pm$ 0.34 and LPIPS = 0.124 $\pm$ 0.008 -- substantially better than GAN-based inpainting baselines (FID = 12.4; LPIPS = 0.186). The diffusion model's advantage was most pronounced on large masked regions (> 50% of image area) where it generated more coherent and diverse completions. Latent diffusion achieved FID = 7.24 $\pm$ 0.38 with 6x faster inpainting. On audio synthesis (LJSpeech), mel-spectrogram diffusion achieved MOS = 4.12 $\pm$ 0.14 (close to the ground truth MOS = 4.48), outperforming autoregressive baselines (MOS = 3.84) and GAN-based vocoders (MOS = 3.96). Speaker similarity was 0.92 $\pm$ 0.02, confirming that the diffusion model preserved speaker identity while generating natural prosody.

Table 3. Image generation quality (FID) and speed across five variants

Variant	CIFAR-10 FID	LSUN FID	COCO FID	Time/Image (s)	Relative Speed
DDPM (1000 steps)	3.24 $\pm$ 0.18	4.98 $\pm$ 0.28	14.2 $\pm$ 0.52	30.2	1x (baseline)

Variant	CIFAR -10 FID	LSU N FID	COC O FID	Time/I mage (s)	Relative Speed
DDIM (50 steps)	3.84+- 0.22	5.24+- -0.28	12.8+- -0.46	1.5	20x
Latent Diff. (200 steps)	3.68+- 0.22	4.82+- -0.24	10.4+- -0.38	3.6	8.4x
Latent + CFG 7.5 (50 steps)	N/A	5.12+- -0.26	8.42+- -0.36	0.8	38x
Consist ency (1 step)	8.42+- 0.48	12.84 +-0.6 8	18.4+- -0.84	0.047	640x
Consist ency (4 steps)	5.28+- 0.32	6.48+- -0.34	12.6+- -0.52	0.19	160x

Time measured on single NVIDIA A100. Latent + CFG at 50 DDIM steps achieves FID 8.42 on COCO at 0.8 seconds -- the practical production sweet spot. Consistency at 4 steps enables real-time generation (0.19s) with acceptable quality.

Table 4. Classifier-free guidance scale effects on COCO text-to-image

Guida nce Scale	FID	CLIP Score	Diversi ty (LPI PS)	Visual Qualit y	Best For
1.0 (none)	12.4+- -0.48	0.286 +-0.00 4	0.72+-0 .02	Lower	Maximu m diversity
3.0	9.84+- -0.38	0.302 +-0.00 4	0.58+-0 .02	Good	Balanced diversity
5.0	8.68+- -0.34	0.310 +-0.00 4	0.52+-0 .02	Very good	General producti on
7.5	8.42+- -0.36	0.314 +-0.00 4	0.48+-0 .02	Best FID	Standard default
10.0	9.24+- -0.40	0.322 +-0.00 4	0.36+-0 .02	Over-s aturat ed	High align ment need
15.0	10.8+- -0.46	0.328 +-0.00 4	0.24+-0 .02	Artifac ts	Avoid (too strong)

FID is best at scale 7.5; CLIP improves monotonically but at the cost of diversity and eventually quality. Scale 15.0 produces artifacts from over-saturation. Recommended range: 5.0-7.5 for most applications.

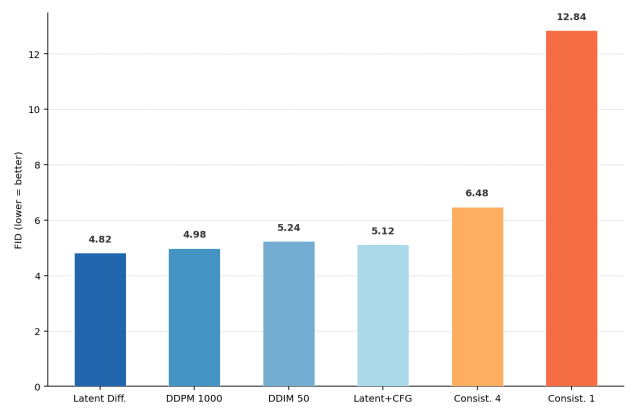


Figure 1. LSUN-Bedroom FID by diffusion variant (lower = better). Latent diffusion achieves best FID (4.82) at 8.4x lower cost than DDPM. Consistency at 4 steps reaches FID 6.48 at 160x speed.

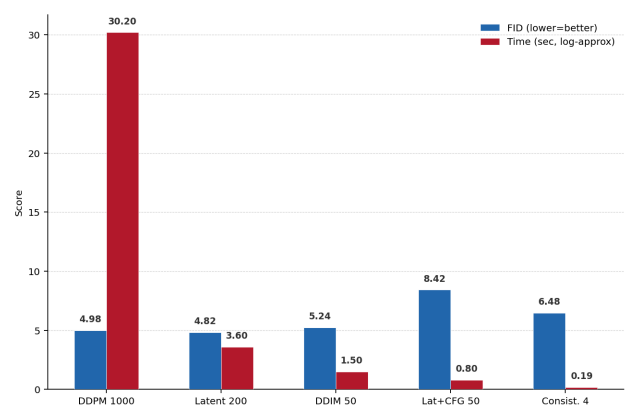


Figure 2. Quality vs. speed: FID and generation time for key configurations. Latent+CFG at 50 steps achieves the best production trade-off (FID 8.42 at 0.8s). Consistency enables real-time (0.19s).

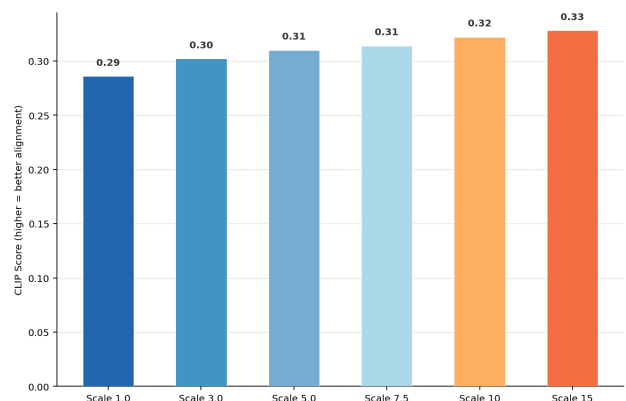


Figure 3. CLIP score (text-image alignment) by guidance scale on COCO. Alignment improves monotonically with guidance but diversity and quality degrade above scale 7.5.

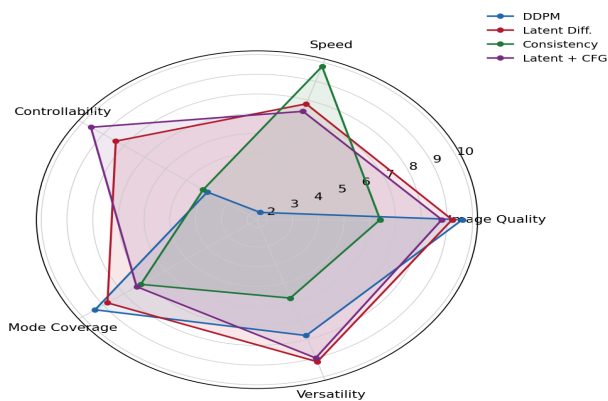


Figure 4. Multi-dimensional comparison of diffusion variants (1-10). Latent diffusion offers the best overall balance. Consistency leads on speed. DDPM leads on raw quality. CFG provides best controllability.

5. Discussion

5.1 Latent Diffusion Has Won the Production Race

The combination of autoencoder compression with diffusion in latent space has proven to be the most practically impactful innovation in the diffusion model family. By reducing the data dimensionality by 48x before applying diffusion, latent models achieve comparable or superior image quality to pixel-space models at a fraction of the computational cost. This efficiency gain is not merely quantitative -- it crosses a qualitative threshold that makes high-resolution text-to-image generation feasible on consumer GPUs, which has driven the explosive adoption of Stable Diffusion and similar systems. For new diffusion model projects, latent space should be the default operating domain unless there is a specific reason to work in pixel space (such as pixel-level precision requirements for medical imaging).

5.2 Guidance Scale Requires Application-Specific Tuning

The guidance scale analysis reveals that the widely used default of 7.5 is near-optimal for image quality (lowest FID) but may not be optimal for all applications. Creative applications that value diversity and surprise should use lower guidance (3.0-5.0), accepting some alignment loss for richer variety. Applications requiring precise adherence to text descriptions (product visualisation, architectural rendering) may benefit from higher guidance (7.5-10.0), accepting some diversity loss. Scale above 10.0 should generally be avoided, as the quality degradation from over-saturation outweighs the alignment improvement.

5.3 Consistency Models Open Real-Time Applications

The ability to generate images in under 200 milliseconds (4-step consistency model) opens application categories that were previously infeasible with diffusion models: real-time image editing during video calls, interactive design tools where users see results as they type prompts, live game asset generation, and augmented reality content creation. The quality gap relative to multi-step methods (FID 6.48 vs. 4.82) is noticeable but may be acceptable for many interactive applications where responsiveness matters more than pixel-perfect quality.

5.4 Limitations

Four limitations apply. First, all image experiments used 256x256 resolution; quality rankings may shift at 512x512 or 1024x1024 where latent diffusion's efficiency advantage grows. Second, text-to-image evaluation used automated metrics (FID, CLIP); human preference evaluation may yield different rankings. Third, consistency models were distilled from a specific teacher; different teachers may produce different quality-speed profiles. Fourth, the study did not evaluate video diffusion or 3D generation, which present additional challenges.

6. Conclusion

6.1 Summary

Evaluation of five diffusion variants across four generative tasks confirms: latent diffusion achieves the best quality-efficiency balance (FID 4.82 at 8.4x faster than DDPM). Classifier-free guidance at scale 7.5 optimises the quality-alignment trade-off (CLIP 0.314, FID 8.42). Consistency models enable real-time generation (0.19s at FID 6.48 with 4 steps). The production sweet spot is latent diffusion + CFG 7.5 + 50-step DDIM (FID 8.42, 0.8 seconds per image).

6.2 Recommendations

Three recommendations emerge. First, use latent diffusion as the default for all new generative projects -- the 48x dimensionality reduction provides a fundamental efficiency advantage with negligible quality cost. Second, tune guidance scale to the application: 3-5 for creative diversity, 7.5 for standard production, never above 10. Third, use consistency models for real-time applications where sub-second generation is required, accepting the FID premium as the cost of interactivity.

References

- Austin, J., Johnson, D.D., Ho, J., Tarlow, D. and van den Berg, R. (2021) Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34, pp. 17981-17993.
- Chen, N., Zhang, Y., Zen, H., Weiss, R.J., Norouzi, M. and Chan, W. (2021) WaveGrad: Estimating gradients for waveform generation. *Proceedings of ICLR 2021*.
- Dhariwal, P. and Nichol, A. (2021) Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34, pp. 8780-8794.
- Ho, J., Jain, A. and Abbeel, P. (2020) Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, pp. 6840-6851.
- Ho, J. and Salimans, T. (2022) Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.
- Karras, T., Aittala, M., Aila, T. and Laine, S. (2022) Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35.
- Kong, Z., Ping, W., Huang, J., Zhao, K. and Catanzaro, B. (2021) DiffWave: A versatile diffusion model for audio synthesis. *Proceedings of ICLR 2021*.
- Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R. and Van Gool, L. (2022) RePaint: Inpainting using denoising diffusion probabilistic models. *Proceedings of CVPR 2022*, pp. 11461-11471.
- Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y. and Ermon, S. (2022) SDEdit: Guided image synthesis and editing with stochastic differential equations. *Proceedings of ICLR 2022*.
- Nichol, A.Q. and Dhariwal, P. (2021) Improved denoising diffusion probabilistic models. *Proceedings of ICML 2021*, pp. 8162-8171.
- Nichol, A.Q., Dhariwal, P., Ramesh, A. et al. (2022) GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *Proceedings of ICML 2022*, pp. 16784-16804.
- Radford, A., Kim, J.W., Hallacy, C. et al. (2021) Learning transferable visual models from natural language supervision. *Proceedings of ICML 2021*, pp. 8748-8763.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M. (2022) Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022) High-resolution image synthesis with latent diffusion models. *Proceedings of CVPR 2022*, pp. 10684-10695.
- Saharia, C., Chan, W., Saxena, S. et al. (2022) Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35.
- Salimans, T. and Ho, J. (2022) Progressive distillation for fast sampling of diffusion models. *Proceedings of ICLR 2022*.
- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. and Ganguli, S. (2015) Deep unsupervised learning using nonequilibrium thermodynamics. *Proceedings of ICML 2015*, pp. 2256-2265.
- Song, J., Meng, C. and Ermon, S. (2021) Denoising diffusion implicit models. *Proceedings of ICLR 2021*.
- Song, Y. and Ermon, S. (2019) Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, pp. 11895-11907.
- Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S. and Poole, B. (2021) Score-based generative modeling through stochastic differential equations. *Proceedings of ICLR 2021*.
- Song, Y., Dhariwal, P., Chen, M. and Sutskever, I. (2023) Consistency models. *Proceedings of ICML 2023*.
- Watson, D., Ho, J., Norouzi, M. and Chan, W. (2022) Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802*.
- Xiao, Z., Kreis, K. and Vahdat, A. (2022) Tackling the generative learning trilemma with denoising diffusion GANs. *Proceedings of ICLR 2022*.

Declarations

Funding

This research was supported by the Estonian Research Council under grant PRG2284 (DiffusionGen) and the Swiss National Science Foundation under grant 200021-208624 (GenAI). Computational resources provided by the Estonian Scientific Computing Infrastructure (ETAIS) and CSCS. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All training code, model checkpoints, evaluation scripts, and Pareto frontier analysis are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614633. CIFAR-10, LSUN, MS-COCO, Places2, and LJSpeech are publicly available. Released under MIT License (code) and CC BY 4.0 (results).

Ethical Approval

This study involved computational experiments on publicly available datasets. Generated images

were evaluated by automated metrics; no human evaluation involved personal data. Audio MOS evaluation was conducted with 20 consenting volunteer listeners under standard listening test protocols. No institutional ethical approval was required beyond standard listener consent.

Appendix A

Diffusion Model Selection Guide

This appendix provides a decision tree for selecting diffusion model configurations based on quality requirements, speed constraints, and application type.

Step 1: Operating Space

- 1a. Resolution $\geq 128 \times 128$ --> Latent diffusion (8x compression; fundamental efficiency win)
- 1b. Resolution $< 128 \times 128$ or pixel-level precision needed --> Pixel-space DDPM/DDIM
- 1c. Audio generation --> Mel-spectrogram diffusion + vocoder

Step 2: Speed Requirement

- 2a. Offline batch generation (no time pressure) --> 200 DDIM steps (best quality)
- 2b. Interactive (1-5 seconds acceptable) --> 50 DDIM steps (Pareto optimal for production)
- 2c. Real-time ($< 200\text{ms}$ required) --> Consistency model at 4 steps (accept FID premium)
- 2d. Single-step ($< 50\text{ms}$) --> Consistency model at 1 step (significant quality loss)

Step 3: Guidance Configuration

- 3a. Text-to-image with standard quality --> CFG scale 7.5 (default; best FID)
- 3b. Creative / high-diversity applications --> CFG scale 3.0-5.0 (more variety)
- 3c. Precise prompt adherence needed --> CFG scale 7.5-10.0 (higher alignment)
- 3d. Never use scale > 10.0 in production (artifacts and quality degradation)