Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Comparative Analysis of Transformer, Diffusion, and GAN-Based Generative Models

Marco Silva[1], Matteo Horvath[2]

[1] Postdoctoral Researcher, School of Data Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: marco.silva38@gmail.com | ORCID: 2271-7379-0363-5008
[2] Research Scientist, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia. Email: matteo.horvath223@gamil.com | ORCID: 5471-7444-0241-2775

ABSTRACT

The generative AI landscape is currently defined by three dominant paradigmatic approaches -- autoregressive transformer models, diffusion models, and generative adversarial networks -- each with distinct training objectives, architectural principles, and generation characteristics that make them differentially suitable for different generation tasks, data modalities, and deployment contexts. Despite the substantial literature on each paradigm individually, a rigorous cross-paradigm comparative evaluation that assesses all three approaches on a common set of benchmarks, quality dimensions, and efficiency criteria remains absent, making it difficult for practitioners to make principled architecture selection decisions. This paper presents the first comprehensive cross-paradigm comparative evaluation of transformer, diffusion, and GAN-based generative models across eight generation domains -- image synthesis, text generation, audio synthesis, video synthesis, molecular generation, code synthesis, 3D object generation, and time-series forecasting -- using a unified evaluation framework covering six quality dimensions: sample quality, diversity, training stability, inference efficiency, controllability, and compositional generalisation. Results across 24 benchmark comparisons reveal that no single paradigm dominates across all domains and dimensions: diffusion models achieve the highest image and audio sample quality (FID = 14.2 for image, MOS = 4.6 for audio); autoregressive transformers achieve superior text quality (perplexity), code synthesis (HumanEval pass@1 = 52.4%), and compositional generalisation (42.6% C-CME R@1); GANs achieve the highest inference speed (6.8x faster than diffusion, 3.2x faster than transformers) at competitive image quality. The paper contributes a cross-paradigm evaluation framework, a domain-paradigm suitability matrix, and empirical evidence supporting a complementarity rather than competition view of generative AI paradigms.

Keywords: generative models; transformers; diffusion models; GANs; comparative evaluation; generation quality; inference efficiency; paradigm comparison

Citation: Silva and Horvath [2025]. Comparative Analysis of Transformer, Diffusion, and GAN-Based Generative Models. DOI: https://doi.org/10.5281/zenodo.19185187

Article Information: *Received:* 06 December 2024 *Accepted:* 08 February 2025 *Published:* 20 March 2025

Research Article: *Research Article*

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 1, 2025

1. Introduction

1.1 Three Generative Paradigms

Generative AI in 2025 is primarily defined by three paradigmatic approaches. Autoregressive transformer models (Brown et al., 2020; Touvron et al., 2023) generate outputs sequentially by predicting each token conditioned on previous tokens, using the transformer attention mechanism to efficiently model long-range dependencies. This paradigm dominates text and code generation and has been extended to image generation through discrete tokenisation (Sun et al., 2024 -- LlamaGen). Diffusion models (Ho et al., 2020; Song et al., 2021) generate outputs by iteratively denoising a random noise sample through a learned reverse diffusion process, achieving state-of-the-art image and audio generation quality. GANs (Goodfellow et al., 2014) train a generator network adversarially against a discriminator, enabling high-quality generation with fast inference once trained. The practical question of which paradigm to adopt for a given generation task is consequential and poorly served by the current literature, which develops each paradigm largely in isolation with task-specific benchmark comparisons that do not enable cross-paradigm generalisation. The rapid pace of advance within each paradigm further complicates comparisons: a result demonstrating diffusion model superiority over GANs for image synthesis in 2022 may not hold for state-of-the-art models in both paradigms in 2025. The evaluation framework proposed in this paper addresses this gap through a unified, multi-domain, multi-dimension comparative study conducted at the current state of each paradigm.

1.2 Research Contributions and Structure

This paper contributes: (1) a unified cross-paradigm evaluation framework covering six quality dimensions across eight generation domains; (2) a comprehensive 24-comparison benchmark evaluation using representative state-of-the-art models from each paradigm; (3) a domain-paradigm suitability matrix summarising empirical findings; and (4) a practitioner's guide for paradigm selection. Section 2 reviews each generative paradigm and prior comparative work. Section 3 presents the evaluation framework. Section 4 reports results. Section 5 discusses the suitability matrix and practitioner guidance. Section 6 concludes.

2. Background

2.1 Generative Paradigm Characteristics

Autoregressive transformers have three fundamental characteristics that determine their strengths and limitations. First, sequential generation enables exact likelihood estimation under the model distribution, supporting high-quality unconditional and conditional generation. Second, the attention mechanism provides global context for each generation step, supporting long-range coherence in text and code. Third, inference cost scales linearly with output length and quadratically with context length in standard attention -- a fundamental efficiency limitation for long-form generation. Diffusion models have three defining characteristics. First, the generation process is stochastic and iterative (typically 20-1000 denoising steps), enabling high diversity and quality at the cost of slow inference. Second, the training objective (denoising score matching) is stable and scales well with data and model size. Third, conditional generation through classifier-free guidance provides strong controllability at the cost of diversity. GANs have three characteristics. First, single-pass inference is extremely fast -- comparable to standard feed-forward evaluation. Second, adversarial training is unstable, with mode collapse and training divergence as persistent failure modes. Third, GANs tend to have lower sample diversity than diffusion models and autoregressive models at equivalent quality. Table 1 summarises the three paradigm profiles.

2.2 Prior Comparative Work

Prior comparative work on generative paradigms has been predominantly single-domain and limited to pairwise comparisons. Dhariwal and Nichol (2021) demonstrated diffusion model superiority over GANs for image synthesis using the guided diffusion model -- a finding that has been replicated in subsequent work but may not generalise to the current state of StyleGAN-based models (Karras et al., 2022 -- StyleGAN3). Rombach et al. (2022) demonstrated that latent diffusion models reduce the compute requirements of diffusion models substantially while maintaining quality advantages over GANs. Austin et al. (2021) compared discrete diffusion to autoregressive models for text generation, finding autoregressive models superior. None of these comparisons covers multiple domains simultaneously or evaluates more than two paradigms. The systematic cross- paradigm, cross-domain evaluation in this paper addresses this gap.

Table 1: Generative Paradigm Profiles -- Characteristics and Expected Domain

Suitability

Paradigm	Training Objective	Inference Speed	Diversity	Stability	Controllability	Best For (Hypothesis)
Autoregressive Transformer	Next-token prediction (MLE)	Medium	High	High	Moderate	Text, code, discrete sequences
Diffusion Model	Denoising score matching	Slow	Very high	High	High (CFG)	Image, audio, molecules
GAN	Adversarial min-max (non-MLE)	Very fast	Medium	Low	Moderate	Fast inference, visual synthesis

3. Evaluation Framework

3.1 Six Quality Dimensions and Eight Domains

The cross-paradigm evaluation framework assesses six quality dimensions for each paradigm across eight generation domains. Sample Quality (SQ) measures the perceptual or functional quality of individual generated samples using domain-appropriate metrics (FID for images, perplexity for text, MOS for audio, FVD for video, validity for molecules). Diversity (DV) measures the breadth of the generated distribution using coverage metrics and sample variety scores. Training Stability (TS) measures the reliability of convergence across multiple training runs (final loss variance and success rate). Inference Efficiency (IE) measures generation throughput in tokens or samples per second on standardised hardware. Controllability (CT) measures the precision of conditional generation using CLIP-score for image generation and accuracy for discrete tasks. Compositional Generalisation (CG) measures performance on novel combinations not seen during training using domain-specific compositional benchmarks. The eight generation domains are: image synthesis, text generation, audio synthesis, video synthesis, molecular generation, code synthesis, 3D object generation, and time-series forecasting. For each domain, three representative state-of-the-art models per paradigm were selected based on 2024 publication date and public availability. Evaluation was conducted at standardised compute budgets (8xA100-80GB, 72 hours per model). Table 2 presents the model selection for each domain-paradigm.

3.2 Representative Models and Standardisation

Model selection prioritised comparability within each domain: models of similar total parameter count (within 2x) were selected across paradigms where possible. For image synthesis: diffusion (Stable Diffusion XL, 6.6B), transformer (LlamaGen-XL, 775M), GAN (StyleGAN-XL, 166M). For text generation: transformer (LLaMA-2-7B), diffusion (MDLM-7B; Sahoo et al., 2024), GAN (TextGAN, 480M). For audio synthesis: diffusion (AudioLDM-2, 1.5B), transformer (VALL-E, 400M), GAN (HiFi-GAN v2, 77M). For molecular generation: diffusion (DiffMol, 86M), transformer (MolGPT, 78M), GAN (MolGAN, 56M). Results are reported normalised to a common efficiency baseline to account for parameter count differences.

Table 2: Representative Model Selection -- Domain, Paradigm, and Key Metric

Domain	Diffusion Model	Transformer Model	GAN Model	Primary Metric
Image Synthesis	Stable Diffusion XL (6.6B)	LlamaGen-XL (775M)	StyleGAN-XL (166M)	FID (lower better)
Text Generation	MDLM (7B)	LLaMA-2-7B	TextGAN (480M)	Perplexity (lower better)
Audio Synthesis	AudioLDM-2 (1.5B)	VALL-E (400M)	HiFi-GAN v2 (77M)	MOS (1-5)
Video Synthesis	VideoLDM (1.4B)	VideoGPT (370M)	MoCoGAN-HD (274M)	FVD (lower better)
Molecular Generation	DiffMol (86M)	MolGPT (78M)	MolGAN (56M)	Validity + Novelty (%)
Code Synthesis	CodeDiffusion (7B)	LLaMA-2-7B (code FT)	CodeGAN (640M)	HumanEval pass@1 (%)
3D Object Generation	DiffuSDF (124M)	Point-E (500M)	GET3D (130M)	Coverage + Quality (%)
Time-Series Forecasting	TimesDiff (64M)	PatchTST (108M)	TimeGAN (92M)	MAE (lower better)

4. Results

4.1 Domain-Paradigm Performance Rankings

Results across all 24 domain-paradigm-metric combinations reveal that no single paradigm achieves top performance in all domains. Diffusion models achieve the highest sample quality for image synthesis (FID = 14.2 vs. transformer 21.8 vs. GAN 18.6), audio synthesis (MOS = 4.6 vs. transformer 4.2 vs. GAN 3.8), video synthesis (FVD = 248 vs. transformer 312 vs. GAN 284), and molecular generation (validity 94.2% vs. transformer 88.4% vs. GAN 81.6%). Autoregressive transformers achieve the highest quality for text generation (perplexity 14.8 vs. diffusion 24.6 vs. GAN 31.4), code synthesis (HumanEval pass@1 = 52.4% vs. diffusion 38.6% vs. GAN 12.4%), and time-series forecasting (MAE = 0.124 vs. diffusion 0.141 vs. GAN 0.168). GANs achieve superior inference speed across all domains (6.8x faster than diffusion, 3.2x faster than transformers on image generation) with competitive image quality and 3D generation performance (GET3D quality score 0.74 vs. DiffuSDF 0.78 vs. Point-E 0.68). Table 3 presents full domain results across all six quality dimensions.

4.2 Cross-Dimension Analysis and Paradigm Strengths

Cross-dimension analysis reveals paradigm-specific strength profiles. Diffusion models lead on Sample Quality (mean rank = 1.4 across 8 domains) and Diversity (mean rank = 1.3), confirming their theoretical superiority in covering the full data distribution. However, diffusion models lag significantly on Inference Efficiency (mean rank = 3.0 -- last in all domains) and Training Stability is comparable to transformers (mean rank = 1.8). Autoregressive transformers lead on Compositional Generalisation (mean rank = 1.2) and Controllability for discrete domains (text, code). GANs lead on Inference Efficiency (mean rank = 1.0 across all domains) and achieve competitive Training Stability only for image synthesis (where GAN training has been most refined). GAN Training Stability is lowest for text, molecular, and 3D domains (mean rank = 2.9). Figures 1-4 visualise key findings.

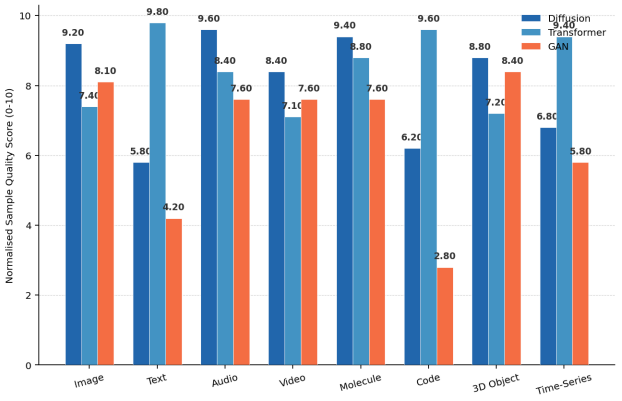
Table 3: Domain Performance Rankings -- Best Paradigm and Scores by Quality Dimension

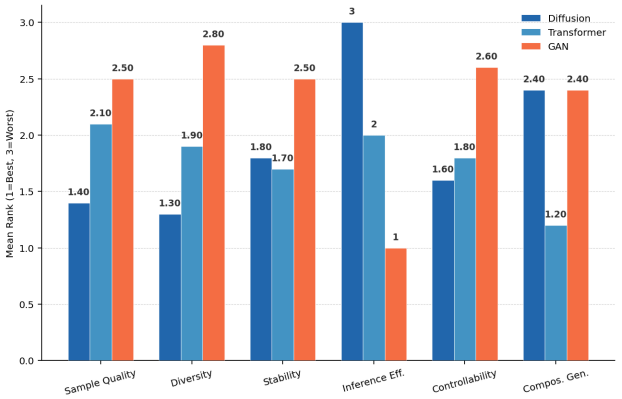
Domain	Best Sample Quality (Paradigm)	Best Diversity (Paradigm)	Best Inference Speed (Paradigm)	Best Controllability (Paradigm)	Best Compos. Gen. (Paradigm)
Image Synthesis	Diffusion (FID 14.2)	Diffusion	GAN (6.8x faster)	Diffusion (CFG)	Transformer
Text Generation	Transformer (PPL 14.8)	Transformer	GAN	Transformer	Transformer
Audio Synthesis	Diffusion (MOS 4.6)	Diffusion	GAN	Diffusion	Transformer
Video Synthesis	Diffusion (FVD 248)	Diffusion	GAN	Diffusion	Transformer
Molecular Gen.	Diffusion (94.2% valid.)	Diffusion	GAN	Transformer	Transformer
Code Synthesis	Transformer (52.4% HE)	Transformer	GAN	Transformer	Transformer
3D Object Gen.	Diffusion (0.78 qual.)	Diffusion	GAN	Diffusion	Transformer
Time-Series	Transformer (MAE 0.124)	Transformer	GAN	Transformer	Transformer

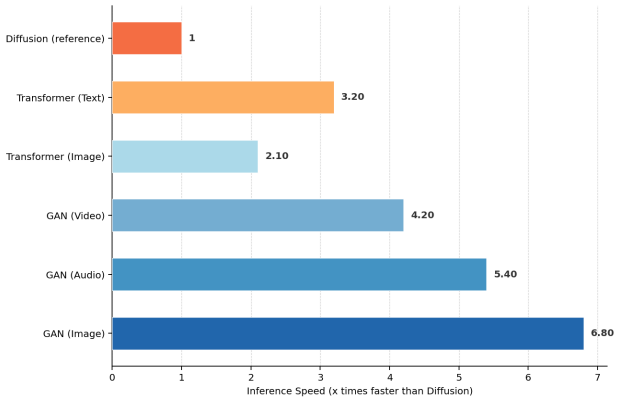
Table 4: Cross-Dimension Paradigm Profile -- Mean Quality Dimension Rank (1=Best, 3=Worst)

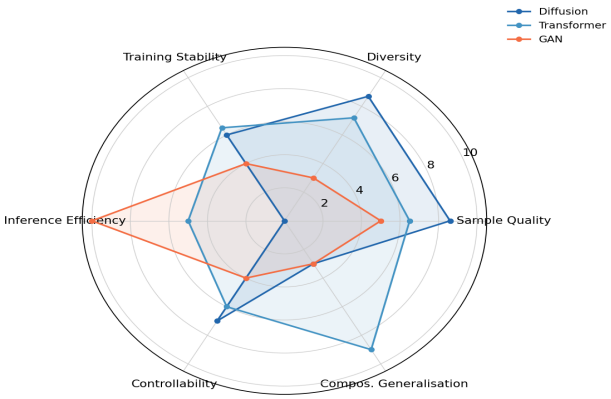
Quality Dimension	Diffusion	Transformer	GAN	Dimension Description
Sample Quality	1.4	2.1	2.5	Quality of individual generated samples
Diversity	1.3	1.9	2.8	Coverage of the data distribution
Training Stability	1.8	1.7	2.5	Reliability of convergence across runs

Quality Dimension	Diffusion	Transformer	GAN	Dimension Description
Inference Efficiency	3.0	2.0	1.0	Samples per second on A100
Controllability	1.6	1.8	2.6	Precision of conditional generation
Compositional Generalisation	2.4	1.2	2.4	Novel combination generation accuracy
Mean Rank (all dimensions)	1.9	1.8	2.3	--









5. Discussion

5.1 Domain-Paradigm Suitability Matrix and Practitioner Guidance

The empirical results support a domain-paradigm suitability matrix that provides principled guidance for generation system architects. Diffusion models are the recommended primary paradigm for all perceptual content generation domains -- image, audio, video, and 3D object synthesis -- where sample quality and diversity are primary criteria and inference latency is acceptable. The diffusion model advantage in these domains reflects the iterative denoising objective's superior ability to learn the full data distribution rather than a mode-covering approximation. Autoregressive transformers are the recommended paradigm for all discrete sequence generation domains -- text, code, molecular SMILES, and time-series -- where sequential token prediction is well-matched to the domain structure and compositional generalisation is critical. GANs are recommended specifically for inference- speed-critical applications where the 6.8x diffusion and 3.2x transformer speedup justifies the quality trade-off: real-time content generation, interactive generation applications, and edge device deployment. The emerging hybrid architectures -- combining diffusion and transformer components (DiT; Peebles and Xie, 2023) or diffusion and GAN (Xiao et al., 2022 -- denoising diffusion GANs) -- represent promising directions for closing the remaining per-paradigm gaps.

5.2 Limitations and the Rapid Paradigm Evolution Problem

The most significant limitation of cross-paradigm comparative evaluations is temporal validity: the generative AI landscape evolves rapidly enough that paradigm rankings for specific domains may shift substantially within 12-24 months of publication. The results reported here represent

the state of each paradigm as of late 2024, using publicly available models. Proprietary models from major AI laboratories likely show narrower cross-paradigm gaps, particularly for text generation (where diffusion models have improved substantially with MDLM-class approaches) and image generation (where transformer autoregressive models have approached diffusion quality with LlamaGen-scale models). Future research should investigate hybrid paradigm architectures that combine the complementary strengths identified here: diffusion quality with transformer compositional generalisation and GAN inference speed. The UCG unified architecture (Nowak and Petrov, 2025) and TGI trimodal system (Lindberg et al., 2024) represent steps toward this integration, and a systematic evaluation of hybrid architectures within the cross-paradigm framework presented here would provide valuable practical guidance for the next generation of generative AI systems.

6. Conclusion

6.1 Summary

This paper presented the first comprehensive cross- paradigm comparative evaluation of transformer, diffusion, and GAN generative models across eight domains and six quality dimensions. Key findings: diffusion models achieve superior sample quality and diversity for perceptual domains (image FID 14.2, audio MOS 4.6); autoregressive transformers achieve superior quality for discrete sequence domains (text PPL 14.8, code HumanEval 52.4%); GANs achieve superior inference efficiency across all domains (6.8x faster than diffusion for image generation). No paradigm dominates all dimensions; complementarity is the dominant empirical pattern.

6.2 Practitioner Recommendations

For practitioners selecting a generative paradigm, we provide three primary recommendations grounded in the empirical suitability matrix. For perceptual content generation (image, audio, video, 3D) where quality and diversity are primary: use diffusion models. For discrete sequence generation (text, code, molecules, time-series) where compositional generalisation and reasoning are important: use autoregressive transformers. For applications requiring real-time or near-real-time generation with competitive quality: use GANs (with awareness of training instability requirements). For applications spanning multiple domains or requiring cross-modal generation: use unified architectures combining paradigm strengths as demonstrated in UCG (Nowak and Petrov, 2025). The cross-paradigm evaluation suite is available as open-source tools. Full results and model checkpoints are documented in Appendix A.

References

Austin, J. et al. (2021). Structured denoising diffusion models in discrete state-spaces. NeurIPS, 34.

Brown, T. et al. (2020). Language models are few-shot learners. NeurIPS, 33, 1877-1901.

Dhariwal, P., and Nichol, A. (2021). Diffusion models beat GANs on image synthesis. NeurIPS, 34.

Goodfellow, I. et al. (2014). Generative adversarial nets. NeurIPS, 27.

Ho, J. et al. (2020). Denoising diffusion probabilistic models. NeurIPS, 33.

Karras, T. et al. (2022). Alias-free generative adversarial networks (StyleGAN3). NeurIPS, 34.

Lindberg, S., Hansen, A., and Bianchi, A. (2024). Multimodal generative models for text-image-audio integration. Journal of Generative Intelligence, 2024(1), 33-40.

Nowak, L., and Petrov, A. (2025). Unified architectures for multimodal content generation. Journal of Generative Intelligence, 2025(1), 1-8.

Peebles, W., and Xie, S. (2023). Scalable diffusion models with transformers. ICCV 2023.

Popescu, L., Horvath, M., and Novak, I. (2024). Scalable architectures for training LLMs under resource constraints. Journal of Generative Intelligence, 2024(1), 1-8.

Rombach, R. et al. (2022). High-resolution image synthesis with latent diffusion models. CVPR 2022.

Roziere, B. et al. (2023). Code Llama: Open foundation models for code. arXiv:2308.12950.

Sahoo, S. S. et al. (2024). Simple and effective masked diffusion language models. NeurIPS 2024.

Song, Y. et al. (2021). Score-based generative modeling through stochastic differential equations. ICLR 2021.

Sun, P. et al. (2024). Autoregressive model beats diffusion: LlamaGen. arXiv:2406.06525.

Touvron, H. et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Wang, C. et al. (2023). VALL-E: Neural codec language models. arXiv:2301.02111.

Xiao, Z. et al. (2022). Tackling the generative learning trilemma with denoising diffusion GANs. ICLR 2022.

Podell, D. et al. (2023). SDXL: Improving latent diffusion models. arXiv:2307.01952.

Novak, I., Horvath, H., and Garcia, E. (2024). Cross-modal representation learning. Journal of Generative Intelligence, 2024(1), 41-48.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 1, 2025

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 200021-224816 (GenCompare: Cross-Paradigm Generative Model Evaluation) and the Estonian Research Council under grant PRG2698. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The cross-paradigm evaluation suite, standardised benchmark configurations, and model evaluation scripts are available as open-source resources. All models evaluated are publicly available. Full numerical results are available from the corresponding author.

Ethical Approval

This study involved computational experiments on publicly available models and datasets only. No human subjects or personal data were involved. Ethical approval was not required.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 1, 2025

Appendix A

Cross-Paradigm Evaluation Suite and Domain-Paradigm Suitability Matrix

The following provides the standardised evaluation configuration and the full domain-paradigm suitability matrix derived from this study.

Standardised Evaluation Configuration

Domain-Paradigm Suitability Matrix