A TECHNICAL COMPARISON OF GANS, VARIATIONAL AUTOENCODERS, AND DIFFUSION MODELS IN MULTIMODAL CONTENT GENERATION

Basu Dev Shivhare[1], Radha Raman Chandan[2]

[1] Professor, School of Computer Science & Engineering, Galgotias University, Greater Noida, India, ,basu.shivhare@galgotiasuniversity.edu.in
[2]Professor, Department of Computer Science, School of Management Sciences (SMS), Varanasi, India, rrchandan@smsvaranasi.com

Abstract

In this study, we compare and contrast Generative Adversarial Networks, Variational Autoencoders and Diffusion Models in three central evaluation metrics: sample fidelity, controllability and training stability. It will cover the adversarial minimax formulation that has been the foundation for Generative Adversarial Networks, the training instabilities such as mode collapse which have historically hindered their reliability and the architectural and loss function modifications that have been made to improve the issues associated with such instabilities.

Based on the analysis, Diffusion Models currently have the best trade-off between sample fidelity and conditional controllability, with high inference latency but low training instability, Variational Autoencoders have the best sample fidelity and a stable training process with an interpretable latent space but lower inference speed, and Generative Adversarial Networks have the lowest training instability and inference speed but a less stable training process due to persistent training instability issues.

Keywords: *Generative Adversarial Networks, Variational Autoencoders, Diffusion Models, Generative Modelling, Sample Fidelity, Training Stability, Controllable Generation, Multimodal Generation, Latent Space, Denoising.*

Corresponding Author: Basu Dev Shivhare, Professor, School of Computer Science & Engineering, Galgotias University, Greater Noida, India, basu.shivhare@galgotiasuniversity.edu.in
How to cite this article: : Basu Dev Shivhare., Radha Raman Chandan (2026). A Technical Comparison Of Gans, Variational Autoencoders, And Diffusion Models In Multimodal Content Generation, Confluence: A Journal of Technology, Business & Society, Vol. I,(1) ,1-10
Source of support: Nil
Conflict of Interest: None
Received: 28-05-2026 Accepted:25-06-2026 Published: 22-07-2026

I. Introduction

Generative Modelling aims at learning the underlying distribution of a training corpus to generate new and realistic samples that are similar to the distribution. In the last decade, this task has been addressed in many different architectural paradigms, which have different ways of posing this generative problem, and thus different trade-offs in terms of sample fidelity, controllability, and training stability. These trade-offs are key for choosing the right generative modelling technique for a specific multimodal content generation system.

GANs incorporated a totally new training paradigm where both the generator network is trained to generate more and more realistic samples and the discriminator network is trained to tell if the samples are generated or real, with both networks continually improving as a result of this adversarial training. This strategy was able to generate very realistic samples but was often found to be very hard to train reliably, with the generator often suffering from "mode collapse", that is, it generates limited variation in output, and/or there was training instability due to the adversarial optimization process.

The alternative approach based on probabilistic inference was Variational Autoencoders, which aimed to learn a compact, structured latent code for the training data by optimizing a variational lower bound on the likelihood of the data. In most cases, this formulation results in more stable training dynamics and a more interpretable latent space, allowing for controllable generation by direct manipulation of the latent space, but also with lower sample sharpness than adversarial based methods.

The third paradigm, called Diffusion Models, is newer and uses the idea of a gradual diffusion of noise over training data. The diffusion-based approaches have shown to be able to achieve comparable to or better sample fidelity than adversarial methods, and they don't suffer from the same instability during training as found in adversarial methods, although they generally need significantly more computation during inference because they follow an iterative sampling procedure. This work integrates the three different paradigms and presents a technical comparison between their fidelity, controllability and stability properties in the context of multimodal content generation.

Research Objectives

1. To discuss the adversarial training formulation for Generative Adversarial Networks (GANs) and the stability problems that come with it.
2. To examine the probabilistic latent-variable formulation of Variational Autoencoders and implications for controllable generation.
3. To consider the iterative denoising formulation of Diffusion Models and its connection with sample fidelity.
4. To assess fidelity, controllability and training stability for all three of the generative modeling paradigms.
5. To perform computational trade-offs for deployment of the generative model for multimodal content generation applications.

Table 1. Core Generative Modeling Paradigms and Their Formulations

Paradigm	Core Formulation	Primary Strength
GAN	Minimax game between generator and discriminator	High sample fidelity
VAE	Variational lower bound on data likelihood via latent encoding	Stable training and interpretable latent space
Diffusion Model	It is an iterative process that reverses a gradual noising process	High Fidelity and Strong Controllability

II. Literature Review

2.1 Adversarial Generative Modeling: Principles and Training Dynamics

The framework of Generative Adversarial Networks (GANs) was introduced where a generator network maps random noise vectors to realistic data samples and a discriminator network is trained to classify generated samples as real or fake based on the ability of the generator to fool it, with the generator learning improved by gradient signals generated from the discriminator [1]. This oppositional formulation exhibited a remarkable power to yield sharp realistic samples, without having to explicitly invoke a likelihood formulation, unlike earlier generative formulations.

More recent deep convolutional variants of the adversarial framework also illustrated the significant value of architectural choices to training stability and quality of samples in image generation applications, leading to a number of architectural features that were adopted and are now common in research with adversarial generative models such as strided convolutions and batch normalization [2].

2.2 Mode Collapse and Stability Challenges in GAN Training

Although adversarial networks were observed to be very sample-fidel, two failure modes were seen: First, the networks often exhibited mode collapse, that is, the generator would converge to generating only a subset of the data distribution; second, they are generally observed to be more prone to training instability compared to standard supervised loss minimization [3] because of the min-max optimization dynamic. Switching to a different distance metric between the real and generated data distribution was demonstrated to greatly increase the training stability and to yield a more informative, and continuous, training signal which directly mitigates several of the instability problems occurring in the original adversarial formulation [3].

To enhance the training stability and the quality of the generated samples, progressive training strategies were also demonstrated; they involve gradually increasing the generator and discriminator networks from low to high resolution over the course of training, and in addition, later style-based generator architectures added explicit controls for the attributes of the generated samples at various levels of visual detail [4], [5].

2.3 Variational Autoencoders and Probabilistic Latent Representations

Variational Autoencoders (VAE) [6] were proposed as a probabilistic generative approach in which an encoder network is trained to map input data onto a distribution over a lower dimensional latent space, and a decoder network is trained to reconstruct the data samples based on the latent representations, with the entire model trained to maximize an analytically tractable variational lower bound on the true data

likelihood. This resulted in a principled probabilistic framework for generative models, and allowed sampling from the learned latent distribution.

Later work showed that by tuning the relative strength of the reconstruction and regularization terms in the variational objective, the learned latent space can achieve more disentanglement, meaning that the learned latent space dimensions are more likely to represent semantically meaningful and independently controllable factors of variations in the data generated in that space and thus make the generator more controllable with Variational Autoencoder-based generation [7].

2.4 Diffusion Models and Iterative Denoising Generation

The conceptual background for diffusion-based generative modelling is laid out in a formulation based on nonequilibrium thermodynamics, where a continuous and tractable noising process is applied to the training data, and a neural network is trained to learn this reverse process that can denoise random noise to reconstruct realistic data samples through iterative denoising [8]. This initial formulation was subsequently simplified and empirically proven to be an effective training objective for denoising diffusion probabilistic models to generate samples of comparable quality to those generated by adversarial training on popular image generation datasets [9].

Subsequent comparative evaluation showed that, with appropriately tuned diffusion models, it was possible to outperform the sample fidelity of adversarial networks on several standard benchmarks, while at the same time enjoying substantially more stable and predictable training dynamics, thanks to the well-defined denoising objective instead of an adversarial minimax game [10], [11].

2.5 Controllability and Conditioning Mechanisms Across Paradigms

In all three paradigms, controllable generation was demonstrated, where sample attributes were controlled by auxiliary conditioning information, like class labels or textual descriptions. Adversarial networks with conditional variants directly integrated class/attribute information into both the generator and discriminator networks, allowing attribute-controlled generation of samples from early in the adversarial network development [12] and [16]. The idea of adversarial conditioning was then generalized to structured input-output mappings in the image-to-image translation framework, both with paired and unpaired settings [13] and [14].

Iterative denoising approaches then showed great text-conditioned generation potential by introducing rich conditioning signals at each denoising step, resulting in high fidelity and strong alignment between conditioning input and generated output, thus making diffusion-based conditioning a particularly powerful mechanism that allows for user-friendly multimodal content generation tasks with fine-grained control over the content [15].

2.6 Research Gap

While they have all been studied extensively on their own, few works directly compare all three paradigms with respect to fidelity, controllability and training stability in the same shared framework of evaluation,

with a focus on the computational cost and benefits for practical use in multimodal content generation deployment. This review fills this gap by systematically comparing all three paradigms, from a technical perspective.

III. Comparative Technical Framework for Generative Model Evaluation

3.1 Adversarial Objective Formulation and Discriminator Dynamics

The proposed comparative approach is based on the adversarial minimax objective that the generator and discriminator networks are trained in opposition against each other instead of against a common, jointly minimized loss. This allows for good incentive towards sample realism, but without the stability guarantees of standard gradient descent on a fixed objective, this might be why adversarial training has been found to suffer from mode collapse and oscillatory, non-convergent training behaviour [18].

One of the common issues in this approach is ensuring a proper balance in terms of ability between the generator and discriminator during training. If the discriminator is too good too soon, the gradient signal it passes on to the generator can become too weak for a meaningful improvement of the generator, for a discriminator which perfectly separates real from generated samples provides little informative gradient for the improvement of the generator. On the other hand, if the discriminator's output is too far ahead of the generator's output, then it will not generate a useful training signal [20]. One of the major practical issues in adversarial training is the balance of these two relationships, usually achieved by ensuring that one learning rate is not too high relative to the other, by swapping between different update schedules or by regularizing the architecture.

Specific regularization methods, such as gradient penalty terms that limit the change in the discriminator's output with respect to small changes in its input, have been seen to make a meaningful contribution to this balance by ensuring that the discriminator does not form decision boundaries that become too sharp and "kill" the gradient information available to the generator in the initial and intermediate phases of training [17].

3.2 Variational Inference and the Evidence Lower Bound

Variational Autoencoders are defined in the framework by using evidence lower bound that is easily tractable as an approximation to the data likelihood, which is usually intractable [21]. Given that this is a fixed objective to optimize with gradients and is not a game of the adversary, training is very stable, although the reconstruction term in the objective has a tendency to induce an averaging behaviour that can make the generated samples less extreme than adversarial alternatives.

The evidence can be decomposed into a term that reconstructs the input data, encouraging a decoder to recover the input data from its latent encoding, and a term that regularizes the learned latent distribution towards a simple prior distribution over the latent space, typically a standard normal distribution [22], [18].

By choosing larger weights on one term or the other, the trade-off between these terms provides a way of balancing the trade-off between representation regularity and reconstruction fidelity, meaning that a more regular latent space is generated while sacrificing reconstruction fidelity on the one hand, and vice versa on the other, depending on the application.

3.3 Forward and Reverse Diffusion Processes

The Diffusion Models are characterized by a two-stage process inside the framework: the forward process is fixed and adds noise to the training data in many steps, and the reverse process is learned and is used to remove the noise step by step, and finally generate a realistic data sample from a random noise sample by iterative refinement. Unlike with Generative Adversarial Networks where the training objective for the reverse process is adversarial, the training objective for the reverse process is stable and well behaved as the forward process is fixed and mathematically well-defined [23].

The forward noising will have a corresponding reverse denoising, where the number of steps performed in the forward operation will correspond to the granularity of the reverse denoising, more steps in the forward process will lead to smoother, more gradual transitions that are easier for the reverse process to learn, and will also increase the sampling time of the forward process by an equal amount. The relationship between steps, learnability, and inference cost has driven a lot of research into reducing the number of sampling steps needed during the inference process without compromising the sample quality benefits of a finely discretized forward process [24].

The choice of noise schedule, the way noise variance changes between the successive steps of the forward process, has a significant impact on training stability and the quality of the reverse-process learning at the intermediate steps as well; schedules that introduce noise too harshly in early steps, or too slowly in late steps give measurably worse learning of the reverse process [25].

3.4 Latent Space Structure and Manipulation for Controllable Generation

Variational Autoencoder provides an explicit low dimensional latent space that can be directly manipulated to control output attribute of generated outputs. While Diffusion Models gain control over the process by using external conditioning signals at every denoising step, adversarial networks can have either explicit conditioning inputs or learned latent space manipulation, depending on the architecture [26].

Diffusion-based conditioning acts at each step of an extended iterative process, providing a unique controllability over conditioning strength where the strength of a conditioning signal can be fine-tuned throughout the generation process such that it can influence the generation of the final output more or less than the model's unconditioned generative prior [24]. This step-wise conditioning flexibility is not offered directly by single pass architectures, where conditioning information needs to be completely included in one forward computation.

3.5 Computational Cost and Inference Efficiency Trade-offs

Generative Adversarial Networks and Variational Autoencoders both produce a set of samples by making just a single forward pass, which makes them fast to run for real-time or interactive applications. In comparison, Diffusion Models need multiple denoising steps to produce one sample, leading to significantly

longer inference latency, which needs to be carefully considered against the merits of fidelity and controllability depending on the specific deployment scenario [19].

Iterative denoising has led to various methods of acceleration, such as reduced-step sampling and distillation, which involves training a student model with a few steps to imitate the results of a teacher model with many steps. The fidelity and controllability benefits of diffusion-based generation don't necessarily stem from the computational cost of the original formulation and it is an open question whether this gap will decrease over time as methods are refined to match inference efficiency of single-pass generative strategies, and whether these benefits will be further developed in the future [25].

IV. Comparative Fidelity and Stability Analysis

Combining the evidence from all three paradigms suggests that no single generative modeling approach dominates on any of the evaluation dimensions. While GANs provide good sample quality and quick inference, they are somewhat more difficult to reliably train, with careful architectural and hyperparameter tuning needed to prevent mode collapse [27]. VAEs have the consistently best training dynamics, and the best interpretable latent structure, but can generally generate fewer crisp samples (due to the smoothing effect of the reconstruction objective used in training).

The three paradigms differ in the degree of fidelity of the samples they can generate and in the degree of conditioning of those samples; Diffusion Models are also fairly expensive to train, and have well-defined denoising objective, which leads to training stability, but are the more expensive of the three approaches at inference time [29].

Table 2. Comparative Summary of Generative Modelling Paradigms

Model Type	Training Stability	Sample Fidelity	Controllability	Inference Speed
GAN	Low to Moderate	High	Moderate	Fast
VAE	High	Moderate	High	Fast
Diffusion Model	High	High	High	Slow

V. Implications for Multimodal Content Generation Applications

These comparative results have implications for selecting a generative modelling technique that can be used for a specific multimodal content creation system. Applications that demand fast, real-time inference of acceptable sample quality, such as interactive design tools, might still be more likely to benefit from adversarial or variational methods because of their single-pass inference. As inference acceleration methods make it possible to bridge the gap between diffusion-based and other methods, applications that require

absolute fidelity to the sample and fine-grained conditional control, such as high-quality text-to-image or text-to-audio generation, are increasingly leaning towards diffusion-based methods, despite their higher computational cost [28].

While this explicit structure of the latent space is suitable for many types of applications, it remains welcome for applications that involve interpretable, directly manipulable latent representations, for example-controlled attribute editing or data compression tasks, where it is still used in spite of its disadvantages [30].

VI. Conclusion

In this review, the technical performance of Generative Adversarial Networks, Variational Autoencoders, and Diffusion Models in fidelity, controllability, and training stability have been compared for generating multimodal contents. The evidence suggests that adversarial networks are robust, have quick inference times and are not very stable to train, Variational Autoencoders are stable to train but have a less sharp sample and have a less controllable output, and Diffusion Models have the best current balance of fidelity and controllability, but at the cost of slower inference times.

It also examines the probabilistic latent-variable formulations of Variational Autoencoders and iterative denoising formulations of Diffusion Models, and their respective advantages and disadvantages in terms of sample quality, interpretability of the latent space, and the cost of the inference. The two trade-offs suggest that generative model selection must be based on the desired fidelity, control, and computational requirement of the specific target multimodal application and not on one dominant paradigm.

VII. Future Scope

Future research can focus on further refining and assessing sampling acceleration methods that approach single-pass generative techniques with a comparable level of inference delay, exploring hybrid architectures that integrate adversarial, variational, and diffusion-based components to synergistically optimize fidelity, controllability, and efficiency, and broaden comparative analysis to cover a wider spectrum of tasks beyond static image generation to audio, video, and cross-modal generation [20]. This is a continuous benchmarking process that will continue to be crucial in the future, as generative modelling methods evolve and diversify.

References:

[1] S. Joshi, "Introduction to diffusion models, autoencoders and transformers: Review of current advancements," 2025.

[2] S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni, "Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers," IEEE Access, vol. 12, pp. 69812–69837, 2024.

[3] Z. Sordo, E. Chagnon, Z. Hu, J. J. Donatelli, P. Andeer, P. S. Nico, T. Northen, and D. Ushizima, "Synthetic scientific image generation with VAE, GAN, and diffusion model architectures," J. Imaging, vol. 11, no. 8, p. 252, 2025.

[4] A. Jlidi and L. Kovács, "Applying deep generative models to portrait art generation: A comparative study of GANs and VAEs," Production Systems and Information Engineering, vol. 12, no. 1, pp. 64–76, 2024.
[5] X. Li, Y. Peng, and H. Zheng, "Research and analysis of VAE, GAN, and diffusion generation models," Science and Technology of Engineering, Chemistry and Environmental Protection, vol. 1, no. 1, 2025.
[6] M. Islam, T. Huang, E. Ahn, and U. Naseem, "Multimodal generative AI with autoregressive LLMs for human motion understanding and generation: A way forward," arXiv preprint arXiv:2506.03191, 2025.
[7] E. Palumbo, L. Manduchi, S. Laguna, D. Chopard, and J. E. Vogt, "Deep generative clustering with multimodal diffusion variational autoencoders," in Proc. 12th Int. Conf. Learning Representations (ICLR), 2024.
[8] Y. Houkpati, "Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers," arXiv preprint, 2023.
[9] G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, T. Nolte, S. Nebelung, and J. N. Kather, "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis," Sci. Rep., vol. 13, no. 1, p. 12098, 2023.
[10] X. Wang, Z. He, and X. Peng, "Artificial-intelligence-generated content with diffusion models: A literature review," Mathematics, vol. 12, no. 7, p. 977, 2024.
[11] X. Wang, Y. Zhou, B. Huang, H. Chen, and W. Zhu, "Multi-modal generative AI: Multi-modal LLMs, diffusions and the unification," IEEE Trans. Circuits Syst. Video Technol., 2025.
[12] K. Chen and L. Ramsey, "Deep generative models for 3D content creation: A comprehensive survey of architectures, challenges, and emerging trends," 2024.
[13] F. Nazarieh, Z. Feng, M. Awais, W. Wang, and J. Kittler, "A survey of cross-modal visual content generation," IEEE Trans. Circuits Syst. Video Technol., vol. 34, no. 8, pp. 6814–6832, 2024.
[14] Z. Li, X. Liu, and Y. Lu, "Deep generative models for image synthesis: GANs, VAEs, and diffusion approaches," Highlights in Science, Engineering and Technology, vol. 160, pp. 205–211, 2025.
[15] J. Zhou, "Text to image generation: A literature review focus on the diffusion model," in ITM Web Conf., vol. 73, 2025, p. 02037.
[16] S. Yazdani, N. Saxena, Z. Wang, Y. Wu, and W. Zhang, "A comprehensive survey of image and video generative AI: Recent advances, variants, and applications," ResearchGate preprint, doi: 10.13140/RG.2.30721.63842, 2024.
[17] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P. A. Heng, and S. Z. Li, "A survey on generative diffusion models," IEEE Trans. Knowl. Data Eng., vol. 36, no. 7, pp. 2814–2830, 2024.
[18] Y. J. Kim and S. P. Lee, "A generation of enhanced data by variational autoencoders and diffusion modeling," Electronics, vol. 13, no. 7, p. 1314, 2024.
[19] Y. Hu, L. Wang, X. Liu, L. H. Chen, Y. Guo, Y. Shi, C. Liu, A. Rao, Z. Wang, and H. Xiong, "Simulating the real world: A unified survey of multimodal generative models," IEEE Trans. Pattern Anal. Mach. Intell., 2026.
[20] S. Yazdani, A. Singh, N. Saxena, Z. Wang, A. Palikhe, D. Pan, U. Pal, J. Yang, and W. Zhang, "Generative AI in depth: A survey of recent advances, model variants, and real-world applications," J. Big Data, vol. 12, no. 1, p. 230, 2025.
[21] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT," arXiv preprint arXiv:2303.04226, 2023.
[22] P. K. Pativada, R. Karne, and A. Dudhipala, "Exploring GANs and VAEs: Advances and challenges in image synthesis and editing," Multidisciplinary Int. J. Res. Dev., vol. 4, no. 4, pp. 131–140, 2025.
[23] A. Khani, A. Rampini, B. Roy, L. Nadela, N. Kaplan, E. Atherton, D. Cheung, and J. Bibliowicz, "Motion generation: A survey of generative approaches and benchmarks," arXiv preprint arXiv:2507.05419, 2025.
[24] S. Adilkhan, M. Alimanova, L. Shi, and A. Soltiyeva, "Advances in artificial intelligence-driven 3D model generation: A review of GAN and VAE methodologies," Bull. Electr. Eng. Inform., vol. 14, no. 6, pp. 4988–5011, 2025.

[25] R. Po, W. Yifan, V. Golyanik, K. Aberman, J. T. Barron, A. Bermano, E. Chan, T. Dekel, A. Holynski, A. Kanazawa, and C. K. Liu, "State of the art on diffusion models for visual computing," Computer Graphics Forum, vol. 43, no. 2, p. e15063, 2024.

[26] Z. Xing, Q. Feng, H. Chen, Q. Dai, H. Hu, H. Xu, Z. Wu, and Y.-G. Jiang, "A survey on video diffusion models," ACM Comput. Surv., vol. 57, no. 2, pp. 1–42, 2024.

[27] A. El Saddik, J. Ahmad, M. Khan, S. Abouzahir, and W. Gueaieb, "Unleashing creativity in the metaverse: Generative AI and multimodal content," ACM Trans. Multimedia Comput. Commun. Appl., vol. 21, no. 7, pp. 1–43, 2025.

[28] A. Ramisa, R. Vidal, Y. Deldjoo, Z. He, J. McAuley, A. Korikov, S. Sanner, M. Sathiamoorthy, A. Kasrizadeh, S. Milano, and F. Ricci, "Multi-modal generative models in recommendation system," arXiv preprint arXiv:2409.10993, 2024.

[29] T. Sakirin and S. Kusuma, "A survey of generative artificial intelligence techniques," Babylonian J. Artif. Intell., pp. 10–14, 2023.

[30] J. Xiang, R. Huang, J. Zhang, G. Li, X. Han, and Y. Wei, "VersVideo: Leveraging enhanced temporal diffusion models for versatile video generation," in Proc. Int. Conf. Learning Representations (ICLR), vol. 2024, 2024, pp. 1964–1982.