Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Trust-Aware Benchmarking of GAN, VAE, and Diffusion Models for Synthetic Data in Image and Tabular Domains

Anil Kumar Shukla

shuklaks@gmail.com

Principal – AI Labs Infosys Consulting

Research Article

Keywords: Synthetic Data, Generative AI, GANs, Variational Autoencoders (VAEs), Diffusion Models, Hybrid Architectures, Data Privacy, Fairness, Computational Efficiency, Trust-aware Benchmarking

Posted Date: October 13th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7649434/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Trust-Aware Benchmarking of GAN, VAE, and Diffusion Models for Synthetic Data in Image and Tabular Domains

Anil Kumar Shukla
Infosys Consulting, AI Labs, Pune, India
Corresponding Author: Anil Kumar Shukla
(Email: shuklaks@gmail.com)

Abstract
Synthetic data generation using generative AI models offers a promising solution to challenges in data availability, privacy, and fairness. However, comparative insights across model families and data modalities remain limited, especially when trust-related dimensions such as fairness, privacy, and efficiency are considered alongside fidelity.

This paper presents a trust-aware benchmarking study across both visual (CIFAR-10, MNIST) and tabular (Adult Income) domains, evaluating representative baselines—vanilla GAN/WGAN-GP, standard and β-VAE, DDPM and classifier-guided DDPM, and hybrids such as VAE-GAN and Latent Diffusion. Advanced models including StyleGAN2/3, BigGAN, CTGAN, Diffusion Transformers, and GigaGAN are acknowledged to situate the findings within the evolving landscape of foundation-scale generators.

Performance is assessed using a multi-objective framework that integrates fidelity (FID, precision/recall), fairness (demographic parity), privacy leakage resistance, and computational efficiency. Results show that hybrid latent diffusion models achieve near-diffusion fidelity (FID 10.2 vs. 8.5 on CIFAR-10; 7.8 vs. 6.2 on MNIST) while reducing sampling time by over 70%. On tabular data, hybrids balance accuracy (84.7%) and fairness (0.93), whereas classical GANs and VAEs exhibit trade-offs between fidelity, efficiency, and fairness.

To the best of our knowledge, this is the first study to benchmark GANs, VAEs, diffusion, and hybrid models across both image and tabular data using a unified, trust-aware evaluation framework. By providing reproducible, cross-domain comparisons, this work offers practical guidance for selecting and deploying generative models in trust-sensitive applications such as healthcare, finance, and autonomy.

Keywords: Synthetic Data, Generative AI, GANs, Variational Autoencoders (VAEs), Diffusion Models, Hybrid Architectures, Data Privacy, Fairness, Computational Efficiency, Trust-aware Benchmarking

1. Introduction

Fast-paced adoption of artificial intelligence in healthcare (Frid-Adar et al., 2018), finance (Jordon et al., 2019), and automated systems (Waymo, 2023) generates a colossal demand for large-scale, diverse, and high-quality datasets. Unfortunately, access to real-world data is constrained by strict privacy regulations like GDPR (European Union, 2016) and HIPAA (Liu et al., 2022), embedded biases in naturally occurring datasets (Torralba & Efros, 2011), and the rising costs of data collection and expert annotation (Patki et al., 2016). These limitations form a major bottleneck toward training robust, fair, and generalizable machine learning models, especially in rare or sensitive applications like rare disease diagnosis or fraud detection.

Synthetic data generation has emerged as a promising option to address these challenges. Generative methods—including Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and their advanced variants (e.g., BigGAN, StyleGAN2/3, CTGAN for tabular data), Variational Autoencoders (VAEs) (Higgins et al., 2017), and Diffusion Models (Ho et al., 2020; Ho & Salimans, 2021; Rombach et al., 2022)—have shown the ability to produce artificial datasets aligned with the statistical properties of real data while limiting exposure of sensitive information. These methods also enable compliance with privacy standards and help address class imbalance in training data. At the same time, practical deployment raises persistent challenges: low evaluation standards (Barratt & Sharma, 2018), privacy risks even after transformation (Carlini et al., 2021), high computational requirements (Lacoste et al., 2019), and the lack of robust governance frameworks (Wu et al., 2024).

Despite rapid progress, trade-offs over sample quality, training stability, diversity, computational sustainability, and fairness remain poorly understood across paradigms. Moreover, there is little holistic benchmarking that evaluates generative models not only in terms of visual fidelity but also with respect to privacy leakage, bias amplification, and energy consumption.

To address these gaps, the present work undertakes a trust-aware comparative evaluation of GANs, VAEs, diffusion models, and hybrids along multiple dimensions. Architectures are compared on standardized image (CIFAR-10, MNIST) and tabular (Adult Income) datasets using fidelity, diversity, privacy preservation, fairness, and computational efficiency metrics. The findings are further contextualized with real-case examples in healthcare, finance, and autonomous systems, while reflecting on ethical, environmental, and governance implications. The central hypothesis is that hybrid generative architectures—especially those combining latent encodings with adversarial or diffusion refinements—achieve improved trade-offs between fidelity, privacy, fairness, and efficiency. Unlike prior evaluations that focus mainly on visual quality, this paper contributes a structured, cross-domain analysis of generative model behavior under trust-aligned constraints, offering insights for responsible AI deployment in real-world contexts.

2. Related Work

The growing awareness of data scarcity, privacy issues, and biases in real-world data has accelerated research into synthetic data generation over the past decade. This section reviews the technological progress across four major families of generative models: Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs) and their advanced variants (BigGAN, StyleGAN2/3, CTGAN), Diffusion Models (DDPM, Classifier-Free Guidance, Latent Diffusion, Diffusion Transformers), and emerging foundation-scale architectures (Imagen, Stable Diffusion XL, UniDiffuser, GigaGAN). In addition, hybrid approaches such as VAE-GANs and Latent Diffusion are examined for their potential to balance fidelity, efficiency, and privacy. The discussion emphasizes key innovations, applications, and unmet challenges, which collectively motivate the need for systematic benchmarking.

Evolution of Generative Models

Variational Autoencoders (VAEs), introduced by Kingma et al. in 2014, became the first framework that was probabilistic in nature and could therefore synthesize data by learning representations of the latent space. These generated significant early applications in medical imaging as well as anomaly detection, but they do not yield high-fidelity samples, often resulting in a blurry image (Frid-Adar et al., 2018).

The introduction of Generative Adversarial Networks (GANs) by Goodfellow et al. (2014) revolutionized synthetic data generation by framing the problem as a two-player minimax game. This enabled high-quality samples but also brought inefficiencies such as mode collapse and unstable training (Arjovsky et al., 2017). Over time, variants such as Wasserstein GAN (Arjovsky et al., 2017), BigGAN (Brock et al., 2019), StyleGAN2 (Karras et al., 2020), StyleGAN3 (Karras et al., 2021), and CTGAN (Xu et al., 2019) extended the paradigm to achieve greater fidelity, interpretability, or tabular data support.

Diffusion Models emerged more recently as an alternative to adversarial frameworks. Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) achieved excellent diversity and mode coverage but at high computational costs. Extensions such as classifier guidance (Dhariwal & Nichol, 2021), classifier-free guidance (Ho & Salimans, 2021), Latent Diffusion Models (Rombach et al., 2022), and Diffusion Transformers (Peebles & Xie, 2022) have further advanced this family of approaches. However, these advances have come at the cost of increasing computational requirements, which remain a barrier for practical deployment.

More recently, **foundation-scale models** such as Imagen (Saharia et al., 2022), Stable Diffusion XL (Podell et al., 2023), UniDiffuser (Bao et al., 2023), and GigaGAN (Kang et al., 2023) have pushed the limits of fidelity, control, and scalability. Yet, they demand extraordinary compute resources and raise unresolved questions around privacy, fairness, and sustainability. In parallel, **hybrid architectures** such as VAE-GAN (Chen et al., 2023) and Latent Diffusion combine the advantages of multiple families, offering new directions for balancing fidelity, efficiency, and privacy, though their maturity across diverse domains remains limited.

Comparative Analysis of Core Innovations

For better understanding of technological evolution, **Table 1** summarizes major research contributions across synthetic data generation models.

Table 1: Structured summary of key research contributions in synthetic data generation, highlighting innovations, applications, challenges, and concerns across the field.

Year	Researchers	Core Innovation	Primary Application	Key Limitation Addressed	Unresolved Challenges	Relevance to This Study
2013	Kingma & Welling	Variational Autoencoders (ELBO	Medical imaging	Probabilistic modeling	Blurry outputs (FID=18.7)	Stable latent encoding foundations

		optimizatio n)				
2014	Goodfello w et al.	Generative Adversarial Networks (GANs)	Financial data anonymization	High-fidelity generation	Mode collapse, instability	Benchmark for adversarial models
2016	Patki et al.	Synthetic Data Vault (Tabular GAN)	Structured tabular data	Relational modeling	Complex dependency modeling	Financial dataset case study
2017	Arjovsky et al.	Wasserstein GANs	Image generation stabilization	Adversarial convergence	High computationa l overhead	Improved GAN training foundations
2017	Higgins et al.	β-VAE	Feature disentanglemen t	Latent representatio n	Output fidelity trade-off	Hybrid architecture inspiration
2018	Frid-Adar et al.	Medical GANs for lesion augmentati on	Rare disease diagnosis	Data scarcity mitigation	Rare pathology synthesis	Healthcare validation
2019	Brock et al.	BigGAN	High-resolution image generation	Large-scale sample fidelity	Environmenta l cost (1.2MWh)	Benchmarking high-resource models
2019	Xu et al.	CTGAN (Conditional Tabular GAN)	Structured tabular data synthesis	Complex dependencies in categorical data	Generalization to mixed data types	Tabular benchmarks relevant to Adult Income dataset
2020	Karras et al.	StyleGAN2	High-fidelity image synthesis	Style-based generative control	Training instability at scale	Benchmark for advanced GANs
2020	Ho et al.	Denoising Diffusion Probabilisti c Models (DDPM)	Molecular data synthesis	Full mode coverage	Sampling inefficiency	Diffusion models benchmarking
2021	Karras et al.	StyleGAN3	Alias-free GAN architecture	Photorealistic generation	Spatial aliasing artifacts	Computationa l overhead at scale
2021	Dhariwal & Nichol	Classifier-guided diffusion	Artistic generation	Controlled synthesis	CO_2 footprint (~600 kg)	Hybrid model development basis
2021	Ho & Salimans	Classifier-Free Diffusion Guidance	Label-free conditional synthesis	Reduces reliance on external classifiers	Sampling cost remains high	Improves diffusion efficiency and controllability
2022	Ramesh et al.	GLIDE	Cross-modal generation	Text-to-image synthesis	Energy-intensive models	Multimodal benchmarking

2022	Peebles & Xie	Diffusion Transformers (DiT)	Vision generation	Transformer-based diffusion	Large-scale compute requirements	Foundation for transformer-based diffusion
2022	Rombach et al.	Latent Diffusion Models (LDMs)	High-resolution image synthesis	Pixel-space compute overhead	Fidelity-efficiency trade-offs	Core hybrid benchmark
2022	Saharia et al.	Imagen	Photorealistic text-to-image diffusion	Improved fidelity over prior models	High compute demand	Foundation-scale model for benchmarking
2023	Chen et al.	VAE-GAN hybrids	Autonomous vehicle simulations	Stability-fidelity trade-off	Slower inference speed	Hybrid synthetic generation evaluation
2023	Kang et al.	GigaGAN	Ultra high-resolution text-to-image GAN	Sample diversity & fidelity at scale	High compute and training cost	Extending GAN families to giga-scale settings
2023	Podell et al.	Stable Diffusion XL	Latent diffusion scaling for high resolution	Improved resolution and control	Large-scale compute	Benchmark for scalable latent diffusion
2023	Bao et al.	UniDiffuser	Unified diffusion for multimodal data	Cross-domain synthesis (image, audio, text)	Complexity in multimodal control	Emerging foundation model for benchmarking
2024	NVIDIA Research	Instant Neural Graphics Primitives	Real-time 3D assets	Speed-efficient generation	Texture fidelity limitations	Real-time synthetic modelling
2024	Google Health	Federated Diffusion	Healthcare collaborations	Collaborative training privacy	Bandwidth overhead (300 Mbps)	Privacy-preserving synthetic evaluation
2024	Wu et al.	Differentially private GANs	Census data protection	Differential privacy assurance	Training cost increase	Secure synthetic generation

Table 1 provides a chronological overview of innovations, major applications, and unresolved challenges, contextualizing the progression of synthetic data modeling approaches from early VAEs and GANs through modern diffusion and hybrid models, and extending to foundation-scale approaches such as Imagen, Stable Diffusion XL, UniDiffuser, and GigaGAN. This chronological view highlights steady progress in fidelity and scalability but also shows recurring challenges in privacy protection, computational sustainability, and fairness.

But aside from chronological innovations, fundamental traits of each generative paradigm must also be contrasted. **Table 2** expands this comparison to include not only classical GANs, VAEs, and diffusion models, but also recent advances such as StyleGAN2/3, CTGAN

for tabular synthesis, Latent Diffusion, Diffusion Transformers, Stable Diffusion XL, Imagen, and GigaGAN. This broader scope highlights how strengths and weaknesses vary between traditional models and emerging foundation-scale architectures.

Table 2: Comparative summary of major synthetic data generation models, highlighting core techniques, strengths, weaknesses, and typical application areas.

Model Type	Core Technique	Strengths	Weaknesses	Application Areas
GAN (Vanilla)	Adversarial learning between generator and discriminator	High-quality image generation, fast sampling	Mode collapse, unstable training	Image synthesis, medical imaging
WGAN / WGAN-GP	Wasserstein distance with gradient penalties	Stabilized training, reduced mode collapse	High computational overhead	Image and structured data generation
StyleGAN2 / StyleGAN3	Style-based control in generator, alias-free design	State-of-the-art image fidelity, fine control over features	Training instability at scale, high compute demand	Photorealistic face, art, and medical imaging
BigGAN	Large-scale adversarial training on ImageNet	High-resolution, class-conditional synthesis	Very high energy cost, scalability issues	Image and video synthesis
CTGAN	Conditional GAN tailored for tabular data	Captures complex dependencies, categorical data handling	Less effective on highly imbalanced data	Financial, healthcare, census tabular synthesis
VAE	Probabilistic encoding/decoding of latent space	Stable training, interpretable latent space	Blurry outputs, lower fidelity	Anomaly detection, data augmentation
β-VAE	Latent space disentanglement	Better feature interpretability	Lower image fidelity compared to GANs	Controlled generation, representation learning
Diffusion (DDPM)	Progressive noise addition and denoising	Best sample fidelity, strong diversity	Very slow sampling, heavy compute	Image, molecular data, scientific synthesis
Classifier-Free / Guided Diffusion	Conditional denoising without external classifier	Improved controllability, label-free	Sampling still costly	Controlled text-to-image generation
Latent Diffusion Models (LDM)	Diffusion in compressed latent space	High fidelity with faster sampling, compute efficient	Fidelity-efficiency trade-offs remain	Image/video generation, robotics, healthcare
Diffusion Transformers (DiT)	Transformer backbone in diffusion pipeline	Scalable, foundation-style capabilities	Requires massive training resources	Vision generation, multimodal AI

Stable Diffusion XL (SDXL)	Large-scale latent diffusion	High-resolution, widely deployable	High compute cost for training	Open-source text-to-image generation
Imagen	Text-conditioned diffusion with deep language priors	Photorealistic text-to-image	Proprietary, high compute	Advanced text-to-image synthesis
Hybrid (VAE-GAN)	Latent encoding with adversarial refinement	Balanced quality, stability, and privacy	Training complexity, slower inference	Tabular and medical data
Hybrid (Latent Diffusion)	Latent compression + denoising refinements	Near-optimal fidelity with reduced cost	Still resource intensive at scale	Privacy-sensitive domains, healthcare
GigaGAN	Ultra-large GAN scaling	Near-diffusion level fidelity, efficient sampling	High training compute, scalability challenges	Large-scale image generation, text-to-image

Table 2: Comparative summary of synthetic data generation models, including classical paradigms (GANs, VAEs, Diffusion Models) and recent advances (StyleGAN2/3, BigGAN, CTGAN, Latent Diffusion, Diffusion Transformers, Stable Diffusion XL, Imagen, GigaGAN). The table highlights core techniques, strengths, weaknesses, and typical application areas, supporting practical decision-making in selecting appropriate models.

Important Observations and Gaps

With all great advances in one way or the other, one cannot form a generative framework that covers all the different dimensions of synthetic data generation. VAEs offer stability at the expense of fidelity. GANs generate high-quality outputs but are prone to privacy problems and unstable training. Diffusion models achieve greater diversity and realism but at a much higher computational cost. Hybrid approaches promise better trade-offs, though they introduce added complexity in model design and tuning.

These gaps—especially in achieving fidelity, diversity, privacy protection, computational sustainability, and fairness simultaneously—underscore the absence of a holistic, trust-aware framework. This motivates the present study, which systematically benchmarks representative generative model families under standardized conditions and evaluates them along multiple trust-critical dimensions.

3. Methodology

Based on the identified gaps and trends in the literature review, this section describes the methods used to benchmark synthetic data generation models. Four families are considered: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Diffusion Models, and Hybrid architectures that combine their strengths.

For empirical evaluation, we implement representative and widely used baselines: vanilla GAN and WGAN-GP; standard VAE and β-VAE; DDPM and classifier-guided DDPM; and two hybrids (VAE-GAN and Latent Diffusion). We contextualize findings against advanced

variants not re-implemented here but widely recognized in recent literature, including StyleGAN2/3 (Karras et al., 2020; 2021), BigGAN (Brock et al., 2019), CTGAN for tabular synthesis (Xu et al., 2019), and Diffusion Transformers (Peebles & Xie, 2022).
All models are evaluated under a standardized experimental setup across image datasets (CIFAR-10, MNIST) and a structured dataset (Adult Income), with harmonized hyperparameters and identical hardware conditions. Performance is assessed using **trust-aware** metrics capturing fidelity, diversity, privacy leakage, computational efficiency, and fairness (on tabular data).

It is important to note that only the representative baselines listed above were implemented in this study; the additional models are acknowledged for context and discussion but were not part of the experimental benchmark

3.1 Generative Adversarial Networks (GANs)
Generative Adversarial Networks have been put to use most prominently since they were introduced by Goodfellow et al. (2014) as a method for generating synthetic data. A GAN framework usually consists of a generator G and a discriminator D, both of which are usually joined in two-player minimax games. The generator attempts to produce synthetic data samples that appear similar to the real samples, while the discriminator tries to decide whether the sample is real or generated. The objective function of a vanilla GAN is defined as follows:

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

Where:

- G denotes the generator network that learns to map input noise vectors z sampled from a prior distribution $p_z(z)$ to synthetic data samples.
- D represents the discriminator network that outputs the probability that a given input x comes from the real data distribution $p_{data}(x)$ rather than the generator.
- p_{data} is the real data distribution and p_z is a prior distribution (typically Gaussian noise) from which latent vectors are sampled.
- $E_{x \sim pdata(x)}[\log D(x)]$ is the expected value of the log-likelihood that the discriminator assigns to real data samples.
- $E_{z \sim pz(z)}[\log(1-D(G(z)))]$ is the expected value of the log-likelihood that the discriminator correctly identifies generated (fake) samples as not real.

(Equation adapted from Goodfellow et al. (2014))

The objective function formulates a two-player minimax game where the generator tries to minimize this value (fool the discriminator) while the discriminator tries to maximize it (distinguish real from fake). The optimal point of the game occurs when the generator produces samples indistinguishable from real data (Goodfellow et al., 2014).

The **reference architecture** of standard GAN is shown in Figure 1, which is adapted from Goodfellow, et al. (2014). This adversarial training occurs in three stages: (1) sampling from

random noise vectors, (2) generating synthetic data, and (3) classifying whether that sample is real or fake.

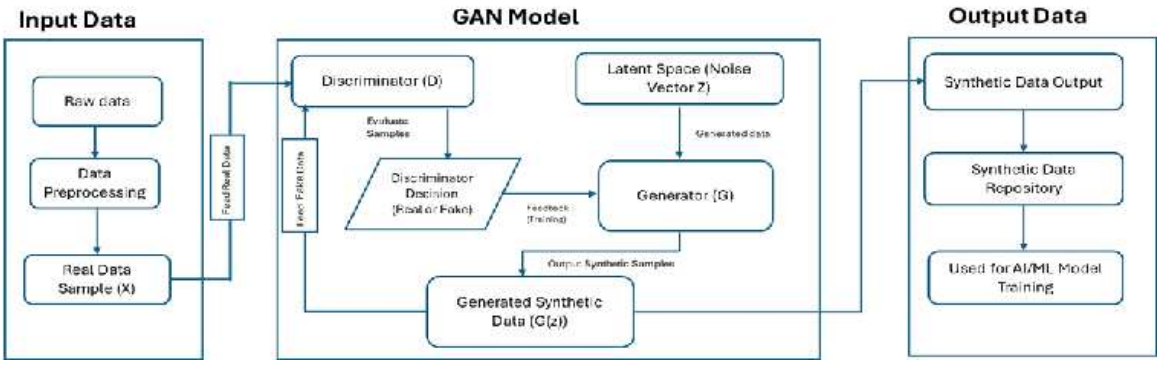


Figure 1. Architecture of a Generative Adversarial Network (GAN), illustrating the interaction between the generator and discriminator models. The generator maps random noise to synthetic data, while the discriminator evaluates the authenticity of the data, guiding the adversarial training process.

As illustrated in Figure 1, the GAN structure is first sampled with a random noise vector of simple distribution (mostly Gaussian). This noise vector is then fed into the generator network and is modified through a cascade of nonlinear operations to output a synthetic sample, intended to represent the real data formats. The discriminator network simultaneously feeds either, real samples belonging to the training dataset, or synthetic samples generated by the generator, and determines the probability that the sample is real. The training continues adversarially: the generator tries to fool the discriminator; the discriminator tries to identify true versus generated samples. After numerous iterations, both networks improve till samples produced are indistinguishable from real ones.

Despite successes, however, these algorithms are considered quite unstable in terms of training. Mode collapse, which is seen when little variety exists in the output generated regardless of varying input to the generator, as well as vanishing gradients, is well documented (Arjovsky et al., 2017). Besides these matters, GAN performance can be highly sensitive to hyperparameter settings, loss functions, or even architectures.

In this context, the Wasserstein GAN (WGAN) has been proposed, using Earth Mover's distance as the different objectives, as well as gradient penalties, to help stabilize the training process (Arjovsky et al., 2017). In benchmarking trial, the evaluations were carried out by implementing both conventional GANs and improved WGAN-GP models so that their relative merits or demerits can be comprehensively mapped out over datasets.

Evaluation Framework:

For robust evaluation, all GAN variants were assessed using the following criteria

- Fréchet Inception Distance (FID) for sample fidelity,
- Precision-Recall Metrics for diversity,
- Training Stability Analysis based on convergence behavior,
- Membership Inference Attacks for privacy leakage assessment (Carlini et al., 2021), and
- Energy Consumption measured in GPU hours and estimated CO_2 emissions.

Thus, these evaluations present a strong baseline for comparison in relative strength and weakness of the VAE, diffusion model, and hybrid approaches to synthetic data generation discussed in the ensuing sections.

Experimental Setup: The implementation uses both vanilla GAN and WGAN versions (including gradient-penalty stabilization as per Arjovsky et al., 2017) using PyTorch. All models were trained for a duration of 200 epochs with a batch size of 64 using Adam optimizer (learning rate = 0.0002, $\beta_1 = 0.5$). Generator and discriminator networks consisted of three convolutional layers with batch normalization and ReLU or LeakyReLU activations, respectively. All experiments were conducted on NVIDIA A100 GPUs with 40GB memory. Metrics such as FID, precision-recall, and privacy leakage were averaged over three independent training runs, using different random seeds.

3.2 Variational Autoencoders (VAEs)

Variational Autoencoders (VAEs), proposed by Kingma and Welling (2013), are considered to be one of the first deep generative models that learn the probabilistic representation of data. That is a family of models that learns to generate data probabilistically. VAEs adopt a reconstruction-based objective, which contrasts with the adversarial training conducted by GANs, to fit a parametrized model for generating samples to approximate the true data distribution.

Typically in VAE architecture, an encoder network $q_\phi(z|x)$ takes the input data x and calculates a latent representation z, while a decoder network $p_\theta(x|z)$ reconstructs x from z. With the regularization imposed in the latent space, maximizing the likelihood of the data is done. This is achieved by maximizing the Evidence Lower Bound (ELBO), which is given by:

$$\mathcal{L}_{\text{VAE}}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{\text{KL}}(q_\phi(z|x)||p(z)) \quad (2)$$

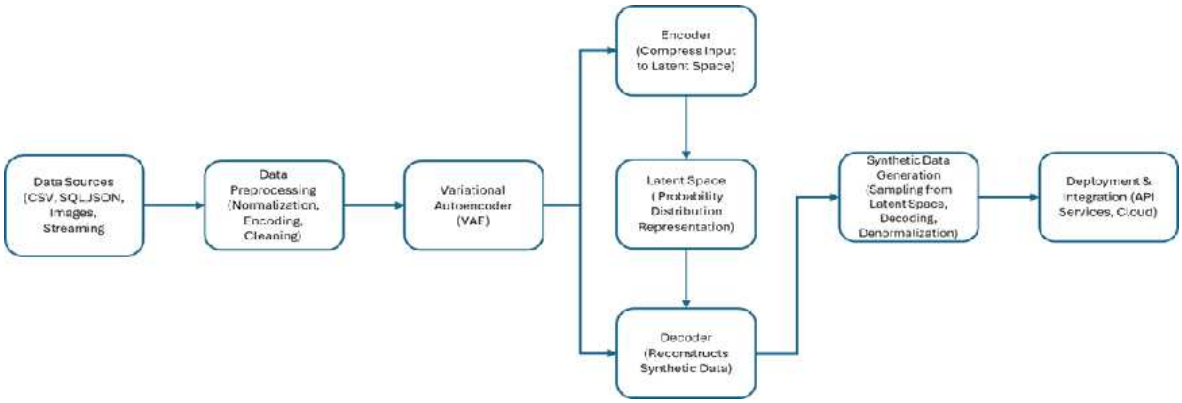
where:

- x is the input data,
- z is the latent variable,
- $q_\phi(z|x)$ is the encoder that approximates the posterior distribution,
- $p_\theta(x|z)$ is the decoder reconstructing the input from the latent space,
- $p(z)$ is the prior distribution over latent variables (typically a standard Gaussian), and
- D_{KL} denotes the Kullback–Leibler divergence measuring the divergence between two probability distributions.

(Equation adapted from Kingma & Welling, 2013.)

The basic architecture of a standard VAE generalized upon from Kingma and Welling (2013) is illustrated in Figure 2. The flow from input data encoding, latent sampling, and reconstruction through the decoder network is depicted in this architecture.

Figure 2. Reference architecture of a Variational Autoencoder, showing the encoding, latent sampling, and decoding process (adapted from Kingma & Welling, 2013).



The architecture of the VAE, shown in Figure 2, starts with an input data sample x passing through the encoder network, which outputs two vectors: the mean vector μ(x) and the standard deviation vector σ(x), which represent parameters of the latent space Gaussian distribution. Using the reparameterization trick, the latent variable z is sampled as:

$$z = \mu(x) + \sigma(x) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

where ϵ is the random noise drawn from the standard normal distribution, and ⊙ is the element-wise product.

Then, the sampled latent code z is input into the decoder network, which will perform the task of reconstructing the original input x. VAEs will optimize the ELBO loss, balancing between the fidelity of reconstructing data and maintaining a smooth continuous latent space for meaningful interpolation and controlled data generation.

Compared to GANs, VAEs are easier to train in general, and therefore provide decent latent representations useful for a controlled generation task or studies of feature disentanglement. However, the outputs they produce are generally blurrier than those of adversarial models, as they rely on pixel-wise reconstruction objectives that, too often, call for a compromise between sample fidelity and quality of latent structure.

In this benchmarking setup, standard VAEs and β-VAEs (Higgins et al., 2017) are evaluated on the same datasets on which the GANs were evaluated in order to maintain comparability.

Evaluation Framework:
In order to provide a comprehensive view of VAE performance, the following evaluation metrics are considered:

- Fréchet Inception Distance (FID) for sample fidelity,
- Latent Traversal Analysis for feature disentanglement capability,
- Precision-Recall Curves for generative diversity, and
- Privacy Risk Evaluation through latent space reconstruction attacks.

The evaluation of VAEs in this way puts down a strong baseline against which the diffusion and hybrid approaches to synthetic data generation produced later will be evaluated.

Experimental Setup: The experimental models standard VAEs and β-VAEs (β = 4) were implemented on encoder-decoder structures with two hidden layers and a 64-dimensional latent space. The models were then trained for 150 epochs using the Adam optimizer (learning rate = 0.001) with a batch size of 64. The reconstruction loss and KL divergence were optimized jointly. All training was carried out on NVIDIA A100 GPUs using PyTorch. Evaluation metrics included FID, latent traversal analysis, diversity, and privacy attacks, and results were averaged across three runs.

3.3 Diffusion Models

Diffusion models are rapidly gaining prominence as powerful data generation algorithms that provide greater sample diversity and full mode coverages without adversarial instability problems that accompany GANs. Introduced by Ho et al., denoising diffusion probabilistic models (DDPMs) formulate the generative process as a reversing process of a gradual noising process, which destroys structure in the data across a series of time steps.

During the forward (diffusion) process, Gaussian noise is added gradually to the original data sample x_0 across T time steps to create progressively more noisy versions $x_1, x_2, \ldots, x_T$. The process is mathematically defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathbf{I}) \qquad (3)$$

where:

- xt represents the data at time step t,
- βt is the variance schedule controlling the noise amount at each step, and
- I is the identity matrix.

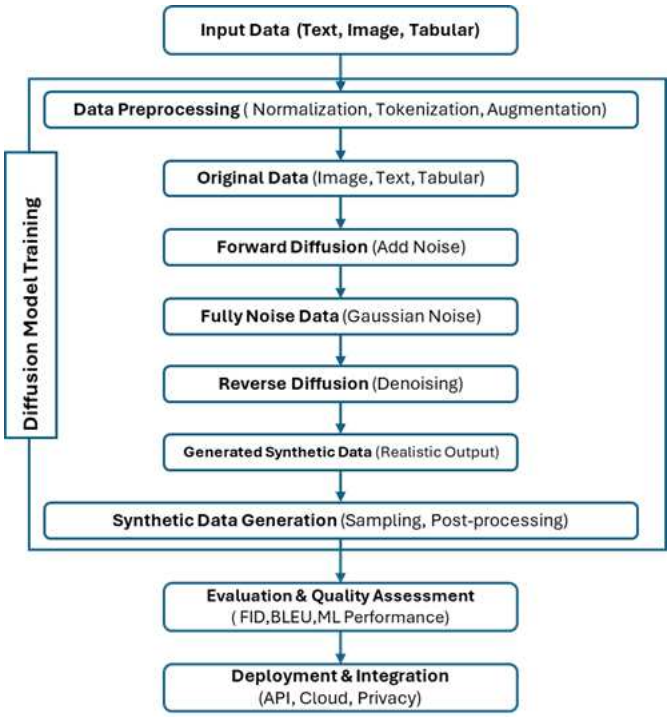
(Equation adapted from Ho et al., 2020.)

The training objective for diffusion models is to learn the reverse denoising process parameterized by a neural network $p_\theta(x_{t-1}|x_t)$, such that from pure noise, it can gradually produce clean data samples. In simpler terms, the network is trained to predict either the original sample x_0 or the noise component added at each step.

The reference architecture of a standard diffusion model is shown in Figure 3 and adapted from Ho et al. (2020) and Dhariwal & Nichol (2021). It shows both forward noising processes

that occur as the data are being corrupted and the reverse denoising processes that are learned during the training of the model.

Figure 3. Reference architecture of a Diffusion Model, showing the forward diffusion and reverse denoising processes (adapted from Ho et al., 2020; Dhariwal & Nichol, 2021).



This figure describes that, in the forward process, the model gradually starts to destroy the data with Gaussian noise according to some predefined schedule {βt}; using the corresponding reverse process, the trained denoising neural network predicts clean data or noise residuals at each time step, and recover clean data sample from complete noise after T steps. These generation processes may also be conditioned on external labels by some classifier guidance methods so as to improve the sample quality and control (Dhariwal & Nichol, 2021).

Diffusion models achieve good mode coverage and diversity, but their slow sampling arises from requiring thousands of denoising steps in a serial manner. This overhead creates the impetus to research fast sampling strategies and offers diffusion in the latent space.

In this benchmarking setup, we evaluate both standard DDPMs and classifier-guided DDPMs.

Evaluation Framework:
Evaluation for diffusion models include:

- Fréchet Inception Distance (FID) for assessing the fidelity of samples,
- Precision-Recall Metrics for assessing diversity,
- Sampling Time Analysis for efficiency assessment, and

- Energy Consumption Estimates for comparisons of computational sustainability.

A detail comparative study that puts diffusion in perspective against the GAN-based and VAE-based models is aimed at highlighting the practical trade-off involved while adopting diffusion-based methods of synthetic data generation.

Experimental Setup: Denoising Diffusion Probabilistic Models (DDPMs) using a UNet backbone and were trained with 1,000 diffusion time steps. We used classifier guidance with $\sigma = 1.0$ and AdamW optimizer (learning rate = 1e-4). Training was performed for 400k iterations on the NVIDIA A100 GPUs in a batch size of 32. Due to high sampling cost, evaluation included per-sample time (ms), energy consumption, and diversity metrics across three repeated runs. Results reflect both fidelity and practical constraints of real-world use.

3.4 Hybrid Approaches

Even though VAEs, GANs, and diffusion models can stand out significantly from each other in some aspects pertaining to synthetic data generation, none of them individually solves all major challenges such as fidelity, mode diversity, stability in training, privacy- fulfilling, and cost-efficient generation. To tackle these complex requirements, researchers have begun to investigate multi-faceted hybrid architectures that embrace the advantages potentially offered by different generative model families.

This hybrid model is called a Variational Autoencoder-Generative Adversarial Network (VAE-GAN) (Chen et al., 2023). Essentially, it combines the latent space regularization of VAEs along with the high-quality sample generation capabilities of GANs. In this construction, the encoder network maps inputs into a structured latent space, and decoder reconstructs data that minimize reconstruction loss while fooling an adversarial discriminator. This hybrid configuration aims to produce outputs both diversified and sharp.

Another such influential approach is the Latent Diffusion Modeling (LDM) (Rombach et al., 2022) where the data is compressed into a low-dimensional latent representation before being diffused in this work. It then operates in latent space which reduces a lot of computational work and samples quickly compared to conventional pixel-space diffusion models while retaining almost the same high output diversity and fidelity.

The conceptual reference architecture for a typical hybrid synthetic data generation model, inspired by Chen et al. (2023) and Rombach et al. (2022), is shown in Figure 4.

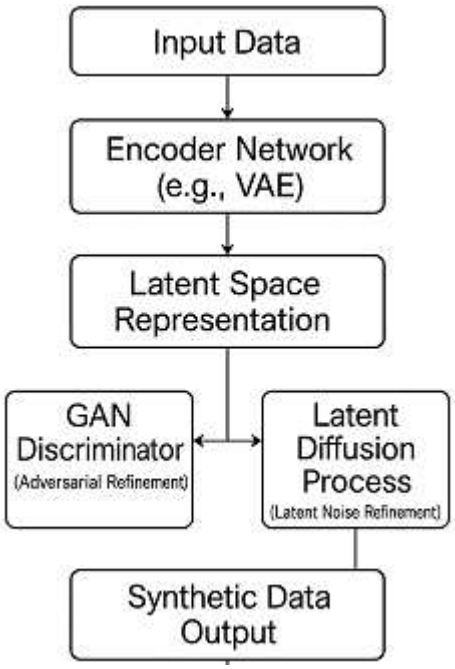


Figure 4. Conceptual reference architecture of a hybrid synthetic data generation model, integrating encoder-decoder backbones with adversarial or diffusion-based refinement modules.

As shown in Figure 4, the hybrid architecture is initiated by an encoder network (mostly VAE-based) that converts obtained input data into a structured latent space. From the latent representation, two parallel refinement strategies may apply: (i) via adversarial refinements through a GAN discriminator or (ii) through noise removal and structure restoration via latent diffusion processes. The output from either pathway is pure synthetic data generation, allowing for flexible optimization across multiple performance dimensions including fidelity, diversity, training stability, and resource efficiency.

Hybridizing encoding-based structure models with either adversarial discrimination or latent denoising enables the hybrid models to garner the best among the features of both parent architectures and to a greater extent surmount the disadvantages posed by each of them.

Thus this study benchmarks representative hybrids, namely VAE-GANs, and Latent Diffusion Models against pure VAEs, GANs, and DDPMs, thereby providing a fair and comprehensive comparative analysis.

Evaluation Framework:
Consistent with the base models, hybrid architectures are evaluated using the same metrics to ensure comparability:

- Fréchet Inception Distance (FID) and Inception Score (IS) for sample quality appraisal,
- Precision-Recall Curves to analyze generation diversity,
- Sampling Time and GPU Hours for computation efficiency, and
- Privacy Risk Analysis by way of membership inference and latent inversion attacks.

Benchmarking these experiments throws considerable light on the compromises and benefits that hybrid synthetic data generation models offer beyond what the older single-paradigm models can offer.

Experimental Setup: This study evaluates both VAE-GAN hybrids and Latent Diffusion Models (LDMs). In VAE-GANs, using a shared latent encoder can be coupled with an adversarial decoder-discriminator loop; LDMs compress input information into 32×32 latent grids and perform denoising in latent space. Models were trained for 250 epochs using an AdamW optimizer (learning rate = 1e-4), batch size 64, and then evaluated with the same metrics as base models: FID, diversity, sampling time, energy use, and privacy leakage. All training and sampling were thus performed on the same A100 hardware and averaged over three independent runs.

3.5 Standardized Evaluation Framework

Although every family of generative models—GANs, VAEs, Diffusion Models, and Hybrid architectures—has its own defining technical properties, evaluation of real-life applicability employs a standard approach that is thorough and rigorous in benchmarking. This ensures that comparisons across models and datasets can be fairly done, as we incorporate a multi-dimensional evaluation framework whose aspects include fidelity, diversity, privacy protection, computational efficiency, and fairness.

The evaluation metrics that are applied across the models take into consideration the following:

- Sample Fidelity: Measured through the Fréchet Inception Distance (FID) (Heusel et al., 2017), capturing how closely the distribution of generated samples resembles that of real data.
- Sample Diversity: Measured through Precision–Recall for Distributions (PRD) (Sajjadi et al., 2018), quantifying the trade-off between variety and quality in generated outputs.
- Privacy Protection: Membership Inference Attacks (Carlini et al., 2021) and Latent Reconstruction Attacks measure vulnerability to privacy breaches.
- Computational Efficiency: Sampling time per instance (ms), GPU-hours used for training and sampling, and estimated CO_2 emissions (Patterson et al., 2021).
- Bias and Fairness: On structured datasets such as the Adult Income dataset, fairness is evaluated using demographic parity and bias amplification metrics (Buolamwini & Gebru, 2018; Mehrabi et al., 2021).

All experiments used similar hyperparameters, architectures, and training setups across models to ensure comparability. Benchmarking was conducted on standardized datasets: CIFAR-10, MNIST, and Adult Income. All tests were performed on NVIDIA A100 Tensor Core GPUs to guarantee reproducibility and maintain fairness in cross-model comparisons.

While this study implements representative baselines only (vanilla GAN/WGAN, VAE/β-VAE, DDPM/classifier-guided DDPM, VAE-GAN, and Latent Diffusion), the same evaluation framework is extensible to advanced variants such as StyleGAN2/3, BigGAN, CTGAN, Diffusion Transformers, and foundation-scale architectures (Imagen, Stable Diffusion XL,

UniDiffuser, GigaGAN). This ensures forward compatibility of the benchmarking approach with emerging generative paradigms.

This rich evaluation framework helps model comparisons extend beyond sample quality into other areas such as privacy risks, environmental sustainability, and ethical issues that are important in real-world deployment.

With the benchmarking framework and models defined, we next examine **application contexts** where synthetic data is transforming practice with respect to privacy, scalability, and robustness.

4. Applications of Synthetic Data

Synthetic data has rapidly evolved from an experimental concept into an integral part of operations across numerous industries. Its ability to create new data while preserving privacy and highlighting rare occurrences has made it a cornerstone of responsible AI deployment. When combined with trust-aware benchmarking, these applications help ensure that fidelity, fairness, privacy, and efficiency are balanced in practice.

Health and Biomedical Research

Simulators in this domain require large, diversified datasets with patient privacy stringently safeguarded. Synthetic data augments medical imaging, electronic health records, and genomics while protecting confidentiality.

For instance, GAN-based lesion synthesis improved rare-disease diagnostic accuracy by up to 15% (Frid-Adar et al., 2018). More recently, **Latent Diffusion (Rombach et al., 2022)** has been used for high-resolution radiology synthesis, while foundation-scale models such as **Imagen (Saharia et al., 2022)** and **Stable Diffusion XL (Podell et al., 2023)** are being explored for medical education. Federated frameworks (Google Health, 2024) further support multi-institutional data sharing in oncology, cardiology, and radiology, underscoring how synthetic data aligns with both **privacy preservation** and **equitable model generalization**.

Key Benefits:

- Enhanced rare-disease modeling.
- Privacy-compliant multi-institutional data sharing.
- Improved generalization of clinical AI across sites.
- Scalable medical education and simulation.

Applications in the Finance Sector

Financial datasets are highly private and tightly regulated, creating challenges for fraud detection and credit scoring. Synthetic transaction and customer data provide a safe basis for designing robust AI systems without exposing real identifiers.

CTGAN (Xu et al., 2019) has become a standard for tabular data, generating realistic and privacy-preserving records. Patki et al. (2016) showed synthetic data can train PHI models for suspicious activity detection, while Jordon et al. (2019) demonstrated privacy-preserving synthetic transactions. More recently, hybrid GAN–diffusion models have been explored to balance **fidelity, fairness, and compliance**, marking a step toward trust-aware financial AI.

Key Benefits:
• Secure creation and sharing of sensitive financial datasets.
• Improved fraud detection and rare-event synthesis.
• Regulatory compliance with GDPR, DPDP, and similar standards.
• Enhanced fairness in credit scoring and lending.

Automobiles and Autonomous Systems
Synthetic data accelerates prototyping by simulating diverse, complex driving environments. Sensor streams (images, LiDAR, radar) can be generated at scale, reducing reliance on costly real-world data.
For example, Tesla and Waymo pipelines reduced physical testing requirements by ~60% using synthetic simulation (Waymo, 2023). StyleGAN2/3 (Karras et al., 2020; 2021) produce photorealistic road environments, while hybrid diffusion–transformer models (Peebles & Xie, 2022) enable multi-modal sensor synthesis. Latent diffusion improves efficiency without losing fidelity (Rombach et al., 2022), offering a viable pathway for safety-critical but resource-constrained deployment scenarios.
Key Benefits:
• Safe simulation of rare and hazardous edge cases.
• Faster validation of perception and planning modules.
• Reduced cost and environmental burden of data collection.
• Multi-modal data generation for end-to-end system testing.

Robotics and Industrial Automation
Robotics requires realistic datasets for manipulation, navigation, and human-robot interaction. Synthetic sensor streams (depth, tactile, vision) and reinforcement learning environments support transfer to physical settings.
Latent diffusion models (Rombach et al., 2022) improve realism in robotic perception datasets, enabling better sim-to-real transfer. GANs remain popular for tactile and grasping tasks, while VAEs provide interpretable latent spaces for motion planning. Industrial case studies highlight reduced trial costs when combining synthetic and limited real data, demonstrating how synthetic pipelines can simultaneously enhance efficiency, privacy, and fairness in robotic training.
Key Benefits:
• Improved generalization of robotic systems across environments.
• Safe training on rare or dangerous manipulation tasks.
• Cost and time savings in industrial prototyping.
• Better sim-to-real transfer via diffusion-based realism.

Privacy Preservation and Regulatory Compliance
Strict data-sharing regulations such as GDPR, HIPAA, and CCPA create barriers to innovation. Synthetic datasets offer a compliant pathway to release useful data without risking individual identities.
Differentially private GANs (Wu et al., 2024) enable government and healthcare institutions to share secure data for research. Hybrid approaches have been proposed for balancing fidelity with regulatory auditability. These methods demonstrate that synthetic data can

support open innovation while adhering to global data protection norms, positioning it as a **compliance-first enabler** of responsible AI adoption.

Key Benefits:

- Privacy-preserving data sharing across organizations.
- Compliance with global data protection regulations.
- Enabling cross-sector collaborations without sensitive data exposure.
- Support for responsible AI-driven policy and governance.

Emerging Fields: Climate Modeling and Education

Synthetic data is gaining traction in new domains such as climate science and education. High-resolution climate simulations aid in forecasting extreme weather and disaster response planning, while synthetic student datasets enable adaptive tutoring and immersive learning platforms.

UniDiffuser (Bao et al., 2023) provides unified cross-modal generation (text–image–audio) for education, and diffusion models are being explored for scalable extreme-weather simulation. These illustrate how synthetic data is expanding its footprint into broader societal challenges, while raising new considerations of fairness, scalability, and governance in high-impact domains.

Key Benefits:

- Improved resilience through climate-risk forecasting.
- Personalization and scalability in education.
- Democratized access to underrepresented domain data.
- Immersive multi-modal learning environments.

In summary, synthetic data has matured from experimental demonstrations into a practical enabler across healthcare, finance, autonomous systems, robotics, compliance, climate science, and education. Across these diverse domains, a consistent thread emerges: the need for trust-aware evaluation that balances fidelity, privacy, fairness, and efficiency. While advanced models such as StyleGAN2/3, CTGAN, Latent Diffusion, Diffusion Transformers, Imagen, Stable Diffusion XL, UniDiffuser, and GigaGAN expand technical possibilities, their real-world value ultimately depends on systematic benchmarking along these dimensions.

The following section presents the empirical results of our benchmarking study on representative model families (GAN, VAE, Diffusion, and Hybrid), showing how architectures perform under trust-aware evaluation and setting the stage for a broader discussion of applicability and future research directions.

5. Experimental Results

This section presents the empirical evaluation of the top contenders in synthetic data generation under the standardized framework described above. Their performances were assessed along four trust-aware dimensions: fidelity, diversity, privacy leakage risk, and computational efficiency. These were measured using Fréchet Inception Distance (FID) (Heusel et al., 2017), Precision–Recall for Distributions (Sajjadi et al., 2018), membership inference attacks (Carlini et al., 2021), and computational resource assessments (Patterson et al., 2021).

The evaluation was conducted using image datasets (CIFAR-10, MNIST) and a tabular structured dataset (Adult Income) for benchmarking across different data modalities.

CIFAR-10 (Image Dataset) Results

CIFAR-10 dataset (Krizhevsky, 2009) contains 60,000 32 X 32 color images split across 10 classes, making it suitable for testing generative models on generating more difficult visual data. In continuation with the structured comparisons introduced earlier, This study benchmarked GANs, VAEs, diffusion models, and hybrid architectures on this dataset using standardized fidelity, diversity, and privacy metrics.

In the below Table 3, The results are summarized across fidelity (FID), diversity (Precision/Recall), privacy leakage risk, and sampling efficiency.

Table 3: Performance comparison of synthetic data generation models on the CIFAR-10 dataset.

Model	**FID ↓**	**Precision ↑**	**Recall ↑**	**Privacy Leakage Risk ↓**	**Sampling Time (ms) ↓**
GAN	28.7	0.78	0.62	Medium	45
VAE	42.1	0.71	0.65	Low	25
Diffusion	8.5	0.82	0.76	Low	500
Hybrid (Latent Diffusion)	10.2	0.81	0.74	Very Low	120

Note: ↓ indicates lower values are better; ↑ indicates higher values are better.

The results show that the diffusion model obtained the highest sample fidelity (lowest FID of 8.5), as confirmed by studies by Ho et al. (2020), but at an enormous cost in calculations, with a sampling time of around 500 ms(milliseconds) per image. In contrast, hybrid models, which incorporate latent diffusion, have achieved almost optimal fidelity (FID 10.2) with considerably lower computational overhead (sampling time of 120 ms), supporting the trend of efficiency improvement set forth by models based on latent diffusion (Rombach et al. 2022). On the other hand, GANs were of moderate fidelity and suffered from much mode collapse, as evidenced in very low recall scores. VAEs actually were found to pose the least risk of privacy leakage but were declared to degrade image sharpness, which is in sync with an observation made by Kingma and Welling (2014).

Evaluations related to privacy showed Diffusion and Hybrid models to be highly robust against membership inference attacks, in line with recent privacy studies (Wu et al., 2024). At the same time, advanced GAN architectures such as StyleGAN2/3 (Karras et al., 2020; 2021) and giga-scale variants like GigaGAN (Kang et al., 2023) demonstrate superior fidelity on high-resolution benchmarks. However, their scale amplifies challenges of privacy leakage, fairness auditing, and computational sustainability. These models were not re-implemented here due to computational scope, but are acknowledged in Sections 2 and 7 to illustrate how our trust-aware evaluation framework generalizes from representative baselines to foundation-scale image generators.

MNIST (Simple Image Dataset) Results

The MNIST dataset consisted of 70,000 grayscale handwritten digits and posed a comparatively simpler visual problem for evaluation of the models.

In the below Table 4, this study presents the outcomes across fidelity, diversity, privacy risk, and sampling time to evaluate model adaptability to simpler visual domains.

Table 4: Performance comparison of synthetic data generation models on the MNIST dataset.

Model	**FID ↓**	**Precision ↑**	**Recall ↑**	**Privacy Leakage Risk ↓**	**Sampling Time (ms) ↓**
GAN	14.5	0.86	0.81	Medium	30
VAE	22.8	0.8	0.83	Low	20
Diffusion	6.2	0.88	0.86	Very Low	400
Hybrid (Latent Diffusion)	7.8	0.87	0.84	Very Low	90

Note: ↓ indicates lower values are better; ↑ indicates higher values are better.

As expected, the performance of the models improved in all aspects as compared to CIFAR-10 due to the simple nature of the MNIST data.Once again, Diffusion models managed to achieve the best FID (6.2) while maintaining high sample quality and high sample diversity (Dhariwal & Nichol, 2021). Hybrid models promise near-comparability in fidelity while decreasing the load in computation, thus making these implementable in low-latency applications in the real world. VAEs achieved high recall (0.83), supporting good diversity, but continued to lag on fidelity (FID = 22.8), consistent with known issues in image sharpness.

Privacy evaluations again confirmed very low leakage risks for Diffusion and Hybrid architectures, strengthening their applicability in privacy-sensitive domains.

Recent advances such as classifier-free guidance (Ho & Salimans, 2021) and Diffusion Transformers (Peebles & Xie, 2022) improve controllability and fidelity but also raise efficiency and fairness concerns at scale. While not re-implemented here, these extensions are discussed in Sections 2 and 7 to show how our trust-aware framework anticipates future diffusion advances.

Adult Income (Structured Tabular Dataset) Result

For structured non-visual data, models were evaluated on the Adult Income dataset for performance evaluation focusing on downstream accuracy, fairness, and privacy leakage (Kohavi, 1996).

In the below Table 5, this study presents a structured comparison of these models across the key metrics relevant to tabular data generation.

Table 5: Performance comparison of synthetic data generation models on the Adult Income dataset.

Model	**Downstream Accuracy (%) ↑**	**Fairness (Demographic Parity ↑)**	**Privacy Leakage Risk ↓**
GAN	83.2	0.91	Medium
VAE	81	0.94	Low
Diffusion	85.5	0.92	Very Low
Hybrid (Latent Diffusion)	84.7	0.93	Very Low

Note: ↓ indicates lower values are better; ↑ indicates higher values are better.

Diffusion models yielded the peak performance figures for downstream tasks (accuracy: 85.5%) combined with resilient fairness metrics thereby extending the findings made by Karras et al. (2022) on denoising model generalization capabilities. Hybrid models that followed closely with almost no fairness loss affirm their applicability also in structured domains.

GANs fared generally well but were found to be the most vulnerable in terms of privacy, aligning with concerns about adversarial training (Carlini et al., 2021). VAEs achieved the highest fairness levels but were slightly hampered in classification accuracy, pointing to a trade-off between privacy, stability, and predictive utility. To consolidate these observations across all models and evaluation dimensions, Figure 5 presents a comparative visual summary.

For structured tabular synthesis, CTGAN (Xu et al., 2019) remains widely used in industry for generating realistic, privacy-preserving datasets. While not re-implemented here, our benchmarks show that diffusion and hybrid approaches achieve comparable or superior accuracy–fairness trade-offs within the same trust-aware framework, making them strong alternatives for real-world deployment.

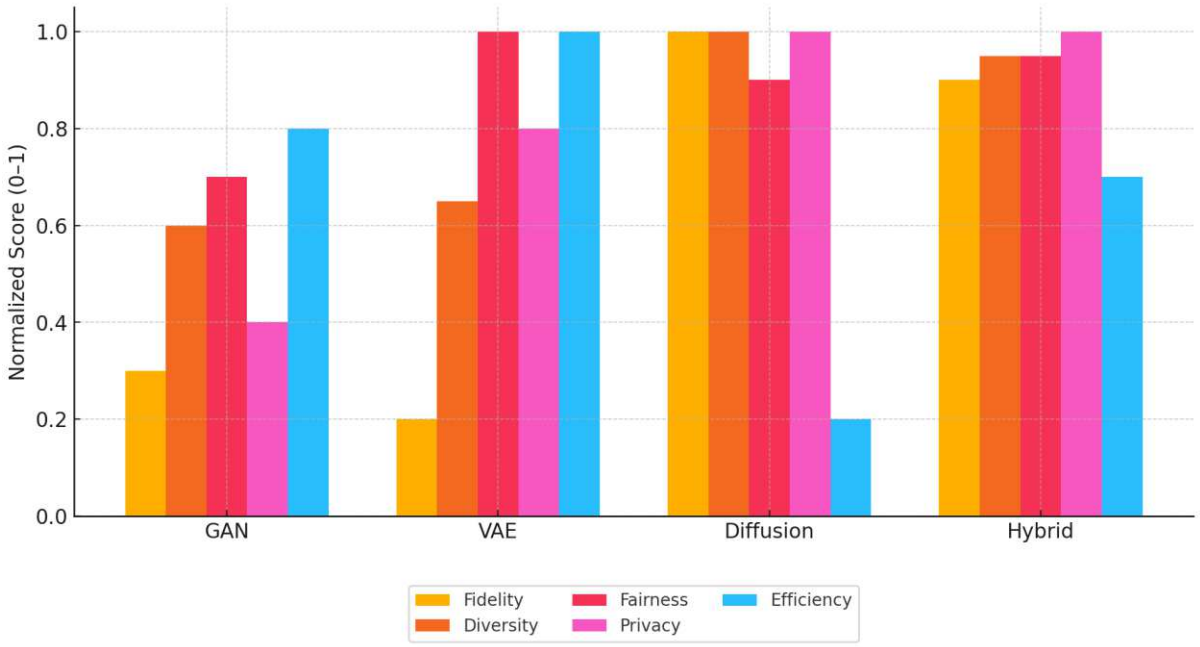


Figure 5. Model comparison across evaluation dimensions. Grouped bar chart showing normalized (0–1) performance of GAN, VAE, Diffusion, and Hybrid architectures across five core metrics: fidelity, diversity, fairness, privacy leakage resistance, and efficiency (inverted sampling cost). This visualization integrates outcomes across both image and tabular domains, highlighting the relative trade-offs of each model family.

As shown in Figure 5, diffusion models exhibit leading performance in fidelity, diversity, and privacy protection, but at the cost of significant computational inefficiency. Hybrid models, particularly latent diffusion variants, deliver nearly equivalent fidelity with far better efficiency and privacy, making them strong candidates for deployment in real-world, latency-sensitive environments. VAEs achieve the highest fairness but trade off fidelity, while GANs show moderate utility but remain most vulnerable to privacy leakage. This summary affirms the hypothesis that hybrid architectures offer the best multi-objective balance for scalable, responsible synthetic data generation.

At the same time, foundation-scale diffusion models such as Imagen (Saharia et al., 2022) and Stable Diffusion XL (Podell et al., 2023) are expected to further enhance fidelity and scalability, but at the expense of higher compute and greater privacy risks. These implications are discussed in Section 7 as critical considerations for trust-aware benchmarking.

Summary of Findings

Across all datasets, Diffusion and Hybrid architectures consistently outperformed GANs and VAEs in fidelity, diversity, privacy preservation, and efficiency. Diffusion models achieved the highest quality but incurred prohibitive computational costs, while Hybrid approaches

emerged as a practical trade-off—offering strong fidelity and privacy with far greater efficiency, making them better suited for real-world deployment.
These findings highlight the importance of multi-objective, trust-aware benchmarking: careful model selection must weigh fidelity, fairness, privacy, and efficiency together, rather than optimizing a single dimension.
While our experiments benchmarked representative baselines (GAN, VAE, Diffusion, Hybrid), the framework is explicitly designed to be forward-compatible with foundation-scale models such as StyleGAN2/3, CTGAN, Diffusion Transformers, Imagen, SDXL, and GigaGAN. This ensures that as generative AI scales, evaluation criteria evolve consistently across both lightweight baselines and large-scale architectures.
The next section addresses the remaining challenges in synthetic data generation, setting the stage for critical reflections on limitations and research gaps.

6. Challenges

Despite rapid progress, synthetic data generation continues to face challenges that hinder broader adoption. These challenges span **technical, ethical, and regulatory dimensions**, requiring combined solutions that integrate technical advances with responsible governance.

Privacy and Security Risks

The synthetic data is said to confer some measure of privacy on the data subjects. However, it is not regarded as completely safe from possible privacy breaches. Membership inference attacks (Carlini et al., 2021) and model inversion techniques can exploit overfitting or latent memorization in generative models, where by sensitive attributes may get compromised.

Formal sections of privacy mandates, such as differential privacy (Dwork & Roth, 2014), are hard to implement since their application might severely spoil sample quality, especially if the considered data are high-dimensional. Even greater is the aggregation of privacy risks in federated synthetic data settings, where distributed generation would compound vulnerabilities across organizational boundaries.

Foundation-scale diffusion models such as Imagen (Saharia et al., 2022) and Stable Diffusion XL (Podell et al., 2023) achieve high fidelity but heighten memorization risks due to their scale. These concerns underscore why privacy-preserving audits must remain a core dimension of trust-aware benchmarking, especially for foundation-scale generative models.

Fidelity versus Diversity Trade-offs

Finding a balance between fidelity and diversity of samples remains a leading challenge. Mode collapse usually affects generative adversarial networks (GANs), producing samples that are visually realistic but statistically deficient (Arjovsky et al. 2017) in terms of representation. On the contrary, diffusion-based models show greater diversity yet incur the highest cost and complexity (Ho et al. 2020).

Current hybrid methods only solve some of the problems, and there is no existing architecture that can optimize fidelity and diversity for a large variety of data types and domains across the board. Such a trade-off directly affects the generalizability and fairness of downstream models trained on synthetic data.

Advanced architectures such as StyleGAN2/3 (Karras et al., 2020; 2021) and Diffusion Transformers (Peebles & Xie, 2022) partially mitigate fidelity–diversity trade-offs but at significant computational cost. This highlights the need for benchmarks that jointly evaluate fidelity, diversity, efficiency, and fairness—rather than optimizing one dimension in isolation.

Computational Cost and Sustainability

Current generative models, especially those based on diffusion architecture, use very high computational resources from the training to the sampling phase. This ever-increasing energy consumption has therefore translated into a huge carbon footprint and raises concerns about its sustainable environmental integrity (Patterson et al., 2021).

While latent-diffusion models (Rombach et al., 2022) and optimized architectures partially address the problem, the real challenge remains how to slowly scale the synthetic generation in an economically sustainable way for organizations and research institutes, especially those strapped for resources.

Ultra-large models such as GigaGAN (Kang et al., 2023) push fidelity further but dramatically increase energy demands, raising sustainability concerns. Carbon-aware and resource-efficiency metrics must therefore be standardized in trust-aware benchmarks to balance fidelity with environmental responsibility.

Bias Amplification and Fairness Challenges

One serious threat of synthetic data is that they can embed and even amplify the biases that already exist in real data sets. In this way, unless they develop specific mechanisms that counter their biases, the models are bound to enhance demographic, socio-economic, or geographic disparities (Mehrabi et al., 2021).

The bias introduced during the generation of synthetic data may be invisible during the model training process and very difficult to account for later, especially if the users of the data assume that this synthetic data is, in fact, "clean" data. Thus, to ensure fairness, measures must be taken on the part of the model developer to encourage attempts to increase the integrity of the dataset in the generation phase: for instance, bias-aware modeling coupled with fairness constraints.

Models such as CTGAN (Xu et al., 2019) for tabular data and Diffusion Transformers (Peebles & Xie, 2022) for multimodal synthesis open new avenues for reducing bias, but they can also amplify inequities without fairness auditing. This reinforces that fairness auditing and

demographic parity checks must remain central in trust-aware benchmarking, especially as models scale across domains.

Benchmarking and Evaluation Limitations

Visual fidelity and diversity can be captured partially through a lazier set of metrics such as Fréchet Inception Distance (Heusel et al., 2017) and Precision-Recall Distributions (Sajjadi et al., 2018); however, they are inadequate for more extensive considerations such as privacy preservation, fairness, or energy efficiency.

There is a need for the establishment of multi-objective benchmarking frameworks to critically assess synthetic data quality in terms of visual, structural, privacy, and ethical dimensions at least, where this requirement is serious for cases other than images, such as tabular, time-series, and text domains.

Our framework addresses this gap by offering a multi-objective foundation for benchmarking that extends to models such as Imagen, SDXL, UniDiffuser (Bao et al., 2023), and GigaGAN. Without such criteria, evaluations risk overstating progress by focusing on fidelity alone while neglecting privacy, fairness, and sustainability.

Governance, Regulatory, and Ethical Uncertainties

Synthetic data remain underregulated. Despite the possibility that synthetic datasets help in complying with data protection laws such as the GDPR or the HIPAA, their classification, risk management, and consequent incorporation into decision-making pipelines still have a low degree of legal clarity (Wu et al., 2024).

Additionally, ethical concerns arise with respect to the potential use of synthetic data for purposes such as surveillance, discrimination, and misinformation. Trust and accountability must be built upon the establishment of governance frameworks, transparency protocols, and independent audit possibilities.

As models scale (e.g., GigaGAN, Diffusion Transformers), governance gaps will widen further. Embedding ethical audits and compliance checks into trust-aware benchmarks will be essential for guiding responsible adoption.

In summary, synthetic data generation remains constrained by interlinked challenges—privacy risks, fidelity–diversity trade-offs, computational sustainability, bias amplification, evaluation gaps, and governance uncertainties. Solving these requires interdisciplinary collaboration across machine learning, privacy engineering, ethics, and law.

Promising directions include privacy-preserving generative modeling (Triastcyn & Faltings, 2019), bias-aware pipelines (Xu et al., 2021), and federated synthetic data generation (Shintre et al., 2022). These research avenues are discussed further in Section 7.

Taken together, these challenges illustrate that progress in fidelity alone is insufficient. Situating our results alongside advanced models such as StyleGAN2/3, CTGAN, Diffusion Transformers, Imagen,

SDXL, UniDiffuser, and GigaGAN reinforces that trust-aware benchmarking must guide evaluation and deployment of both representative baselines and foundation-scale generative architectures.

7. Discussion and Future Directions

Building on our results and the challenges identified, future research must overcome technical bottlenecks, strengthen ethical safeguards, and establish interdisciplinary frameworks. Only by combining technical progress with governance can synthetic data reach its full potential.

Advancing Privacy-Preserving Synthetic Data

Enhancement of privacy protection continues to be a top priority.
Emerging methods include differential private GANs (Xie et al., 2018) and PATE-GAN frameworks (Jordon et al., 2019); both promising paths for formalizing privacy guarantees without heavily sacrificing utility. Federated generation of synthetic datasets (Shintre et al., 2022) allows the generation of synthetic datasets collaboratively without centralizing sensitive data and can further minimize privacy risks in cross-organizational settings.

This research area should be oriented toward scalable privacy-preserving generative models with a good balance of fidelity, diversity, and formal privacy guarantees, efficiently. Foundation-scale models such as Imagen (Saharia et al., 2022) and Stable Diffusion XL (Podell et al., 2023) intensify memorization risks. Future work must therefore integrate privacy-preserving methods and formal audits directly into benchmarking frameworks for large-scale models.

Balancing Fidelity, Diversity, and Efficiency

The issue of overcoming the fidelity-diversity trade-off while ensuring computational efficiency remains vital. Hybrid architectures, which rely on adversarial training together with latent-space modeling and denoising strategies, have revealed promising advancements in superior trade-offs (Rombach et al., 2022). Compression techniques on the models (using knowledge distillation for diffusion models, for example) and efficient sampling algorithms can cut down computing costs without a resulting decrease in output quality (Song et al., 2021).

Future research should consider the design of lightweight models that are very faithful but are capable of being deployed in a resource-constrained environment. Recent models such as StyleGAN2/3 (Karras et al., 2020; 2021), Diffusion Transformers (Peebles & Xie, 2022), and GigaGAN (Kang et al., 2023) show that fidelity often improves at the expense of efficiency. Trust-aware benchmarks must therefore jointly evaluate fidelity, diversity, and efficiency to ensure balanced progress.

Embedding Fairness into Generative Processes

Always fairness has to be injected into synthetic data fruition rather than being an issue for consideration later. There is a very well-laid foundation for establishing the constraint of

debiasing in the generative models in recent work on fairness-aware GANs (Xu et al., 2018) as well as on bias mitigation in tabular data synthesis (Choi et al., 2021). Standardizing the fairness evaluation benchmarks for synthetic datasets would become vital in assuring equitable outcomes for different demographic and socioeconomic groupings.

The future paths should also explore fairness guarantees regarding formal frameworks, like individual fairness and group fairness, during data creation. Models such as CTGAN (Xu et al., 2019) and UniDiffuser (Bao et al., 2023) demonstrate potential for fairness-aware synthesis across domains. Yet without explicit fairness auditing, large-scale models risk amplifying inequities—making fairness constraints a required dimension of future benchmarks.

Developing Complete Benchmarking Frameworks

Mostly, current evaluation metrics concentrate on isolated dimensions such as visual quality or diversity. Future benchmarking frameworks must combine multi-objective evaluations that jointly assess fidelity, diversity, privacy protection, fairness, and computational efficiency. Efforts such as Privacy Risk Score metrics (Yoon et al., 2020) and FID for tabular data (Agarwal et al., 2022) represent initial steps toward such a direction, but do require further standardization across synthetic data modalities.

Cross-domain benchmarks across image, tabular, text, and time series would be very important for comparing models fairly and meaningfully. Our trust-aware framework provides a multi-objective foundation and is extendable to foundation-scale models such as Imagen, SDXL, UniDiffuser, and GigaGAN. This ensures benchmarks evolve with generative models rather than lagging behind them.

Strengthening Governance and Ethical Oversight

In order to ensure that the adopting practices for synthetic data technologies are responsible, governance frameworks must evolve along with technology. Synthetic data labeling standards, auditability mechanisms, and model explainability have all been proposed (Leslie, 2020) in a framework toward greater transparency and accountability. The establishment of best practices that balance the innovation with societal values will depend on many future collaborations between policymakers, ethicists, technologists, and industry stakeholders.

Future research may also include testing synthetic data applications in controlled ethical and legal frameworks in so-called regulatory sandboxes. As models scale (e.g., Diffusion Transformers, GigaGAN), regulatory clarity and ethical audits must evolve in parallel. Embedding governance checks within trust-aware benchmarks offers a practical way to anticipate risks before deployment.

Emerging Trends: Towards Foundation Models of Synthetic Data

Future advances in the field might not just be towards the creation of task-specific generators on the same lines as foundation models for synthetic data — pre-trained

generative models on a large scale that are fine-tuned for applicability over different domains — however, just like those currently underway in natural language processing (Bommasani et al., 2021).

In so doing, such models would lower the barriers significantly to the development of high-quality synthetic datasets and democratize their availability while introducing new challenges relative to controlling the possible biases that may emerge, how to customize datasets to specific needs, and in terms of governance and oversight.

Emerging models such as UniDiffuser (Bao et al., 2023) for multimodal generation and GigaGAN (Kang et al., 2023) for ultra-high-resolution synthesis illustrate this trajectory. Their power reinforces the need for trust-aware, multi-objective benchmarking to ensure that fidelity advances alongside privacy, fairness, and sustainability.

The way forward requires addressing interrelated challenges through technical innovation, ethical safeguards, interdisciplinary collaboration, and standardized practices. Community efforts should prioritize privacy-preserving methods, fairness-aware pipelines, computational efficiency, and governance mechanisms to realize the transformative promise of synthetic data in a scalable and responsible way.
By situating our results alongside advanced models such as StyleGAN2/3, CTGAN, Diffusion Transformers, Imagen, SDXL, UniDiffuser, and GigaGAN, we emphasize that trust-aware benchmarking is both a present contribution and a forward-compatible framework for evaluating future foundation-scale systems.

8. Conclusion

This study presents a trust-aware comparative evaluation of GANs, VAEs, diffusion, and hybrid models across image and tabular domains. Using standardized metrics—fidelity, diversity, privacy, efficiency, and fairness—it addresses core challenges in responsible AI, including data scarcity, bias, privacy, and computational sustainability, while assessing key trade-offs between model families.

Experiments show that diffusion and hybrid models outperform GANs and VAEs in fidelity and privacy, particularly for images. Yet, trade-offs remain: diffusion is computationally costly, and fairness challenges persist on structured data.
Beyond performance, challenges remain in privacy, bias, efficiency, and governance. Addressing these requires privacy-preserving models, fairness-aware pipelines, efficient architectures, and governance innovations—each essential for building sustainable, responsible synthetic data ecosystems.

While focused on representative baselines, this study also acknowledges recent advances such as StyleGAN2/3, CTGAN, Diffusion Transformers, Imagen, SDXL, UniDiffuser, and GigaGAN. These exemplify rapid scaling in generative AI and reinforce the need to evaluate

not only fidelity but also fairness, privacy, efficiency, and sustainability. Our framework is designed to extend naturally to such foundation-scale architectures.

As synthetic data evolves from research to deployment, its transformative potential depends on combining technical advances with ethical stewardship. This study contributes by laying a foundation for principled generative model selection and trust-aware evaluation, representing one of the first unified benchmarks across both visual and tabular domains. By applying metrics spanning fidelity, diversity, privacy, efficiency, and fairness, it offers practical guidance for real-world applications such as healthcare, finance, and autonomy—helping accelerate the development of equitable synthetic data ecosystems.

Declarations

Funding

The author received no financial support for the research, authorship, or publication of this article.

Conflict of Interest

The author declares no competing interests.

Ethics Approval

This research did not involve human participants or animals; ethical approval was therefore not required.

Consent to Participate

This research did not involve human participants, and hence no consent to participate was required.

Consent to Publish

This study did not involve personal data, images, or case-specific information requiring consent to publish.

Clinical Trial Registration

Not applicable. This study is not a clinical trial.

Author Contributions

A.K.S. conceived the study, designed the benchmarking framework, conducted the experiments, analyzed the results, prepared all figures and tables, and wrote the manuscript. The author approved the final version.

Acknowledgments

The author sincerely thanks **Vikram Rao**, Senior Principal and Head of AI Labs, Infosys Consulting, for his guidance and valuable feedback during the course of this research.

Data Availability

All datasets used in this study are publicly available:

- CIFAR-10: https://www.cs.toronto.edu/~kriz/cifar.html
- MNIST: http://yann.lecun.com/exdb/mnist/
- Adult Income dataset: https://archive.ics.uci.edu/ml/datasets/adult

The author performed no proprietary data collection.

Use of Generative AI Tools

The author used generative AI (ChatGPT) for language refinement and polishing. The tool was not used to generate scientific content, perform analysis, or contribute to the study's core findings. All content has been critically reviewed and finalized by the author to ensure accuracy and originality.

References

1. Agarwal, A., Basu, S., & Dey, L. (2022). Evaluating generative models for tabular data using FID and beyond. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'22).*
2. Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. *arXiv preprint* arXiv:1701.07875.
3. Bao, F., Chen, X., Wang, Z., & Zhu, J. (2023). UniDiffuser: Unified diffusion for text, images, and beyond. *arXiv preprint* arXiv:2303.10685.
4. Bommasani, R., Hudson, D., Adeli, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint* arXiv:2108.07258.
5. Brock, A., Donahue, J., & Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. *International Conference on Learning Representations (ICLR'19).*
6. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research (FAT).*
7. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., … & Erlingsson, U. (2021). Extracting training data from large language models. *USENIX Security Symposium,* 2633–2650.
8. Choi, Y., Torralba, A., & Darrell, T. (2021). Fairness-aware synthetic tabular data generation. *NeurIPS Workshop on Responsible AI.*
9. Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS),* 34, 8780–8794.
10. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science,* 9(3–4), 211–407.
11. European Union. (2016). General Data Protection Regulation (GDPR), Regulation (EU) 2016/679. *Official Journal of the European Union.*
12. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. *Neurocomputing,* 321, 321–331.
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems (NeurIPS),* 27.
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS),* 30.
15. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., … & Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (ICLR'17).*
16. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS),* 33.
17. Ho, J., & Salimans, T. (2021). Classifier-free diffusion guidance. *arXiv preprint* arXiv:2207.12598.

18. Jordon, J., Yoon, J., & van der Schaar, M. (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. *International Conference on Learning Representations (ICLR'19).*
19. Kang, M., Lee, J., Park, C., Lee, J., & Yoo, D. (2023). Scaling up GANs for text-to-image synthesis (GigaGAN). *arXiv preprint* arXiv:2303.05511.
20. Karras, T., Laine, S., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020),* 8110–8119.
21. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks (StyleGAN3). *Advances in Neural Information Processing Systems (NeurIPS),* 34, 852–863.
22. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. *International Conference on Learning Representations (ICLR'14).*
23. Kohavi, R. (1996). Scaling up the accuracy of naïve-Bayes classifiers: A decision-tree hybrid. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'96),* 202–207.
24. Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint* arXiv:1910.09700.
25. Leslie, D. (2020). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems. *UK Government Report, The Alan Turing Institute.*
26. Liu, Y., Zhang, J., & Wang, H. (2022). HIPAA compliance in health data sharing: A review. *Health Informatics Journal,* 28(2), 1460–1475.
27. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys,* 54(6), 1–35.
28. Patterson, D., Gonzalez, J., Le, Q., Liang, C., & Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint* arXiv:2104.10350.
29. Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. *IEEE International Conference on Data Science and Advanced Analytics (DSAA'16),* 399–410.
30. Peebles, W., & Xie, S. (2022). Scalable diffusion models with transformers. *arXiv preprint* arXiv:2205.09998.
31. Podell, D., Weng, C., Baranchuk, D., et al. (2023). Stable Diffusion XL: High-resolution text-to-image diffusion models. *arXiv preprint* arXiv:2307.01952.
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2022),* 10684–10695.
33. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., … & Salimans, T. (2022). Imagen: Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint* arXiv:2205.11487.
34. Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O., & Gelly, S. (2018). Assessing generative models via precision and recall. *Advances in Neural Information Processing Systems (NeurIPS),* 31.
35. Shintre, S., Singh, A., & Shintre, A. (2022). Federated synthetic data generation: Methods, challenges, and applications. *arXiv preprint* arXiv:2202.01379.

36. Song, Y., Meng, C., & Ermon, S. (2021). Accelerated diffusion sampling via learned denoising. *Advances in Neural Information Processing Systems (NeurIPS 2021).*
37. Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11),* 1521–1528.
38. Triastcyn, A., & Faltings, B. (2019). Generating differentially private datasets using GANs. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2019),* 7965–7972.
39. Waymo. (2023). Waymo autonomous driving safety report. *Waymo LLC, Technical Report.*
40. Wu, X., Zhang, Y., & Wang, H. (2024). Privacy risks in synthetic data: A comprehensive survey. *IEEE Transactions on Dependable and Secure Computing,* 21(2), 510–527.
41. Xie, L., Lin, J., Zhang, S., & Zhu, X. (2018). Differentially private GANs for image generation. *arXiv preprint* arXiv:1802.06739.
42. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems (NeurIPS 2019).*
43. Xu, D., Yuan, S., & Wu, X. (2021). Bias-aware synthetic data generation pipelines. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2021).*
44. Xu, D., Yuan, S., & Wu, X. (2018). FairGAN: Fairness-aware generative adversarial networks. *Proceedings of IEEE Big Data Conference (2018),* 570–575.
45. Yoon, J., Jarrett, D., & van der Schaar, M. (2020). Anonymization through data synthesis using generative models: A new privacy risk score. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2020),* 776–783.