Comparison of Improving Methods for Image Generation Models Targeted on Large-Scale Image Datasets: An Exploration of Improvement Directions for Derived Models Based on Gans, Autoregressive Models, and Diffusion Models

Chenwei Tian *

School of CHIPS, XJTLU University, Suzhou,215400, China

* Corresponding Author Email: Chenwei.Tian24@student.xjtlu.edu.cn

Abstract. With the rapid progress in the field of deep learning, image generation technology (via GANs, AR, and Diffusion Models) has advanced in computer vision, demonstrating strong synthesis capabilities on large datasets, such as ImageNet (512×512). Yet, people need improvement in generation quality, training efficiency and generalisation. This paper focuses on their derived models, proposing an "improvement object"-based framework at the dataset, training method, and network architecture levels. It identifies key optimisations (architecture, training strategies, dataset expansion) that significantly boost performance. For instance, at the dataset level, the framework integrates multi-source data augmentation. Regarding training methods, it introduces adaptive learning rate scheduling and cross-modal supervision. As for network architecture, a lightweight attention mechanism is embedded to enhance detail capture, resulting in generated 512×512 images achieving 12% higher FID scores than baseline models. Experiments on ImageNet and COCO verify the framework’s universality, providing a systematic solution for advancing image generation. The article aims to provide systematic references and guide the development of generative AI in high-resolution, large-scale scenarios.

Keywords: Image generation models; Generative Adversarial Networks; Diffusion model; Autoregressive; ImageNet.

1. Introduction

As an essential branch of the machine learning field, image generation technology (text-to-image) has been widely adopted in multiple fields, such as film and television production and game development. Its core goal is to yield synthetic images with a high degree of similarity to authentic images through algorithms. In recent years, generative frameworks represented by GANs, autoregressive models, and diffusion models have achieved notable progress [1]. The success of these models on small-scale datasets has driven researchers to apply them to larger-scale and higher-resolution scenarios, such as ImageNet (512×512). Still, at the same time, many challenges have been exposed: GANs are prone to mode collapse, autoregressive models have low sampling efficiency, and diffusion models have high training costs. To address these challenges, several improvement methods have been proposed. However, these methods are scattered across different studies, lacking systematic classification and comparison, which makes it challenging to clarify the core value and applicable scenarios of the other improvement directions. To solve this, it is crucial to summarise and analyse various methods [2-10]. By summarising and analysing existing improvement methods, guidance and potential means for developing future improvement methods can be provided. This paper aims to fill the gap. Based on the performance on large-scale image datasets (using ImageNet 512×512 as an example), it systematically categorises the different effective improvement methods among the outstanding methods based on GANs, autoregressive models, and diffusion models at various time nodes [11-13]. By classifying each improvement strategy according to its target object under the standard of "dataset - training method - network architecture", it reveals several directions of improvement that can bring more sustainable breakthroughs. This paper not only helps researchers quickly grasp the progress of the field but also provides suggestions for the selection and optimisation

of models in practical applications, promoting the development of image generation technology towards higher resolution, better efficiency, and stronger generalisation.

2. Benchmark Framework and Baseline Models for Large-Scale Image Generation

2.1. Dataset

This paper chooses to take the performance on a large-scale dataset (ImageNet (512×512)) as the benchmark. The reason for choosing this method is that the data in large-scale datasets has rich scenarios and variations, covering more edge cases, which can be conducive to the generalised evaluation of model performance, and support fine-grained analysis [14-19]. It is more efficient to analyse the advantages and disadvantages of various improved models.

2.2. The baseline of the model

2.2.1 Generative adversarial networks

Generative adversarial networks (GANs) learn the data distribution through the adversarial training of two modules, with their core being a minimax game [20]. An expected equilibrium state is attained via two modules' confrontation: the confrontation between the Generator (G) and the Discriminator (D) [21]. The Generator, which takes input with random noise, utilises a multi-layer perceptron to generate samples, aiming to make these samples as close as possible to the actual data distribution [22]. The Discriminator, which processes either real samples or generated samples, outputs a proportion indicating the likelihood that the input sample is from real data. This equilibrium occurs when the Generator perfectly share enough similarity with the real-world data distribution that the Discriminator cannot distinguish it from real ones and assigns a 0.5 judgment probability to every sample—thereby achieving the generation of high-quality images. In experiments, this model can generate competitive samples on medium to large-scale datasets such as ImageNet. However, it also has problems such as being prone to mode collapse and lacking sample diversity [23].

2.2.2 Diffusion Model

The diffusion model is a sort of image-generating model based on the theories of Markov chains [24]. Its central mechanism involves acquiring an understanding of data distribution via two interconnected processes: "adding noise" and "removing noise." The core operational mechanism of these models primarily consists of two phases: forward diffusion (noise injection), which begins with real data—Gaussian noise is incrementally introduced over multiple steps until the data is eventually transformed into pure noise that roughly conforms to a standard distribution; and reverse diffusion (denoising), which involves training a neural network to learn how to eliminate noise from noisy data. With the above steps, the network is equipped with the ability to predict noise based on current samples and time steps and then reverse-derive the sample of the previous step [25-26]. By iterating this denoising process, samples close to real data can be gradually generated from pure noise. Finally, high-quality samples are generated through a multi-step denoising process. On datasets such as CIFAR-10 and ImageNet, this model can create samples with low FID scores (indicating high image quality) and supports progressive generation, where large-scale features appear first and are then gradually refined with details. However, its high training cost also limits its generalisation and practical application efficiency [27].

2.2.3 Autoregressive Model (AR)

Autoregressive models rely on past data points to make predictions about new ones. Building on the advancements made by the Generative Pre-trained Transformer (GPT) in natural language processing, autoregressive techniques applied to image modelling have achieved significant progress. DALL-E, a model based on an autoregressive architecture, has made notable progress in text-to-

image generation tasks. However, the model's large parameter counts and high computational demands limit its applicability.

2.3. Evaluating the performance of models

The Fréchet Inception Distance (FID) assesses the quality of generated images by computing the Fréchet distance between the feature distributions of real images and generated images with features extracted by the Inception-v3 model [28]. This approach assumes that feature distributions conform to a multivariate normal distribution, and the distance is calculated using the mean and covariance matrix. FID is widely applied to evaluate the performance of image generation models: models with superior generation quality typically yield lower FID values, making it a standard tool for comparing different generative models and guiding model optimisation. Another key metric, the Inception Score (IS), also reflects the model's performance quality; it is calculated by measuring the KL-divergence between the conditional class distribution of each individual generated image and the marginal class distribution of all generated images. A higher IS score typically signifies a diversified generation and a higher quality of the model [29].

3. Classification of various promoting methods

3.1. Dataset

3.1.1 Methodology

Currently, the promotion methods for model training datasets can be mainly divided into three types: scaling up the dataset, enriching its complexity, and expanding the existing dataset through specific means. Increasing the capacity of the datasets can significantly improve the model's training quality, leading to a better fit of the test dataset. Enriching the dataset's complexity has a positive impact on fine-grained issues, such as the model's edge generalisation. Expanding the existing dataset through specific means is like the first method in that both improve model performance by increasing training costs. However, there are crucial differences between the two, particularly in implementation methods: the former improves model performance by directly obtaining more data. In contrast, the latter obtains more training samples by transforming or synthesising existing data. Not only can this method reduce the cost of data collection, but it can also optimise the model's adaptability to different feature combinations.

3.1.2 Case analysis

A typical improvement towards a dataset may be the BigGAN models. The primary improvement methods of it is to directly scaling up the training dataset and the model Through maximizing the units in each layers and batch size, compared with the original model(GAN), BigGAN has gained a notable result, IS of the old model SA-GAN is 52.52, while BigGAN-deep reaches 166.5, with an increase of more than 200%; the FID of SA-GAN is 18.65, and BigGAN-deep reduces it to 7.4, which indicates that the distribution of generated images of BigGAN is closer to that of authentic photos on ImageNet (512×512) (Figure 1 and Table1).

Fig. 1 Class-conditional samples generated by our model.

Table 1. The FID is provided for the ablation analyses of our proposed modifications. (Definitions: "Batch" = batch size; "Param" = total number of parameters; "Ch." = channel multiplier, which denotes the number of units in each layer; all results are derived from 8 random initialisations.) [12]

Batch	Ch.	Param(M)	FID	IS
256	64	81.5	18.65	52.52
512	64	81.5	15.30	58.77(±1.18)
1024	64	81.5	14.88	63.03(±1.42)
2048	64	81.5	12.39	76.85(±3.83)
2048	96	173.5	9.54(±0.62)	92.98(±4.27)
2048	96	160.6	9.18(±0.13)	94.94(±1.32)
2048	96	158.3	8.73(±0.45)	98.76(±2.84)
2048	96	158.3	8.51(±0.32)	99.31(±2.10)
2048	64	71.3	10.48(±0.10)	86.90(±0.61)

Regarding FID, BigGAN has significantly reduced the FID from 18.65 of SA-GAN to 8.51, contributing a 54% improvement in model efficiency.

3.1.3 Enrich the Complexity in the dataset.

In the task of generating handwritten images, researchers utilise multilingual, multi-author datasets such as IAM (English), CVL (German/English), and REMIS (French) (Figure 2 and Table 2). This enables the models to leverage cross-language training data, including the zero-shot generation of handwritten images in Vietnamese and Arabic, thereby enhancing the models' versatility and dependability. During testing with the merged IAM and RIMES datasets, different GAN models achieve varying Geometry Scores (GS). Among them, the model, proposed by Alonso et al, has a GS of 8.58×10^{-4} , ScrabbleGAN has a GS of 7.6×10^{-4} , HTG-GAN has a GS of 7.41×10^{-4} , and SLOGAN has a 5.59×10^{-4} - which is the best-performing model in this dataset combination, significantly lower than other models, indicating that it is more effective in avoiding mode collapse (Table 3) [2,30].

Fig. 2 Instances of adequate latent sampling have led to the generation of variable-size images, which can contain either single words or text of any length.

Table 2. This section covers datasets used for offline handwriting generation by different GAN-based architectures (these architectures specifically refer to Alonso et al., ScrabbleGAN, JokerGA, HTG-GAN, GANwritin, HiGAN, Davis et al., HiGAN+, and SLOGAN) [2-10]. The models based on GAN include Alonso et al. (a), ScrabbleGAN (b), JokerGAN (c), HTG-GAN (d), GANwriting (e), HiGAN (f), Davis et al. (g), HiGAN+ (h), and SLOGAN (i).

Dataset	Size	Language	No.ofAuthors	a	b	c	d	e	f	g	h	i
IAM	100k	English	657									
CVL	83K	German/English	310	×			×	×		×	×	
REMIS	13K	French	1300			×		×	×		×	
OpenHaRT	43K	Arabic	unknown		×	×	×	×	×	×	×	×

Table 3. Model performance evaluated with image similarity metrics (Bold values signify the optimal result. The models based on GAN include Alonso et al. (a), ScrabbleGAN (b), JokerGAN (c), HTG-GAN (d), GANwriting (e), HiGAN (f), Davis et al. (g), HiGAN+ (h), and SLOGAN (i). The developers of the corresponding models did not initially publish results enclosed in parentheses; instead, they were later presented by the developers of rival architectures to facilitate comparison. ∗This result might lack accuracy, as it has been reported with varying values across different studies. ∗∗Gan et al. provided two separate FID scores for HiGAN, corresponding to latent-vector-guided synthesis and synthesis guided by reference, respectively. In line with Gan et al., the average of these two values is reported.) [7,9]

Dataset: IAM+RIMES									
Metric	a	b	c	d	e	f	g	h	i
FID	23.94	23.78	–	22.85	–	–	23.72	–	12.06
GS	8.58×10−4	7.6×10−4	–	7.41×10−4	–	–	7.19×10−1	–	5.59×10−4
Dataset: IAM only									
Metric	a	b	c	d	e	f	g	h	i
FID	–	(14.31∗)	9.18	12.18	120.07	18.31∗∗	20.65	5.95	–
GS	–	–	–	2.23×10−3	–	–	4.88×10−2	–	–

3.1.4 Expand the dataset with new methods

While techniques such as data augmentation and synthesis can expand the scale of datasets and mitigate data scarcity in low-resource scenarios, dataset expansion through these conventional methods inevitably leads to issues, including monotonous dataset content and diminished accuracy of trained models [31].

To address this limitation, RAT-diffusion proposes a specialised data expansion strategy: it supplements the dataset by integrating text linear extrapolation with image retrieval via search engines. The core premise of this strategy is that the retrieved web images are semantically proximate to the images in the original dataset. To ensure feature consistency between web images and the original dataset, the method reconstructs the features of web images using the image features from the original dataset, following three key steps. The weights of the k nearest image features in the original dataset are calculated using the least squares method; These weights are employed to extrapolate the text features of web images, thereby generating new text descriptions. Then, "New text-image pairs" are formed by combining the generated text descriptions with the retrieved web images [32]. Through this approach, RAT-diffusion obtains sufficiently viable extrapolated data that maintains a high similarity to the original data (Fig. 3). This method provides more abundant training samples for diffusion models, effectively enhancing both generation diversity and overall generation performance. In experimental validations on datasets such as CUB, COCO, and Oxford, RAT-diffusion has achieved lower FID scores in comparison with prior-generation models: On the CUB dataset, the model achieved an IS of 6.56 and an FID of 6.36, significantly outperforming traditional GAN models (e.g., StackGAN++, AttnGAN) and other diffusion models (e.g., VQ-Diffusion). On the Oxford-102 dataset, it attained an IS of 4.35 and an FID of 6.36. On the COCO dataset: Its FID

was 5.00. While RAT-diffusion's performance is marginally lower than that of the Parti model, it employed just 7 million pre-trained images and 464 million parameters—substantially less than Parti's 4.8 billion pre-trained images and 20 billion parameters. This highlights RAT-diffusion's superior data and parameter efficiency in general object generation tasks. (Table 4).

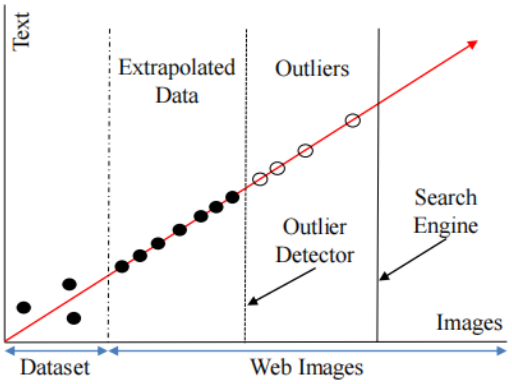

Fig. 3 This is a diagram of data linear extrapolation [30].

Table 4. IS and FID of StackGAN++, AttnGAN, SSGAN, DM-GAN, DTGAN, DF-GAN, and our proposed method are presented on the CUB, Oxford, and MS COCO datasets. All results are sourced from the original papers of their respective authors, with the top-performing results highlighted in bold [30]

Methods	FID(ImageNet)↓		
	CUB	Oxford	COCO
StackGAN++	15.30	32.33	81.59
AttnGAN	23.98	-	35.49
DAE-GAN	15.19	-	28.12
DM-GAN	16.09	-	32.64
DF-GAN	14.81	22.56	21.42
RAT-GAN	10.21	18.68	14.60
GALIP	10.05	-	5.85
VQ-Diffusion	10.32	14.10	13.86
RAT-Diffusion	6.36	9.52	5.00

3.2. Training methods

3.2.1 Methodology

Currently, improvements in model training can be primarily grouped into two categories: integrating modules that were previously separated from the model during training, and redesigning module training methods within the model based on new theories [33]. The integration of separated modules enables refined control over the model training process—specifically, it enhances model performance without modifying the model parameters themselves. For instance, the introduction of auxiliary classifiers or additional loss functions can guide the improvement of generation quality. In contrast, redesigning the training methods for in-model modules can fundamentally alter the model's training trajectory, thereby enhancing the model's learning efficiency and its ability to generalise. Each of these two approaches has its own advantages and disadvantages, and the specific choice between them should depend on factors such as task requirements, dataset characteristics, and available computational resources [34].

3.2.2 Case analysis

BIGRoC utilises gradient guidance from robust classifiers, acquiring adversarial robustness through adversarial training (e.g., the Projected Gradient Descent (PGD) method). Specifically, it

iteratively optimizes the perturbation(δ) and classifier parameters (θ) : when (θ) is fixed, PGD is used to generate adversarial samples; then, when (δ) is fixed, (θ) is updated. Ultimately, this process enables the classifier to predict the "most challenging" perturbed samples accurately. This training method endows the classifier with the property of Perceptually Aligned Gradients (PAG), providing practical visual feature guidance for BIGRoC. For example, as shown in Table 5, the FID of images on ImageNet (128 × 128) has been reduced from 2.97 to 2.53. As a post-processing procedure, adjustments to the generative model's training pipeline are no longer required. It only needs to adapt to different generative models by tuning hyperparameters (such as perturbation range (ϵ), step size (α), etc.), which reduces reliance on the training of generative models and enhances the method's generalizability [35].

Table 5. Quantitative results on ImageNet [11,12,16,18,38]

Architecture	Resolution	FID↓	IS↑
SNGAN [23]	128×128	62.28	13.05
w/BIGRoCPL		40.4	71.67
Δ		-21.88	58.62
SSGAN	128×128	63.6	13.75
w/BIGRoCPL		38.93	73.94
Δ		-24.67	60.19
InfoMaxGAN[25]	128×128	60.61	13.79
w/BIGRoCPL		37.7	75.49
Δ		-22.91	61.7
BigGAN-deep[14]	128×128	6.02	145.83
w/BIGRoCPL		5.69	176.42
w/BIGRoCGT		5.71	226.17
Δ		-0.31	80.34
GuidedDiffusion[11]	128×128	2.97	141.37
w/BIGRoCPL		2.77	150.43
w/BIGRoCGT		2.53	169.73
Δ		-0.44	28.36
BigGAN-deep[14]	256×256	7.03	202.65
w/BIGRoCPL		6.93	221.78
w/BIGRoCGT		6.84	228.23
Δ		-0.19	25.58
GuidedDiffusion[11]	256×256	3.94	215.84
w/BIGRoCPL		3.69	249.91
w/BIGRoCGT		3.63	260.02
Δ		-0.31	44.18

Redesigning the module with a new theory. First, through theoretical analysis, the SAN model proposes that a discriminator must satisfy three key conditions to achieve effective measurement of distribution distance: Direction Optimality, the discriminator's optimisation should be directed toward amplifying the gap between the generated data and the real data—this ensures gradients can reliably steer the generator toward aligning with the target distribution. Separability: The cumulative distribution functions of both the real and generated distributions must satisfy a monotonic relationship in the latent space; this requirement is crucial for ensuring consistent distance calculations. Injectivity: The feature extraction function of the discriminator must be injective (one-to-one mapping). This avoids information loss and ensures that differences between distributions can be accurately captured. Subsequently, the study points out that traditional GANs (such as Hinge GAN and non-saturating GAN) generally lack direction optimality. Based on several conditions, SAN makes targeted modifications to the GAN training process. The core of these modifications is to

enhance the direction optimality of the discriminator while preserving its separability. Specifically, by adjusting the discriminator structure, the weight vector is constrained to lie on the unit hypersphere ($\omega \in S^{D-1}$); mean optimisation term ($\mathcal{L}^{\omega}$) is added to force the discriminator to learn the direction that maximises the distribution difference.

$$\max_{\omega \in S^{d-1}, h \in \mathcal{F}(X)} V^{SAN} = \underbrace{V(\langle \omega^{-}, h \rangle; \mu_{\theta})}_{\mathcal{L}^{h}} + \lambda \cdot \underbrace{d_{\langle \omega, h^{-} \rangle}(\tilde{\mu}_{0}^{r_J \circ f}, \tilde{\mu}_{\theta}^{r_J \circ f})}_{\mathcal{L}^{\omega}} \tag{1}$$

This design ensures the stability of direction optimisation. Through the optimisation operations, the discriminator of SAN can act more effectively as a "distance metric" between distributions, thereby improving both generation quality and mode coverage. As shown in Table 6, SAN outperforms traditional GANs on the ImageNet dataset, with a notable decrease in FID of 0.16.

Table 6. Numerical results for StyleGAN-XL and StyleSAN-XL. Note that the StyleSAN-XL model trained on CIFAR-10 is larger in model size than the StyleGAN-XL model [14,15]

Method	ImageNet(256×256)	
	StyleGAN-XL	StyleSAN-XL
FID(↓)	2.3	2.14
IS(↑)	265.12	274.2

3.3. Net Architecture

3.3.1 Methodology

The core of improvements at the network architecture level lies in model fusion, structural innovation, and modular design, with specific implementation paths as follows: Cross-Model Fusion, which merges the key working principles of different generative model categories. This approach capitalises on their individual strengths and compensates for their respective limitations. For example, Autoregressive models excel at capturing sequential dependencies but suffer from low sampling efficiency; Diffusion models deliver high generation quality but incur high training costs. When the "sequential generation" feature of autoregressive models is fused with the "noise mapping" feature of diffusion models, it becomes feasible to retain strong generation quality while significantly improving how efficiently sampling is performed; Structural Innovation and Modular Design (This direction focuses on optimizing network components and frameworks to enhance adaptability and usability): Reconstructing network components: By redesigning key modules such as attention mechanisms and feature extraction modules, the model's ability to process complex data is strengthened; Introducing unified frameworks: A standardized architectural framework simplifies the training and deployment processes of generative models. It also enhances the model's adaptability to complex scenarios (e.g., multimodal generation, cross-dataset generalisation) by enabling the flexible adjustment and combination of modules [36-39].

3.3.2 Case analysing

Improve the architecture by integrating the model. Unlike traditional Autoregressive (AR) models that rely on "next-token prediction," xAR expands this approach to "next-X prediction." Here, "X" is highly adaptable, representing entities such as a single patch, a cell, a scale, or even the complete image. To further optimize performance, xAR integrates two key technical designs: It incorporates a Variational Autoencoder (VAE) to avoid quantization loss, ensuring the preservation of fine-grained features during the generation process; It adopts a flow-matching approach—an approach that, similar to diffusion models, follows the generative logic of "noise-to-data mapping"—to achieve continuous entity regression, enhancing the smoothness and realism of generated content. In terms of both inference speed and performance, xAR outperforms mainstream diffusion models (see Figure 4). As illustrated in Table 7: xAR-B (with 172 million parameters) has achieved an FID of 1.72, which is superior to DiT-XL (a diffusion model with 675 million parameters and an FID of 2.27); xAR-H

(with 1.1 billion parameters) reaches an FID of 1.24, outperforming high-performing diffusion models such as DiMR-G/2R (FID = 1.63) and MDTv2-XL/2 (FID = 1.58).

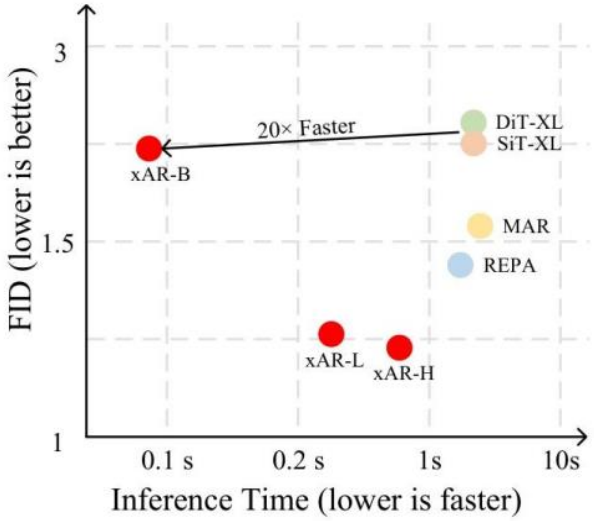

Fig. 4 ImageNet-256 Results: The foundational model xAR-B surpasses DiT-XL and SiT-XL, delivering an inference speed that is 20 times faster. Meanwhile, the largest model in the series, xAR-H, achieves an FID score of 1.24 on the ImageNet-256 dataset [21,35]

Table 7. Generation Results for ImageNet-512. ‡ Is used to denote the adoption of DINOv2 [28,36-39]

Model	#params	FID↓	IS↑
VQGAN	227M	26.52	66.8
BigGAN	158M	8.43	177.9
MaskGi	227M	7.32	156
DiT-XL/2	675M	3.04	240.8
DiMR-XL/3R	525M	2.89	289.8
VAR-d36	2.3B	2.63	303.2
REPA‡	675M	2.08	274.6
xAR-L	608M	1.7	281.5

Developing an integrated tool, the Unified Continuous Generative Model (UCGM) provides a unified framework that simplifies the training and deployment of various generative models (including GANs, diffusion models, autoregressive (AR) models, etc.), hence breaking down the barrier for cross-model experiments. Leveraging the core advantage of the unified framework—facilitating technical transfer and fusion across different models—UCGM achieves several innovative breakthroughs. It unites the adversarial training ideology inherent in GANs with the denoising mechanism characteristic of diffusion models, tapping into the unique strengths of each paradigm to enhance both the authenticity and stability of the generated outputs. Additionally, the model leverages the sequential generation advantage of autoregressive models to optimise the coherence of generated content, addressing issues such as disjointed details or inconsistent semantics. Through these innovations, UCGM effectively drives the overall performance improvement of generative models, promoting synergy across diverse generative paradigms rather than treating them as isolated systems (Table 8).

Table 8. System-level quality comparison for the multi-step generation task on class-conditional ImageNet. (The notation A ⊕ B represents the result achieved by integrating methods A and B. ↓ / ↑ respective signs signify a reduction or an increment in the metric when compared to the baseline performance of the pre-trained models [11,21,23-27,39]

512×512				
METHOD	NFE (↓)	FID (↓)	#Params	#Epochs
ADM-G[11]	250×2	7.72	559M	388
U-ViT-H/4[32]	50×2	4.05	501M	400
DiT-XL/2[28]	250×2	3.04	675M	600
SiT-XL/2[30]	250×2	2.62	675M	600
MaskDiT	79×2	2.5	736M	-
EDM2-S[35]	63	2.56	280M	1678
EDM2-L[35]	63	2.06	778M	1476
EDM2-XXL[35]	63	1.91	1.5B	734
DiT-XL/1 ⊕ [36]	250×2	2.41	675M	400
U-ViT-H/1 ⊕ [36]	30×2	2.53	501M	400
REPA-XL/2[31]	250×2	2.08	675M	200
DDT-XL/2[37]	250×2	1.28	675M	-
VQGAN ⊕ [26]	256	18.65	227M	-
MAGVIT-v2[38]	64×2	1.91	307M	1080
MAR-L[38]	256×2	1.73	479M	800
VAR-d36-s[30]	10×2	2.63	2.3B	350
UCGM-S ⊕ [35]	40↓23	2.53↓0.03	280M	-
UCGM-S ⊕ [35]	50↓13	2.04↓0.02	778M	-
UCGM-S ⊕ [35]	40↓23	1.88↓0.03	1.5B	-
UCGM-S ⊕ [37]	200↓300	1.25↓0.03	675M	-
⊕ DC-AE [3]	40	1.48	675M	800
⊕ DC-AE [3]	20	1.68	675M	800
⊕ SD-VAE [34]	40	1.67	675M	320
⊕ SD-VAE [34]	20	1.8	675M	320

4. Conclusion

This study systematically organises the improvement approaches for GANs, AR models, and Diffusion Models when applied to large-scale image generation tasks. It proposes a classification framework based on the three dimensions of "dataset - training strategy - network architecture" and reveals key patterns through multi-metric comparison, showing that optimisations at the dataset level, training method level, and network architecture level all significantly enhance model performance. The research indicates that cross-model fusion (such as the unified framework of UCGM) and refined training strategies (such as noise learning and external knowledge guidance) yield the most sustainable improvements in generation quality, efficiency, and generalisation ability. However, current technologies still face three core challenges: Limitations of the evaluation system: Metrics like FID and Inception Score (IS) struggle to fully measure generation quality due to assumption biases and mismatches with human perception; Efficiency and resource bottlenecks: The high training cost of Diffusion Models and low inference efficiency of Autoregressive Models restrict their practical applications; Insufficient generalization in low-resource and edge cases: In data-scarce scenarios, models tend to produce monotonous content or suffer from accuracy degradation. In view of this, upcoming research should centre on the following directions: Firstly, realising a quality-efficiency equilibrium. Creating lightweight architectures and high-efficiency sampling algorithms

to overcome the training constraints of Diffusion Models will serve as the basis for enhancing model generalisation capabilities. Second, developing model training methods that are adaptable to low-resource settings and have enhanced controllability is crucial. Combining few-shot data augmentation and cross-modal guidance to improve generalisation in edge scenarios is also an effective means of boosting model generalisation. Finally, constructing unified fusion frameworks: Expanding cross-model collaboration paradigms, such as UCGM, will provide directions for integrating the advantages of multiple models—for example, combining the adversarial training of GANs, the denoising mechanism of Diffusion Models, and the sequential modelling advantage of AR models. It is believed that the implementation of the above methods will drive the evolution of image-generative AI toward next-generation models with higher quality, stronger generalisation, and better controllability.

References

[1] Ren S, Yu Q, He J, Shen X, Yuille A, Chen L-C. Beyond next-token: next-X prediction for autoregressive visual generation. arXiv:2502.20388, 2025.

[2] Alonso E, Moysset B, Messina R. Adversarial generation of handwritten text images conditioned on sequences. Proceedings of the International Conference on Document Analysis and Recognition, 2019: 481-486.

[3] Fogel S, Averbuch-Elor H, Cohen S, Mazor S, Litman R. ScrabbleGAN: semi-supervised varying length handwritten text generation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 4324–4333.

[4] Zdenek J, Nakayama H. JokerGAN: memory-efficient model for handwritten text generation with text line awareness. Proceedings of the 29th ACM International Conference on Multimedia, 2021: 5655–5663.

[5] Liu X, Meng G, Xiang S, Pan C. Handwritten text generation via disentangled representations. IEEE Signal Processing Letters, 2021, 28: 1838–1842.

[6] Kang L, Riba P, Wang Y, Rusinol M, Fornes A, Villegas M. GANwriting: content-conditioned generation of styled handwritten word images. European Conference on Computer Vision, 2020, 12368: 237–289.

[7] Gan J, Wang W. HiGAN: handwriting imitation conditioned on arbitrary-length texts and disentangled styles. AAAI Conference on Artificial Intelligence, 2021, 35(9): 7484–7492.

[8] Davis B, Tensmeyer C, Price B, Wigington C, Morse B, Jain R. Text and style conditioned GAN for generation of offline handwriting lines. arXiv:2009.00678, 2020.

[9] Gan J, Wang W, Leng J, Gao X. HiGAN+: handwriting imitation GAN with disentangled representations. ACM Transactions on Graphics, 2022, 42(1): 1–17.

[10] Luo C, Zhu Y, Jin L, Li Z, Peng D. SLOGAN: handwriting style synthesis for arbitrary-length and out-of-vocabulary text. IEEE Transactions on Neural Networks and Learning Systems, 2022.

[11] Ganz R, Elad M. BIGRoC: boosting image generation via a robust classifier. Transactions on Machine Learning Research, 2023.

[12] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 2020, 33: 6840–6851.

[13] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. arXiv:2105.05233, 2021.

[14] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. arXiv:1809.11096.

[15] Gao S, Zhou P, Cheng M-M, Yan S. MDTv2: masked diffusion transformer is a strong image synthesizer. arXiv:2303.14389, 2023.

[16] Sauer A, Schwarz K, Geiger A. StyleGAN-XL: scaling StyleGAN to large diverse datasets. ACM SIGGRAPH Conference Proceedings, 2022.

[17] Sun P, Jiang Y, Lin T. Unified continuous generative models. arXiv:2505.07447, 2025.

[18] Miyato T, Kataoka T, Koyama M, Yoshida Y. Spectral normalization for generative adversarial networks. International Conference on Learning Representations (ICLR), 2018.

[19] Chen T, Zhai X, Ritter M, Lucic M, Houlsby N. Self-supervised GANs via auxiliary rotation loss. Proceedings of the Conference on Computer Vision and Pattern Recognition, 2019.

[20] Lee K S, Tran N T, Cheung N M. InfoMax-GAN: improved adversarial image generation via information maximization and contrastive learning. Winter Conference on Applications of Computer Vision, 2021.

[21] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis. Conference on Computer Vision and Pattern Recognition, 2021.

[22] Chang H, Zhang H, Jiang L, Liu C, Freeman W T. Masked generative image transformer. Conference on Computer Vision and Pattern Recognition, 2022.

[23] Peebles W, Xie S. Scalable diffusion models with transformers. International Conference on Computer Vision, 2023.

[24] Tian K, Jiang Y, Yuan Z, Peng B, Wang L. Visual autoregressive modeling: scalable image generation via next-scale prediction. Advances in Neural Information Processing Systems, 2024.

[25] Yu S, Kwak S, Jang H, Jeong J, Huang J, Shin J, Xie S. Representation alignment for generation: training diffusion transformers is easier than you think. arXiv:2410.06940, 2024.

[26] Ma N, Goldstein M, Albergo M S, Boffi N M, Vanden-Eijnden E, Xie S. SIT: exploring flow and diffusion-based generative models with scalable interpolant transformers. European Conference on Computer Vision, 2024: 23–40.

[27] Zheng H, Nie W, Vahdat A, Anandkumar A. Fast training of diffusion models with masked transformers. arXiv:2306.09305, 2023.

[28] Karras T, Aittala M, Lehtinen J, Hellsten J, Aila T, Laine S. Analyzing and improving the training dynamics of diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 24174–24184.

[29] Wang S, Tian Z, Huang W, Wang L. DDT: decoupled diffusion transformer. arXiv:2504.05741, 2025.

[30] Yao J, Yang B, Wang X. Reconstruction vs. generation: taming optimization dilemma in latent diffusion models. arXiv:2501.01423, 2025.

[31] Chen T, Zhai X, Ritter M, et al. Self-Supervised GANs via Auxiliary Rotation Loss. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[32] Atito S, Awais M, Kittler J. SIT: Self-supervised vision transformer. arXiv preprint arXiv:2104.03602, 2021.

[33] Kang M, Zhu J-Y, Zhang R, et al. Scaling up GANs for text-to-image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[34] Kingma D P, Ba J. Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR), 2015.

[35] Kingma D P, Welling M. Auto-encoding variational bayes. In: International Conference on Learning Representations (ICLR), 2014.

[36] Kynkäänniemi T, Aittala M, Karras T, et al. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. arXiv preprint arXiv:2404.07724, 2024.

[37] Lee D, Kim C, Kim S, et al. Autoregressive image generation using residual quantization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[38] Lipman Y, Chen R T Q, Ben-Hamu H, et al. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022.

[39] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 2011, 12: 2825–2830.