UDC 004.8
DOI https://doi.org/10.36994/2788-5518-2025-01-09-10

COMPARATIVE ANALYSIS OF ADVERSARIAL-DIFFUSION HYBRID ARCHITECTURES OF GENERATIVE MODELS

Bondar V. V., *Ph.D. student, Cherkasy State Technological University, Cherkasy, Ukraine. https://orcid.org/0009-0002-3230-5979, v.v.bondar.asp24@chdtu.edu.ua*

Babenko V. H., *D.Scs., Prof., Cherkasy State Technological University, Cherkasy, Ukraine. https://orcid.org/0000-0003-2039-2841, v.babenko@chdtu.edu.ua*

Abstract. *This paper examines the emerging integration of adversarial objectives into diffusion-based generative frameworks, with the purpose of systematically analyzing how hybrid architectures can overcome the limitations of individual approaches while combining their respective strengths. We identify and compare three distinct categories of hybrid models: distillation-based approaches that compress diffusion models using adversarial feedback, in-loop methods that incorporate discriminators during diffusion training, and diffusion-assisted generative adversarial networks that leverage noise processes to stabilize adversarial learning. Our comparative analysis reveals critical trade-offs across generation speed, sample fidelity, training stability, and application suitability. Distillation-based hybrids achieve near-instant high-quality generation but require intensive training processes, while in-loop adversarial diffusion methods offer superior image quality with moderate sampling efficiency. Diffusion-assisted generative adversarial networks maintain single-pass generation while addressing traditional generative adversarial networks instabilities through diffusion mechanisms. Empirical evidence demonstrates that these hybrid approaches consistently outperform pure diffusion or adversarial methods in quality and efficiency metrics. We identify significant research gaps, particularly regarding the theoretical unification of adversarial and score-matching objectives, scaling challenges for high-resolution generation, and expansion beyond image synthesis to other modalities. This analysis provides a foundation for understanding how complementary strengths of diffusion and adversarial techniques can be leveraged to develop more effective generative frameworks combining distribution coverage with perceptual realism.*

Key words: *diffusion models, generative adversarial networks, generative models, training stability, sampling efficiency, hybrid architectures, deep learning.*

ПОРІВНЯЛЬНИЙ АНАЛІЗ ГІБРИДНИХ ГЕНЕРАТИВНИХ МОДЕЛЕЙ З ДИФУЗІЙНИМИ ТА ГЕНЕРАТИВНО-ЗМАГАЛЬНИМИ ЕЛЕМЕНТАМИ

Віталій Бондар, *здобувач ступеня доктор філософії, Черкаський державний технологічний університет, Черкаси, Україна. https://orcid.org/0009-0002-3230-5979, v.v.bondar.asp24@chdtu.edu.ua*

Віра Бабенко, *д.т.н., проф., Черкаський державний технологічний університет, Черкаси, Україна. https://orcid.org/0000-0003-2039-2841, v.babenko@chdtu.edu.ua*

Анотація. *Робота присвячена дослідженню інтеграції принципів генеративно-змагальних мереж у генеративні фреймворки на основі дифузії з метою систематичного аналізу щодо того, як гібридні архітектури можуть подолати обмеження окремих підходів, поєднуючи їхні переваги. Виокремлено та порівняно три категорії гібридних моделей: підходи на основі дистиляції, які стискають дифузійні моделі за допомогою змагального зворотного зв'язку; методи поєднання в циклі навчання, що включають дискримінатори під час тренування дифузії; генеративно-змагальні мережі з підтримкою дифузії, які використовують шумові процеси для стабілізації змагального навчання. Порівняльний аналіз дав змогу виявляти критичні компроміси між швидкістю генерації, точністю зразків, стабільністю навчання й придатністю для застосування. Гібриди на основі дистиляції досягають майже миттєвої високоякісної генерації, але потребують інтенсивних процесів навчання, тоді як додавання дискримінатора в цикли дифузійних методів пропонують вищу якість зображень з помірною ефективністю семплювання. Генеративно-змагальні мережі з підтримкою дифузії зберігають одноетапну генерацію, вирішуючи традиційні проблеми нестабільності генеративно-змагальної мережі через дифузійні механізми. Емпіричні дані демонструють, що ці гібридні підходи послідовно перевершують чисті дифузійні або змагальні методи за показниками якості й ефективності. Подано опис виявлених значних прогалин у дослідженнях, зокрема щодо теоретичного об'єднання функцій утрат змагального та дифузійних підходів, проблем масштабування для високороздільної генерації й розширення за межі синтезу зображень на інші модальності. Проведений у роботі аналіз закладає основу для розуміння того, як можна викори-*

стати комплементарні переваги дифузійних і змагальних технік для розроблення ефективніших генеративних фреймворків, що поєднують покриття розподілу з покращеним реалізмом.

Ключові слова: *дифузійні моделі, генеративно-змагальні мережі, генеративні моделі, стабільність навчання, ефективність семплювання, гібридні архітектури, глибоке навчання.*

Introduction

Integrating adversarial objectives into diffusion models has significantly advanced the generation of high-fidelity samples while preserving the theoretical guarantees of score-based approaches. Diffusion models have emerged as powerful generative frameworks that learn to reverse a predefined noise process through score matching, ensuring convergence to the true data distribution [1]. Despite their theoretical soundness, these models often produce overly smooth or blurred outputs when sampling with limited denoising steps or in compressed latent spaces. Integrating adversarial losses overcomes this fundamental limitation by introducing a discriminator that distinguishes between real and generated samples, effectively compelling the generator to capture fine-grained details that pure score matching might miss [2].

Adversarial-diffusion hybrids can be broadly categorized into several architectural approaches, each with distinct advantages. Distillation-based methods represent a new category in which adversarial training enhances the process of compressing a large "teacher" model into a more efficient "student". For instance, Adversarial Diffusion Distillation (ADD) optimizes both score-distillation and adversarial losses for a few-step diffusion process to preserve sharp details that might otherwise be lost during compression [3]. Building on this foundation, progressive schemes such as SDXL-lightning implement sequential adversarial refinement across multiple distillation stages, systematically improving the student's output fidelity [4].

Beyond distillation, researchers have developed techniques that embed adversarial components directly within the diffusion training process. Constant Acceleration Flow introduces an innovative flow discriminator that evaluates whether intermediate denoising trajectories align with an accelerated process, thereby enforcing more realistic dynamics and faster convergence [5]. Similarly, Structure-Guided Adversarial Training employs a manifold discriminator operating in latent space to ensure that learned score fields maintain consistency with the underlying data geometry [6].

The purpose of this paper is to provide a comprehensive analysis of emerging adversarial-diffusion hybrid architectures, examining their respective strengths and limitations in terms of sample quality, generation speed, and training stability. We aim to identify critical research gaps and promising directions for future work, particularly regarding the theoretical unification of adversarial and score-matching objectives and their potential applications to complex multimodal datasets.

1. Background

1.1.Diffusion Models

Diffusion models represent a class of generative approaches that learn data distributions by reversing a gradual noise addition process through score matching techniques. These models operate by adding Gaussian noise to data across multiple timesteps, then training a neural network to estimate the gradient of the log data density at each noise level [1]. The generative process begins from pure noise and progressively denoises the sample by following the estimated score function, typically implemented via discretized stochastic differential equation (SDE) or ordinary differential equation (ODE). The training objective minimizes the error between network predictions and the actual noise added to the data, providing theoretical guarantees on distribution matching and ensuring robust mode coverage. Despite these advantages, diffusion models often require numerous sampling steps during generation, resulting in computational inefficiency and potentially blurred outputs when using fewer steps [7]. Recent research has focused on accelerating sampling procedures while maintaining generation quality through techniques such as distillation and improved ODE solvers [8].

1.2.Adversarial Training

Adversarial training frameworks employ a competitive learning paradigm between generator and discriminator networks to produce high-fidelity samples. In the canonical generative adversarial network (GAN) architecture, a generator transforms random noise into synthetic samples while a discriminator learns to distinguish between real and generated data [9]. The generator optimizes an adversarial loss designed to maximize the probability of the discriminator misclassifying generated samples as real, thereby encouraging the production of increasingly realistic outputs. This adversarial dynamic enables GANs to capture fine-grained perceptual details and high-frequency components that enhance visual quality beyond what traditional likelihood-based models achieve. However, despite their ability to generate sharp, realistic images, adversarial training suffers from fundamental limitations, including training instability, mode collapse, and sensitivity to hyperparameter selection [10]. These challenges have motivated extensive research into stabilization techniques, such as spectral normalization, and alternative objective functions, such as the Wasserstein distance [11].

2. Adversarially-Refined Diffusion Models

Integrating adversarial training techniques with diffusion-based generation significantly improves sample fidelity and sampling efficiency in generative modeling. Adversarially-refined diffusion models utilize discriminator networks to critique diffusion outputs, thereby creating a training dynamic where the diffusion model learns to produce increasingly realistic samples. This section examines two distinct approaches: distillation-based hybrids that compress diffusion models into faster generators and in-loop adversarial methods that directly incorporate adversarial feedback during diffusion training.

2.1. Distillation-based Hybrids

Knowledge distillation combined with adversarial supervision enables the compression of diffusion models while maintaining high-quality outputs. Distillation-based hybrids employ a teacher-student paradigm where a pre-trained diffusion model (teacher) transfers its capabilities to a simplified student model and adds the ability of accelerated generation [5]. The student receives dual supervision through a distillation loss matching the teacher's denoising behavior and an adversarial loss pushing outputs toward the real data distribution. This adversarial guidance effectively addresses the quality degradation typically observed when aggressively compressing diffusion models to fewer steps.

Adversarial Diffusion Distillation (ADD) demonstrates how score-based distillation and GAN-style training can be effectively combined to train one-step generative models [3]. In ADD, a student network distills knowledge from a large diffusion model like Stable Diffusion while simultaneously learning to fool a discriminator judging the realism of generated images. The discriminator provides crucial feedback that forces the student to capture fine details often missed by pure distillation approaches. This combination enables ADD to achieve remarkable image fidelity even when generating samples in just 1–4 denoising steps, effectively combining GAN-like speed with diffusion model quality.

Latent Adversarial Diffusion Distillation (LADD) extends this approach by applying adversarial distillation in latent space to enable high-resolution generation. LADD operates in the compressed latent domain of an autoencoder but applies adversarial loss to decoded outputs, ensuring the preservation of high-frequency details. By shifting adversarial distillation to latent feature space, LADD dramatically reduces computational requirements for large-resolution synthesis compared to pixel-space ADD. The latent-domain discriminator provides effective feedback on both global structure and local texture, allowing models like SD3-Turbo to match state-of-the-art image quality using only four denoising steps [12].

SDXL-Lightning implements a progressive multi-stage distillation process with adversarial guidance at each stage. Rather than directly distilling to a one-step model, SDXL-Lightning creates a sequence of intermediate models with progressively fewer sampling steps. Each distillation stage employs adversarial loss to maintain image quality while increasing generation speed. This staged approach preserves diversity in earlier phases while focusing on detail refinement in later stages, producing a one-step generator that achieves high image quality at 1024×1024 resolution without the fidelity loss typical of aggressive distillation [4].

Diffusion2GAN takes a different approach by framing distillation as a supervised image translation problem. This method generates paired examples from a diffusion teacher (random noise vectors and corresponding images) and trains a GAN to replicate this mapping in a single step. The adversarial loss ensures the GAN's outputs match both the diffusion model's results and real images. Using pre-trained discriminators and multiscale adversarial feedback, Diffusion2GAN stabilizes training and successfully captures the complex distribution learned by the diffusion model, creating a one-step generator with diffusion-level quality and GAN-like interactivity [13].

2.2. In-loop Adversarial Diffusion

Incorporating adversarial loss directly into the diffusion training process enhances model capabilities without requiring separate distillation. In-loop adversarial diffusion methods train diffusion models from scratch with discriminators guiding the learning process. Unlike distillation approaches, the discriminator evaluates intermediate denoising states or sample sets during training, transforming diffusion optimization into a minimax game similar to GANs. This direct integration helps enforce global realism and structural consistency that standard diffusion losses might not capture.

Constant Acceleration Flow (CAF) employs discriminators to guide ODE-based diffusion trajectories during training. CAF models the diffusion sampling with an acceleration term and trains this acceleration predictor using adversarial objectives. The discriminator distinguishes between learned model trajectories and ground truth diffusion trajectories from training data, effectively correcting the generation path when it deviates from realistic patterns. This continuous adversarial guidance helps CAF avoid flow-crossing issues and quality degradation observed in other fast-generation techniques, resulting in more accurate few-step sampling [5].

Structure-Guided Adversarial Training (SADM) uses specialized discriminators to enforce structural relationships present in real data. Rather than evaluating individual samples, SADM's structure discriminator assesses the organization of multiple samples or latent representation consistency. This approach applies adversarial loss to the diffusion model's score field in latent space, forcing the model to produce denoising directions that align with the real data distribution structure. By adding manifold-level supervision to traditional instance-level diffusion loss, SADM significantly improves generation fidelity and diversity, achieving better FID scores on challenging benchmarks like high-resolution ImageNet [6].

Adversarially-refined diffusion represents a powerful synthesis of diffusion models' diversity with GANs' realism and efficiency. By strategically incorporating adversarial feedback through distillation or in-loop training, these approaches overcome the traditional limitations of both frameworks, delivering high-fidelity, high-speed generative models that represent the current state-of-the-art in image synthesis.

3. Diffusion-Assisted GANs

Diffusion-assisted GANs integrate the denoising diffusion process within GAN frameworks to address the inherent instability of adversarial training. These hybrid models improve sample quality and training stability by combining structured noise processes with GAN discriminator feedback. The integration occurs through two primary mechanisms: noise-based regularization of GAN training dynamics and diffusion processes operating in compressed latent spaces. Empirical results demonstrate that this combination of diffusion and adversarial learning produces more realistic outputs while mitigating typical GAN training issues, such as mode collapse and convergence difficulties.

3.1.Noise-Smoothing GANs

Noise-smoothing GANs use diffusion processes to inject controlled noise during training, effectively regularizing adversarial learning dynamics. Diffusion-GAN uses this approach by applying a forward diffusion chain to both real and generated images, progressively adding Gaussian noise. A timestep-dependent discriminator then distinguishes between diffused real and diffused generated images at multiple noise levels. This mechanism provides smoother, more consistent feedback to the generator rather than emphasizing high-frequency details. The generator optimizes its parameters by backpropagating gradients through the diffusion process, learning to create images that remain realistic under significant noise perturbations. This technique stabilizes training and improves both image quality and data efficiency compared to conventional GANs [14].

Adversarial Noise Scheduling represents another noise-smoothing strategy that optimizes the diffusion noise schedule itself through adversarial training. Rather than using fixed schedules like linear or cosine decay, this method trains a noise schedule generator alongside the diffusion model. The schedule adjusts to maximize sample realism as determined by an adversarial criterion. The diffusion process learns optimal noise injection or removal at each step based on discriminator feedback comparing generated outputs with real images. By treating the noise schedule as a learnable component in the adversarial framework, the model adapts its denoising trajectory to produce outputs that better deceive the discriminator. This adversarially optimized scheduling leads to smoother denoising and higher-fidelity samples while sometimes improving sampling efficiency compared to hand-crafted schedules [15].

3.2.Latent-Space Hybrids

Latent-space hybrid methods perform diffusion in lower-dimensional latent spaces while applying adversarial losses either on output images or latent features. Latent Denoising Diffusion GAN (LDDGAN) utilizes a pre-trained autoencoder to compress images into compact latent codes, where a diffusion model then generates new samples. Since denoising operates on smaller latent representations, the model samples significantly faster than pixel-space diffusion while maintaining distribution coverage. After decoding latent samples back to images, a GAN discriminator evaluates them against real images, enforcing an adversarial loss that enhances realism. This division of labor allows the diffusion model to handle coarse structure and diversity in latent space while the discriminator refines fine details in image space. LDDGAN achieves faster sampling without sacrificing image quality, reporting improved fidelity and diversity compared to standard diffusion models while maintaining competitive generation speed [16].

LaDiffGAN represents an alternative latent-space hybrid that employs diffusion-based denoising as supervision for GAN training in latent space. Unlike LDDGAN, LaDiffGAN does not perform diffusion during inference; instead, it uses diffusion noise during training to guide the GAN's learning process. The generator operates on latent codes that are randomly perturbed with Gaussian noise of varying magnitudes during training. The model learns to denoise corrupted latent inputs while producing corresponding output images. A special latent diffusion GAN loss aligns generated image features with real image features under noise perturbations, effectively teaching the generator to reverse diffusion steps in latent space. Simultaneously, a standard adversarial loss ensures visual realism by comparing generated and real images. This dual training signal, combining diffusion-style denoising in latent space with image-level adversarial loss, stabilizes the GAN and improves its ability to model complex distributions. By learning to recover clean representations from noisy latent inputs, the generator gains diffusion-like strengths, such as noise robustness and mode coverage, while maintaining the crisp details characteristic of adversarial training. In unsupervised image-to-image translation tasks, LaDiffGAN outperforms conventional GAN baselines and approaches or exceeds pure diffusion models' visual quality while maintaining efficient one-step generation [17].

Diffusion-Assisted GANs demonstrate that incorporating diffusion processes into GANs through adversarial objectives helps to address fundamental challenges in GAN training. Whether smoothing discriminator learning with noise or using diffusion as a guiding mechanism in latent space, these techniques effectively combine strengths from both approaches. The integration of adversarial loss with diffusion stabilizes training dynamics and enhances sample diversity and fidelity. This promising direction suggests that further exploration of adversarially guided diffusion could continue advancing generative modeling capabilities.

4. Comparative Analysis

Hybrid adversarial-diffusion models fall into three distinct categories, each balancing fidelity, speed, and complexity differently. These include distillation-based hybrids that compress diffusion models into fast samplers using adversarial feedback, in-loop adversarial diffusion approaches that integrate discriminators during diffusion training, and diffusion-assisted GANs that incorporate diffusion processes into GAN frameworks. All three categories aim to combine diffusion models' diversity and stability with GANs' sharpness and speed through different architectural and algorithmic trade-offs.

Generation speed versus image fidelity represents a key comparison dimension across hybrid approaches. Distillation-based hybrids prioritize speed by compressing generation to as few as 1-4 denoising steps while maintaining high fidelity. Adversarial Diffusion Distillation achieves comparable quality to its diffusion teacher in just four steps [12], while SDXL-Lightning demonstrates state-of-the-art 1024×1024 image synthesis in 1–8 steps through progressive distillation of Stable Diffusion XL [4]. The adversarial component in these methods significantly enhances visual sharpness in low-step regimes. In contrast, in-loop adversarial approaches vary in focus: Structure-Guided Adversarial Training prioritizes fidelity by enforcing manifold-level realism at standard diffusion sampling speeds [6], and Constant Acceleration Flow emphasizes speed through optimized diffusion dynamics [5]. Diffusion-assisted GANs offer single-pass generation speed while using diffusion-based techniques to enhance image quality. Diffusion-GAN maintains one-step generation but produces more realistic images than standard GANs through diffusion-based noise scheduling [14]. For example, LaDiffGAN operates in one step with improved stylization quality by incorporating latent diffusion guidance [17]. All hybrid approaches leverage adversarial losses to enhance fidelity when generation steps are limited, though distillation and GAN-based methods specifically target minimal-step synthesis for speed.

Training stability versus methodological complexity presents another important consideration. Adversarial objectives introduce minimax dynamics that can destabilize training, requiring different stabilization strategies across approaches. Distillation-based hybrids benefit from pre-trained teacher models that provide strong baseline signals via score or feature distillation, anchoring training even as discriminators fine-tune output distributions. This architecture, especially with progressive distillation, has proven relatively stable. For instance, SDXL-Lightning gradually reduces diffusion steps while carefully designing discriminators to prevent divergence [4]. The trade-off appears in added complexity: training requires optimizing a student model against both teacher and discriminator, incurring substantial computational costs and hyperparameter tuning. In-loop adversarial diffusion methods eliminate the external teacher, simplifying dependencies but potentially complicating training. SADM directly integrates a discriminator into diffusion training to align with real data manifold structures, achieving state-of-the-art fidelity but requiring novel “structure” discriminators and careful training strategies to ensure convergence [6]. Diffusion-assisted GANs address traditional GAN stability challenges through diffusion-based mechanisms. Diffusion-GAN stabilizes discriminator training by gradually increasing noise-effectively creating a curriculum that prevents discriminator overfitting [14]. LaDiffGAN employs pre-trained diffusion models in latent space as supervisory signals, reducing mode collapse and improving convergence [17]. Each hybrid type must navigate adversarial training instability: some leverage existing diffusion models for stability, others modify training procedures through structural losses or noise schedules.

Application suitability varies significantly across hybrid categories depending on deployment priorities. Distillation-based hybrids excel when pre-trained diffusion models are available and low-latency deployment is required. Methods like ADD or SDXL-Lightning demand intensive one-time training but subsequently provide near-instant high-quality generation, making them ideal for real-time synthesis or on-device applications [4; 12]. For the maximum generative quality or adaptation to new domains from scratch, in-loop adversarial diffusion approaches may be the most appropriate. SADM offers superior image quality and cross-domain generalization when fidelity is critical and additional training complexity is acceptable [6]. If the goal is to reduce sampling steps without external teacher models, techniques like CAF provide an efficient path by integrating acceleration directly into diffusion training [5]. Diffusion-assisted GANs offer advantages when instantaneous generation is essential or in situations when classical diffusion models underperform. For tasks like image-to-image translation or small-data generation, GANs with diffusion guidance can outperform pure approaches. LaDiffGAN demonstrates superior stylization while avoiding overfitting on limited data [17], and Diffusion-GAN provides sharper outputs than standard GANs with improved stability [14]. These approaches are particularly fit for interactive applications or video synthesis where generation speed is critical and slight diversity compromises are acceptable.

Table 1
Comparison of hybrid adversarial-diffusion approaches across key criteria

Approach Category	Sample Quality	Generation Speed	Training Stability	Resource Efficiency
Distillation-Based Hybrids (nADD, SDXL-Lightning)	Very high fidelity (near teacher model quality even with few steps)	Very fast sampling (1–4 diffusion steps)	Relatively stable training (teacher guidance with adversarial fine-tuning)	Heavy training cost (requires large pre-trained model and discriminator); light inference load (few steps)
In-Loop Adversarial Diffusion (SADM, CAF)	High fidelity (improved generative quality; SADM achieves state-of-the-art)	Moderate speed (standard diffusion steps; faster if using accelerated flows like CAF)	Moderate stability (minimax training requires careful design; no external teacher)	High training cost (training from scratch with added objectives); inference cost varies (CAF reduces steps, others do not)
Diffusion-Assisted GANs (Diffusion-GAN, LaDiffGAN)	High fidelity for GAN standards (sharper details, better coverage than vanilla GANs)	Instant generation (single-step GAN-like inference)	Improved stability vs. standard GANs (noise or diffusion guidance stabilizes training)	Complex training (GAN + diffusion components); very efficient inference (generator forward pass only)

In summary, combining adversarial loss with diffusion yields diverse hybrid techniques with distinct trade-offs. Distillation-based hybrids deliver diffusion-quality images with drastically reduced sampling time but require complex distillation processes. In-loop adversarial diffusion methods integrate advantages during training without external teachers but demand careful design for stability. Diffusion-assisted GANs achieve the fastest generation and use diffusion to enhance quality and stability, though inheriting GAN training complexities. The selection among these approaches depends on specific priorities: maximum quality, minimum latency, training simplicity, or performance under limited data. Table \ref{tab:comparison} summarizes how these three categories compare across key evaluation criteria.

5. Conclusions and Future Directions

Adversarial-diffusion hybrid models substantially improve generative modeling by leveraging complementary advantages of both approaches [6; 12]. These hybrid architectures achieve better image quality, faster sampling speed, and more robust training dynamics than either method alone. Empirical evidence consistently shows that adversarial components contribute to finer details and enhanced realism, while diffusion elements provide structural coherence and prevent mode collapse. For instance, Sauer et al. demonstrated efficient one-step generation through adversarial distillation without sacrificing quality [3; 12], and Yang et al. reported record-breaking image fidelity by incorporating discriminators into diffusion training pipelines [6].

Current hybrid frameworks fall into three distinct categories with unique integration strategies. Distillation-based approaches compress multi-step diffusion into single-step generators by using pre-trained diffusion models as teachers and incorporating adversarial losses to preserve details [12]. In-loop integration methods embed discriminators directly within the diffusion process, often using noise-conditioned discriminators to guide each denoising step. Diffusion-assisted GANs instead enhance traditional GAN architectures with diffusion components, employing score-based generation for initialization or guidance. Despite methodological differences, all these approaches effectively combine diffusion's noise removal process with the adversarial training mechanism.

The emergence of these hybrid models reveals that diffusion and adversarial objectives serve as complementary rather than competing paradigms. Diffusion models excel at stable density estimation and comprehensive mode coverage, while adversarial training provides powerful feedback for photorealistic details and adaptive learning. The fusion of these properties enables adversarial-diffusion hybrids to achieve results previously unattainable, effectively bridging likelihood-based and adversarial approaches. This convergence suggests a broader trend toward unified generative frameworks that leverage multiple learning objectives simultaneously.

Despite empirical successes, the theoretical foundation for combining adversarial and score-based training remains undeveloped. Most hybrid architectures rely on experimental validation rather than rigorous mathematical justification. The formal reconciliation between diffusion objectives (typically optimizing likelihood or score matching) and GAN objectives (minimizing implicit divergences through minimax games) requires further investigation. Understanding how these different training signals interact (affecting convergence properties, mode coverage, and sample quality) would enable a more principled design of hybrid models. Such theoretical frameworks could guide the development of specialized loss functions that ensure stability when combining adversarial and diffusion criteria.

Balancing image quality, generation speed, and training stability presents a significant challenge, particularly at scale. Current hybrid approaches involve inevitable trade-offs: reducing diffusion steps improves speed but may compromise quality, while optimizing for maximum fidelity often requires more complex discriminators that complicate training. These challenges intensify when scaling to higher resolutions or more complex data modalities like video and 3D content. Early research demonstrates promising results for real-time high-resolution image generation [4] and short video sequences, but generating extended videos or detailed 3D environments at comparable quality levels remains unexplored. Advancing these capabilities will require innovations in both model architecture and training methodologies, which may simultaneously achieve fidelity, speed, and stability.

Extending adversarial-diffusion frameworks beyond image generation represents a promising research direction. Current hybrid models predominantly focus on image synthesis, leaving their applicability to other domains largely unexplored. Adapting these techniques to text or audio generation involves unique challenges due to different data structures and evaluation metrics. While diffusion models have shown success in text-to-image and audio synthesis tasks, integrating adversarial components for these modalities may require fundamentally different approaches. Multimodal generation presents another opportunity, potentially enabling hybrid models that simultaneously generate correlated content across multiple modalities, using diffusion for cross-modal consistency and adversarial feedback for realism enhancement. Addressing these challenges could improve generative performance across diverse domains, combining the best aspects of diffusion and adversarial approaches to create content with superior fidelity, diversity, and generation speed.

References

1. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, 6840–6851. https://doi.org/10.48550/arXiv.2006.11239.

2. Chen, J., Cai, H., Chen, J., Xie, E., Yang, S., Tang, H., ... & Han, S. (2024). Deep compression autoencoder for efficient high-resolution diffusion models. arXiv preprint arXiv:2410.10733. https://doi.org/10.48550/arXiv.2410.10733.

3. Sauer, A., Lorenz, D., Blattmann, A., & Rombach, R. (2023). Adversarial diffusion distillation. arXiv. arXiv preprint arXiv:2311.17042. https://doi.org/10.48550/arXiv.2311.17042.

4. Lin, S., Wang, A., & Yang, X. (2024). Sdxl-lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:2402.13929. https://doi.org/10.48550/arXiv.2402.13929.

5. Park, D., Lee, S., Kim, S., Lee, T., Hong, Y., & Kim, H. J. (2024). Constant Acceleration Flow. arXiv preprint arXiv:2411.00322. https://doi.org/10.48550/arXiv.2411.00322.

6. Yang, L., Qian, H., Zhang, Z., Liu, J., & Cui, B. (2024). Structure-guided adversarial training of diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7256–7266). https://doi.org/10.48550/arXiv.2402.17563.

7. Nichol, A. Q., & Dhariwal, P. (2021, July). Improved denoising diffusion probabilistic models. In International conference on machine learning (pp. 8162–8171). PMLR. https://doi.org/10.48550/arXiv.2102.09672.

8. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456. https://doi.org/10.48550/arXiv.2011.13456.

9. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27. https://doi.org/10.48550/arXiv.1406.2661.

10. Arjovsky, M., Chintala, S., & Bottou, L. (2017, July). Wasserstein generative adversarial networks. In International conference on machine learning (pp. 214–223). PMLR. https://doi.org/10.48550/arXiv.1701.07875.

11. Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957. https://doi.org/10.48550/arXiv.1802.05957.

12. Sauer, A., Boesel, F., Dockhorn, T., Blattmann, A., Esser, P., & Rombach, R. (2024, December). Fast high-resolution image synthesis with latent adversarial diffusion distillation. In SIGGRAPH Asia 2024 Conference Papers (pp. 1–11). http://dx.doi.org/10.1145/3680528.3687625.

13. Kang, M., Zhang, R., Barnes, C., Paris, S., Kwak, S., Park, J., ... & Park, T. (2024, September). Distilling diffusion models into conditional gans. In European Conference on Computer Vision (pp. 428–447). Cham: Springer Nature Switzerland. https://doi.org/10.48550/arXiv.2405.05967.

14. Wang, Z., Zheng, H., He, P., Chen, W., & Zhou, M. (2022). Diffusion-gan: Training gans with diffusion. arXiv preprint arXiv:2206.02262. https://doi.org/10.48550/arXiv.2206.02262.

15. Hang, T., Gu, S., Geng, X., & Guo, B. (2024). Improved noise schedule for diffusion training. arXiv preprint arXiv:2407.03297. https://doi.org/10.48550/arXiv.2407.03297.

16. Trinh, L. T., & Hamagami, T. (2024). Latent Denoising Diffusion GAN: Faster sampling, Higher image quality. IEEE Access. http://dx.doi.org/10.1109/ACCESS.2024.3406535.

17. Liu, X., Zeng, B., Gao, S., Li, S., Feng, Y., Li, H., ... & Zhang, B. (2024). Ladiffgan: Training gans with diffusion supervision in latent spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1115–1125). https://doi.org/10.1109/CVPRW63382.2024.00118.

Дата першого надходження рукопису до видання: 23.05.2025
Дата прийнятого до друку рукопису після рецензування: 20.06.2025
Дата публікації: 25.07.2025