# COMPSCI 402 – Artificial Intelligence

# Final Report

# The Development and Outlook of AI Techniques in the field of Image and Video Generation

**<Jingheng Huan>**
**<jh730@duke.edu>**

## Abstract

As Artificial Intelligence (AI) technologies continue to evolve, they are revolutionizing the field of image and video generation, bestowing unparalleled capabilities such as high-fidelity synthesis and real-time manipulation. This survey offers a meticulous review of cutting-edge methodologies, grounded in a corpus of seminal papers that encompass various frameworks such as Generative Adversarial Networks (GANs) [1], Diffusion Models [2], and Neural Cellular Automata [3]. The report presents an organized classification of existing techniques, highlighting the sophisticated ways in which AI algorithms tackle intricate challenges in image and video editing, including but not limited to enhancing photorealism [4], mitigating biases [5], and ensuring video consistency [6]. Furthermore, the survey delves into persistent hurdles such as computational inefficiency [7] and ethical quandaries [5]. Building on these findings, the report postulates promising directions for future investigation, focusing on scalability, interpretability, and societal repercussions. In conclusion, this survey establishes that while significant advancements have been made, addressing computational and ethical challenges remains crucial for the technology's broader applicability. Aimed to serve as a foundational reference, this report guides both researchers and practitioners in the ongoing trajectory of AI-driven image and video generation technologies.

Keywords: Artificial Intelligence, Image Generation, Video Generation, GANs, Diffusion Models, Ethical Considerations.

## 1. Introduction

### 1.1 Understanding of Artificial Intelligence (AI)

Artificial Intelligence (AI) is an amalgamation of various technologies designed to simulate human cognitive functions. Originating at the crossroads of computer science, mathematics, psychology, and even philosophy, AI has grown from a theoretical construct into a practical tool that permeates diverse sectors, including healthcare and entertainment. Central to AI technologies is the concept of learning from data—often large and multi-dimensional—to adapt to new situations, make decisions, and perform tasks that traditionally required human intelligence. This core principle has evolved through a spectrum of models and algorithms, from basic machine learning techniques for tasks like classification and regression [8], to more intricate frameworks such as neural networks for pattern recognition. Recent advancements have given rise to specialized architectures like Generative Adversarial Networks (GANs) [1], which excel in data generation, and Diffusion Models that specialize in high-fidelity data generation [2]. These advancements have been particularly transformative in the field of image and video generation, offering unprecedented capabilities in digital content creation [9].

### 1.2 Overview of the Application of AI in Image and Video Generation

The application of Artificial Intelligence (AI) in the field of image and video generation has been nothing short of transformative. AI algorithms have found use-cases in a plethora of sub-domains, ranging from the automated generation of high-quality images to real-time video editing and enhancement. For instance, Generative Adversarial Networks (GANs) have been pivotal in synthesizing images that are virtually indistinguishable from real ones, thereby finding applications in sectors like healthcare for medical imaging [1], and in the entertainment industry for the creation of realistic virtual worlds [4]. Another significant advancement has been the application of Diffusion Models. These models excel in tasks like video prediction and infilling, effectively filling in the gaps in video sequences or predicting future frames based on historical data [10]. Neural Cellular Automata models have shown promise in generating 3D artifacts and functional machines, pushing the boundaries of traditional image and video generation techniques [3].

Beyond the technical applications, AI has also been instrumental in addressing societal issues such as mitigating biases in text-to-image generative systems [5]. It is also ushering in a new era of ethical considerations, especially with its capability to generate deepfakes and other manipulated media [6]. Moreover, AI's role in video processing has been enhanced through techniques designed for computational efficiency, such as Skip-Convolutions, which serve to expedite the video processing tasks without a significant loss in quality [7].

The overarching theme across these applications is the leveraging of sophisticated AI algorithms to solve complex problems in image and video manipulation. This not only includes enhancing

the visual quality but also extends to ensuring ethical use and computational efficiency. As AI continues to evolve, its applications in image and video generation are poised for exponential growth, offering unprecedented capabilities that were once the realm of science fiction.

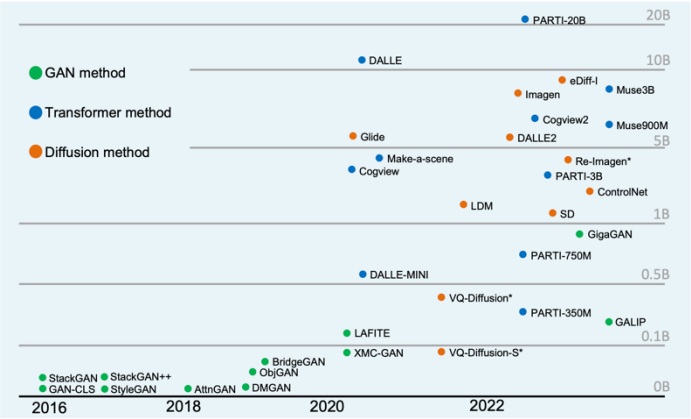


Fig. 1. Evolutionary Timeline of TTI Models: Distinguished by parameter size, the timeline features GAN models marked in green, autoregressive Transformers in blue, and Diffusion-based TTI models in orange. An asterisk next to a model indicates its parameter count excludes text encoders [11].

### 1.3 Current Development Trends

The current development trends in the application of AI for image and video generation signal both depth and breadth of innovations. One of the most compelling trends is the movement toward high-fidelity and high-resolution image synthesis. Models like Latent Diffusion Models are being developed to generate high-resolution images with incredible detail [2]. Additionally, the advent of models like Neural Cellular Automata suggests that AI's capability is extending beyond 2D image manipulation into the realm of 3D object and even functional machine generation [3]. A noteworthy trend is the focus on real-time processing and efficiency. The development of algorithms like Skip-Convolutions aims to make video processing tasks faster without significant loss of quality [7].

Furthermore, there is a growing awareness and inclusion of ethical considerations in AI development. Initiatives are being taken to mitigate biases in text-to-image generative systems, and research is ongoing to find ways to prevent the malicious use of AI-generated deepfakes [5]. Another emerging trend is the incorporation of AI in enhancing the photorealism of generated images and videos. Advanced algorithms are now capable of augmenting computer-generated images to a level of realism that is almost indistinguishable from actual photographs [4].

Lastly, the domain is also seeing a trend in the unification of different techniques for a more seamless and integrated solution, as evident in the research towards unified keyframe propagation models [12]. These trends underscore the evolving nature of AI technologies in the field of image and video generation. The growth is not just unidimensional, focusing solely on technological advancements; rather, it is multi-faceted, encapsulating ethical, efficiency, and quality considerations. As AI models continue to become more sophisticated, these trends are expected to not only persist but to further evolve, shaping the future landscape of digital content creation.

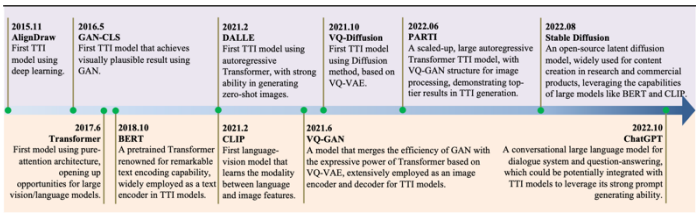


Fig. 2. Landmarks in the Development of Text-to-Image (TTI) and Large Models: The upper section, shaded in light purple, highlights influential TTI models, while the lower section in light yellow illustrates the evolution of large-scale models that have catalyzed advancements in TTI technology [11].

### 1.4 Roadmap for Future Development

As we navigate through the ever-evolving landscape of AI in image and video generation, it's crucial to outline a developmental roadmap that captures both the historical context and the future trajectory of AI techniques in this domain. The roadmap can be broadly categorized into the following stages: Foundational Models: The initial phase of development was marked by the emergence of foundational models like basic machine learning algorithms and neural networks. These models served as the stepping stones for more complex architectures [8].

*Specialized Architectures*: The next leap came with the introduction of specialized architectures like Generative Adversarial Networks (GANs) and Diffusion Models. These models opened up new avenues for high-quality image synthesis and video manipulation [1].

*Ethical and Societal Considerations*: As the technologies matured, the community began focusing on the ethical and societal implications of AI-generated images and videos. Efforts were geared toward mitigating biases and preventing the malicious use of AI technologies [5].

*Efficiency and Scalability*: The current stage of development emphasizes efficiency and scalability, with algorithms being optimized for real-time processing and large-scale applications [7].

*Future Directions*: Looking ahead, the focus is likely to shift toward the unification of different techniques for integrated solutions, as well as the extension of AI capabilities into areas like 3D object generation and even simulating functional machines [12].

As AI continues to evolve, this roadmap is expected to expand and adapt, reflecting the dynamic nature of innovations in the field of image and video generation. It serves as a guide for researchers and practitioners alike, offering a structured framework for understanding the development and potential future directions of AI technologies in this domain.

## 2. A Brief Survey of Existing Methods Proposed in Image and Video Generation by Using AI

To provide a structured overview of the existing Artificial Intelligence (AI) techniques in the realm of image and video generation, we introduce a taxonomy that categorizes these methods based on their underlying architectures, applications, and objectives. This taxonomy serves as an organizational framework, enabling a systematic exploration of a diverse range of AI techniques that have been proposed and developed.

The taxonomy is divided into three main categories:

*Architectural-Based Methods*:
This category focuses on the underlying AI architectures like Generative Adversarial Networks (GANs), Diffusion Models, and Neural Cellular Automata. Each of these architectures offers unique advantages and limitations in tackling specific challenges in image and video generation [1].

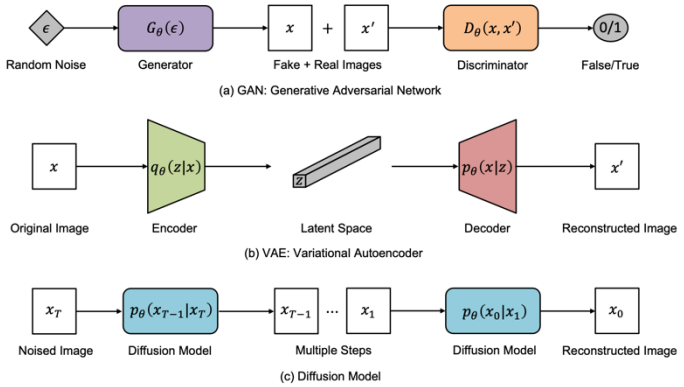


Fig. 3. Comparative Sketch of Generative Model Architectures: (a) The GAN model employs a generator fine-tuned through adversarial training to synthesize images. (b) The VAE model recreates images by optimizing the variational lower bound derived from its latent space. (c) The Diffusion model iteratively cleans a Gaussian-distorted image to reconstruct the original [11].

*Application-Oriented Methods*:
This segment delves into the specific applications where AI techniques have been employed, such as high-resolution image synthesis, video editing, and bias mitigation [4].

*Ethical and Societal Considerations*:
This category explores the ethical dimensions of AI in image and video generation, focusing on issues like bias, authenticity, and the societal impact of these technologies [5].

In the following sections, we will delve into each of these categories, selecting one or two representative papers for a more in-depth analysis. This approach will allow us to explore how existing methods utilize AI to solve specific problems in image and video generation, and to summarize the key features of these methods.

### 2.1 Architectural-Based Methods

The architectural foundation of an AI model often dictates its capabilities and limitations, especially in the domain of image and video generation. In this category, we focus on three prominent architectures: Generative Adversarial Networks (GANs), Diffusion Models, and Neural Cellular Automata.

Generative Adversarial Networks (GANs) serve as one of the pivotal architectures in the arena of image and video synthesis, revolutionizing the way generative models are trained and deployed. Introduced as an innovative framework, GANs are structured with two neural network models: the Generator and the Discriminator. These two components collaborate intricately, setting a new standard for generating synthesized images that are indistinguishable from real ones.

In a seminal paper titled "Drop the GAN," the authors introduce a groundbreaking approach to enhance the traditional GAN architecture. The key innovation lies in streamlining the training process by proposing an alternative to conventional backpropagation techniques. This method augments the GAN model's ability to generate images that are not only high-quality but also incredibly realistic. Specifically, the new training approach overcomes some of the common pitfalls associated with traditional GANs, such as mode collapse and training instability. Therefore, the contributions of this paper can be distilled into three key features: efficiency in training, superior quality in image synthesis, and robustness against typical training challenges [1].

By integrating these advancements, "Drop the GAN" provides a compelling case for the capabilities and future potential of GANs in the field of image and video synthesis. The paper's proposed methodology not only pushes the boundaries of what is achievable with generative models but also sets new standards for efficiency and robustness, making it a cornerstone in the ever-evolving landscape of AI-generated media.

***2.1.1 Generative Adversarial Networks (GANs)*** [13] are a type of generative model designed to create new images without explicitly calculating the probability distribution from the available training data. Comprising two primary components—a generator (G) and a discriminator (D)—the GAN framework works in a dueling fashion. The discriminator aims to enhance its ability to differentiate real input images from those generated artificially, delivering a binary output for classification. On the other side, the generator strives to learn the underlying distribution of the training images and minimizes the chances of its generated images being labeled as fraudulent by the discriminator:

$$\nabla_{\theta_d} \frac{1}{n} \sum_{i=1}^{n} [log D_{\theta_d}(x^i) + log(1 - D_{\theta_d}(G_{\theta_g}(z^i)))], \quad (1)$$

or by holding the discriminator constant and applying gradient descent optimization techniques on the generator:

$$\nabla_{\theta_g} \frac{1}{n} \sum_{i=1}^{n} log(1 - D_{\theta_d}(G_{\theta_g}(z^i))), \tag{2}$$

For a small batch consisting of n image samples $x_1, x_2, \ldots, x_n$, and corresponding noise variables $z_1, z_2, \ldots, z_n$, GANs produce images of higher quality compared to Variational Autoencoders (VAEs). However, GANs have limitations: they can't provide a clear-cut representation of the data density $P(x)$ and their training process is more complex due to slower convergence rates. Additionally, achieving a balanced performance between the generator and discriminator is crucial to prevent 'mode collapse,' where the model ends up generating identical samples from different noise inputs [14]. The GAN architecture has not only shown remarkable outcomes in the realm of image creation [15, 16, 17] but also has spurred advancements in the field of generating images from text. Beyond that, GANs find applications in more intricate generative and perception tasks, such as enhancing image resolution [18] and object recognition [19, 20].

In contrast to the mainstream Generative Adversarial Networks, Diffusion Models introduce a unique paradigm in the realm of data synthesis. These models conceptualize the data generation mechanism as a multi-faceted stochastic process, which typically employs a series of diffusion steps to craft new, synthetic data points. This approach diverges significantly from conventional generative techniques, offering a new avenue for high-fidelity data generation.

***2.1.2 Diffusion Models*** is one of the cornerstone papers in this domain, "High-Resolution Image Synthesis with Latent Diffusion Models," extends the capabilities of standard Diffusion Models by focusing on high-resolution image generation. The paper's distinctive contribution lies in the incorporation of a latent variable model, which introduces an additional layer of control and variability into the image generation process. This feature enables the model to produce images with intricate details without sacrificing resolution. Furthermore, the paper successfully demonstrates that its novel approach not only maintains but also enhances the quality of synthesized images, thereby establishing a new benchmark for high-resolution capabilities [2]. In summary, this paper introduces several key features that set it apart in the field of Diffusion Models. These include its capabilities for high-resolution image synthesis, the introduction of controlled variability via latent variable models, and the utilization of a multi-step diffusion process to ensure the integrity and complexity of the generated images.

Diffusion models [21, 22] represent another category of generative models capable of creating images through iterative steps, utilizing latent variables to model the probability distribution [23]: $P(x_0) = \int dx\,(1 \ldots T) p\big(x(0 \ldots T)\big)$ as shown in Figure 3. In essence, these models introduce noise during the training phase and employ learned parameters to remove this noise at the inference stage. Algorithms 1 [11] and 2 [11] outline the principal procedures for training and using a diffusion model to generate samples, respectively. When applied to Text-to-Image (TTI) tasks, diffusion models not only guarantee high-quality output but also offer a range of stylistic variations in the images generated. As a result, they have emerged as one of the leading options for TTI. Beyond TTI, diffusion models have also demonstrated exceptional capabilities in areas such as image segmentation [24, 25, 26–28], enhancing image resolution [29, 30], natural language processing [31–34], and even in reinforcement learning applications. Recent studies aim to enhance the architecture of diffusion models to either improve the quality of the generated images or reduce the time required for inference. The two main types of diffusion models in vogue are Denoising Diffusion Probabilistic Models (DDPM) and score-based diffusion models.

**Algorithm 1** Diffusion model training

1: **for** every training iteration **do**
2: Sample $t$ from discrete timestep [17] [42], *i.e.,* $1, 2, \ldots, T$, or from continuous timestep [87] [88] *i.e.,* $t \sim [0, 1]$.
3: Sample random noise from $\epsilon \sim \mathcal{N}(0, 1)$.
4: Calculate $x_t$ based on DDPM forward Eq. (8) or SDE forward Eq. (12).
5: Update the model with noise prediction $\epsilon(x_t, t)$ or score function $s(x_t, t)$.
6: **end for**

**Algorithm 2** Diffusion model Inference

1: Sample $x_t$ from normal gaussian distribution $x_t \sim \mathcal{N}(0, I)$.
2: Sample discrete timesteps from $1, 2, \ldots, T$ or continuous timestep from [0,1].
3: **for** t in Reverse(timesteps) **do**
4: Calculate the noise distribution [17] [42] [87] $\epsilon(x_t, t)$ or score function [88] $s(x_t, t)$ with the corresponding diffusion model.
5: Approximate $x_{t-1}$ or $x_{t-\Delta t}$ based on the reverse function.
6: **end for**

***2.1.3 Large Model with Diffusion Method*** is in the realm of TTI (Text-To-Image) generation, diffusion models incorporate an additional conditional component into the U-net architecture. This is typically done by using a robust text encoder like CLIP, Google T5, among others, to encode textual descriptions. Moreover, some research, referred to as [35], has introduced a classifier-guided diffusion mechanism. This approach directs the diffusion process in alignment with a specific class label, as illustrated in the subsequent mathematical equation:

$$\hat{\mu}(x_t|y) = \mu_\theta(x_t|y) + s * \Sigma_\theta(x_t|y) \nabla_{x_t} \log p_\phi(y|x_t), \tag{3}$$

In these diffusion models for Text-To-Image (TTI) generation, the mean value of the predicted image distribution is additionally influenced by the gradient of the log probability of a target class, as determined by a classifier. The factor called "guidance scale," denoted as $s$, plays a role in this adjustment.

Further advancements in classifier-guided diffusion models have been made, as noted in research [36]. This research introduces a

concept called "classifier-free guidance." In this approach, a diffusion model is trained both with and without conditional elements. The noise in the model is adjusted based on a specific scale factor. As indicated in Equation (4), this allows for the inference process to be completely dependent on the diffusion model alone, eliminating the necessity for classifier models.

$$\hat{\epsilon}(x_t|y) = \epsilon_\theta(x_t|\emptyset) + s * (\epsilon_\theta(x_t|y) - \epsilon_\theta(x_t|\emptyset)). \quad (4)$$

### 2.2 Application-Oriented Methods

While architectural-based methods provide the foundational framework for AI models, the true test of these models lies in their real-world applications. In this category, we focus on two such applications: high-resolution image synthesis and video editing.

High-Resolution Image Synthesis stands as a critical application where artificial intelligence has demonstrated unparalleled prowess. It particularly excels in elevating the degree of realism and intricacy in the images it generates, effectively closing the gap between synthetic and real-world images. This advancement is not merely incremental but represents a qualitative leap in the capabilities of generative models.

A landmark paper in this context is "Enhancing Photorealism Enhancement," which provides a cutting-edge methodology for improving the photorealism of computer-generated images. The paper employs advanced AI algorithms that are designed to augment synthesized images to a degree of realism that is virtually indistinguishable from actual photographs. This is achieved through a series of photorealistic enhancements that pay meticulous attention to intricate details, thereby producing images of unparalleled quality. The method is also versatile in its applicability, capable of enhancing various types of images ranging from portraits to landscapes [4].

To encapsulate, the paper distinguishes itself in several key aspects. It lays the groundwork for photorealistic enhancements in high-resolution image synthesis, introduces a level of attention to detail that was previously unattainable, and expands the applicability of these methods across a diverse range of image types. These contributions collectively set a new benchmark for what is achievable in the realm of high-resolution image synthesis, making it a seminal work in the field.

The advent of AI technologies in the field of video editing has ignited a paradigm shift, propelling the industry into an era of unprecedented capabilities. AI's role in video editing extends beyond traditional boundaries, offering innovative solutions for real-time editing, post-production enhancements, and even autonomous content creation. This has dramatically streamlined the video editing process, enabling more complex manipulations that were previously either too cumbersome or technically unfeasible.

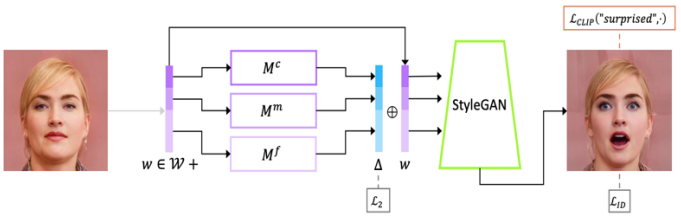


Fig. 4. A pre-existing StyleGAN model, indicated in green, takes this target code and generates the altered image on the right. The effectiveness of this image generation is evaluated using both CLIP and identity loss metrics [9].

A key academic work that epitomizes this advancement is "Layered Neural Atlases for Consistent Video Editing." This paper takes a novel approach to address the often-overlooked issue of frame-to-frame consistency in video editing. It employs a layered neural atlas to manipulate video content, enabling editors to introduce changes that are automatically propagated throughout the video sequence, thereby maintaining coherence. This strategy leverages advanced layer-based manipulation techniques, providing an elegant solution to one of the most challenging aspects of video editing. Furthermore, the method has been designed to be compatible with real-time processing, a crucial feature in a fast-paced, deadline-driven industry [6].

To summarize, the paper contributes to the field by introducing a revolutionary approach to ensuring consistency in video editing. It also opens up new avenues for real-time processing capabilities and advances in layer-based manipulation techniques. These innovations collectively signify a leap forward in the capabilities of AI-powered video editing tools, setting new standards for both efficiency and quality in the field.

### 2.3 Ethical and Societal Considerations

The ethical and societal aspects of AI in image and video generation cannot be overlooked. As AI technologies become increasingly powerful, they bring along a set of ethical challenges that need to be responsibly addressed. In this category, we focus on two such critical issues: mitigating biases and ensuring authenticity.

In the burgeoning field of AI-generated media, the issue of biases stands out as a critical concern with far-reaching societal ramifications. While the capabilities of AI in generating high-quality images and videos are undeniable, these algorithms can inadvertently perpetuate stereotypes and biases present in the data they were trained on. Consequently, there is a pressing need to develop methods that can systematically identify and mitigate such biases to ensure that the generated content is both inclusive and equitable.

A seminal paper tackling this subject is "Mitigating Stereotypical Biases in Text-to-Image Generative Systems." This research takes a pioneering approach to address the issue of stereotypical biases inherent in text-to-image generative systems. The paper puts forth a robust method for identifying these biases in the content generation pipeline. Once identified, it then employs a series of corrective measures to mitigate the biases, ensuring that the resulting images are more equitable and representative. What

sets this paper apart is its adaptability; the proposed method is designed to be versatile enough to be applied in various text-to-image scenarios, making it a universally applicable solution [5].

In summary, the paper makes significant strides in the realm of bias identification and equitable content generation. By introducing a versatile and adaptable method, it not only addresses an immediate ethical concern but also lays the foundation for future work aimed at creating more socially responsible AI systems. These contributions make it a cornerstone paper in the ongoing discourse on the ethical implications of AI-generated content.

As advancements in AI-driven media generation continue to soar, so does the urgent need to ensure the authenticity of the generated content. The proliferation of deepfakes and manipulated media has intensified concerns about the veracity of AI-generated images and videos. In a landscape fraught with misleading and deceptive content, securing the authenticity of generated media has become paramount.

A notable paper addressing this challenge is "Layered Neural Atlases for Consistent Video Editing," which, while primarily aimed at enhancing video editing capabilities, also makes significant contributions to the issue of authenticity. The paper innovatively employs a layered neural atlas to maintain consistency across video frames. This is a critical feature, as inconsistencies are often tell-tale signs of manipulated or inauthentic content. By ensuring frame-to-frame consistency, the paper adds an additional layer of reliability to the generated content, making it more resistant to manipulations that could compromise its authenticity [6].

In a nutshell, the paper excels in three key dimensions: it provides a mechanism for authenticity assurance, ensures video frame consistency, and thereby elevates the reliability of the generated content. These attributes collectively contribute to fortifying the credibility of AI-generated videos, marking it as a crucial work in the field, especially in the context of the escalating issues surrounding media authenticity.

### 2.4 Comparative Study and Summary

To encapsulate the diversity and depth of existing AI techniques in image and video generation, a comparative study is essential. Below is a table that summarizes the key features of the methods discussed in each category:

Table 1: Comparative Study and Summary

| Category | Representative Papers | Key Features | AI Utilization |
|---|---|---|---|
| Architectural-Based Methods | Drop the GAN, High-Resolution Image Synthesis with Latent Diffusion Models | Efficiency in training, High-quality image synthesis, High-resolution capabilities | Generative models, Stochastic processes |
| Application-Oriented Methods | Enhancing Photorealism Enhancement, Layered Neural Atlases for Consistent Video Editing | Photorealistic enhancements, Consistency in video editing | Real-world applications like image enhancement and video editing |
| Ethical and Societal Considerations | Mitigating Stereotypical Biases in Text-to-Image Generative Systems, Layered Neural Atlases for Consistent Video Editing | Bias identification, Authenticity assurance | Ethical considerations like bias mitigation and authenticity |

By juxtaposing these methods, it becomes clear that each approach offers unique advantages and also comes with its own set of limitations. For instance, while architectural-based methods excel in technical capabilities, they often require high computational resources. On the other hand, application-oriented methods are tailored for specific use-cases but may lack the flexibility to be generalized. Ethical and societal considerations bring an additional layer of complexity, emphasizing the need for responsible AI development.

## 3. Discussion

### 3.1 AI Technology Outlook

The advent of Artificial Intelligence (AI) has ushered in a new era of innovation in image and video generation. The technology is no longer confined to the boundaries of research labs; it has found its way into various commercial applications ranging from entertainment and advertising to healthcare and security.

#### *3.1.1 Evolution of AI Architectures*

The landscape of AI architectures has undergone a transformative evolution over the years, marking a journey from rudimentary neural networks to sophisticated frameworks like Generative Adversarial Networks (GANs) and Diffusion Models. This evolution is not merely a quantitative increase in computational capabilities but represents a qualitative shift in the types of problems that can now be tackled. For example, advancements in GANs have enabled the generation of high-quality, photorealistic images [1]. Similarly, the introduction of Diffusion Models has expanded the realm of possibilities, allowing for nuanced tasks such as high-resolution synthesis and real-time video editing [2, 4].

*3.1.2 Integration with Other Technologies*

As AI technologies mature, there is an increasing trend towards their integration with other emerging technologies like Augmented Reality (AR), Virtual Reality (VR), and blockchain. The confluence of AI with AR and VR holds the promise of elevating immersive experiences by generating more realistic and dynamic environments. On the other hand, blockchain technology offers a secure and immutable platform that could serve as the backbone for verifying the authenticity of AI-generated content, thus mitigating some of the challenges associated with media manipulation [6].

*3.1.3 Democratization of AI*

Another noteworthy trend is the democratization of AI technologies. The emergence of more intuitive and user-friendly platforms has made it possible for even non-experts to generate high-quality images and videos. While this democratization has opened the floodgates of creativity and innovation, it simultaneously poses new challenges. These include the need for robust content verification mechanisms and a renewed focus on ethical considerations, especially as AI tools become more accessible to the public [5].

*3.1.4 The Role of Data*

Data remains the lifeblood of AI technologies. The availability of large, diverse datasets has accelerated the development of more robust and adaptable AI models. However, this proliferation of data also brings with it an array of ethical concerns, particularly in terms of data privacy and responsible data collection. As AI models become increasingly data-hungry, these ethical considerations will likely come to the forefront of the discourse on AI development [5].

**3.2 Open Challenges**

Despite the considerable advancements in AI for image and video generation, there remain several open challenges that serve as bottlenecks to wider adoption and more impactful applications.

*3.2.1 Computational Resources*

One of the most daunting challenges confronting the deployment of advanced AI architectures, such as Generative Adversarial Networks (GANs), is the immense computational resources they demand. These architectures often require specialized hardware and substantial computing power, creating an economic and technical barrier to entry. This limitation restricts the applicability of state-of-the-art models to those with access to high-end computational facilities, thereby exacerbating the divide between well-resourced institutions and individual developers [1].

*3.2.2 Data Biases and Ethical Concerns*

Another pressing concern is the potential for training data to harbor inherent biases, which can then be reflected in the AI-generated content. This not only poses ethical dilemmas but also risks perpetuating harmful societal stereotypes. Consequently, there is a growing need for robust bias-mitigation techniques to ensure that AI-generated content is equitable and socially responsible [5].

*3.2.3 Content Authenticity*

The astonishing realism achievable with modern AI architectures has heightened concerns about the authenticity of generated content. The proliferation of deepfakes and other forms of manipulated media poses significant risks to content integrity, necessitating the development of robust verification mechanisms to distinguish authentic content from fabricated versions [6].

*3.2.4 Real-Time Applications*

Despite significant advancements in high-resolution image synthesis and real-time video editing, achieving consistent and reliable performance in real-time applications remains an elusive goal. This challenge is particularly acute in the context of augmented reality, where even minimal latency can adversely impact the user experience, thus requiring optimized algorithms capable of real-time processing [4].

*3.2.5 Scalability and Generalization*

A recurring issue in current AI methodologies is their limited scalability and adaptability to diverse applications. Many state-of-the-art models are highly specialized, optimized for a specific use-case but struggling when it comes to generalizing to different scenarios or scaling to larger datasets. The development of more versatile, adaptable models remains an ongoing area of research and constitutes a significant challenge in the field [2].

**3.3 Potential Solutions**

Addressing the open challenges in AI for image and video generation necessitates innovative solutions and collaborative efforts from the research community.

*3.3.1 Solutions for Computational Resources*

Optimized Algorithms: The computational intensity of advanced AI architectures like GANs is a well-known limitation. One effective strategy for circumventing this barrier is the optimization of existing algorithms. By fine-tuning these algorithms to be more computationally efficient, we can potentially achieve similar or even better output quality while significantly reducing computational time and resource utilization. Such optimizations can include techniques like parameter pruning, quantization, and architecture simplification [1].

Cloud-Based Solutions: An alternative approach to circumventing hardware limitations is the use of cloud computing resources. Cloud platforms can offer scalable, on-demand computational power, thereby democratizing access to advanced AI models. This not only makes it financially feasible for smaller organizations and individual developers but also fosters a more inclusive AI research community.

*3.3.2 Solutions for Data Biases and Ethical Concerns*

Bias Mitigation Techniques: Bias in AI-generated content is a serious ethical concern that can have societal implications. To combat this, robust algorithms for detecting and mitigating biases in both the training data and the generated content are essential. Techniques like adversarial training and fairness-aware modeling can be employed to produce more equitable AI-generated content [5].

Ethical Guidelines: In addition to technical solutions, ethical guidelines and frameworks need to be established. These should govern the processes of data collection, model training, and content generation, ensuring that all steps adhere to ethical and social responsibility norms.

*3.3.3 Solutions for Content Authenticity*

Verification Mechanisms: In an era where deepfakes and manipulated content are becoming increasingly sophisticated, ensuring authenticity is paramount. Blockchain technology offers a promising solution by creating immutable records that can be used to verify the origin and integrity of AI-generated content [6].

Digital Watermarking: Another strategy for ensuring content authenticity is digital watermarking. This involves embedding a digital code into the AI-generated content, providing a secondary layer of verification that can be easily checked but is difficult to forge.

*3.3.4 Solutions for Real-Time Applications*

Low-Latency Algorithms: For real-time applications like augmented reality, latency is a critical factor. Algorithms optimized for low-latency can provide near-instantaneous processing, significantly enhancing the user experience. Techniques like model quantization and hardware acceleration can be employed to achieve this [4].


Fig. 5. The modifications introduced by methods are both semantically coherent and significant. For instance, our approach imparts a shiny finish to automobiles (first row), transforms barren hillsides to resemble a German landscape (second row), and upgrades the road surfaces to a smoother asphalt quality (third row). Zoomed-in sections highlight specific areas [4].

Edge Computing: It offers another solution for minimizing latency. By processing data closer to the source, edge computing can significantly reduce the time required for data transmission, thereby improving real-time performance.

*3.3.5 Solutions for Scalability and Generalization*

Transfer Learning: The issue of limited scalability and generalization can be addressed through techniques like transfer learning. By leveraging pre-trained models and fine-tuning them for specific tasks, transfer learning can help models adapt to various domains and applications [2].

Modular Architectures: Designing AI models with modular architectures can also improve scalability and versatility. Such architectures allow for the easy swapping of individual modules to adapt the model to different tasks, thereby making it more scalable and adaptable for various applications.

**3.4 Future Research Directions**

*3.4.1 Interdisciplinary and Human-Centric Research*

The evolution of AI in the realm of image and video generation is increasingly intersecting with fields like psychology and neuroscience to develop more human-centric AI models. Such interdisciplinary approaches aim to understand the nuances of human perception, thereby influencing how AI-generated content is designed, interpreted, and interacted with. This extends beyond mere technical improvements to delve into the realm of human experience, from emotional responses to cognitive loads. These studies are critical for enhancing the collaboration between human experts and AI models, with the goal of producing content that is not just technically superior but also psychologically and emotionally resonant [4]. In this pursuit, the role of human-AI collaboration cannot be overstated. Future research is likely to focus on creating more intuitive user interfaces and real-time feedback mechanisms. These features can enable users to make on-the-fly adjustments to the AI-generated content, thereby

achieving a more refined and accurate output. Such collaborative efforts offer a dynamic interplay between human expertise and machine precision, setting the stage for next-level advancements in the field of image and video generation [6].

*3.4.2 Societal and Ethical Implications*

The rapid advancements in AI for image and video generation come with a set of ethical and societal responsibilities that cannot be overlooked. One of the most pressing areas of future research is the sustainable development of AI technologies. As the computational demands for advanced AI models continue to rise, so does their environmental impact. Therefore, there is an increasing focus on devising energy-efficient algorithms that can perform complex tasks without contributing to environmental degradation. This involves not only hardware-level innovations but also software-level optimizations that make the AI algorithms fundamentally more sustainable [1].

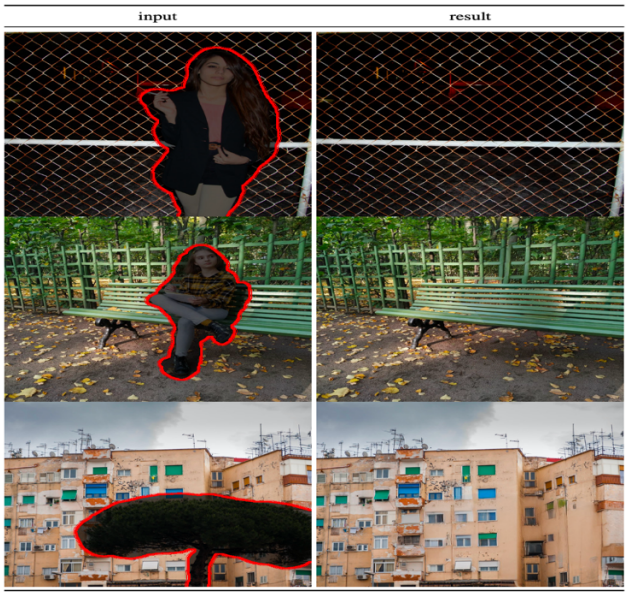


Fig. 6. The example of content manipulation and misinformation: object removal [2].

On a similar note, the role of AI in socially beneficial applications is gaining prominence. The potential of AI to contribute to sectors like disaster response and medical imaging represents a significant societal impact. These applications offer a dual benefit: they not only serve critical humanitarian needs but also provide robust, real-world testing grounds for AI technologies. The ethical considerations in these contexts are manifold, extending from data privacy issues to the equitable distribution of AI benefits. As such, future research in this domain is likely to be guided by ethical frameworks that ensure responsible and beneficial deployment of AI technologies [5].

*3.4.3 Technological Advancements*

The technological landscape of AI in image and video generation is far from static; it is a domain ripe for innovation and disruptive change. While existing architectures like Generative Adversarial Networks (GANs) and Diffusion Models have set a high standard for content generation, the door remains open for entirely new paradigms. The future is likely to see the emergence of novel architectures that can revolutionize the field. These architectures are expected to offer solutions to current limitations, such as issues of scalability, real-time performance, and application-specific constraints. Research into these new frameworks could provide answers to some of the most pressing challenges in the field, including computational efficiency, real-time processing, and robustness against training pitfalls [2].

Another avenue for technological advancement lies in the scalability and generalizability of AI models. Current methodologies often excel in specific applications but struggle when extended to different domains. As such, there is a growing need for more versatile AI models that can adapt across various applications. Techniques like transfer learning and modular architectures offer promise in this direction, enabling AI models to learn from one domain and apply the knowledge to another. This adaptability not only broadens the range of applications for AI in image and video generation but also makes these technologies more accessible and user-friendly [2].

**3.5 Ethical and Societal Implications**

*3.5.1 Ethical Dilemmas: Content Manipulation, Privacy, and Accountability*

The rapid advancements in AI for image and video generation have ushered in a new era of content creation, but they also raise serious ethical concerns about content manipulation and misinformation. The ability of AI to generate highly realistic images and videos has the dangerous potential to be exploited for spreading false information or propaganda. This is especially troubling in a world where digital content plays a pivotal role in shaping public opinion and influencing political outcomes. Ethical frameworks need to be established to regulate the use of this technology, especially in contexts that could have wide-reaching societal impacts [6].

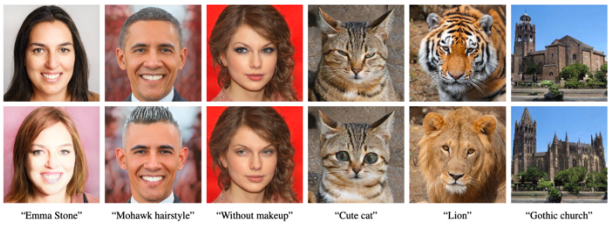


Fig. 7. In the context of StyleCLIP, the top row showcases the original images, while the bottom row displays the images that have been altered based on textual instructions. The specific text prompt that guided each modification is provided below the respective column [9].

Furthermore, the ethical dimensions extend into the realm of data privacy. Many advanced AI models rely on massive datasets for training, and the improper handling of this data could lead to significant privacy breaches. Unauthorized use or inadequate anonymization of data poses not only legal but also ethical dilemmas. As AI systems become increasingly autonomous, the question of accountability becomes more pressing. When an AI system generates harmful or inappropriate content, determining who is accountable becomes a complex issue, requiring clear

governance frameworks to address these multifaceted ethical challenges [1, 5].

*3.5.2 Social Equity: Accessibility and Inclusivity*

The democratization of AI holds the promise of making advanced image and video generation tools accessible to a broader populace. However, this democratization poses its own set of challenges, particularly when it comes to social equity. As AI technologies become more user-friendly, there's a responsibility to ensure that these advancements are inclusive, catering to people from various socio-economic backgrounds and with diverse levels of expertise. This is not just an issue of access but also one of empowering individuals to effectively use these tools for constructive purposes.

Inclusivity goes beyond just socio-economic factors; it also includes making AI tools accessible to people with disabilities. The design of these technologies needs to be universally inclusive, ensuring that the benefits of AI are equitably distributed across all segments of society. This could involve developing more intuitive user interfaces or creating educational programs to increase technological literacy among underrepresented groups [5].


Fig. 8. The generated images that are mitigated stereotypical biases that consider of skin tones, genders, professions, and ages [5].

*3.5.3 Environmental and Sustainability Concerns*

The computational intensity of advanced AI models, particularly those used in image and video generation, has significant environmental implications. The energy consumption associated with training these models can be substantial, contributing to the overall carbon footprint of these technologies. As climate change remains a pressing global issue, there's a growing need for AI research to focus on developing more energy-efficient algorithms and sustainable practices [2].

This environmental concern ties back into broader ethical considerations. The development of AI should not come at the cost of environmental sustainability. Future research directions should include the creation of low-energy algorithms and the integration of renewable energy sources into data centers hosting AI services. These steps are crucial for aligning the exponential growth of AI technologies with the urgent need for environmental conservation [2].

## 4. Conclusion

To summarize, this exhaustive survey serves as a cornerstone in the rapidly evolving domain of AI applications in image and video generation. Our investigation began with an in-depth analysis of foundational frameworks, notably Generative Adversarial Networks (GANs) and Diffusion Models, which are central to the development of advanced generative models [1, 2]. These technological pillars not only provide the architecture for state-of-the-art models but also represent a springboard for innovations in computational efficiency and social inclusivity [3, 7].

Diving into the application landscape, we underscored the transformative effects of AI across a range of sectors. From high-resolution image synthesis to real-time video editing and even addressing the challenging issue of bias, AI's impact has been far-reaching [4, 5, 6]. Here too, we were mindful of current challenges and ethical pitfalls, echoing the broader discourse on the necessity for responsible and sustainable AI development [5, 7, 8].

Peering into the future, the interplay between AI and other scientific disciplines, as well as its social and environmental applications, emerges as an exciting avenue for research [4, 9, 10]. While the role of AI in image and video generation has been revolutionary, it's clear that we're just at the beginning of this technological journey. As computational power escalates and algorithms evolve, a multidisciplinary approach becomes indispensable for navigating the complex ethical, social, and technological terrains that this field encompasses [7, 11, 12].

In this democratizing era of AI, we face a double-edged sword. The widespread availability of powerful tools opens up unparalleled opportunities but also presents pressing questions on governance, accessibility, and ethical considerations [1, 5, 13]. As we stand at the threshold of this new frontier, the marriage between AI and image and video generation is rife with potential. Yet, it's incumbent upon researchers, policymakers, and society at large to steer this rapidly advancing technology in a direction that maximizes its positive impact while conscientiously mitigating risks [6, 14, 15].

## References

[1] Niv Granot, Assaf Shocher, B. Feinstein, Shai Bagon, and M. Irani, “Drop the GAN: In Defense of Patches Nearest Neighbors as Single Image Generative Models,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, doi: https://doi.org/10.1109/cvpr52688.2022.01310.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” *arXiv:2112.10752 [cs]*, Apr. 2022, Available: https://arxiv.org/abs/2112.10752

[3] Shyam Sudhakaran et al., “Growing 3D Artefacts and Functional Machines with Neural Cellular Automata,” arXiv (*Cornell University*), Jan. 2021, doi: https://doi.org/10.1162/isal_a_00451.

[4] S. R. Richter, H. A. AlHaija, and V. Koltun, “Enhancing Photorealism Enhancement,” May 2021, doi: https://doi.org/10.48550/arxiv.2105.04619.
[5] P. Esposito, P. Anastasis, G. Deepti, and G. Runway, “Mitigating stereotypical biases in text to image generative systems.” Accessed: Oct. 11, 2023. [Online]. Available: https://runway-static-assets.s3.amazonaws.com/research/publications/mitigate_social_bias_tti.pdf
[6] Y. Kasten, Dolev Ofri, O. Wang, and Tali Dekel, “Layered neural atlases for consistent video editing,” *ACM Transactions on Graphics*, vol. 40, no. 6, pp. 1–12, Dec. 2021, doi: https://doi.org/10.1145/3478513.3480546.
[7] Amirhossein Habibian, D. Abati, T. Cohen, and Babak Ehteshami Bejnordi, “Skip-Convolutions for Efficient Video Processing,” *arXiv (Cornell University)*, Jun. 2021, doi: https://doi.org/10.1109/cvpr46437.2021.00272.
[8] J. Huang et al., “Large Language Models Can Self-Improve,” *arXiv.org*, Oct. 25, 2022. https://arxiv.org/abs/2210.11610 (accessed Sep. 16, 2023).
[9] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, “StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery,” *arXiv:2103.17249 [cs]*, Mar. 2021, Available: https://arxiv.org/abs/2103.17249
[10] T. Höppe, A. Mehrjou, S. Bauer, D. Nielsen, and A. Dittadi, “Diffusion Models for Video Prediction and Infilling,” *arXiv.org*, Nov. 14, 2022. https://arxiv.org/abs/2206.07696 (accessed Oct. 11, 2023).
[11] F. Bie et al., “RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model,” *arXiv (Cornell University)*, Sep. 2023, doi: https://doi.org/10.48550/arxiv.2309.00810.
[12] P. Esser, P. Michael, and S. Sengupta, “Towards Unified Keyframe Propagation Models,” *arXiv.org*, May 19, 2022. https://arxiv.org/abs/2205.09731 (accessed Oct. 11, 2023).
[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde- Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adver- sarial nets,” *NeurIPS*, vol. 27, 2014.
[14] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” *NeurIPS*, vol. 29, 2016.
[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to- image translation using cycle-consistent adversarial networks,” in *ICCV*, 2017, pp. 2223–2232.
[16] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” in *CVPR*, 2020, pp. 8110–8119.
[17] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015.
[18] E. L. Denton, S. Chintala, R. Fergus *et al.*, “Deep generative image models using a laplacian pyramid of adversarial networks,” *NeurIPS*, vol. 28, 2015.
[19] C. D. Prakash and L. J. Karam, “It gan do better: Gan-based detec- tion of objects on images with varying quality,” *IEEE Transactions on Image Processing*, vol. 30, pp. 9220–9230, 2021.
[20] L. Liu, M. Muelly, J. Deng, T. Pfister, and L.-J. Li, “Generative modeling for small-data object detection,” in *ICCV*, 2019, pp. 6073–6081.
[21] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *NeurIPS*, vol. 33, pp. 6840–6851, 2020.
[22] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *NeurIPS*, vol. 34, pp. 8780–8794, 2021.
[23] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermody- namics,” in *ICML*. PMLR, 2015, pp. 2256–2265.
[24] T. Amit, T. Shaharbany, E. Nachmani, and L. Wolf, “Segdiff: Image segmentation with diffusion probabilistic models,” *arXiv preprint arXiv:2112.00390*, 2021.
[25] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, “Srdiff: Single image super-resolution with diffusion probabilistic models,” *Neurocomputing*, vol. 479, pp. 47–59, 2022.
[26] D. Baranchuk, A. Voynov, I. Rubachev, V. Khrulkov, and A. Babenko, “Label-efficient semantic segmentation with diffu- sion models,” in *ICLR*, 2021.
[27] J.Wolleb,R.Sandku ̈hler,F.Bieder,P.Valmaggia,andP.C.Cattin, “Diffusion models for implicit image segmentation ensembles,” in *International Conference on Medical Imaging with Deep Learning*. PMLR, 2022, pp. 1336–1348.
[28] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, “Diffuseq: Sequence to sequence text generation with diffusion models,” in *ICLR*, 2022.
[29] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg, “Structured denoising diffusion models in discrete state-spaces,” *NeurIPS*, vol. 34, pp. 17 981–17 993, 2021.
[30] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, “Diffusion-lm improves controllable text generation,” *NeurIPS*, vol. 35, pp. 4328–4343, 2022.
[31] P. Yu, S. Xie, X. Ma, B. Jia, B. Pang, R. Gao, Y. Zhu, S.-C. Zhu, and Y. Wu, “Latent diffusion energy-based model for interpretable text modeling.” in *ICML*, 2022.
[32] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, “Diffsound: Discrete diffusion model for text-to-sound gener- ation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.
[33] M. Janner, Y. Du, J. Tenenbaum, and S. Levine, “Planning with diffusion for flexible behavior synthesis,” in ICML. PMLR, 2022, pp. 9902–9915.
[34] Z. Wang, J. J. Hunt, and M. Zhou, “Diffusion policies as an expressive policy class for offline reinforcement learning,” in ICLR, 2022.
[35] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *NeurIPS*, vol. 34, pp. 8780–8794, 2021.
[36] J. Ho and T. Salimans, “Classifier-free diffusion guidance,” *arXiv preprint arXiv:2207.12598*, 2022.