A Survey of Video Generation Models and Embodied Intelligence Techniques

Abstract

The survey paper provides a comprehensive exploration of advanced artificial intelligence and machine learning techniques, focusing on video generation models, embodied intelligence, generative adversarial networks (GANs), diffusion models, reinforcement learning, neural networks, and autonomous agents. Video generation models, at the forefront of AI, facilitate the creation of dynamic visual content, crucial for applications across various fields, including medical education and multimedia editing. The integration of GANs and diffusion models has propelled the synthesis of realistic visual content, although challenges in evaluation and control persist. Embodied intelligence emphasizes AI's interaction with physical environments, enhancing adaptability and decision-making. The survey highlights the pivotal role of reinforcement learning in training autonomous agents for complex decision-making, with innovations like DiffPoGAN and hierarchical frameworks enhancing policy exploration and learning efficiency. Neural networks, foundational to AI, continue to evolve, with advancements in energy-efficient computing and visual perception. Despite significant progress, challenges remain in computational efficiency, model generalization, and dataset constraints. Future research directions include optimizing model architectures, enhancing integration techniques, and exploring new applications to address these limitations. The survey underscores the transformative potential of these technologies in shaping the future landscape of AI and machine learning.

1 Introduction

1.1 Significance of Video Generation Models

Video generation models are pivotal in artificial intelligence (AI) and machine learning, enabling the creation of dynamic visual content that transcends traditional static imagery. These models advance visual intelligence, provide extensive data sources for AI systems, and enhance user interaction across applications [1]. The capacity to generate high-quality, temporally coherent videos is essential for video manipulation, content creation, and multimedia editing, as exemplified by the RL-V2V-GAN model [1]. Their implications extend to real-world decision-making processes, particularly in biomedical video generation, which significantly enhances medical education, patient care, and public health [2].

The emergence of deep generative models, notably Generative Adversarial Networks (GANs), has enabled the synthesis of complex data through learned representations, marking a significant advancement in realistic visual content creation [3]. However, challenges in evaluating generated content persist, especially in areas like video game level generation, where traditional metrics are inadequate. This highlights the necessity for improved evaluation methodologies to enhance the fidelity and control of generated videos [4].

Recent advancements in foundation models have further enhanced video generation capabilities, raising questions about the potential to learn fundamental physical laws from visual data alone [5].

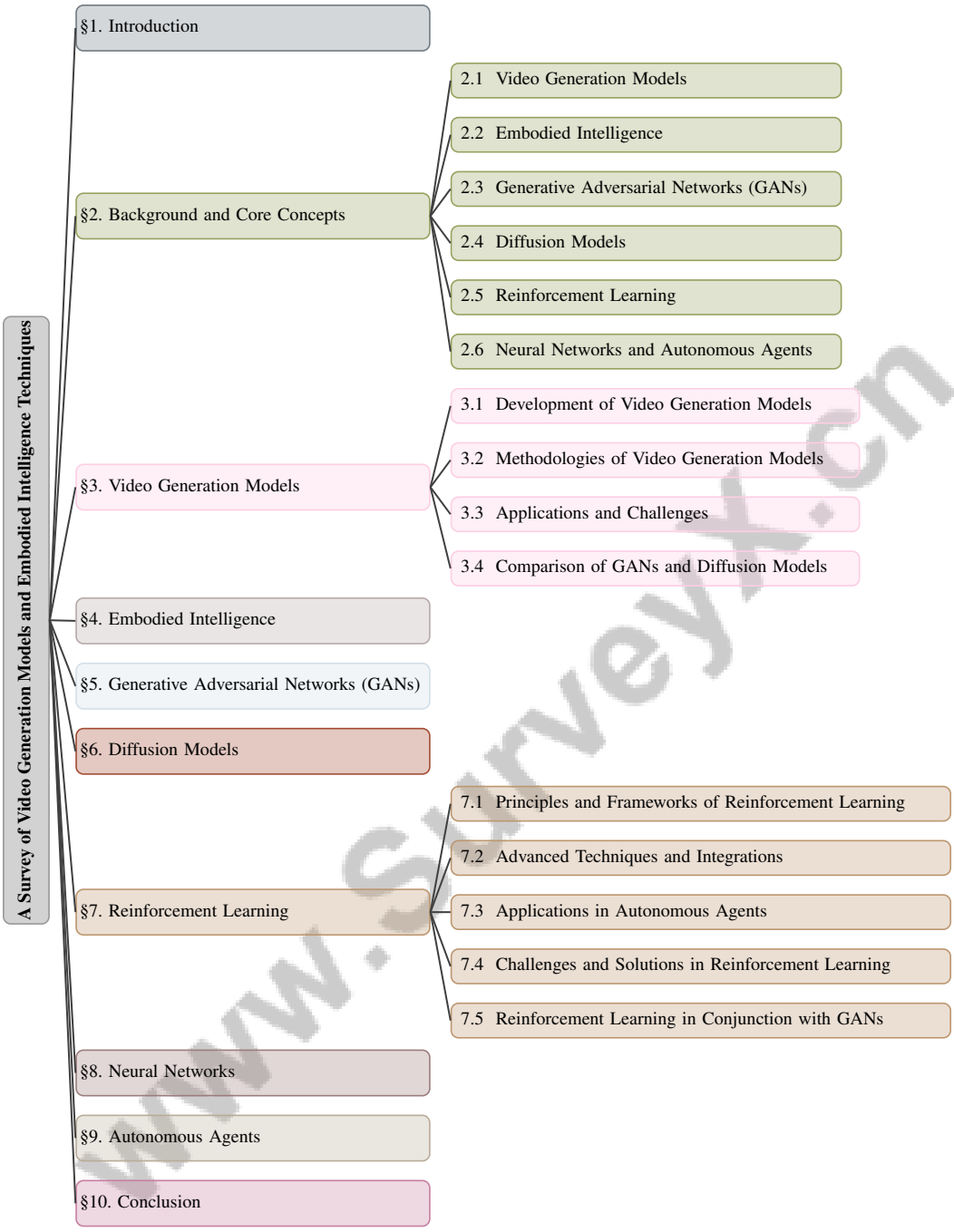


Figure 1: chapter structure

These developments are crucial for synthesizing realistic neural population activity patterns, essential for understanding neural information processing [6]. The integration of video generation models with multimodal AI systems is a significant area of interest, enhancing the understanding and generation of visual content by merging textual and visual modalities [7]. Nonetheless, achieving controllable video generation remains challenging, as existing methods often struggle with effective motion and appearance control [8]. The inherent complexity of video content also poses hurdles for effective editing and generation, necessitating further research to address these limitations [9].

The generation of high-quality 4D content has gained traction, underscoring the importance of video generation models in producing spatial-temporal content efficiently [10]. Challenges in text-to-video

generation, particularly in ensuring high-quality and motion-consistent outputs, further emphasize the need for innovation in this domain [11]. In AI-generated art, advancements in image generation via deep neural networks highlight the broader significance of these models [12]. Additionally, machine learning-assisted image generation serves as a catalyst for engaging public discourse on complex socioscientific issues, such as the Sustainable Development Goals, showcasing the societal impact of these technologies [13].

Video generation models are critical to advancing AI and machine learning, fostering innovation in content creation and broadening the scope of AI-generated media applications. Techniques like the VQGAN-CLIP model facilitate personalized video production methodologies, with integration into sectors such as media, education, and entertainment, allowing the synthesis of high-quality video content tailored to diverse needs, from abstract visual arts to over-the-top (OTT) streaming services. As technology evolves, AI-driven video production techniques are expected to proliferate, addressing existing challenges and expanding creative possibilities in the video domain [14, 15]. Continued research positions these models to shape the future of visual intelligence and interactive media.

1.2 Overview of Key Technologies

The realm of video generation and embodied intelligence is shaped by various advanced technologies, each contributing significantly to the field's evolution. Generative Adversarial Networks (GANs) have been essential in synthesizing realistic visual content through a competitive framework of generator and discriminator networks. This architecture not only facilitates high-quality video content creation but also reduces computational costs in complex simulations, such as computational fluid dynamics (CFD), showcasing its versatility beyond traditional image and video generation [16].

Diffusion models represent another cornerstone of generative technology, characterized by their iterative refinement of data samples to improve quality. These models have been utilized in innovative frameworks like Diffusion4D, which efficiently generates spatial-temporal consistent 4D content, demonstrating their capacity for handling complex generative tasks [10]. In manufacturing systems, diffusion models support complex decision-making processes, as seen in Generative Manufacturing Systems (GMS), optimizing operations and enhancing productivity [17]. Furthermore, diffusion models have been applied to generate laparoscopic videos interactively, conditioned on text prompts and surgical tool positions, advancing medical training and procedural planning [18].

Transformers, particularly in the form of diffusion models, have increasingly been integrated into video generation systems, such as Imagen Video, which employs a cascade of video diffusion models to produce high-definition videos from text prompts. This approach enhances the fidelity and controllability of generated content while addressing challenges in accurately rendering complex motion dynamics through techniques like search-based generation pipelines and motion prior distillation. These advancements pave the way for more realistic and diverse video synthesis, expanding creative possibilities across various artistic styles and applications [19, 20, 15, 21]. This integration exemplifies the transformative potential of combining diffusion processes with transformer architectures, resulting in enhanced video generation capabilities.

In the domain of embodied intelligence, neural networks, including convolutional and other deep architectures, are fundamental for pattern recognition and data classification, forming the backbone of intelligent systems that interact with the physical world [12]. These networks are complemented by multi-modal learning techniques, which facilitate the integration of diverse data types, enhancing the system's understanding and generation of complex content [22]. Techniques such as Multimodal Chain of Thought (M-COT) and Multimodal Instruction Tuning (M-IT) further augment these systems' capabilities by enabling more nuanced and context-aware interactions [7].

The benchmark for evaluating video generation models across various scenarios, including in-distribution, out-of-distribution, and combinatorial generalization, underscores the significance of robust evaluation metrics in advancing these technologies [23]. Additionally, the development of text-to-video generation models, with diverse building blocks and auxiliary features, exemplifies ongoing innovation in creating more sophisticated and versatile generative systems [5].

Ultimately, the convergence of these technologies—GANs, diffusion models, transformers, and neural networks—along with advanced evaluation frameworks, is propelling the field of video generation and embodied intelligence toward new frontiers of capability and application [24].

1.3 Structure of the Survey

This survey is systematically organized to comprehensively explore video generation models and embodied intelligence techniques, emphasizing their significance and technological foundations. The introductory section discusses the importance of video generation models in AI and machine learning, followed by an overview of key technologies such as GANs, diffusion models, reinforcement learning, neural networks, and autonomous agents. The background and core concepts section delves deeper into these technologies, elaborating on their development, methodologies, and interrelationships in advancing AI capabilities.

Subsequent sections provide detailed analyses of each core technology. The video generation models section examines their evolution, methodologies, applications, and challenges, with a specific focus on comparing GANs and diffusion models. The embodied intelligence section discusses the integration of AI with physical entities and the role of autonomous agents in real-world interactions. The survey then offers an in-depth analysis of GANs, highlighting architectural innovations, training challenges, and applications in video and image generation.

Diffusion models are explored regarding their mechanisms, innovations, and performance, with comparisons to GANs. Reinforcement learning is discussed in the context of its principles, advanced techniques, and applications in autonomous agents, including integration with other AI models. The neural networks section covers their structure, function, and role in visual perception and energy-efficient computing.

The concept of autonomous agents is analyzed, focusing on their adaptability and integration with video generation models. The conclusion synthesizes essential findings, innovative advancements, and prevailing challenges identified in the research, while outlining prospective directions for future investigation aimed at furthering the development of generative video modeling technologies. This includes addressing the intricate interplay between temporal dynamics, dataset quality, and evaluation metrics, as well as the socio-technical relationships that shape trust in AI-generated content. By emphasizing the need for interdisciplinary collaboration, the conclusion advocates for comprehensive studies that enhance the technical capabilities of text-to-video generation and explore its broader implications in fields such as biomedicine and democratic engagement [25, 5, 2, 26]. This structured approach ensures a thorough understanding of the current landscape and future potential of video generation models and embodied intelligence.The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Video Generation Models

Video generation models are pivotal in AI, enabling the synthesis and manipulation of visual content across domains. These models face challenges due to the high-dimensional nature of video data and the limitations of current methods, which often restrict training data [27]. The complexity of video generation requires architectures that learn spatial and temporal dependencies without oversimplifying data [27].

Innovative methodologies have emerged to address these challenges, such as tracklet-conditioned frameworks that model nuanced object movements, tackling issues like occlusion and overlapping. However, ensuring instance-level consistency across frames remains a significant obstacle, crucial for capturing complex temporal dynamics in natural scenarios [28]. Generative models, including GANs and diffusion models, have shown promise in creating realistic synthetic images and videos. For instance, 'GANalyze' uses GANs to generate images with varying cognitive attributes by navigating latent space, highlighting nuanced content generation potential. Recent advancements in diffusion models have improved their capacity to generate high-quality videos with motion and content consistency across different camera trajectories. Models like Control-A-Video integrate motion and content priors to enhance coherence and reduce flickering, while the Collaborative Video Diffusion framework employs cross-video synchronization to ensure consistency between frames from various angles, thus improving multi-video generation quality [10, 29, 30].

The applications of video generation models are extensive, from simulating crowd dynamics to generating human motion sequences. These tasks require models to replicate data distributions while

adhering to domain-specific objectives, such as maintaining the quality and structure of generated samples [31]. However, generating realistic human motion sequences remains challenging due to reliance on small, homogeneous datasets, limiting the production of high-quality multi-person interactions [32].

Recent advancements include generating long-duration animated videos from static images while ensuring motion consistency [33]. The creation of realistic talking head videos from a single identity frame and audio sequence further illustrates ongoing efforts to address realistic representation and temporal coherence issues [34]. Moreover, multimodal diffusion models have been integrated to generate high-quality videos of humans speaking and moving based on audio input from a single image, showcasing the potential for complex content generation [35].

Despite these advancements, challenges persist. The difficulty in generating high-quality videos from textual descriptions demands larger datasets and greater computational resources than text-to-image generation [11]. Additionally, generating high-fidelity and controllable motions for virtual characters in real-time is critical for interactive applications [36]. The high costs and resource requirements associated with training models on large-scale video datasets further constrain progress [19].

Video generation models also grapple with mode collapse in GANs, where the model produces limited variations, curbing content diversity [37]. The reliance on training data restricts these models' ability to generate unseen actions, as they struggle with contextual learning [38]. Furthermore, the absence of a unified multimodal model and challenges in achieving coherent reasoning across modalities remain significant hurdles [7].

The ongoing development of video generation models aims to overcome data scarcity and model limitations, with applications spanning crowd simulation to human motion generation. As research progresses, these models are poised to play an increasingly integral role in diverse domains, fostering innovation and application. The integration of cognitive architectures and generative models can enhance intelligent agents' capabilities, enabling them to emulate human-like cognition and creativity more effectively [39].

2.2 Embodied Intelligence

Embodied intelligence represents a paradigm shift in AI, emphasizing the interaction of systems with physical environments to execute complex tasks. This approach is crucial for developing adaptive AI systems that respond effectively to dynamic and unpredictable environments, addressing the limitations of traditional machine learning methods that often struggle to generalize across novel tasks and settings [40]. Embodied intelligence requires purposeful exchanges of energy and information with the physical world, presenting unique challenges not typically encountered in conventional machine learning applications [40].

A critical aspect of embodied intelligence involves integrating large-scale models capable of processing multi-view images and text inputs to enhance decision-making in complex environments. LARM, a large-scale embodied AI model, exemplifies this integration, addressing the shortcomings of existing methods that rely on large language models (LLMs) for real-world interactions [41]. Traditional LLM-based approaches often suffer from ambiguous action outputs and slow response times due to multiple inference operations, underscoring the need for more efficient models that can operate seamlessly in real-time scenarios [41].

The development of embodied intelligence is further challenged by the necessity for systems to adapt to diverse and evolving environments. Current machine learning techniques often exhibit brittleness and inefficiency when deployed in real-world scenarios, highlighting the importance of designing AI systems that can learn and adapt continuously [40]. This adaptability is essential for creating AI systems capable of functioning across a wide range of scenarios, thereby driving advancements in machine learning technology beyond mere application domains [40].

2.3 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) have transformed generative modeling by introducing a framework comprising two adversarial components: a generator and a discriminator. The generator aims to create data samples indistinguishable from real data, while the discriminator seeks to

accurately differentiate between real and generated samples. This adversarial process fosters iterative improvements in the quality of generated data [42].

Despite their transformative impact, GANs encounter challenges regarding training stability and data diversity. Unstable training can lead to mode collapse, where the generator produces limited sample variations. To mitigate these issues, improved training techniques and architectural innovations, such as new loss functions and features, have been proposed to enhance GAN performance [42]. The integration of Transformer networks into GAN architectures offers a promising avenue for overcoming the limitations of traditional convolutional approaches, as Transformers excel at capturing global relationships within data, unlike Convolutional Neural Networks (CNNs), which focus on local relationships [20].

The versatility of GANs extends to various applications, including scenarios with incomplete conditioning vectors. Partially Conditioned Generative Adversarial Networks (PCGANs) have been developed to generate samples under such conditions without estimating missing entries, broadening GAN applicability in real-world scenarios [43]. Additionally, GANs' use in creative domains, such as character design, showcases their potential to augment human creativity by generating novel visual concepts beyond mere realism.

Since their introduction in 2014, GANs have established themselves as a foundational technology in generative modeling, driving significant advancements in generating realistic and diverse synthetic data across fields like computer vision and artificial intelligence. Their unique architecture, involving a generative and a discriminative network engaged in a Minimax game, has led to numerous variants—such as conditional GANs, Wasserstein GANs, CycleGANs, and StyleGANs—that enhance sample quality and training stability. GANs are recognized as a leading breakthrough technology and have integrated with modern deep learning frameworks like Transformers and Diffusion models, paving the way for innovative applications in image synthesis, video prediction, and virtual environment creation. As research continues to address challenges in training efficiency and model interpretability, GANs remain at the forefront of synthetic data generation, continually expanding their influence and capabilities [44, 45, 46, 47]. The ongoing development of GAN architectures, including Transformer integration, underscores their critical role in advancing artificial intelligence through realistic and diverse data generation across various applications.

2.4 Diffusion Models

Diffusion models signify a significant advancement in generative modeling, characterized by their structured noise addition and denoising process. They begin by introducing noise to data samples in a forward process and then reverse this process to reconstruct the original data distribution, exemplified by Denoising Diffusion Probabilistic Models (DDPMs) [48]. This probabilistic framework has proven instrumental in achieving high realism and fidelity in both image and video generation tasks, enhancing data quality [49].

The versatility of diffusion models is further demonstrated through their applications in text-to-video (T2V) and image-to-video (I2V) generation, enabling the creation of visually coherent and temporally consistent content. Innovations like Diffusion4D have integrated spatial-temporal consistency within a single model, contrasting with previous methods that treated these aspects separately, thereby enhancing the coherence of generated video content [10].

Despite their advantages, diffusion models are computationally intensive, particularly in video generation, often requiring thousands of model evaluations, which can be impractical for real-time applications [50]. To alleviate these computational demands, methodologies combining diffusion models with GANs through knowledge distillation have been proposed, aiming to reduce sampling time while maintaining high-quality outputs [50]. Additionally, the application of diffusion models in generating high-definition, text-conditional videos through a cascade of video diffusion models exemplifies their capability to integrate spatial and temporal super-resolution techniques, thereby enhancing video quality [21].

Innovative approaches have also reframed denoising as a multi-step decision-making task, leading to algorithms like denoising diffusion policy optimization (DDPO), which aligns diffusion models with reinforcement learning principles, broadening their applicability [51]. However, the intractability of exact likelihood computation in diffusion models presents challenges for conventional reinforcement learning applications [51].

The integration of diffusion models with other AI technologies, such as transformer-based architectures, highlights ongoing efforts to expand their creative potential and application scope [20]. These advancements demonstrate significant strides in diffusion-based video generation models, offering enhanced user control and improved temporal coherence [52].

Diffusion models continue to advance generative modeling, providing robust solutions for creating and refining high-quality data across diverse applications. Their ongoing development, particularly in enhancing computational efficiency and integrating with other AI technologies, promises to expand their effectiveness and applicability in various fields, including complex, multi-dimensional content generation [53].

2.5 Reinforcement Learning

Reinforcement learning (RL) is a pivotal paradigm in AI, focusing on training agents to make sequential decisions by maximizing cumulative rewards through environment interactions. The foundational principle of RL involves an agent navigating a state-action space, where the environment provides feedback in the form of rewards or penalties, facilitating the learning of optimal policies [54]. This framework has significantly advanced autonomous decision-making across various complex domains.

A major challenge in reinforcement learning is the dependence on high-dimensional observations and the necessity for reward or action labels, which often limits the applicability of traditional RL methods in real-world scenarios [54]. To address these challenges, innovative approaches such as the Diffused Value Function (DVF) have been proposed, leveraging diffusion models to learn environment dynamics and estimate value functions without requiring reward or action data during training, thereby broadening RL applicability [54].

Recent advancements in RL emphasize integrating generative models to enhance decision-making capabilities. For instance, the Actor-Critic under Adversarial Learning (ACtuAL) framework reframes the GAN paradigm by employing a learned critic instead of relying solely on gradients from the discriminator, enabling effective training of generative models on discrete sequences [55]. Additionally, the use of GANs in distributional reinforcement learning has been explored to improve performance in complex Markov Decision Processes (MDPs), demonstrating GANs' potential to address computational challenges in high-dimensional settings [56].

Furthermore, integrating diffusion models with RL frameworks, such as the Temporally-Composable Diffuser (TCD) method, has shown significant improvements in generating temporal conditions from interaction sequences, enhancing the overall generation process in RL [57]. The incorporation of diffusion policies with GANs in offline RL settings enables learning optimal policies from pre-collected datasets without direct environmental interaction, reducing the need for extensive and costly data collection [58].

Hierarchical reinforcement learning algorithms, such as the Two-Stage Hierarchical Training (TSHT), have been developed to manage complex control tasks by utilizing a hierarchical action space representation. This approach effectively controls full-body musculoskeletal models, showcasing hierarchical RL's potential in managing high-dimensional action spaces [59].

In decision-making applications, RL is increasingly integrated with multi-modal data inputs, including text, images, and sensor data, to enhance the robustness and adaptability of autonomous agents. This integration poses challenges in reducing latency for real-time applications and improving training efficiency without compromising performance, as highlighted in recent studies on embodied navigation [60].

Moreover, combining RL with behavior cloning techniques, supported by generative models, has been shown to enhance the robustness of learned behaviors, providing provable guarantees for improved performance in complex environments [61]. This integration is crucial in scenarios where high-quality demonstrations are scarce or costly to obtain, necessitating innovative value function estimation methods to ensure effective learning [54].

The field of reinforcement learning is advancing through the incorporation of advanced generative models, such as Generative Diffusion Models (GDMs) and Generative Adversarial Networks (GANs), alongside hierarchical frameworks, effectively addressing critical challenges in decision-making applications across various domains, including network optimization, robotics, and self-driving

technologies [62, 56, 63]. These advancements pave the way for more robust and efficient autonomous systems capable of operating in diverse and dynamic environments.

2.6 Neural Networks and Autonomous Agents

Neural networks, inspired by the human brain's structure and function, are foundational to developing intelligent systems capable of pattern recognition, decision-making, and data classification. These networks, particularly deep learning architectures, have revolutionized various domains by enabling machines to learn complex patterns from vast datasets. Neural network architecture typically involves layers of interconnected nodes, or neurons, processing inputs to produce outputs through weighted connections. This layered structure allows for hierarchical feature extraction, making neural networks particularly effective in tasks such as image and speech recognition [64].

Recent advancements in neural network architectures focus on enhancing efficiency and adaptability. The integration of Spiking Neural Networks (SNNs) into neuromorphic computing exemplifies this trend, offering a biologically-inspired approach that mimics the brain's energy-efficient information processing capabilities. SNNs are particularly promising for robotics applications, where real-time processing and low power consumption are critical [65]. These networks leverage spike timing for information encoding, providing a robust framework for developing embodied intelligence systems that interact with the physical world.

Autonomous agents, capable of operating independently to perform tasks in various environments, rely on neural networks to process sensory inputs, make decisions, and execute actions, often in dynamic and unpredictable settings. The development of world models, such as PolyGRAD, represents a significant advancement in this area, enabling agents to generate on-policy trajectories by diffusing initially random states and actions into coherent sequences through learned denoising models and policy gradients [66]. This approach enhances the agent's ability to anticipate future states and adapt its behavior accordingly, improving performance in complex environments.

The convergence of neural networks with autonomous agents has facilitated the integration of natural language processing capabilities, as demonstrated by systems like 'Miron.' This system unifies natural language understanding and generation through a shared structural model, enhancing the agent's ability to interact with humans intuitively and contextually [67]. Furthermore, categorizing research into autoregressive methods and Joint Embedding Predictive Architecture (JEPA) methods highlights ongoing efforts to develop efficient world models that underpin the decision-making processes of autonomous agents [7].

In recent years, video generation models have garnered significant attention within the field of artificial intelligence, particularly due to their diverse applications and evolving methodologies. As illustrated in Figure 2, the hierarchical structure of these models is categorized into four key areas: development, methodologies, applications, and a comparative analysis between Generative Adversarial Networks (GANs) and diffusion models. This figure not only highlights the generative techniques employed in video generation but also elucidates the challenges and innovations that accompany these methodologies. Furthermore, it provides a comprehensive overview of the advantages and limitations inherent to different models, thereby enhancing our understanding of the current landscape in video generation research. Such visual representation serves to enrich the narrative by offering a concise summary of the complexities involved in this rapidly advancing field.

3 Video Generation Models

3.1 Development of Video Generation Models

Video generation has evolved from procedural methods to sophisticated frameworks leveraging deep learning and generative techniques. Early procedural content generation (PCG) methods, primarily used in gaming, faced limitations in content diversity due to evolutionary algorithm constraints [68]. The advent of Generative Adversarial Networks (GANs) marked a pivotal shift, enabling realistic and varied data synthesis across domains, from synthetic log data to artistic video enhancements [69].

Diffusion models have further advanced video generation, offering robust frameworks for unified representation learning that facilitate both classification and generation tasks. TrackDiffusion exemplifies this by generating continuous video sequences from tracklets, enhancing video quality and

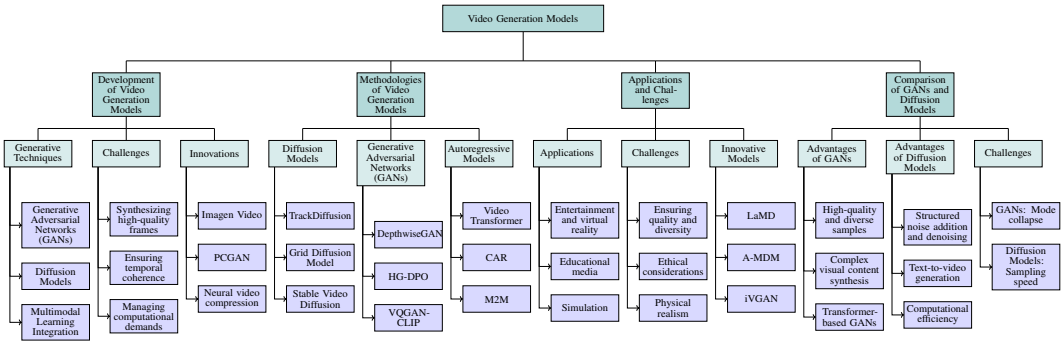


Figure 2: This figure illustrates the hierarchical structure of video generation models, categorizing them into development, methodologies, applications, and comparisons between GANs and diffusion models. It highlights generative techniques, challenges, and innovations in video generation, as well as the advantages and limitations of different models.

Method Name	Generative Techniques	Challenges Addressed	Applications
CGAN[68]	Conditional Gans	Piece Distribution	Game Levels
HCGANs[69]	Cross-modal Gans	Complex Data Distributions	Semi-autonomous Ground Vehicles
TD[28]	Diffusion Models	Temporal Coherence	Animated Storytelling
GDM[11]	Grid Diffusion Model	Temporal Consistency	Text-to-video
VC[70]	Diffusion Models	High-quality Frames	Video Generation
SG[52]	Diffusion Models	Temporal Coherence	Surgical Education
AICL[38]	Video Diffusion Models	Diverse Actions Generation	Open-domain Scenarios
DepthwiseGAN[71]	Depthwise Separable Convolutions	Extensive Computational Resources	Image Generation
CGAN[72]	Cyclegan	Paired Data	Remote Sensing
VIDM[73]	Diffusion Models	Temporal Coherence	Video Generation
iVGAN[27]	Wasserstein Gan	Temporal Coherence	Video Colorization
IV[21]	Diffusion Models	Temporal Coherence	Creative Applications
PCGAN[43]	Conditional Gans	Incomplete Conditioning Information	Digit Synthesis
NVC-GAN[74]	Gans	High-quality Frames	Video Compression

Table 1: This table presents an overview of various video generation methods, highlighting the generative techniques employed, the challenges addressed, and their respective applications. It includes a range of models such as Conditional GANs, Diffusion Models, and CycleGAN, showcasing their utility in diverse fields like game level design, video generation, and surgical education. The table underscores the evolution and versatility of these methods in overcoming specific challenges such as temporal coherence and computational resource demands.

temporal consistency [28]. The Grid Diffusion Model (GDM) improves efficiency by using grid images, reducing computational costs [11]. VideoCrafter employs two diffusion models to generate high-quality videos from text or image inputs while maintaining structural integrity [70]. SurGen exemplifies the potential of diffusion models by producing high-resolution surgical videos without real data inputs [52].

Multimodal learning integration has further propelled video generation. AICL (Action In-Context Learning) enables video diffusion models to generate actions using reference videos, enhancing generative performance [38]. DepthwiseGAN improves training efficiency with depthwise separable convolutions while maintaining image quality [71]. CycleGAN's application in remote sensing imagery illustrates GANs' versatility [72].

Despite these advancements, challenges persist in synthesizing high-quality frames, ensuring temporal coherence, and managing the computational demands of video generation [14]. The Video Implicit Diffusion Model (VIDM) addresses these by modeling content and motion separately with two diffusion models [73]. iVGAN generates videos from latent codes without separating background and foreground, effectively handling moving elements [27].

Innovations like Imagen Video, a text-conditional video generation system using a cascade of video diffusion models, facilitate high-definition video creation from textual prompts [21]. PCGAN addresses traditional conditioning method limitations by generating synthetic data from incomplete conditioning information [43]. Techniques in neural video compression further illustrate advancements in video generation and compression [74].

As research progresses, video generation models, particularly those leveraging GANs, are set to become increasingly vital across media, education, and entertainment sectors. Innovations in generative techniques enable the production of personalized, high-quality videos, with VQGAN-CLIP's text-to-video applications showcasing potential for OTT platforms and visual arts. Enhancements in GAN architectures expand video generation capabilities, allowing robust applications like video colorization and inpainting. The continuous evolution of these models aims to overcome existing challenges while broadening applications in gaming, entertainment, simulation, and training [14, 27, 15].

Figure 3 illustrates the hierarchical structure of video generation models, categorizing them into generative techniques, challenges, and applications. This figure highlights key advancements in GANs, diffusion models, and multimodal learning, while also addressing challenges such as frame quality, temporal coherence, and computational demands. Additionally, applications in media, surgical education, and remote sensing are depicted. Table 1 provides a comprehensive overview of the development of video generation models, detailing the generative techniques they utilize, the challenges they address, and their applications across various domains.

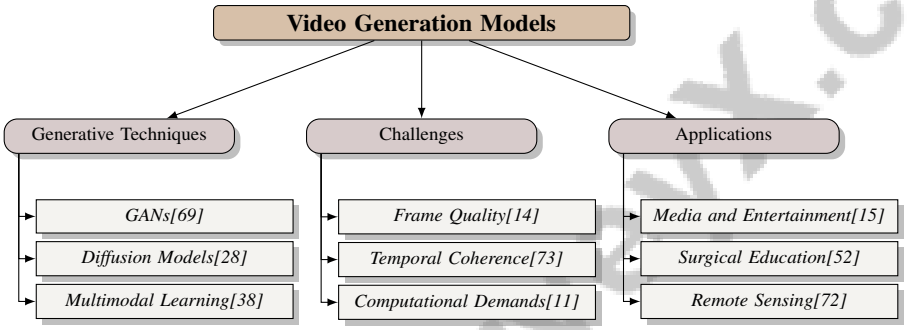


Figure 3: This figure illustrates the hierarchical structure of video generation models, categorizing them into generative techniques, challenges, and applications. It highlights key advancements in GANs, diffusion models, and multimodal learning, while also addressing challenges such as frame quality, temporal coherence, and computational demands. Applications in media, surgical education, and remote sensing are also depicted.

3.2 Methodologies of Video Generation Models

Method Name	Generative Frameworks	Temporal Consistency	Architectural Design
TD[28]	Diffusion Models	Temporal Instance Enhancer	Gated Cross-attention
GDM[11]	Grid Diffusion Model	Interpolating Between Frames	Grid Image Representation
DepthwiseGAN[71]	Depthwise Separable Convolutions	-	Depthwise Separable Convolutions
HG-DPO[75]	-	-	Lora Weight
NTS[76]	Progressivegan	-	Locally Linear Approximations
VT[77]	Autoregressive Model	Smooth Transitions	Block-local Self-attention
CAR[78]	Autoregressive Models	-	Multi-scale Latent
M2M[79]	Diffusion Model	Maintain Coherence	Autoregressive Approach
MM[80]	Diffusion Models	Temporal Attention Maps	Temporal Camera Module
iVGAN[27]	Wasserstein Gan	Spatial Temporal Dependencies	One-stream Framework
CGAN-PSG[81]	Conditional Gans	Paired Sample Generation	Additional Training Step
VQGAN-CLIP[15]	Vqgan-CLIP Model	Coherent Video Sequences	Image Generation Evaluation

Table 2: Overview of various video generation models, detailing their generative frameworks, temporal consistency techniques, and architectural designs. This table highlights the diversity in methodologies, ranging from diffusion models and generative adversarial networks (GANs) to autoregressive models, each contributing uniquely to advancements in video generation technology.

Video generation methodologies leverage advanced generative frameworks to achieve high-quality, temporally consistent outputs. Diffusion models have emerged as a prominent approach, with TrackDiffusion employing diffusion models to generate videos conditioned on object trajectories, ensuring frame-to-frame consistency [28]. The Grid Diffusion Model generates key grid images from textual descriptions, interpolating between them to create smooth video frames [11].

Generative Adversarial Networks (GANs) remain foundational, with techniques like DepthwiseGAN utilizing depthwise separable convolutions to enhance training efficiency while maintaining output

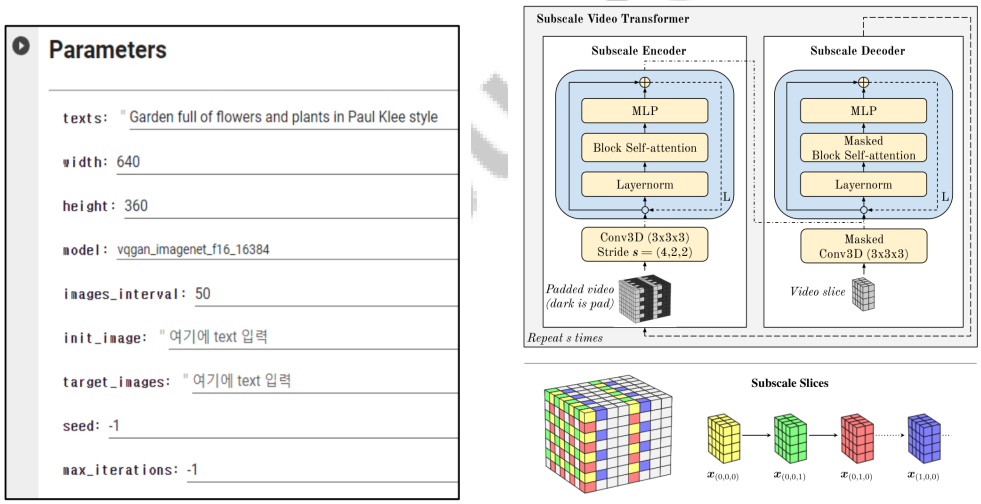
quality [71]. The HG-DPO method aligns model outputs with human preferences, improving generated human image quality [75]. Integrating GANs with reinforcement learning frameworks extends their applicability, as seen in Markov Decision Process (MDP) formulations over pre-trained GAN latent spaces for tasks like age manipulation, ensuring identity preservation [76].

Autoregressive models have significantly advanced video generation. The Video Transformer employs a three-dimensional self-attention mechanism to process video as a spatiotemporal volume, enhancing temporal coherence [77]. CAR introduces multi-scale latent variable modeling into pre-trained autoregressive models, allowing for greater control over the visual generation process [78].

Innovative autoregressive approaches like M2M generate interrelated images from input sets, ensuring coherence across generated frames [79]. MotionMaster disentangles camera from object motions, enhancing realism and dynamism [80].

Generating RGB videos from low-dimensional latent codes exemplifies latent space manipulation potential [27]. Maintaining a consistent input vector while varying conditions to generate multiple samples, which are merged to create a single object representation, illustrates sophisticated video generation techniques [81].

Video generation methodologies continue to evolve, integrating cutting-edge techniques to overcome coherence, quality, and computational efficiency challenges. Innovations enhance capabilities across creative content production, educational media, and realistic simulations. Advancements like VQGAN-CLIP enable personalized AI-generated videos, while Stable Video Diffusion improves visual consistency and natural movement. Systems like Imagen Video leverage text prompts for high-definition videos with diverse artistic styles, highlighting these technologies' versatility in artistic and commercial applications [14, 82, 15, 21].



(a) Parameters for Image Generation[15] (b) Subscale Video Transformer[77]

Figure 4: Examples of Methodologies of Video Generation Models

As illustrated in Figure 4, video generation methodologies include parameter settings for image generation and subscale video transformer architecture. "Parameters for Image Generation" details model configuration, showcasing adjustable parameters like text prompts and dimensions, emphasizing precision in visual content generation. The "Subscale Video Transformer" illustrates sophisticated video generation through its architectural design, involving a subscale encoder and decoder processing padded video inputs via convolutional and self-attention layers, reconstructing video slices through masked convolutional layers. These methodologies exemplify the balance between parameter customization and architectural design in dynamic video content creation [15, 77]. Table 2 presents a comprehensive comparison of video generation methodologies, emphasizing their underlying generative frameworks, temporal consistency strategies, and architectural designs.

Method Name	Application Domains	Technical Challenges	Model Capabilities
N/A[32]	Motion Generation	Limited Training Data	Complex Human Motions
GD[83]	Image Generation	Computational Intensity	Unified Representation Learners
CVD[29]	3D Scene Generation	Multi-view Dynamic Data	Consistent Video Pairs
KUB[84]	Synthetic Video Production	Physical Motion, Lighting	Realistic Animations
TiV-ODE[8]	Controllable Video Generation	Continuous Dynamical System	Flexible Frame Rates
SALAD[85]	3D Shapes	High-resolution Outputs	Intuitive Manipulation
IV[86]	-	Video Quality	Video Quality
CDM[87]	Virtual 3D Environments	Linguistic Ambiguities	Generate Diverse Outputs
GANO[88]	Natural Sciences	Computational Complexity	Generate Function Samples
CCAN[89]	Art Generation	Training Stability	Style-ambiguous Portraits
LaMD[90]	Video Generation Tasks	Temporally Incoherent Outputs	Coherent And Realistic
A-MDM[36]	Interactive Experiences	Unnatural Motions	Real-time Motion
iVGAN[27]	Video Colorization	Scene Dynamics Dependencies	Generate High-quality Videos
VIDM[73]	Entertainment, Simulation, Education	Spatial Temporal Changes	Realistic Continuous Sequences
PCGAN[43]	-	Model Collapse	Generate Samples
CGAN-PSG[81]	3D Models	Model Collapse	Consistent Object Representations

Table 3: This table provides a comprehensive overview of various video generation models, detailing their application domains, technical challenges, and model capabilities. It highlights the diverse range of methods employed in fields such as motion generation, image synthesis, and video production, alongside the specific challenges they address, including computational intensity, training stability, and model collapse. The table serves as a valuable resource for understanding the current landscape of video generation technologies and their potential applications.

3.3 Applications and Challenges

Video generation models are pivotal across entertainment, simulation, and education, enhancing animation and virtual reality by creating immersive environments and lifelike characters. However, achieving functional gameplay elements presents challenges due to approximating game element distributions [32]. Generating realistic human motion sequences is vital for animation and virtual reality, with motion-based priors improving control over movement [91].

Unified video representations offer a cohesive approach to video synthesis, facilitating effective real-world applications [63]. Diffusion models leverage a single pre-trained model for generative and discriminative tasks, achieving competitive performance without extensive retraining [83]. Collaborative Video Diffusion (CVD) maintains consistent content and dynamics across different camera views, crucial for 3D scene generation and video synthesis [29].

Challenges persist in ensuring quality, diversity, and ethical considerations in video generation. Existing methods struggle with physical motion, lighting, and camera dynamics, hindering coherent and visually appealing video production [84]. Techniques like TiV-ODE face limitations due to time-consuming training and potential frame blurriness [8]. SALAD's reliance on initial part-level decomposition quality illustrates challenges in achieving optimal generation [85].

Text-to-video generation remains challenging, particularly in creating visually appealing outputs due to video diffusion model and training data quality limitations [86]. The Camera Diffusion Model generates diverse and realistic camera motions, addressing linguistic ambiguities and user control in cinematography [87]. Ensuring safe and unbiased video generation, accurately depicting complex actions and facial expressions, and maintaining physical realism remain ongoing challenges [92].

Generative adversarial neural operators (GANO) demonstrate the ability to learn and generate function samples in infinite-dimensional spaces, outperforming traditional GANs in controlled and real-world scenarios [88]. However, challenges persist in achieving high-resolution outputs without model collapse, as evidenced by limitations in jointly trained models [93]. The complexity and training time constraints of current implementations often result in reduced detail compared to baseline models [89].

Addressing these challenges is crucial for advancing video generation technologies in real-world scenarios. As research evolves, these models are expected to play an increasingly vital role across domains, offering new avenues for innovation and application. The LaMD model, evaluated on datasets like BAIR Robot Pushing and CATER-GENs, demonstrates potential for image-to-video and text-image-to-video generation tasks, highlighting modern video generation techniques' adaptability [90]. The A-MDM method showcases the capability to generate diverse, high-fidelity motion sequences in real-time, outperforming existing methods in motion quality and responsiveness [36]. Generative models in recommendation systems highlight their potential to outperform traditional

methods by generating more diverse and personalized outputs, especially in data-scarce scenarios [24].

The iVGAN model demonstrates superior performance in generating videos with static and dynamic backgrounds [27]. VIDM creates realistic, continuous video sequences by modeling content and motion separately [73]. PCGAN exhibits robustness to missing conditioning information, producing higher quality samples under challenging conditions [43]. The proposed method generates paired 3D models with conditional GANs, significantly reducing the average difference between generated models and producing noise-free high-resolution instances [81]. Table 3 presents a detailed summary of video generation models, focusing on their application domains, technical challenges, and inherent capabilities, thus providing a clear context for the subsequent discussion on their implementation and performance.

3.4 Comparison of GANs and Diffusion Models

Generative Adversarial Networks (GANs) and diffusion models are foundational frameworks in generative modeling, each with distinct advantages and challenges. GANs, characterized by their adversarial training process involving a generator and a discriminator, excel in producing high-quality and diverse samples. Their architecture allows for complex visual content synthesis, particularly effective in creative applications like character design, enhancing designers' creativity [94]. Transformer-based GANs outperform traditional GANs in generating high-quality images and videos, underscoring their potential in fidelity and diversity-demanding applications [20]. However, GANs often suffer from mode collapse, where the generator fails to capture the data distribution's full diversity. MIC-GANs address this by inferring the number of modes and avoiding mode collapse, improving image quality from complex distributions [95].

Conversely, diffusion models have gained traction for their structured noise addition and denoising process, enabling high realism and fidelity in data generation. These models excel in text-to-video (T2V) generation, maintaining content consistency and coherence across frames. Integrating Gaussian latent spaces and DPM-Encoders enhances diffusion models' applicability in various generative tasks [96]. Diffusion models offer significant computational efficiency, with adversarial knowledge distillation approaches reducing model size and improving sampling speed, making them more practical for real-world applications [50]. The Grid Diffusion Model (GDM) exemplifies diffusion models' effectiveness in video generation, outperforming traditional methods in quantitative and qualitative evaluations [62].

While GANs are renowned for generating high-quality and diverse content, diffusion models excel in maintaining temporal coherence and content consistency in video generation applications. Ongoing development of both frameworks continues to enhance their capabilities, with diffusion models increasingly integrated with other AI technologies to expand their creative potential and application scope [20]. As research progresses, integrating GANs and diffusion models into various application areas is expected to broaden, offering new opportunities for innovation in generative modeling.

4 Embodied Intelligence

4.1 Integration of AI with Physical Entities

The integration of artificial intelligence (AI) with physical entities marks a significant stride in embodied intelligence, enhancing interaction capabilities in real-world contexts. Systems like VidMan leverage video diffusion models to analyze environmental dynamics, thereby optimizing robotic interactions [97]. Generative Adversarial Networks (GANs) enhance learning processes in semi-autonomous ground vehicles, improving navigation in complex environments [69].

Efficient domain augmentation techniques in Autonomous Driving Systems (ADS) enhance Operational Design Domain (ODD) diversity and realism, crucial for failure detection and safe AI deployment [98]. The Curious Representation Learning (CRL) framework generates diverse data for representation learning models, enhancing robustness and transferability [99]. The Action In-Context Learning (AICL) method demonstrates AI's flexibility in open-domain scenarios, generating complex actions beyond initial training data [38]. The LARM model enhances decision-making through multi-modal inputs, improving response times in dynamic environments [41].

Embodied intelligence benefits from diverse morphologies and control tasks facilitated by frameworks like DERL, promoting seamless AI integration with physical entities [100]. Studies emphasize the need for embodied agents to learn within physical constraints and adapt to non-stationary environments [40]. This integration enhances AI systems' perception, understanding, and interaction with their environments. By combining cognitive architectures with generative models, these systems navigate complex scenarios more adeptly. The incorporation of Large Language Models (LLMs) refines decision-making and environmental perception, enabling efficient task execution. This approach broadens AI's operational capabilities, addressing challenges related to real-time interaction and learning in dynamic settings [40, 65, 39, 101, 60]. These developments pave the way for sophisticated, adaptive AI systems capable of complex tasks in dynamic environments.

4.2 Autonomous Agents and Interaction

Autonomous agents are pivotal in interacting with environments, executing complex tasks, and adapting to dynamic scenarios. They utilize advanced generative models and multimodal frameworks to enhance decision-making and efficiency. Integrating AI with full-body musculoskeletal models improves human-robot interaction, fostering natural collaboration across applications [59]. Multimodal Environment Memory (MEM) modules enhance this integration by combining visual and linguistic memories, enabling actionable plans from diverse inputs [101].

Autonomous agents' ability to generate diverse and realistic trajectories in real-time is crucial, facilitating efficient task performance. This adaptability is vital for applications like crowd simulation, where agents navigate dynamic environments while maintaining realistic behavior [31]. Generative Diffusion Models (GDMs) demonstrate high-quality data generation, enhancing autonomous systems through flexibility and simplicity [62]. Recent advancements in interactive character control showcase agents' capacity for diverse motions in real-time, adapting to various strategies without additional training, essential for virtual reality and gaming [36]. Integrating text-guided creative generation and image editing broadens applicability, enabling more context-aware interactions [102].

By integrating cognitive architectures and generative models, autonomous agents enhance interaction capabilities, adaptively performing tasks in complex environments. They leverage multimodal perception to translate high-level tasks into actions, improving decision-making and navigation. Evolutionary design principles enhance learning and adaptation, making them effective in diverse scenarios [39, 100, 60, 101]. The development of multimodal frameworks and generative models drives innovation, paving the way for sophisticated autonomous systems.

4.3 Human-AI Collaboration

Human-AI collaboration in embodied intelligence systems enhances AI capabilities and adaptability. Systems like Miron exemplify this by offering scalability and design through a unified template system, resulting in intuitive user interactions [67]. Neuromorphic systems, emulating brain processing, enhance robotic applications, improving energy efficiency, adaptability, and learning capabilities [65]. These systems foster human-AI collaboration, enabling real-time learning and adaptation in unpredictable environments.

Integrating reasoning and generative capabilities enhances AI effectiveness in collaboration [39]. Universal perceptual value representations (PVRs) generalize across AI tasks, offering a holistic approach to collaboration [103]. In creative domains, AI systems like GANterpretations facilitate dynamic content creation linked to audio, fostering new artistic expression mediums and collaboration between AI and designers [104]. This approach augments creativity and establishes collaborative design processes, as seen in benchmarks promoting synergy between AI and designers [94].

The LARM model exemplifies AI's potential in long-horizon scheduling tasks, such as crafting in Minecraft, showcasing robust scheduling capabilities and potential for collaboration in complex processes [41]. Future research should focus on developing architectures and evaluation methods tailored to embodied agents in dynamic environments, ensuring effective collaboration across diverse scenarios [40]. Human-AI collaboration in embodied intelligence systems catalyzes advancements across fields. This synergy leverages cognitive architectures and generative models to create capable agents, opening avenues for innovation, particularly in navigation and robotics, where AI's understanding of complex environments is crucial. By integrating machine learning techniques and

neuromorphic computing, these systems become adept at real-time decision-making and perception, driving AI technologies' evolution and applications [40, 65, 39, 60, 25].

5 Generative Adversarial Networks (GANs)

5.1 Architectural Innovations in GANs

Recent developments in Generative Adversarial Networks (GANs) have significantly improved their architecture and performance, expanding their use across multiple domains. Key advancements include feature matching, minibatch discrimination, and virtual batch normalization, which enhance stability and output quality by addressing mode collapse and convergence issues [42]. Partially Conditioned GANs (PCGANs) represent a notable innovation, enabling the use of incomplete data via a feature extraction network, unlike traditional models that require full conditioning [43]. Extending the Wasserstein GAN framework to video generation has further improved convergence and stability, allowing for high-quality video production by effectively managing complex dynamics [27].

In finance, GAN architectures like Deep Convolutional GAN (DCGAN) are evaluated for capturing complex time series characteristics, while others like Self-Attention GAN (SAGAN) and Wasserstein GAN with Gradient Penalty (WGAN-GP) face challenges in generating realistic clusters, highlighting the need for ongoing innovation [105]. These architectural innovations have propelled GANs to greater heights, enhancing their performance and versatility. Future research aims to address current limitations and improve sample quality, expanding GAN applications in computer vision, imaging science, and virtual environments. The integration of GANs with emerging deep learning frameworks promises realistic dataset generation, image translations, and customized virtual worlds, fostering new innovations [44, 46, 47].

5.2 Training Stability and Convergence

Training stability and convergence are critical challenges for Generative Adversarial Networks (GANs), affecting their ability to produce high-quality outputs. The adversarial setup, involving a generator and a discriminator, often results in oscillatory behavior and mode collapse, where the generator fails to capture data diversity. This instability is compounded by the sensitivity of GAN training to parameter changes, hindering convergence [42]. To address these issues, techniques such as feature matching, minibatch discrimination, historical averaging, one-sided label smoothing, and virtual batch normalization have been introduced, stabilizing training and improving convergence [42]. Neuroscore, a benchmark for evaluating GANs based on human perceptual quality, aids in assessing output quality [106].

Despite these advancements, the lack of a unified metric for evaluating generated data quality, especially in specialized domains like finance, remains a challenge [105]. The complexity of manipulating latent spaces in GANs requires careful tuning to achieve desirable outputs without sacrificing stability [107]. Ongoing research is vital for developing robust frameworks to tackle persistent challenges in GAN training and to gain deeper insights into their internal mechanisms and convergence behavior. Recent advancements in architectural features and training techniques that enhance image quality and semi-supervised learning performance are crucial for expanding GAN applicability across diverse fields, ensuring reliable production of high-quality outputs [108, 109, 42, 46, 47].

5.3 Applications in Video and Image Generation

Generative Adversarial Networks (GANs) have become pivotal in video and image generation, offering innovative solutions for creating high-quality, diverse content across various applications. In image synthesis, GANs have excelled on datasets like MNIST, CIFAR-10, SVHN, and ImageNet, achieving state-of-the-art results in semi-supervised learning and high-quality image generation through advanced training techniques [42]. They are also effectively used for generating synthetic data in scenarios with scarce real data, which is particularly valuable for financial modeling [105]. This capability is crucial in domains with limited data availability, enabling the creation of realistic models that support decision-making processes.

For video generation, GANs produce dynamic sequences that enhance applications in creative storytelling and content creation. Integration with multimodal datasets facilitates the synthesis of complex video narratives, demonstrating versatility beyond traditional visual domains. GANs have also been used to generate paired 3D models, showing applicability in diverse generative tasks with object classes such as chairs, beds, and sofas [81]. Despite challenges like mode collapse and training instability, ongoing research continues to enhance GAN capabilities. Techniques such as feature matching, minibatch discrimination, and virtual batch normalization improve training stability and output quality [42]. Exploring more parameter-efficient architectures, particularly in DCGANs, could further enhance sample quality and expand GAN applicability to larger datasets.

GANs remain at the forefront of video and image generation, driving innovation and expanding potential applications of AI-generated content. As research progresses, the convergence of GANs with other generative technologies is poised to significantly enhance functionality, paving the way for innovative applications in media, education, and entertainment. This integration supports the generation of high-quality images and videos and facilitates complex tasks such as image-to-image translation, video captioning, and the creation of immersive environments in virtual worlds. Ongoing improvements in GAN frameworks, informed by insights into their internal workings and the resolution of existing limitations, will likely broaden exploration and creativity in content generation [14, 44, 47].

6 Diffusion Models

6.1 Introduction to Diffusion Models

Diffusion models have emerged as a leading class of generative models, excelling in synthesizing high-quality and diverse images by transforming simple prior distributions, typically Gaussian noise, into complex data distributions through iterative denoising processes [51]. The Diffusion of Diffusion (DoD) model enhances this capability by using visual priors from previously generated samples to guide image synthesis [110]. These models are trained using an approximation to the log-likelihood objective, allowing for the generation of highly realistic visual content [51]. Their flexibility is evident in applications such as zero-shot object rearrangement in robotics, eliminating the need for prior dataset-specific training [111].

The theoretical framework of diffusion models is categorized into training-based and sampling-based approaches, essential for understanding their methodologies and applications [48]. Unlike the simpler latent spaces of GANs, diffusion models use sequences of progressively denoised samples as their latent code, providing a more intricate approach to generative modeling [96]. Despite their success in image synthesis, research on detecting diffusion model outputs is less extensive than that for GANs, indicating a gap in understanding and identifying these models' outputs [112]. Thus, diffusion models represent a significant advancement in generative modeling, offering robust solutions for image and video synthesis while presenting new challenges for future exploration [113].

6.2 Mechanisms and Innovations

Diffusion models synthesize high-quality data through iterative noise addition and removal, beginning with noise introduction and followed by a reverse diffusion process to reconstruct the original data distribution. This mechanism is exemplified by Denoising Diffusion Probabilistic Models (DDPM), which generate complex action sequences, showcasing their capacity to learn intricate distributions [51]. Innovations like CycleDiffusion and DPM-Encoder unify the latent space, facilitating effective image-to-image translation and guidance, thereby enhancing content consistency [96]. Temporal conditions have improved video generation applications, as demonstrated by the Temporally-Composable Diffuser (TCD), which enhances temporal coherence through historical, immediate, and prospective conditions [51].

In cinematography, the Camera Diffusion Model generates camera motion sequences from textual descriptions and keyframe constraints, enhancing visual storytelling with controlled camera dynamics, as illustrated by the MotionMaster framework [80]. The ADV-KD method integrates the denoising process into a student model's architecture, enabling high-quality image generation with a single denoising step via adversarial training [50]. The VIDM model separates content and motion modeling, enhancing realistic and continuous video sequence generation while addressing temporal coherence

challenges [73]. Moreover, DDPO, a policy gradient algorithm, optimizes diffusion models for downstream tasks using a black-box reward function, expanding their applicability in reinforcement learning contexts [51].

Despite their capabilities, diffusion models face challenges related to computational efficiency due to their complexity, arising from extensive function evaluations and gradient calculations during training and inference. This overhead limits accessibility and raises concerns about energy consumption and environmental impact, prompting research into more efficient diffusion models [114, 48]. Recent advancements categorize efficient diffusion models based on design choices affecting computational demands, continuing to push the boundaries of generative modeling.

6.3 Applications and Performance

Benchmark	Size	Domain	Task Format	Metric
DMimageDetection[115]	1,000,000	Synthetic Media Forensics	Image Classification	AUC, Accuracy
DM-Deepfake[112]	100,000	Image Synthesis	Deepfake Detection	AUROC, Pd@1PHYLAW[23]
3,000,000	Physics	Video Generation	FVD, SSIM	
CG-Style[93]	32,000	Cartoon Animation	Image Generation	Inception Score, FID
CORTEXBENCH[103]	5,621,987	Robotic Manipulation	Imitation Learning	Mean Success, Mean Rank
SynthMRI[116]	114,749	Medical Imaging	Segmentation	Dice, Hausdorff distance
WGAN-LDW[117]	1,000,000	Causal Inference	Average Treatment Effect Estimation	RMSE, Coverage Rate
Diffusion-Synthetic[118]	5,000	Robotic Manipulation	Keypoint Detection	mAP, AKD

Table 4: This table provides a comprehensive overview of various benchmarks utilized in evaluating diffusion models across different domains. It details the benchmark name, dataset size, application domain, task format, and the specific metrics used for performance assessment. These benchmarks illustrate the diverse applications and evaluation criteria pertinent to diffusion models in fields ranging from synthetic media forensics to medical imaging.

Diffusion models have established themselves as formidable tools in generative modeling, excelling in producing high-quality and diverse data samples across various applications. They have set benchmarks in image synthesis, with innovations like the DoD model achieving an FID-50K score of 1.83 with only 1 million training steps, surpassing other state-of-the-art methods [110]. The CycleDiffusion approach further highlights their potential in unpaired image-to-image translation and zero-shot editing through a unified latent space [96].

In video generation, diffusion models outperform GANs and autoregressive models in generating realistic medical videos [2]. They are capable of creating longer dynamic videos from still images, providing fine-grained motion control and overcoming traditional methods' limitations focused solely on texture objects [33]. Evaluation metrics such as the Fréchet Inception Distance (FID) score assess generated video quality by comparing feature distributions between generated and real videos [1]. Table 4 presents an organized summary of the benchmarks employed to evaluate the performance of diffusion models across multiple domains, highlighting their application scope and assessment metrics.

Beyond visual content generation, diffusion models demonstrate versatility in bioinformatics, addressing challenges in protein design and drug discovery [49]. This adaptability underscores their broad applicability across scientific and industrial domains, where sophisticated data modeling capabilities are essential.

While diffusion models exhibit impressive generative capabilities, challenges related to computational efficiency remain critical. The ADV-KD method addresses these burdens, facilitating diffusion models' use in resource-constrained environments [50]. Improving computational efficiency is vital for wider adoption in practical applications [114]. The choice of diffusion coefficient significantly impacts sample generation quality, with the ODE model outperforming the SDE model when perturbations occur later in the generative process, while the SDE model excels when perturbations happen earlier [119].

Diffusion models continue to advance generative modeling, providing robust solutions for high-quality data synthesis across diverse applications. The ongoing development of generative AI technologies, particularly in computational efficiency and integration with cognitive architectures, holds promise

for enhancing effectiveness in generating complex, multi-dimensional content such as personalized videos and human-like text. These advancements aim to create more capable AI systems by leveraging cognitive models that mimic human processes alongside generative algorithms, ultimately broadening their applicability and impact in real-world scenarios [39, 15].

7 Reinforcement Learning

7.1 Principles and Frameworks of Reinforcement Learning

Reinforcement learning (RL) is a fundamental aspect of artificial intelligence, focusing on training agents to maximize cumulative rewards via sequential decision-making within a Markov Decision Process (MDP). Agents explore a state-action space, receiving feedback as rewards or penalties, which informs the development of optimal policies [54]. RL's core principle is the agent's capacity to learn from both immediate and delayed rewards, fostering strategies that maximize long-term gains.

Recent advancements, such as the Online Decision MetaMorphFormer (ODM), integrate generative models into RL, drawing from cognitive and behavioral psychology to enable agents to learn through a blend of offline demonstrations and online interactions, resulting in adaptable policies [120]. The Diffused Value Function (DVF) innovatively estimates value functions by modeling the discounted state occupancy measure, addressing high-dimensional observation challenges [54]. The fusion of diffusion models with RL frameworks, exemplified by DiffPoGAN, enhances policy exploration and improvement in offline settings by generating diverse action distributions [58].

To address inefficiencies in traditional RL, frameworks like GAN Q-learning enhance state-action value distribution estimation, improving algorithmic stability and efficiency [56]. The Adversarial Environment Design (ADD) algorithm innovates by using diffusion models to create environments based on the agent's regret, promoting policy enhancement through challenging scenarios [121]. The Temporally-Composable Diffuser (TCD) employs historical and prospective behaviors to guide action sequence generation, enhancing decision-making capabilities [57].

RL principles and frameworks are evolving, integrating advanced generative models and cognitive insights to address critical decision-making challenges. Innovations in AI, particularly in text-to-video generation and shared autonomy, significantly enhance autonomous systems' capabilities, enabling applications across fields such as robotics and education. Models like Sora exemplify this potential by simulating realistic scenarios and leveraging cognitive architectures to improve adaptability and decision-making in real-world contexts [92, 39, 63, 5, 122].

7.2 Advanced Techniques and Integrations

Recent advancements in reinforcement learning (RL) emphasize integrating advanced techniques to enhance autonomous agents' decision-making capabilities. Incorporating diffusion models into RL frameworks, as demonstrated by DiffPoGAN, improves the generation of diverse action distributions and policy exploration in offline settings [58]. Hierarchical reinforcement learning techniques, such as the Two-Stage Hierarchical Training (TSHT) method, enable agents to decompose complex tasks into manageable sub-tasks, enhancing efficiency in controlling intricate systems like musculoskeletal models [59].

The ODM framework further exemplifies the integration of cognitive insights, combining offline demonstrations with online interactions to develop robust policies capable of adapting to dynamic environments [120]. Adversarial training techniques, notably the ADD algorithm, create challenging training scenarios guided by agents' regrets, fostering resilience and adaptability in policy development [121].

The ongoing fusion of advanced techniques, including Generative Diffusion Models (GDMs) and video generation, significantly enhances autonomous agents' capabilities. By leveraging video data and iterative learning processes, these frameworks drive innovation in environments like robotics and self-driving technology, improving decision-making and planning through a unified interface that incorporates diverse data modalities [63, 62, 39]. Such advancements pave the way for more robust RL systems capable of addressing real-world challenges.

7.3 Applications in Autonomous Agents

Reinforcement learning (RL) is pivotal in developing autonomous agents, enabling adaptation to complex environments. The integration of advanced RL techniques with multimodal frameworks significantly enhances agents' task performance. For instance, the MEIA framework combines large language models and vision language models to translate high-level instructions into executable actions, improving adaptability in dynamic settings [101].

The VidMan framework exemplifies RL application by utilizing implicit dynamics knowledge from video data to predict actions, allowing agents to anticipate environmental changes and enhance performance in real-world scenarios [97]. Additionally, techniques like DiffPoGAN facilitate the generation of diverse action distributions, enhancing exploration capabilities and policy learning, particularly in scenarios where direct environmental interaction is impractical [58].

The application of RL, coupled with multimodal frameworks, continues to advance the development of robust autonomous agents. By leveraging generative models and multimodal understanding, researchers are creating agents capable of performing complex tasks across various environments, enhancing capabilities in robotics, decision-making, and simulation [63, 22, 7]. These advancements are instrumental in developing sophisticated autonomous systems that effectively address real-world challenges.

7.4 Challenges and Solutions in Reinforcement Learning

Reinforcement learning (RL) encounters several challenges affecting its effectiveness across domains. A primary concern is the reliance on task-specific modules, as seen in the ODM framework, complicating the creation of generalized models applicable to diverse tasks [120]. This limitation necessitates the development of more adaptable architectures requiring minimal customization.

Another challenge is the computational cost associated with self-training schemes, particularly in semi-supervised GANs, which may be impractical for resource-constrained applications [123]. More efficient training methodologies are needed to balance computational demands with performance improvements.

Classic policy optimization methods often result in out-of-distribution (OOD) actions that cannot be accurately estimated by evaluation models, hindering optimal policy learning [58]. Integrating diffusion models with RL frameworks, such as DiffPoGAN, can enhance policy exploration and learning outcomes by generating diverse action distributions.

Challenges persist in optimizing GANs for discrete outputs due to inadequate gradient flow from the discriminator to the generator, leading to unreliable learning signals [124]. Innovative solutions are necessary to facilitate effective gradient transmission and improve GAN training stability.

Moreover, tuning noise schedules for various tasks and explicitly conditioning policies for rollouts present significant limitations in RL [54]. Developing adaptive noise scheduling and flexible conditioning methods can enhance the robustness of RL algorithms.

In representation learning, frameworks like Curious Representation Learning (CRL) depend on exploration policy quality, affecting data diversity and salience [99]. Improving exploration strategies is crucial for effective data collection.

Data scarcity, the need for extensive training datasets, and generalization difficulties across populations and imaging modalities remain persistent challenges in RL applications [125]. Addressing these requires more data-efficient learning algorithms capable of generalizing across contexts.

Despite these challenges, RL presents significant advantages, such as modeling complex distributions and enhanced stability through methods like GAN Q-learning [56]. Approaches like ADD generate diverse environments that strengthen policy robustness and generalization [121].

However, methods such as policy-guided trajectory generation necessitate careful tuning to prevent instability in policy updates and may struggle with low-entropy policies, which can hinder performance [66]. Developing stable and adaptive policy update mechanisms is essential for improving RL effectiveness in complex environments.

7.5 Reinforcement Learning in Conjunction with GANs

The integration of reinforcement learning (RL) with Generative Adversarial Networks (GANs) signifies a substantial advancement in enhancing generative modeling and decision-making capabilities. This synergy leverages both paradigms to produce diverse, high-quality data while facilitating robust policy learning. The DiffPoGAN framework exemplifies this integration by utilizing diffusion models as policy generators to create diverse action distributions, incorporating maximum likelihood estimation (MLE) for behavior policy approximation, thereby enhancing RL agents' exploration and learning [58].

In adversarial environment design, the synergy of RL and GANs is explored through regret-guided diffusion models. The Adversarial Environment Design (ADD) approach directs a diffusion-based environment generator using the agent's regret, creating instructive and challenging environments that enhance training processes and promote policy improvement [121].

Combining GANs with RL frameworks, such as GAN Q-learning, provides a novel perspective on learning state-value distributions. This approach employs GANs to model value function distributions, improving the robustness and adaptability of RL agents in complex scenarios [56]. By utilizing GANs' capacity to model intricate distributions, RL agents can achieve more stable and efficient learning outcomes.

The continued integration of RL with GANs drives innovation in generative modeling and decision-making applications. By combining cognitive architectures and generative models, researchers are developing robust autonomous systems capable of addressing a wide range of real-world challenges. This synergy enhances capabilities and lays the groundwork for achieving general embodied intelligence, which is vital for the evolution of artificial general intelligence (AGI) [25, 39, 5].

8 Neural Networks

8.1 Structure and Function of Neural Networks

Neural networks, inspired by biological systems, are computational models designed to recognize patterns and perform complex tasks through interconnected layers of neurons. These networks consist of an input layer, multiple hidden layers, and an output layer, where each neuron processes inputs via weighted connections to produce outputs [64]. The strength of neural networks lies in their ability to learn by adjusting these weights to minimize prediction errors, a process often guided by backpropagation algorithms that compute the gradient of the loss function relative to the weights [64].

Recent advancements have improved the efficiency and adaptability of neural networks. Spiking Neural Networks (SNNs), integrated into neuromorphic computing, mimic the brain's energy-efficient processing by utilizing spike timing for information encoding, supporting the development of embodied intelligence systems that interact with the physical world [65]. Neural networks are versatile, excelling in applications such as image and speech recognition by leveraging their layered structure for hierarchical feature extraction, making them particularly effective in identifying complex patterns within large datasets [64].

The evolution of neural networks is driven by advancements in computational techniques and biological insights. Developments in deep learning, such as Sora and UMMGAN, are producing adaptable architectures capable of generating realistic video content and synthesizing complex multimodal data distributions. These models address real-world challenges across various industries, including film, education, and healthcare, while also highlighting the need to overcome limitations like data bias and controlling generated outputs [92, 126, 42, 127].

8.2 Neural Networks in Visual Perception and Image Generation

Neural networks play a crucial role in advancing visual perception and image generation by modeling complex patterns from visual data. Convolutional Neural Networks (CNNs) significantly enhance visual perception tasks, enabling hierarchical feature extraction essential for object detection, image segmentation, and classification [64]. Their layered architecture allows for the progressive refinement of visual information, capturing intricate details necessary for accurately interpreting visual scenes.

In image generation, neural networks have transformed synthetic image creation through models like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). These models learn latent representations of image data, facilitating the generation of high-quality images indistinguishable from real ones. GANs excel in producing realistic images using an adversarial framework, where a generator creates images and a discriminator evaluates their authenticity, driving the generator towards increasingly realistic outputs [42].

Recent advancements, including the integration of attention mechanisms found in transformer models, have enhanced neural networks' ability to capture global contexts and dependencies within images, improving performance in tasks requiring detailed visual understanding and synthesis [20]. Additionally, the development of energy-efficient neural networks inspired by neuromorphic computing has enabled the deployment of visual perception systems in resource-constrained environments, such as mobile and embedded devices [65].

Neural networks, particularly GANs, are at the forefront of innovation in visual perception and image generation, offering advanced solutions across various fields. Their applications span autonomous vehicles, enhancing object recognition and navigation; medical imaging, improving diagnostic accuracy through synthesized complex 3D brain volumes; and creative industries, generating high-quality images and videos, including AI-generated art and multimedia content. Furthermore, their impact extends to virtual reality, facilitating immersive environment creation, underscoring deep learning's transformative potential in practical and artistic domains [125, 14, 126, 128, 12]. As research progresses, the integration of neural networks with other advanced AI technologies is expected to further expand their potential, enabling more sophisticated and versatile visual systems.

8.3 Learning Rules and Energy-Efficient Computing

The pursuit of energy-efficient computing in neural networks aims to optimize performance while minimizing power consumption, especially in resource-constrained environments. This is supported by advanced learning rules that enhance the efficiency and adaptability of neural networks. Spiking Neural Networks (SNNs), inspired by the human brain's energy-efficient processing capabilities, utilize spike timing for information encoding, significantly reducing power requirements compared to traditional neural networks [65].

Integrating neuromorphic computing principles into neural network design has resulted in systems capable of real-time processing with low power consumption, making them suitable for applications in robotics and autonomous systems where energy efficiency is crucial [65]. These systems employ learning rules that mimic synaptic plasticity, allowing for adaptive learning and improved performance in dynamic environments.

Recent advancements have focused on optimizing learning algorithms to enhance computational efficiency. Techniques such as weight pruning and quantization reduce the computational load by minimizing the number of active parameters, thereby decreasing energy consumption without compromising accuracy [64]. Additionally, lightweight neural network models designed for deployment on edge devices highlight the importance of balancing computational efficiency with performance [65].

The exploration of learning rules and energy-efficient computing continues to drive innovation, offering robust solutions for a wide range of applications, from mobile devices to large-scale data centers. As research advances, merging biologically inspired computing principles—such as cognitive architectures and neuromorphic computing—with cutting-edge AI technologies, including generative models, is poised to significantly enhance the efficiency and adaptability of neural networks. This integration not only improves AI systems' capabilities to generate novel content and interact intelligently with their environments but also lays the groundwork for developing sustainable and scalable AI solutions that effectively address real-world challenges across various domains, including robotics and biomedicine [2, 65, 39].

9 Autonomous Agents

9.1 Conceptual Framework and Adaptability

Autonomous agents are designed to operate independently, adapt to diverse environments, and perform complex tasks without human intervention. Recent advancements emphasize shared autonomy, where collaborative control between users and agents transcends traditional methods that rely on predefined environmental knowledge and user goals. Modern approaches employ model-free deep reinforcement learning, allowing agents to learn desired behaviors without explicit reward feedback, situating them as active participants in socio-technical networks that foster trust and comprehension of AI-generated content [25, 122]. Reinforcement learning (RL) remains a core mechanism, enabling agents to derive optimal policies through environmental interactions and adapt to dynamic scenarios by refining strategies based on feedback [54].

The integration of generative models, such as GANs and diffusion models, enhances adaptability by enabling agents to generate diverse action distributions and explore broader potential solutions [58]. Hierarchical frameworks, like the Two-Stage Hierarchical Training (TSHT) method, support adaptability by breaking down complex tasks into manageable sub-tasks, improving learning efficiency and performance [59]. Multimodal perception systems, exemplified by the MEIA framework, enhance decision-making by enabling agents to process and integrate various sensory inputs [101].

Frameworks such as the Online Decision MetaMorphFormer (ODM) incorporate cognitive and behavioral insights, allowing agents to learn from offline demonstrations and online interactions, thus enhancing adaptability and robustness in real-world applications [120]. These advanced techniques foster innovation in autonomous agent development, facilitating sophisticated systems capable of addressing a wide spectrum of real-world challenges.

9.2 Integration with Video Generation Models

The integration of autonomous agents with advanced video generation models, such as VQGAN-CLIP architecture and collaborative Vision Large Language Models (VLMs), significantly advances artificial intelligence. This synergy enhances agent functionality and adaptability across applications, addressing challenges related to video quality, consistency, and user instruction adherence, thus paving the way for innovations in personalized video content creation, visual arts, and entertainment [5, 15, 84]. By leveraging video generation models, including GANs and diffusion models, autonomous agents can effectively generate, interpret, and utilize visual data, improving decision-making and interaction capabilities.

GANs excel in creating realistic and diverse visual content, enabling autonomous agents to simulate and anticipate environmental changes, thereby enhancing adaptability and performance in dynamic settings [42]. The generation of synthetic video data using GANs allows agents to refine policies in simulated environments, reducing the need for extensive real-world data collection [105]. Diffusion models, known for iterative refinement, provide high-quality visual inputs, enhancing perception and environmental understanding [50]. By integrating diffusion models, agents benefit from improved temporal coherence and content consistency, crucial for applications requiring precise motion analysis and prediction [73].

Collaboration between video generation models and autonomous agents fosters more interactive and responsive systems. Utilizing video diffusion models enables agents to exploit implicit dynamics knowledge from video data, enhancing anticipation and reaction capabilities, crucial in contexts like autonomous driving and robotics [97]. This integration drives innovation, offering enhanced functionality and adaptability across diverse applications, with future research expected to significantly augment autonomous agents' capabilities in complex environments [92, 39, 63, 122, 25].

9.3 Applications in Real-World Environments

Autonomous agents are increasingly utilized across various real-world environments, employing advanced AI techniques to perform complex tasks with minimal human oversight. In autonomous driving, agents are crucial for developing systems that navigate dynamic and unpredictable road conditions, enhancing safety and reliability through domain augmentation techniques [98]. In robotics, agents execute intricate tasks in manufacturing and assembly lines, optimizing processes and adapting

to environmental changes with minimal downtime [59]. The incorporation of multimodal perception systems enriches their ability to process varied sensory inputs, facilitating effective decision-making [101].

In healthcare, agents contribute to personalized and efficient care delivery by simulating patient-specific scenarios, assisting in diagnosis and treatment planning, particularly in telemedicine [2]. In environmental monitoring, agents equipped with video generation models provide insights into ecological changes, autonomously collecting data in remote locations for real-time analysis [97]. In security and surveillance, agents simulate threat scenarios and optimize strategies, enhancing security measures in urban environments [42].

The deployment of autonomous agents is rapidly increasing, driven by advancements in AI technologies, particularly multi-modal generative models like MLLM and diffusion models like Sora, which excel in understanding and generating complex content across media forms. As researchers explore the integration of cognitive architectures with generative models, the potential for creating sophisticated embodied agents expands, paving the way for innovative applications across diverse fields [22, 39]. These agents are set to transform industries, enhancing efficiency, safety, and adaptability in complex and dynamic settings.

10 Conclusion

10.1 Innovations and Advancements

This survey highlights significant advancements in artificial intelligence, focusing on video generation models and embodied intelligence. Notable progress includes the TrackDiffusion framework, which enhances video synthesis through optimized model training, indicating a leap forward in video generation methodologies. Similarly, GANalyze demonstrates the potential of Generative Adversarial Networks (GANs) in altering cognitive properties within images, offering insights into image memorability and creative expression.

In the realm of video generation, the Grid Diffusion Model (GDM) emerges as a key innovation, improving efficiency and scalability with its reduced memory usage and capability to produce videos with higher frame counts. The MotionMaster framework further exemplifies progress by enabling effective camera motion transfer, resulting in high-quality, diverse video outputs that enhance camera control across applications.

The SurGen method significantly enhances the visual quality and temporal dynamics of surgical videos, proving invaluable for surgical education and training. Additionally, an interactive method for generating laparoscopic videos achieves state-of-the-art fidelity, demonstrating its effectiveness in medical video generation.

In text-to-video generation, Imagen Video successfully extends text-to-image diffusion models to the video domain, achieving high fidelity and temporal consistency while maintaining alignment with text prompts. CreativeGAN showcases the automation of creativity in design synthesis, generating novel designs that push the boundaries of creative applications in AI.

The introduction of Neuroscore as a metric for evaluating GAN performance marks a significant advancement, aligning with human perception and effectively ranking images with fewer samples. Moreover, advancements in neural video compression are setting new benchmarks for subjective visual quality, underscoring ongoing efforts to optimize video data processing.

In 3D modeling, significant progress is demonstrated by approaches that generate consistent and realistic 3D models across varying conditions, highlighting improvements in model consistency without altering standard architectures. Collectively, these innovations represent substantial progress in AI and machine learning, paving the way for more sophisticated systems capable of addressing a wide range of real-world challenges. Future research should focus on refining integration techniques and exploring new applications that leverage the strengths of cognitive architectures and generative models.

10.2 Limitations and Future Directions

Despite remarkable progress, video generation models and embodied intelligence techniques face several limitations that necessitate further research. A primary challenge is the computational intensity associated with training diffusion models, which limits scalability and accessibility across domains. Optimizing model architectures for efficiency is crucial for broader adoption, particularly given the resource demands highlighted by recent models.

Challenges in video generation include dataset constraints, model heterogeneity, and difficulties in producing high-quality, consistent outputs. Future research should explore new model architectures and improve label conditioning to address high variance in video datasets. Enhancing frameworks to handle complex motion patterns and improve video quality remains critical. Additionally, substantial training data requirements in video-to-video translation networks pose challenges for complex engineering scenarios, while limitations in achieving high realism suggest a need for improvements in resolution and realism.

Embodied intelligence frameworks are limited by their reliance on task-specific modules. Future research could explore self-adaptive structures to enhance general applicability, along with refining control algorithms for applications in assistive robotics and rehabilitation. The complexity of developing robust evaluation metrics and reliance on traditional machine learning paradigms pose significant challenges for embodied agents.

Detecting synthetic images generated by GANs and diffusion models remains an area requiring improvement. Current detection methods often struggle to generalize across different generative models, necessitating the development of robust techniques that effectively identify synthetic content. Future research should focus on leveraging specific properties of diffusion models to enhance detection capabilities and establish more comprehensive benchmarks.

In human motion generation, enhancing generalization capabilities and adapting techniques to other generative tasks is crucial. Improving the modeling of group dynamics and system performance in high-density environments is essential for advancing crowd simulation techniques. Additionally, future research should explore enhancing the generalization capabilities of video generation models by integrating additional modalities or refining training approaches.

The implementation of certain data modalities, such as videos and audio, in EDDPMs remains untested, which may limit applicability. Potential performance degradation due to the double regularization effect in diffusion policies warrants further investigation. Future research could focus on strategies to mitigate degeneration in generative models, such as methods for preserving knowledge during training. Addressing inherent biases in datasets that affect the authenticity of generated outputs is also critical.

Extending methods to environments with complex dynamics or where the synthesis oracle is unavailable could enhance the robustness of generative models. Optimizing the training process and exploring additional applications of Generative Adversarial Neural Operators (GANO) in various scientific domains could further expand their utility. Future research should prioritize developing robust models capable of handling diverse datasets, incorporating multi-modal data, and addressing fairness and interpretability issues in GAN applications. Enhancing training stability and exploring the integration of GANs into complex reinforcement learning frameworks are also critical areas for future exploration. Improving the theoretical foundations of regret estimation and ensuring convergence to optimal solutions could significantly advance the field. Additionally, exploring more complex decision scenarios and performance metrics while enhancing integration of human inquiries through advanced embeddings could further improve generative manufacturing systems. Incorporating 3D tool positioning and camera movement dynamics to enhance realism in generated videos is another promising research avenue. Finally, developing reliable perceptual metrics to evaluate video compression methods without extensive user studies could optimize video data processing.

References

[1] Yintai Ma, Diego Klabjan, and Jean Utke. Video to video generative adversarial network for few-shot learning based on policy gradient. *arXiv preprint arXiv:2410.20657*, 2024.

[2] Linyuan Li, Jianing Qiu, Anujit Saha, Lin Li, Poyuan Li, Mengxian He, Ziyu Guo, and Wu Yuan. Artificial intelligence for biomedical video generation, 2024.

[3] Simranjeet Singh, Rajneesh Sharma, and Alan F. Smeaton. Using gans to synthesise minimum training data for deepfake generation, 2020.

[4] Jacob Schrum, Jake Gutierrez, Vanessa Volz, Jialin Liu, Simon Lucas, and Sebastian Risi. Interactive evolution and exploration within latent level-design space of generative adversarial networks, 2020.

[5] Joseph Cho, Fachrina Dewi Puspitasari, Sheng Zheng, Jingyao Zheng, Lik-Hang Lee, Tae-Ho Kim, Choong Seon Hong, and Chaoning Zhang. Sora as an agi world model? a complete survey on text-to-video generation, 2024.

[6] Manuel Molano-Mazon, Arno Onken, Eugenio Piasini, and Stefano Panzeri. Synthesizing realistic neural population activity patterns using generative adversarial networks, 2018.

[7] Xinji Mai, Zeng Tao, Junxiong Lin, Haoran Wang, Yang Chang, Yanlan Kang, Yan Wang, and Wenqiang Zhang. From efficient multimodal models to world models: A survey. *arXiv preprint arXiv:2407.00118*, 2024.

[8] Yucheng Xu, Li Nanbo, Arushi Goel, Zijian Guo, Zonghai Yao, Hamidreza Kasaei, Mohammadreze Kasaei, and Zhibin Li. Controllable video generation by learning the underlying dynamical system with neural ode, 2023.

[9] Moayed Haji Ali, Andrew Bond, Tolga Birdal, Duygu Ceylan, Levent Karacan, Erkut Erdem, and Aykut Erdem. Vidstyleode: Disentangled video editing via stylegan and neuralodes, 2025.

[10] Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N. Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models, 2024.

[11] Taegyeong Lee, Soyeong Kwon, and Taehwan Kim. Grid diffusion models for text-to-video generation, 2024.

[12] Anne-Sofie Maerten and Derya Soydaner. From paintbrush to pixel: A review of deep neural networks in ai-generated art, 2024.

[13] Janet Rafner, Lotte Philipsen, Sebastian Risi, Joel Simon, and Jacob Sherson. The power of pictures: using ml assisted image generation to engage the crowd in complex socioscientific problems, 2020.

[14] Nuha Aldausari, Arcot Sowmya, Nadine Marcus, and Gelareh Mohammadi. Video generative adversarial networks: A review, 2020.

[15] SukChang Lee. Transforming text into video: A proposed methodology for video production using the vqgan-clip image generative ai model. *International Journal of Advanced Culture Technology*, 11(3):225–230, 2023.

[16] Hiromitsu Kigure. Application of video-to-video translation networks to computational fluid dynamics, 2021.

[17] Xingyu Li, Fei Tao, Wei Ye, Aydin Nassehi, and John W. Sutherland. Generative manufacturing systems using diffusion models and chatgpt, 2025.

[18] Ivan Iliash, Simeon Allmendinger, Felix Meissen, Niklas Kühl, and Daniel Rückert. Interactive generation of laparoscopic videos with diffusion models, 2024.

[19] Haoran Cheng, Liang Peng, Linxuan Xia, Yuepeng Hu, Hengjia Li, Qinglin Lu, Xiaofei He, and Boxi Wu. Searching priors makes text-to-video synthesis better, 2024.

[20] Shiv Ram Dubey and Satish Kumar Singh. Transformer-based generative adversarial networks in computer vision: A comprehensive survey, 2023.

[21] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022.

[22] Hong Chen, Xin Wang, Yuwei Zhou, Bin Huang, Yipeng Zhang, Wei Feng, Houlun Chen, Zeyang Zhang, Siao Tang, and Wenwu Zhu. Multi-modal generative ai: Multi-modal llm, diffusion and beyond. *arXiv preprint arXiv:2409.14993*, 2024.

[23] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective, 2024.

[24] Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, Rene Vidal, Maheswaran Sathiamoorthy, Atoosa Kasrizadeh, Silvia Milano, and Francesco Ricci. Recommendation with generative models, 2024.

[25] Piero Polidoro. Strengthening democratic engagement through value-based generative adversarial networks.

[26] Luís Arandas, Mick Grierson, and Miguel Carvalhais. Antagonising explanation and revealing bias directly through sequencing and multimodal inference, 2023.

[27] Bernhard Kratzwald, Zhiwu Huang, Danda Pani Paudel, Acharya Dinesh, and Luc Van Gool. Improving video generation for multi-functional applications, 2018.

[28] Pengxiang Li, Kai Chen, Zhili Liu, Ruiyuan Gao, Lanqing Hong, Guo Zhou, Hua Yao, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Trackdiffusion: Tracklet-conditioned video generation via diffusion models, 2024.

[29] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control, 2024.

[30] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning, 2024.

[31] Javad Amirian, Wouter van Toll, Jean-Bernard Hayet, and Julien Pettré. Data-driven crowd simulation with generative adversarial networks, 2019.

[32] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior, 2023.

[33] Qiang Wang, Minghua Liu, Junjun Hu, Fan Jiang, and Mu Xu. Controllable longer image animation with diffusion models, 2024.

[34] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zięba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation, 2023.

[35] Enric Corona, Andrei Zanfir, Eduard Gabriel Bazavan, Nikos Kolotouros, Thiemo Alldieck, and Cristian Sminchisescu. Vlogger: Multimodal diffusion for embodied avatar synthesis, 2024.

[36] Yi Shi, Jingbo Wang, Xuekun Jiang, Bingkun Lin, Bo Dai, and Xue Bin Peng. Interactive character control with auto-regressive motion diffusion models, 2024.

[37] Amin Heyrani Nobari, Muhammad Fathy Rashad, and Faez Ahmed. Creativegan: Editing generative adversarial networks for creative design synthesis, 2021.

[38] Jianzhi Liu, Junchen Zhu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Aicl: Action in-context learning for video diffusion model, 2024.

[39] Yanfei Liu, Yuzhou Liu, and Chao Shen. Combining minds and machines: investigating the fusion of cognitive architectures and generative models for general embodied intelligence. In *Proceedings of the AAAI Symposium Series*, volume 2, pages 307–314, 2023.

[40] Nicholas Roy, Ingmar Posner, Tim Barfoot, Philippe Beaudoin, Yoshua Bengio, Jeannette Bohg, Oliver Brock, Isabelle Depatie, Dieter Fox, Dan Koditschek, Tomas Lozano-Perez, Vikash Mansinghka, Christopher Pal, Blake Richards, Dorsa Sadigh, Stefan Schaal, Gaurav Sukhatme, Denis Therien, Marc Toussaint, and Michiel Van de Panne. From machine learning to robotics: Challenges and opportunities for embodied intelligence, 2021.

[41] Zhuoling Li, Xiaogang Xu, Zhenhua Xu, SerNam Lim, and Hengshuang Zhao. Larm: Large auto-regressive model for long-horizon embodied intelligence, 2025.

[42] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.

[43] Francisco J. Ibarrola, Nishant Ravikumar, and Alejandro F. Frangi. Partially conditioned generative adversarial networks, 2020.

[44] Soheyla Amirian, Thiab R. Taha, Khaled Rasheed, and Hamid R. Arabnia. Generative adversarial network applications in creating a meta-universe, 2022.

[45] Bo Zhao and Hakan Bilen. Synthesizing informative training samples with gan, 2022.

[46] Tanujit Chakraborty, Ujjwal Reddy K S, Shraddha M. Naik, Madhurima Panja, and Bayapureddy Manvitha. Ten years of generative adversarial nets (gans): A survey of the state-of-the-art, 2023.

[47] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks, 2018.

[48] Melike Nur Yeğin and Mehmet Fatih Amasyalı. Theoretical research on generative diffusion models: an overview, 2024.

[49] Panagiotis Alimisis, Ioannis Mademlis, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, and Georgios Th. Papadopoulos. Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions, 2025.

[50] Kidist Amde Mekonnen, Nicola Dall'Asen, and Paolo Rota. Adv-kd: Adversarial knowledge distillation for faster diffusion sampling, 2024.

[51] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2024.

[52] Joseph Cho, Samuel Schmidgall, Cyril Zakka, Mrudang Mathur, Dhamanpreet Kaur, Rohan Shad, and William Hiesinger. Surgen: Text-guided diffusion model for surgical video generation, 2024.

[53] Zifan Shi, Sida Peng, Yinghao Xu, Andreas Geiger, Yiyi Liao, and Yujun Shen. Deep generative models on 3d representations: A survey, 2023.

[54] Bogdan Mazoure, Walter Talbott, Miguel Angel Bautista, Devon Hjelm, Alexander Toshev, and Josh Susskind. Value function estimation using conditional diffusion models for control, 2023.

[55] Anirudh Goyal, Nan Rosemary Ke, Alex Lamb, R Devon Hjelm, Chris Pal, Joelle Pineau, and Yoshua Bengio. Actual: Actor-critic under adversarial learning, 2017.

[56] Thang Doan, Bogdan Mazoure, and Clare Lyle. Gan q-learning, 2018.

[57] Jifeng Hu, Yanchao Sun, Sili Huang, SiYuan Guo, Hechang Chen, Li Shen, Lichao Sun, Yi Chang, and Dacheng Tao. Instructed diffuser with temporal condition guidance for offline reinforcement learning, 2023.

[58] Xuemin Hu, Shen Li, Yingfen Xu, Bo Tang, and Long Chen. Diffpogan: Diffusion policies with generative adversarial networks for offline reinforcement learning, 2024.

[59] Chenhui Zuo, Kaibo He, Jing Shao, and Yanan Sui. Self model for embodied intelligence: Modeling full-body human musculoskeletal system and locomotion control with hierarchical low-dimensional representation, 2024.

[60] Jinzhou Lin, Han Gao, Xuxiang Feng, Rongtao Xu, Changwei Wang, Man Zhang, Li Guo, and Shibiao Xu. Advances in embodied navigation using large language models: A survey, 2024.

[61] Adam Block, Ali Jadbabaie, Daniel Pfrommer, Max Simchowitz, and Russ Tedrake. Provable guarantees for generative behavior cloning: Bridging low-level stability and high-level behavior, 2023.

[62] Hongyang Du, Ruichen Zhang, Yinqiu Liu, Jiacheng Wang, Yijing Lin, Zonghang Li, Dusit Niyato, Jiawen Kang, Zehui Xiong, Shuguang Cui, et al. Enhancing deep reinforcement learning: A tutorial on generative diffusion models in network optimization. *IEEE Communications Surveys & Tutorials*, 2024.

[63] Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Video as the new language for real-world decision making, 2024.

[64] Y. H. Yseng, F. J. Jiang, and C. Y. Huang. A universal training scheme and the resulting universality for machine learning phases, 2022.

[65] Rachmad Vidya Wicaksana Putra, Alberto Marchisio, Fakhreddine Zayer, Jorge Dias, and Muhammad Shafique. Embodied neuromorphic artificial intelligence for robotics: Perspectives, challenges, and research development stack, 2024.

[66] Marc Rigter, Jun Yamada, and Ingmar Posner. World models via policy-guided trajectory diffusion, 2024.

[67] Frank Joublin, Antonello Ceravola, and Cristian Sandu. Introducing brain-like concepts to embodied hand-crafted dialog management system, 2024.

[68] Andreas Hald, Jens Struckmann Hansen, Jeppe Kristensen, and Paolo Burelli. Procedural content generation of puzzle games using conditional generative adversarial networks, 2023.

[69] Mahdyar Ravanbakhsh, Mohamad Baydoun, Damian Campo, Pablo Marin, David Martin, Lucio Marcenaro, and Carlo S. Regazzoni. Hierarchy of gans for learning embodied self-awareness model, 2018.

[70] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023.

[71] Mkhuseli Ngxande, Jules-Raymond Tapamo, and Michael Burke. Depthwisegans: Fast training generative adversarial networks for realistic image synthesis, 2019.

[72] Christopher X. Ren, Amanda Ziemann, James Theiler, and Alice M. S. Durieux. Deep snow: Synthesizing remote sensing imagery with generative adversarial nets, 2020.

[73] Vidm: Video implicit diffusion models.

[74] Fabian Mentzer, Eirikur Agustsson, Johannes Ballé, David Minnen, Nick Johnston, and George Toderici. Neural video compression using gans for detail synthesis and propagation, 2022.

[75] Sanghyeon Na, Yonggyu Kim, and Hyunjoon Lee. Boost your own human image generation model via direct preference optimization with ai feedback, 2024.

[76] Kumar Shubham, Gopalakrishnan Venkatesh, Reijul Sachdev, Akshi, Dinesh Babu Jayagopi, and G. Srinivasaraghavan. Learning a deep reinforcement learning policy over the latent space of a pre-trained gan for semantic age manipulation, 2021.

[77] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models, 2020.

[78] Ziyu Yao, Jialin Li, Yifeng Zhou, Yong Liu, Xi Jiang, Chengjie Wang, Feng Zheng, Yuexian Zou, and Lei Li. Car: Controllable autoregressive modeling for visual generation, 2024.

[79] Ying Shen, Yizhe Zhang, Shuangfei Zhai, Lifu Huang, Joshua M. Susskind, and Jiatao Gu. Many-to-many image generation with auto-regressive diffusion models, 2024.

[80] Teng Hu, Jiangning Zhang, Ran Yi, Yating Wang, Hongrui Huang, Jieyu Weng, Yabiao Wang, and Lizhuang Ma. Motionmaster: Training-free camera motion transfer for video generation, 2024.

[81] Cihan Öngün and Alptekin Temizel. Paired 3d model generation with conditional generative adversarial networks, 2019.

[82] Elijah Miller, Thomas Dupont, and Mingming Wang. Enhanced creativity and ideation through stable video synthesis, 2024.

[83] Soumik Mukhopadhyay, Matthew Gwilliam, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Srinidhi Hegde, Tianyi Zhou, and Abhinav Shrivastava. Diffusion models beat gans on image classification, 2023.

[84] Liu He, Yizhi Song, Hejun Huang, Daniel Aliaga, and Xin Zhou. Kubrick: Multimodal agent collaborations for synthetic video generation, 2024.

[85] Juil Koo, Seungwoo Yoo, Minh Hieu Nguyen, and Minhyuk Sung. Salad: Part-level latent diffusion for 3d shape generation and manipulation, 2024.

[86] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback, 2023.

[87] Hongda Jiang, Xi Wang, Marc Christie, Libin Liu, and Baoquan Chen. Cinematographic camera diffusion model, 2024.

[88] Md Ashiqur Rahman, Manuel A. Florez, Anima Anandkumar, Zachary E. Ross, and Kamyar Azizzadenesheli. Generative adversarial neural operators, 2022.

[89] Sebastian Hereu and Qianfei Hu. Creative portraiture: Exploring creative adversarial networks and conditional creative adversarial networks, 2024.

[90] Yaosi Hu, Zhenzhong Chen, and Chong Luo. Lamd: Latent motion diffusion for video generation, 2023.

[91] Zuhao Yang, Fangneng Zhan, Kunhao Liu, Muyu Xu, and Shijian Lu. Ai-generated images as data source: The dawn of synthetic era. *arXiv preprint arXiv:2310.01830*, 2023.

[92] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models, 2024.

[93] Eman T. Hassan and David J. Crandall. A study of cross-domain generative models applied to cartoon series, 2017.

[94] Mohammad Lataifeh, Xavier Carrasco, Ashraf Elnagar, and Naveed Ahmed. Augmenting character designers creativity using generative adversarial networks, 2023.

[95] Hui Ying, He Wang, Tianjia Shao, Yin Yang, and Kun Zhou. Unsupervised image generation with infinite generative adversarial networks, 2021.

[96] Unifying diffusion models' latent space with applications to cyclediffusion and guidance.

[97] Youpeng Wen, Junfan Lin, Yi Zhu, Jianhua Han, Hang Xu, Shen Zhao, and Xiaodan Liang. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation, 2024.

[98] Luciano Baresi, Davide Yi Xian Hu, Andrea Stocco, and Paolo Tonella. Efficient domain augmentation for autonomous driving testing using diffusion models, 2025.

[99] Yilun Du, Chuang Gan, and Phillip Isola. Curious representation learning for embodied intelligence, 2021.

[100] Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution, 2021.

[101] Yang Liu, Xinshuai Song, Kaixuan Jiang, Weixing Chen, Jingzhou Luo, Guanbin Li, and Liang Lin. Meia: Multimodal embodied perception and interaction in unknown environments, 2024.

[102] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, In So Kweon, and Junmo Kim. Text-to-image diffusion models in generative ai: A survey, 2024.

[103] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence?, 2024.

[104] Pablo Samuel Castro. Ganterpretations, 2020.

[105] Florian Eckerli and Joerg Osterrieder. Generative adversarial networks in finance: an overview, 2021.

[106] Zhengwei Wang, Qi She, Alan F. Smeaton, Tomas E. Ward, and Graham Healy. Synthetic-neuroscore: Using a neuro-ai interface for evaluating generative adversarial networks, 2020.

[107] Parthak Mehta, Sarthak Mishra, Nikhil Chouhan, Neel Pethani, and Ishani Saha. Face editing with gan – a review, 2022.

[108] Hrishikesh Sharma. A chronological survey of theoretical advancements in generative adversarial networks for computer vision, 2023.

[109] Samuel A. Barnett. Convergence problems with generative adversarial networks (gans), 2018.

[110] Xiaoyu Yue, Zidong Wang, Zeyu Lu, Shuyang Sun, Meng Wei, Wanli Ouyang, Lei Bai, and Luping Zhou. Diffusion models need visual priors for image generation, 2024.

[111] Ivan Kapelyukh, Vitalis Vosylius, and Edward Johns. Dall-e-bot: Introducing web-scale diffusion models to robotics, 2023.

[112] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes, 2024.

[113] Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey, 2025.

[114] Anwaar Ulhaq and Naveed Akhtar. Efficient diffusion models for vision: A survey, 2024.

[115] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models, 2022.

[116] Muhammad Usman Akbar, Måns Larsson, and Anders Eklund. Brain tumor segmentation using synthetic mr images – a comparison of gans and diffusion models, 2024.

[117] Susan Athey, Guido Imbens, Jonas Metzger, and Evan Munro. Using wasserstein generative adversarial networks for the design of monte carlo simulations, 2020.

[118] Thomas Lips and Francis wyffels. Evaluating text-to-image diffusion models for texturing synthetic data, 2024.

[119] Yu Cao, Jingrun Chen, Yixin Luo, and Xiang Zhou. Exploring the optimal choice for generative processes in diffusion models: Ordinary vs stochastic differential equations, 2023.

[120] Luo Ji and Runji Lin. Online decision metamorphformer: A casual transformer-based reinforcement learning framework of universal embodied intelligence, 2024.

[121] Hojun Chung, Junseo Lee, Minsoo Kim, Dohyeong Kim, and Songhwai Oh. Adversarial environment design via regret-guided diffusion models, 2024.

[122] Takuma Yoneda, Luzhe Sun, , Ge Yang, Bradly Stadie, and Matthew Walter. To the noise and back: Diffusion for shared autonomy, 2023.

[123] Alan Do-Omri, Dalei Wu, and Xiaohua Liu. A self-training method for semi-supervised gans, 2017.

[124] Sylvain Lamprier, Thomas Scialom, Antoine Chaffin, Vincent Claveau, Ewa Kijak, Jacopo Staiano, and Benjamin Piwowarski. Generative cooperative networks for natural language generation, 2022.

[125] Rongguang Wang, Vishnu Bashyam, Zhijian Yang, Fanyang Yu, Vasiliki Tassopoulou, Sai Spandana Chintapalli, Ioanna Skampardoni, Lasya P. Sreepada, Dushyant Sahoo, Konstantina Nikita, Ahmed Abdulkadir, Junhao Wen, and Christos Davatzikos. Applications of generative adversarial networks in neuroimaging and clinical neuroscience, 2023.

[126] Anders Eklund. Feeding the zombies: Synthesizing brain volumes using a 3d progressive growing gan, 2020.

[127] Omry Sendik, Dani Lischinski, and Daniel Cohen-Or. Unsupervised multi-modal styled content generation, 2020.

[128] Hanne Carlsson and Dimitrios Kollias. Image generation and recognition (emotions), 2019.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.