A Survey on Video Generation Human Feedback Reinforcement Learning and User Interaction in Computer Vision

www.surveyx.cn

Abstract

Video generation, at the intersection of artificial intelligence and machine learning, is a rapidly evolving field that integrates human feedback, reinforcement learning, generative models, computer vision, and user interaction. This survey explores the multidisciplinary nature of video generation, emphasizing its transformative potential across various domains, including healthcare, autonomous driving, and entertainment. The integration of human feedback is crucial for aligning AI systems with user preferences, enhancing the realism and coherence of generated content. Reinforcement learning, particularly when combined with human feedback, facilitates the development of adaptive and responsive video generation models. Generative models, such as GANs and diffusion models, play a pivotal role in synthesizing high-quality, contextually relevant video sequences, while advancements in computer vision enable machines to interpret and generate complex visual data. User interaction frameworks empower users to exert creative control over video synthesis, fostering personalization and engagement. Despite significant progress, challenges remain, including ensuring temporal consistency, addressing ethical considerations, and improving scalability. The survey highlights the need for robust evaluation metrics and diverse datasets to advance the field. Future research directions include enhancing model adaptability, integrating advanced technologies, and exploring ethical implications to ensure responsible AI-driven content creation. Overall, this survey underscores the importance of continued innovation in video generation technologies to create more realistic, interactive, and user-aligned content.

1 Introduction

1.1 Multidisciplinary Nature of Video Generation and AI

The interdisciplinary nature of video generation is highlighted by its integration with artificial intelligence (AI) and machine learning, forming a cohesive framework for producing dynamic and realistic video content. This integration is pivotal in applications ranging from autonomous driving to biomedical visualization, showcasing the extensive potential of video generation technologies. Advanced generative models, such as those for creating high-dynamic videos with motion-rich actions, exemplify the synergy between AI and machine learning, enhancing both visual effects and realism [1].

Recent advancements in generative AI have introduced sophisticated methods for synthesizing diverse and contextually relevant video content, facilitating real-world decision-making processes similar to the transformative impact of language models in the digital domain [2]. Large-scale multi-modal generative models have further propelled AI capabilities, offering unprecedented performance across various fields [3]. In autonomous systems, these models are essential for creating realistic driving scenarios, thereby advancing autonomous driving through enhanced simulation capabilities [4].

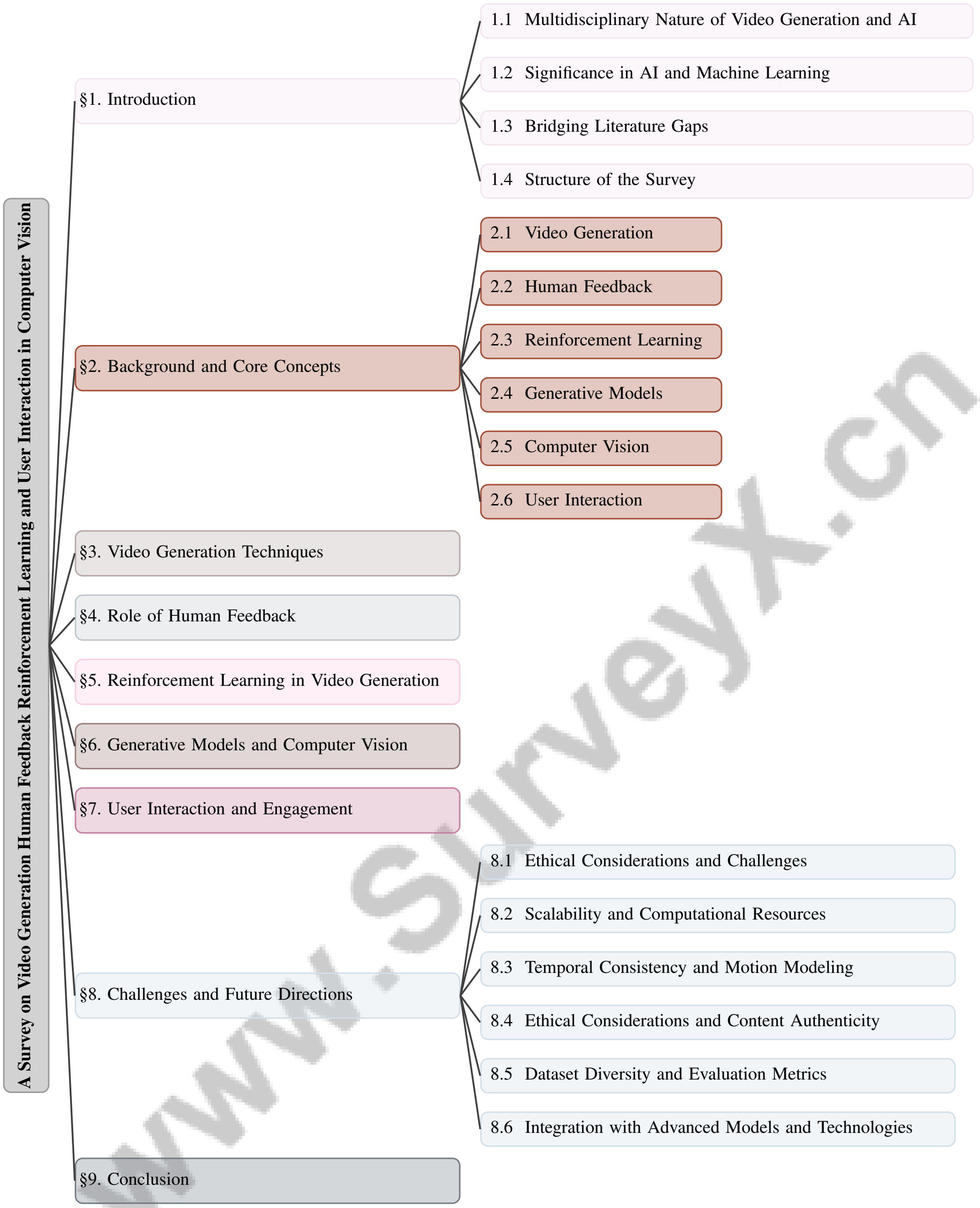
A Survey on Video Generation Human Feedback Reinforcement Learning and User Interaction in Computer Vision
§1. Introduction
1.1 Multidisciplinary Nature of Video Generation and AI
1.2 Significance in AI and Machine Learning
1.3 Bridging Literature Gaps
1.4 Structure of the Survey
§2. Background and Core Concepts
2.1 Video Generation
2.2 Human Feedback
2.3 Reinforcement Learning
2.4 Generative Models
2.5 Computer Vision
2.6 User Interaction
§3. Video Generation Techniques
§4. Role of Human Feedback
§5. Reinforcement Learning in Video Generation
§6. Generative Models and Computer Vision
§7. User Interaction and Engagement
§8. Challenges and Future Directions
8.1 Ethical Considerations and Challenges
8.2 Scalability and Computational Resources
8.3 Temporal Consistency and Motion Modeling
8.4 Ethical Considerations and Content Authenticity
8.5 Dataset Diversity and Evaluation Metrics
8.6 Integration with Advanced Models and Technologies
§9. Conclusion

Figure 1: chapter structure

The role of human feedback in reinforcement learning emphasizes the interdisciplinary nature of video generation, enabling systems to learn complex behaviors from sparse feedback in interactive environments. This approach is crucial for developing interactive agents capable of adapting to novel tasks and environments, thereby expanding the applicability of video generation technologies [5]. The integration of AI-generated content in domains like fluid dynamics highlights the transformative potential of generative AI models, reshaping traditional methods and introducing innovative solutions [6].

Furthermore, the development of interactive systems like Agent AI, which can perceive visual stimuli and language inputs, exemplifies the interdisciplinary approach required for effective video generation [7]. Incorporating Generative AI technologies emphasizes the urgent need to maximize system performance to reflect intricate human knowledge and contextual subtleties [8]. The challenges

AI-generated, for reference only.

of generating high-quality and controllable videos, as highlighted in existing literature, underscore the necessity for innovative solutions to address complex spatio-temporal relationships inherent in video synthesis [9].

The interdisciplinary aspects of video generation underscore its integration with AI and machine learning, fostering innovations that redefine how visual content is created, interpreted, and interacted with across diverse domains [10]. This integration enhances the technical capabilities of video generation systems while broadening their applicability, from entertainment to scientific visualization, paving the way for future advancements in AI-driven content creation.

1.2 Significance in AI and Machine Learning

Video generation plays a pivotal role in advancing AI and machine learning technologies, offering transformative capabilities across various domains. The integration of video generation with AI methodologies, such as Reinforcement Learning with Human Feedback (RLHF), is crucial for aligning AI systems with human preferences and societal needs, particularly in healthcare, where high-quality video content enhances medical education and clinical practice [11]. Traditional reliance solely on text instructions has proven insufficient, necessitating approaches that integrate both image and text instructions for dynamic video generation [1].

The rapid advancements in Artificial Intelligence Generated Content (AIGC) underscore the need for robust evaluation benchmarks to ensure quality and applicability [12]. In autonomous driving, generating diverse and realistic datasets is essential for overcoming limitations of traditional data augmentation methods, thereby enhancing AI training and performance [4]. Moreover, generating multiple long-term future scenarios in video generation advances AI's capabilities in understanding human behavior, providing deeper insights into complex interactions [10].

The significance of video generation is further emphasized by leveraging models trained on web data to enable robot manipulation, expanding AI's applicability in robotics and automation [5]. The development of Agent AI, utilizing large foundation models and multimodal understanding, represents a step towards Artificial General Intelligence (AGI), showcasing video generation's potential in achieving broader AI objectives [7]. Additionally, modeling the inherent continuity in video data is necessary for improving prediction accuracy, emphasizing the importance of robust video generation techniques [13].

Responsive Action-based Video Synthesis (RAVS) empowers users with creative control over video synthesis, fostering user engagement and customization, thereby advancing AI and machine learning technologies [14]. Addressing performance disparities between closed-source and open-source models in video generation is crucial for equitable advancements in AI technologies [15]. Optimizing prompts in Generative AI systems is vital for enhancing the quality of AI-generated outputs, further emphasizing the significance of video generation in AI [8]. Proposed methods demonstrate significant improvements in video quality, successfully encoding temporal styles transferable to unseen targets, thus enhancing the diversity and dynamics of generated content [9]. Collectively, these advancements underscore video generation's role in driving innovation and application across sectors, contributing to the broader advancement of AI and machine learning technologies.

1.3 Bridging Literature Gaps

The current landscape of video generation research is marked by critical gaps that hinder effective application and advancement. A prominent issue is the inadequate integration of video generation models in addressing real-world tasks, despite the abundance of available video data online [2]. This necessitates a more robust framework that effectively incorporates video generation techniques into practical applications. Furthermore, the literature reveals a notable deficiency in the systematic co-development of multi-modal data and generative models, resulting in suboptimal performance and inefficient resource utilization [3].

The absence of a comprehensive classification system for types of human feedback within the context of Reinforcement Learning with Human Feedback (RLHF) limits the effective integration of insights from human-computer interaction and machine learning [16]. This gap is exacerbated by insufficient support for both novice and expert users in existing video editing technologies, which fail to offer a balanced and empowering structure [14]. Additionally, the scarcity of robust open-source foundation models in video generation presents a significant barrier to innovation and progress in the field [15].

The optimization of prompts for Generative AI systems to enhance human-AI interaction remains underexplored, necessitating focused efforts to improve user engagement and satisfaction [8]. Reliance on manually labeled datasets for benchmarking, which are often costly, time-consuming, and lacking in diversity, complicates the evaluation of AI-generated content [17]. Moreover, aggregating potentially diverging human feedback into consistent data about collective preferences presents a complex challenge in AI alignment [18].

Evaluating AI-generated video content is fraught with challenges, particularly due to the intricate spatial and temporal dynamics involved [19]. Existing evaluation metrics often fail to align with human perceptions, underscoring the need for comprehensive and versatile benchmarks [20]. Furthermore, the lack of open-source datasets and precise control over motion patterns in video generation has been identified as a significant gap that needs addressing [21].

This survey aims to bridge these gaps by providing a thorough analysis of current methodologies and proposing new frameworks that facilitate the integration of video generation models into real-world applications. By addressing these critical issues, the survey seeks to enhance the understanding and development of video generation technologies, ensuring alignment with human values and the capability of producing high-quality, reliable content. Additionally, it evaluates the social impacts of RLHF, addressing potential benefits and ethical concerns associated with its application [22]. By doing so, the survey contributes to advancing the field towards more ethical and socially responsible AI technologies.

1.4 Structure of the Survey

This survey is systematically organized to provide a comprehensive exploration of the multidisciplinary field encompassing video generation, human feedback, reinforcement learning, generative models, computer vision, and user interaction. The paper opens with an introduction emphasizing the crucial importance of integrating various domains, such as generative artificial intelligence and large language models, within artificial intelligence (AI) and machine learning frameworks to enhance capabilities in video generation, understanding, and streaming, addressing existing challenges and exploring future research opportunities [23, 24, 25, 26, 27]. Following this, the survey delves into the background and core concepts essential for understanding the interplay between these technologies.

The subsequent sections examine specific aspects of video generation and its applications. Section 2 provides foundational definitions and explanations of core concepts, setting the stage for more detailed discussions. Section 3 explores advanced video generation techniques, including the use of generative adversarial networks (GANs) and diffusion models, and their applications in creating realistic and high-quality video content. Section 4 focuses on the role of human feedback in refining video generation models, discussing methods for incorporating such feedback to enhance model performance and alignment with human preferences.

Section 5 investigates the application of reinforcement learning in video generation, highlighting key algorithms and their effectiveness in enhancing video generation processes. Section 6 offers a comprehensive analysis of the role of generative models, particularly GANs, in computer vision, focusing on their application in video generation. It highlights challenges associated with generating videos from text, detailing a hybrid framework that utilizes both Variational Autoencoders (VAEs) and GANs to effectively extract and synthesize static and dynamic features from textual input. The section further discusses how these models enhance the quality and diversity of generated videos and improve the interpretation of visual information by accurately reflecting the nuances of the input text. Additionally, it reviews the current state of video GANs, categorizing them based on their conditional frameworks and addressing the limitations and challenges that remain in this evolving field [28, 29]. Section 7 explores user interaction and engagement, examining how user input influences video generation and the development of user-centric models.

The survey concludes with Section 8, identifying current challenges and future directions in the field, such as scalability, model alignment with human values, and ethical considerations. The paper emphasizes the need for video generation models to act as planners and agents in real-world scenarios, covering applications in domains like robotics, self-driving, and science [2]. Additionally, the survey introduces a structured framework for AI-Generated Video Evaluation (AIGVE), emphasizing the dual criteria of alignment with human perception and instructions [19]. This structured approach ensures a thorough examination of the field, providing valuable insights and directions for future research and development.The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Video Generation

Video generation involves synthesizing coherent, high-resolution sequences from inputs like images, text, and motion patterns, crucial for AI applications in fields such as autonomous driving and human-computer interaction. The task's complexity arises from the high-dimensional nature of video data, requiring realistic spatial and temporal modeling. Advanced frameworks, such as "Make-Your-Video," utilize textual prompts with motion guidance to enhance output control, while latent path construction addresses challenges in generating videos from text [30, 31].

Current methodologies, including conditional generation and embodied AI, leverage deep generative models [2]. Large language models like VideoPoet showcase zero-shot capabilities, processing multimodal inputs to create high-quality videos [32]. Despite progress, predicting future frames remains challenging due to dynamic transitions, addressed by models like VideoVAE, which integrates variational autoencoders with long short-term memory networks for temporal coherence [13, 33].

Synthesizing videos from natural language requires adherence to storyboards and temporal coherence, complicated by the need for explicit 3D camera control, often lacking in generative models [34, 35]. Innovative methods like C3V and Make-A-Video extend capabilities by decomposing prompts into concepts or integrating spatiotemporal layers [36, 37]. Video generation enhances human-computer interaction and machine learning model interpretation, with applications in predictive modeling and personalized medicine [38]. Overcoming challenges like controlling camera transitions remains critical, with initiatives like HunyuanVideo focusing on architecture design and training strategies to improve capabilities [15].

Video-to-video synthesis aims for photorealistic outputs with smooth transitions, yet predicting future frames is hindered by object dynamic uncertainties [39]. Customized generation, guided by text and reference images, emphasizes high-quality adherence to constraints. Markov decision processes and comprehensive benchmarks like VBench ensure temporal coherence and realism, addressing issues like freezing artifacts [40, 41].

2.2 Human Feedback

Human feedback is vital for refining AI models, aligning outputs with user expectations when training data is insufficient. In reinforcement learning, it provides nuanced critique, facilitating adaptation to complex tasks [42, 21]. Feedback includes implicit cues and explicit instructions, reducing reliance on predefined datasets and enhancing model customization [21]. In video generation, it enables robots to learn task performance from human-generated videos, improving user satisfaction [42].

Benchmarks for evaluating text-to-video models underscore feedback's importance in assessing alignment with prompts, focusing on dynamics and vividness [43]. As AI progresses, incorporating feedback remains crucial for models that are technically proficient and aligned with human values, empowering users to specify content characteristics and enhancing customization and engagement.

2.3 Reinforcement Learning

Reinforcement learning (RL) is crucial in video generation, enabling models to learn optimal strategies for creating coherent sequences through dynamic environment interactions. It allows discovery of strategies maximizing rewards, particularly when explicit programming is impractical [39]. Integrating RL with generative models like GANs and VAEs has significantly advanced technologies, as seen in MoCoGAN-HD, which enhances temporal coherence by discovering motion trajectories in latent space [44].

RL models stochastic future events, reducing blurry predictions [45]. The Make-A-Video approach extends text-to-image models with spatiotemporal layers, enhancing temporal dynamics without paired text-video data [37]. Challenges like handling raw video signals remain significant [46].

AI-generated, for reference only.

Diffusion models address high-quality generation from text prompts, highlighting visual and temporal dynamics [31, 40]. Evaluating generated videos on visual quality and temporal consistency is critical for assessing RL applications [47].

Traditional RL fine-tuning can lead to static motion and poor combinations, necessitating innovative approaches to maintain adaptability [48]. Overcoming these challenges is essential for advancing AI-driven content creation and future innovations in video generation.

2.4 Generative Models

Generative models are foundational in AI, enabling the synthesis of new data resembling original datasets. Models like GANs, diffusion models, and autoregressive models facilitate high-quality content generation [49]. The first open-source Image-to-Video model highlights their transformative potential, animating reference images into videos [49].

Beyond creation, generative models enhance adaptability through human feedback integration, refining outputs in unexplored domains [50]. The UniReal framework exemplifies this, offering a unified approach to image generation tasks [50]. Challenges remain in generating clinically plausible synthetic medical images, underscoring the need for refined architectures [51].

Generative models contribute to AI evolution, enabling diverse content creation, particularly in video production. Recent advancements, including conditional models, VAEs, and GANs, enhance video understanding and streaming through LLMs, transforming video technology and AI applications [28, 24]. Their role encompasses content generation, system adaptability, and enhancing human-AI interaction, contributing to intelligent systems that learn and adapt to human expectations.

2.5 Computer Vision

Computer vision, a fundamental AI field, enables machines to interpret visual information, significantly advancing video generation capabilities. This integration is crucial for simulating object interactions over time, capturing visual and temporal dynamics [52]. Hierarchical datasets like GenVideo provide a foundation for training and evaluating models [53].

Recent advancements include methods like Image Conductor, enhancing precision and realism by controlling camera transitions and object movements [54]. Collaborative diffusion processes, such as CVD, synchronize outputs through cross-view modules, ensuring sequence coherence [55]. Datasets like HumanVid, with diverse human-centric videos, advance integration with video generation, capturing interactions and behaviors [56]. The combination of computer vision and video generation enhances technical capabilities and applicability across domains like entertainment and autonomous systems. Continuous refinement of models and datasets is essential for advancing this integration.

2.6 User Interaction

User interaction is pivotal in developing user-centric AI models, particularly in video generation, where preferences shape output and functionality. Integrating interaction mechanisms allows personalized experiences, enabling control over video generation aspects through multimodal instructions [57]. This approach facilitates customizing outputs, ensuring alignment with individual needs.

Frameworks emphasizing user-specified parameters, like expression length and type, highlight the necessity of incorporating feedback to enhance content relevance [58]. Developing systems for interpreting and generating multimodal information improves real-world interaction, advancing user-centric models [59]. Datasets like HumanVid, with diverse data types, provide a foundation for training models to respond accurately to inputs [38]. Human preference annotations enhance output tailoring, as shown by benchmark suites [60]. These datasets and methods are essential for developing AI models aligned with user values and expectations [19].

3 Video Generation Techniques

Recent advancements in video generation have been propelled by innovative techniques enhancing the quality and coherence of synthesized content. This section explores key methodologies, beginning with Generative Adversarial Networks (GANs), which have become foundational in creating realistic

Category	Feature	Method
Generative Adversarial Networks (GANs)	Training Techniques Output Quality Improvement	MAGVIT[46], TSGAN[61] MoCoGAN-HD[44]
Diffusion Models	Space-Based Diffusion	GR[62], TwoStreamVAN[63], VC1[49]
Text-to-Video Generation	Dynamic Sequence Modeling	MAV[37], MYV[31]
Multimodal and Controllable Video Generation	User Feedback and Customization Aspect Control Input Integration	MORA[64] DAV[65] MMVID[66]
Unsupervised and Autonomous Video Generation	Control Mechanisms	CAGE[67], GenDeF[68]

Table 1: This table provides a comprehensive summary of recent advancements in video generation methodologies, categorized into five main areas: Generative Adversarial Networks (GANs), Diffusion Models, Text-to-Video Generation, Multimodal and Controllable Video Generation, and Unsupervised and Autonomous Video Generation. Each category is further detailed with specific features and methods, highlighting key innovations and frameworks that have contributed to the evolution of video generation techniques.

video sequences through a competitive framework that synthesizes high-quality visual content. Table 1 offers a detailed summary of the current state of video generation methodologies, outlining the categories, features, and methods that define recent advancements in this field. Additionally, Table 4 presents a comprehensive comparison of the leading methodologies in video generation, detailing the core techniques, innovations, and challenges associated with Generative Adversarial Networks (GANs), Diffusion Models, and Text-to-Video Generation. **??** illustrates the hierarchical structure of video generation techniques, categorizing them into five main areas: Generative Adversarial Networks (GANs), Diffusion Models, Text-to-Video Generation, Multimodal and Controllable Video Generation, and Unsupervised and Autonomous Video Generation. Each category is further divided into subcategories highlighting key innovations, frameworks, techniques, and challenges, thereby providing a comprehensive overview of the current landscape in video generation.

3.1 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) are central to video generation, producing high-quality and temporally coherent sequences. The GAN architecture comprises a generator, which creates video content, and a discriminator, which differentiates real from generated sequences, thus capturing complex spatial and temporal dynamics [69, 61]. The Two-step GAN (TSGAN) exemplifies this approach by using a convolutional generator for static frames and a recurrent model for sequencing, improving temporal coherence [61]. Despite progress, GANs face challenges like training instability and mode collapse [61]. Innovations such as the Masked Generative Video Transformer (MAGVIT) demonstrate GANs' adaptability in complex video data handling [46]. Comparative analyses indicate diffusion models often surpass GANs in visual quality and temporal coherence [40]. Novel architectures like MoCoGAN-HD illustrate GANs' potential in generating high-resolution videos with enhanced temporal dynamics [44]. The evolution of GANs continues to advance video generation, providing robust solutions for creating realistic and contextually relevant content.

3.2 Diffusion Models

Diffusion models have emerged as a powerful approach in video generation, recognized for their ability to produce high-quality and temporally coherent sequences by refining noise into structured data [49]. These models often outperform GANs in visual quality, handling complex tasks adeptly [62]. The Endora framework exemplifies their versatility, integrating spatial-temporal transformers for dynamic content synthesis [11]. Innovations like UV-space noise initialization and enhanced self-attention layers ensure consistent rendering across frames [62]. In predictive modeling, the VideoFlow method extends flow-based models for precise and diverse future frame prediction [63]. Diffusion models are increasingly replacing traditional methods, with advancements like Control-A-Video and Make-Your-Video expanding creative possibilities in AI-generated media [70, 31, 71, 49, 72].

3.3 Text-to-Video Generation

Text-to-video (T2V) generation represents a significant AI advancement, enabling the synthesis of video sequences from textual descriptions. The Make-A-Video method exemplifies this by using paired text-image data and unsupervised video footage to generate videos directly from text inputs [37]. Frameworks like compositional 3D-aware video generation decompose textual prompts into sub-prompts for 3D representations, showing the potential of integrated vision and language models [36, 40]. Make-Your-Video addresses T2V synthesis limitations by ensuring contextual relevance and visual coherence [31]. Innovations like Latent Linear Interpolation improve video fidelity and narrative coherence by creating intermediate frames from textual descriptions [73, 36, 74, 30, 43]. Despite advancements, challenges remain in replicating closed-source systems' capabilities, particularly in generating lengthy, coherent videos from text [75, 74, 27, 76, 43]. The integration of advanced generative techniques and user interaction mechanisms continues to advance T2V generation, paving the way for innovative applications across various domains.

3.4 Multimodal and Controllable Video Generation

Method Name	User Control	Integration Techniques	Multimodal Frameworks
MMVID[66]	Textual Prompts	Bidirectional Transformer	Multimodal Video Generation
MORA[64]	Human-in-the-loop	Self-modulation Collaboration	Multi-agent Framework
DAV[65]	Decoupled Control	Attention Mechanisms	Text Prompt

Table 2: Comparison of various multimodal and controllable video generation methods, highlighting user control mechanisms, integration techniques, and multimodal frameworks. The table provides insights into how different methods like MMVID, MORA, and DAV achieve alignment with user preferences and creative intentions through advanced techniques.

Multimodal and controllable video generation enables high-quality, user-responsive content synthesis across different modalities. Table 2 presents a comprehensive comparison of contemporary frameworks in multimodal and controllable video generation, detailing their user control features, integration techniques, and underlying multimodal frameworks. Frameworks like MMVID allow users to specify video content through visual inputs and guide generation with textual prompts, enhancing user control [66]. The Mora framework employs multi-agent systems and human-in-the-loop mechanisms to refine data filtering and align outputs with user preferences [64]. Direct-a-Video provides independent control over camera movements and object motions, offering an immersive, personalized experience [65]. These frameworks integrate advanced techniques like Latent Diffusion Models and bidirectional transformers to achieve alignment with creative intentions, resulting in high-quality, coherent videos [19, 66, 31].

3.5 Unsupervised and Autonomous Video Generation

Method Name	Generative Models	Control Mechanisms	Quality and Consistency
CAGE[67]	Unsupervised Video Generation	Unified Control Format	High Video Realism
GenDeF[68]	Gan-based Framework	Motion-related Deformation	Enhanced Visual Quality

Table 3: Overview of video generation methods highlighting the generative models, control mechanisms, and the resulting quality and consistency. The table compares the CAGE and GenDeF frameworks, emphasizing their respective approaches to unsupervised video generation and control.

Unsupervised and autonomous video generation focuses on synthesizing content without labeled training data, leveraging advanced generative models to capture complex dynamics. Table 3 provides a comparative analysis of two prominent frameworks, CAGE and GenDeF, in the domain of unsupervised and autonomous video generation. MoCoGAN-HD exemplifies this with a pre-trained image generator for high-quality frames, discovering motion trajectories in latent space [77]. The CAGE framework conditions on random visual features for controlled scene composition [67]. Autoregressive models like YODA process motion controls for coherent sequences, while MAGE aligns visual and motion information for dynamic content [78, 79]. GenDeF enhances visual quality and temporal consistency by separating content and motion, facilitating nuanced control [68]. These techniques, including Markov Decision Process frameworks, advance AI's capability to independently produce high-quality video content, addressing challenges in evaluation and alignment with human perceptions [28, 19, 31].

Feature	Generative Adversarial Networks (GANs)	Diffusion Models	Text-to-Video Generation
Core Technique	Adversarial Training	Noise Refinement	Textual Synthesis
Key Innovation	Masked Video Transformer	Uv-space Initialization	Latent Interpolation
Main Challenge	Training Instability	Not Specified	Coherence IN Lengthy Videos

Table 4: This table provides a comparative analysis of three prominent video generation methodologies: Generative Adversarial Networks (GANs), Diffusion Models, and Text-to-Video Generation. It highlights their core techniques, key innovations, and main challenges, offering insights into the distinct features and limitations of each approach. This comparison aids in understanding the current advancements and challenges in the field of video generation.

4 Role of Human Feedback

4.1 Importance of Human Feedback in Video Generation

Human feedback is pivotal in refining video generation models, aligning them with user expectations to enhance realism and coherence. This mechanism allows user-driven control over video actions via various interfaces, effectively capturing both static and dynamic elements to produce contextually relevant sequences. In reinforcement learning, human feedback aligns agent behaviors with human values, refining learning processes and model adaptability [22]. Particularly when training data is insufficient, human insights enhance model precision and adaptability, as shown by the COACH algorithm, which uses policy-dependent feedback to improve learning efficiency [80]. Frameworks like YODA demonstrate controllable video generation without annotations, modeling complex interactions [78].

In creative domains, feedback empowers users to control motion, enhancing expression and output quality. CustomCrafter underscores the role of feedback in refining image generation quality [19]. Models like VideoVAE benefit from efficient feedback, achieving faster generation and improved visuals [33]. Evaluation benchmarks like VBench++ align metrics with human judgment, enhancing model evaluation [20]. The DEVIL protocol focuses on dynamics in T2V generation, addressing overlooked aspects in benchmarks [43]. Recent advancements, such as the LiFT method, utilize human ratings to train reward models, significantly improving video quality. Rich feedback mechanisms, including marking implausible regions or annotating misrepresented words, refine content generation, fostering systems that meet user expectations and enhance user experience [81, 82, 83, 42, 43].

4.2 Methods for Incorporating Human Feedback

Incorporating human feedback into video generation models enhances alignment with user expectations, improving content quality. This involves constructing datasets of human ratings, training reward models to capture preferences, and employing evaluation protocols considering dynamics and temporal consistency [84, 43, 81, 83]. InstructVideo exemplifies refining outputs through human feedback without full regeneration [85]. The COACH algorithm modifies policies based on feedback as an advantage function, enabling interactive learning [80]. RAVS framework empowers users to define actions interactively, enhancing engagement.

Frameworks categorize feedback into dimensions, improving RLHF systems [16]. The HATL creates a dynamic feedback loop, enhancing generative AI models [8]. Self-attention mechanisms in diffusion models incorporate feedback, aligning outputs with user preferences [51]. QDHF learns diversity metrics from feedback, optimizing for diverse solutions without manual metrics. VBench++ provides multi-dimensional evaluation, facilitating feedback integration [20]. RichHF-18K dataset enables effective feedback incorporation in assessments [42].

Integrating human feedback is crucial for user-centric AI systems. By applying innovative methods, researchers can significantly improve AI-generated content, ensuring it meets user preferences. This approach addresses challenges in video generation and evaluation, leveraging advancements in AI to create personalized content, enhancing user engagement in multimedia applications [86, 24, 87, 19, 27].

4.3 Impact of Human Feedback on Model Alignment

Human feedback aligns video generation models with user preferences, ensuring outputs are proficient and relevant. It refines outputs by revealing user expectations not captured by datasets [88]. While feedback can affect trust and accuracy perceptions, it balances performance and satisfaction. Iterative refinement enhances models' ability to replicate intricate interactions. Feedback guides accuracy and relevance, addressing alignment complexities, as shown in studies on fine-grained feedback and human-in-the-loop systems. These insights highlight the need for effective feedback mechanisms [89, 88, 90, 83, 91].

Incorporating feedback into evaluation processes fine-tunes models to user preferences, enhancing effectiveness. This approach incorporates detailed feedback, capturing distinctions in quality and alignment, addressing trust complexities. While feedback corrects errors and adapts to data changes, it can affect trust if mismanaged. Comprehensive evaluation protocols considering dynamics refine model capabilities, ensuring outputs remain vivid and honest [81, 92, 88, 90, 43]. This enhances technical capabilities and fosters trust and engagement, as models align with preferences.

Human feedback enhances T2V model alignment, as evidenced by the LiFT method using Human Rating Annotation datasets. This approach addresses alignment challenges with subjective expectations, improving quality and metrics. Large-scale datasets like VideoFeedback and protocols like DEVIL highlight comprehensive feedback's importance, refining technologies for accurate outputs [47, 43, 81, 83]. Leveraging feedback enhances performance, ensuring high-quality content aligned with expectations, advancing AI-driven content creation.

4.4 Challenges in Utilizing Human Feedback

Utilizing human feedback in video generation faces challenges hindering alignment with expectations. RLHF risks outputs misaligned with ethical standards, necessitating mechanisms reflecting diverse values [22]. High feedback costs hinder adoption, emphasizing cost-effective methods for capturing insights [22]. Lack of robust metrics complicates quality assessment, as benchmarks rely on subjective evaluations, hindering objective comparisons [93]. Temporal consistency and accurate scene representation remain challenging, with studies struggling with motion coherence.

Ethical considerations complicate feedback utilization, concerning authenticity and representation. Biases and implications necessitate continuous adaptation of mechanisms to mitigate misuse [26]. As models evolve, protective measures like VGMShield may diminish, requiring ongoing refinement [94]. A comprehensive strategy integrating dataset advancements, such as LiFT-HRA with 10,000 annotations, and innovative metrics like DEVIL, emphasizing dynamics, is essential. Ethical considerations ensure meaningful feedback use, demonstrated by rich feedback enhancing training. Maximizing insights improves text-to-video models' quality and alignment with user preferences [81, 84, 19, 83, 43].

5 Reinforcement Learning in Video Generation

5.1 Key Algorithms and Frameworks

Reinforcement learning (RL) plays a pivotal role in enhancing video generation through various algorithms and frameworks. The Convergent Actor-Critic by Humans (COACH) algorithm demonstrates this by incorporating policy-dependent human feedback, which facilitates convergence to local optima and refines agent performance according to human preferences [80]. This integration improves model adaptability and output quality.

The VGPNN method sets a benchmark for RL applications by enabling rapid and diverse video generation from a single input, showcasing RL's potential to foster creativity and diversity in video content [61]. This method effectively manages the complexities introduced by high dimensionality and temporal components of videos.

InstructVideo leverages RL by focusing on editing existing videos based on human feedback, rather than generating new content, underscoring the importance of human insights in refining outputs [85]. This approach aligns generated content with user expectations, enhancing video quality and relevance.

AI-generated, for reference only.

The DreaMoving framework utilizes RL by training on high-quality dance videos to learn motion patterns and identity features, demonstrating RL's capacity to capture complex motion dynamics [21]. By leveraging real-world video data, this method achieves a more accurate representation of dynamic scenes [95].

Probabilistic video generation techniques conditionally sample from a structured latent space that incorporates holistic attributes and temporal information, ensuring generated sequences are both diverse and coherent [33]. Additionally, methods that independently manipulate motion and content, such as those using sinusoidal functions to capture periodic motion patterns, highlight RL's adaptability in video generation [9]. These advancements underscore RL's critical role in developing high-quality, contextually relevant content that aligns with human values and preferences. By integrating innovative techniques and human feedback, RL continues to advance AI-driven content creation, paving the way for future innovations.

5.2 Challenges and Optimization Techniques

Applying reinforcement learning (RL) to video generation presents several challenges that necessitate robust optimization techniques. A primary issue is ensuring temporal consistency, especially when user modifications deviate from trained data distributions, leading to inconsistencies in generated content [57]. This challenge is compounded by the inefficiency of methods like per-frame Neural Radiance Fields (NeRFs), which require substantial computational resources, making them impractical for dynamic scenes [96].

Reliance on the last frame of a generated clip as input for the next often results in inconsistencies and a lack of temporal coherence across sequences [97]. Models also struggle to generalize effectively in out-of-distribution scenarios, despite strong performance in in-distribution settings [98]. The inherent complexity of video data and the extensive computational demands of RL further exacerbate these challenges [77].

To address these issues, optimization techniques are essential. Aggregating feedback from diverse trainers and adjusting the RL policy based on each trainer's reliability can enhance model performance [99]. Leveraging universal supervision from video data allows models to learn from natural variations, maintaining consistency across diverse image generation and editing tasks [50]. Nonetheless, challenges remain regarding temporal consistency and visual artifacts, particularly in the Image-to-Video (I2V) model [49].

Current methods often struggle with high-resolution videos, limiting the generation of detailed scenes [100]. While the Markov Decision Process (MDP) formulation is beneficial, it may increase computational complexity, necessitating more resources for training [41]. Future research should focus on refining detection methods to adapt to advancements in video generation technology and exploring defense mechanisms against video modification [94].

To effectively advance RL's application in video generation, it is crucial to tackle inherent challenges through innovative optimization techniques and comprehensive evaluation frameworks. This approach not only enhances the capability to generate high-quality, contextually relevant video content but also addresses critical issues such as generating long-duration videos and controlling content through advanced guidance methods. Recent research emphasizes strategies like joint-conditional video generation and two-stage approaches to improve the diversity and coherence of generated videos, leading to more robust and user-preferred outcomes [75, 76, 31].

5.3 Future Directions in Reinforcement Learning for Video Generation

Future research in reinforcement learning (RL) for video generation is poised to explore critical avenues to enhance the robustness, efficiency, and applicability of generated content. One promising direction involves improving the model's robustness and efficiency during training and inference, essential for advancing RL capabilities in video generation [15]. This includes optimizing algorithms to manage more complex interactions and varying motion inputs, thereby broadening the applicability of video generation models [78].

Another significant area for exploration is integrating textual descriptions to control video generation, enhancing the interactivity and precision of generated content [63]. By developing models that interpret and respond to textual inputs, researchers can create more user-centric video generation systems that align closely with user intentions and preferences.

Additionally, research should focus on developing new architectures that combine the strengths of various modeling approaches to improve the fidelity and quality of video outputs. This includes exploring innovative frameworks that enhance the model's ability to generate realistic and contextually relevant videos, ensuring high-quality outputs across diverse applications [15].

Expanding RL applications in video generation to include more complex physical interactions and refining evaluation methods to minimize subjectivity are crucial steps for advancing the field. By introducing comprehensive assessment frameworks, researchers can ensure the consistency and quality of multi-scene videos, thereby enhancing the overall effectiveness of video generation models [63].

The future of reinforcement learning in video generation lies in addressing these challenges and exploring innovative solutions that enhance model performance, efficiency, and user alignment. Focusing on critical areas such as AI-generated video evaluation and advancements in text-to-video generation technologies will drive significant progress in AI-driven content creation. This focus aims to enhance the effectiveness of evaluation frameworks assessing video quality, semantic coherence, and alignment with human intent while addressing existing challenges in the field. As generative models like Sora evolve, ensuring that video generation technologies adapt to the complexities of diverse applications will be essential for meeting the increasing demands of content production and consumption [27, 19].

6 Generative Models and Computer Vision

6.1 Integration of Generative Models in Computer Vision

The integration of generative models into computer vision represents a significant advancement in AI, enhancing visual information synthesis, interpretation, and manipulation. The MMVID framework exemplifies this by using a bidirectional transformer to generate video representations from text and image inputs, illustrating the power of multimodal integration [66]. Similarly, the StoryAgent framework employs the LoRA-BE method for Image-to-Video conversion, ensuring narrative coherence and visual consistency [101]. The TVP method leverages causal relationships between text and motion to produce coherent video frames, emphasizing the interplay between modalities for high-quality outputs [102].

Frameworks like Conditional FlowGAN and Conditional TextureGAN synthesize realistic video sequences by modeling complex spatial and temporal dynamics [103]. The Endora framework enhances video content by capturing long-range spatial-temporal correlations through an advanced transformer architecture [11]. DirectorLLM integrates generative models in pose interpolation and video generation, refining human-centric content [104]. The HunyuanVideo framework adopts a systematic approach to generative model integration, significantly outperforming existing models [15].

Overall, integrating Generative Adversarial Networks (GANs) and large language models (LLMs) into computer vision tasks enhances synthesis of diverse, high-quality content. Techniques such as conditional generation and noise introduction in diffusion models bridge real-world dynamics with digital creation, effectively interpreting and generating complex visual data. This advancement extends AI's potential applications in multimedia and content creation across various sectors, including media and education [24, 25, 28, 29, 69].

6.2 Techniques and Frameworks for Video Generation

Video generation is enriched by techniques and frameworks leveraging generative models for high-quality, contextually relevant content. The Gen-L-Video framework extends short video diffusion models to manage long video generation and editing, maintaining coherence and quality over extended durations [73]. It addresses long video content complexities, preserving narrative continuity and visual consistency across diverse segments, building on diffusion models and large-scale image-text datasets [73, 28].

Advanced generative techniques, particularly through GANs and hybrid models combining Variational Autoencoders (VAEs), signify a transformative shift in video generation. These advancements enhance the quality and versatility of AI-generated video content, enabling effective application across sectors like media, education, and entertainment. Recent developments demonstrate significant improvements in visual fidelity and diversity, allowing for accurate reflection of textual input [28, 29].

6.3 Quality Assessment and Evaluation Metrics

Benchmark	Size	Domain	Task Format	Metric
SCV[52]	10,000	Video Generation	Video Synthesis	FVD
THG-Bench[93]	4,754,000	Talking-head Video Generation	Video Synthesis	SSIM, FID
PHAV[105]	39,982	Human Action Recognition	Action Recognition	Accuracy, F1-score
GENAI-ARENA[106]	9,000	Image Generation	Preference Voting	Elo Rating
GenVideo[53]	2,294,594	Ai-generated Video Detection	Video Classification	F1, AP
EA[107]	100	Art Generation	Open-Ended Query Evaluation	FID, FVD
HumanVid[56]	20,000	Human Image Animation	Video Generation	PSNR, FID
EvalCrafter[108]	700	Text-to-Video Generation	Video Generation	VQAA, VQAT

Table 5: Table ef presents a comprehensive overview of various benchmarks utilized in the evaluation of video and image generation tasks. Each benchmark is characterized by its size, domain, task format, and the specific metrics employed for quality assessment, highlighting the diversity and scope of evaluation criteria across different applications.

Evaluating generated video quality requires metrics capturing visual fidelity and temporal coherence. Traditional metrics like Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) often inadequately assess overall quality, focusing on static image properties [52]. Advanced metrics like Fréchet Video Distance (FVD) provide comprehensive assessments by effectively capturing video distribution [52].

Comprehensive benchmark suites such as VBench enhance evaluation by introducing metrics measuring video quality aspects, including temporal consistency and aesthetic appeal, reflecting human perceptions [60]. The VideoScore framework emphasizes metrics correlating with human assessments across multiple dimensions, highlighting the need for evaluation criteria encompassing both technical and perceptual aspects [47]. Table 5 provides a detailed overview of representative benchmarks that are employed in the evaluation of video and image generation tasks, illustrating the diversity of domains, task formats, and metrics used to assess quality and performance.

In applications like talking head video generation, metrics are selected for robustness in evaluating visual fidelity and correlation with human perception [93]. This alignment ensures generated content meets desired quality standards. The advancement of video generation technology necessitates sophisticated quality assessment metrics and comprehensive evaluation frameworks addressing video content dynamics, ensuring alignment with text prompts in terms of visual fidelity and coherence. Recent studies, such as the DEVIL evaluation protocol, stress the importance of dynamics in model performance assessment, demonstrating that metrics focused on this dimension enhance evaluation consistency with human ratings [19, 43, 31].

6.4 Innovations in Photorealistic Video Generation

Recent innovations in photorealistic video generation have significantly enhanced realism and quality of synthesized content. The PVG method illustrates this progress by synthesizing plausible scenes over extended time horizons, surpassing existing view generation techniques [109]. This method's ability to maintain visual coherence and detail over longer sequences represents a substantial advancement in photorealistic video generation.

The SPC model innovates by reducing distortion artifacts and preserving object structures in long-term predictions, addressing challenges in maintaining photorealism [110]. Additionally, the TwoStream-VAN model sets a new benchmark by outperforming state-of-the-art methods across various datasets, with its dual-stream architecture improving motion modeling and content synthesis [63]. These advancements underscore the importance of integrating sophisticated motion modeling techniques to achieve photorealism.

Innovative methodologies like Make-Your-Video leverage textual context and motion structure for customized generation. Comprehensive surveys on long video generation delineate critical challenges and methodologies, emphasizing robust evaluation metrics accounting for video content dynamics [75, 43, 31]. These innovations collectively reflect a growing emphasis on improving realism and coherence of generated videos, aligning them more closely with user expectations and textual descriptions.

6.5 Challenges and Defenses in AI-Generated Media

AI-generated media presents challenges and necessitates robust defenses to mitigate misuse. A significant concern is AI-generated content's potential exploitation for malicious purposes, such as deepfakes, which can mislead viewers and propagate misinformation. The rapid advancement in generating highly realistic videos complicates efforts to distinguish genuine media from fabricated content, as AI-generated videos mimic real-life scenarios. Furthermore, the intricate spatial and temporal dynamics of video content necessitate sophisticated evaluation frameworks to assess alignment with human perception, challenging traditional content verification methods and raising implications for misinformation and media integrity [87, 111, 19, 29, 112].

Innovative defense mechanisms are being developed to address these challenges. The VGMShield employs pre-trained video recognition models to detect inconsistencies and anomalies, preventing misuse [94]. This method leverages advanced AI capabilities to scrutinize video content for manipulation signs.

Developing standardized evaluation metrics and frameworks is essential for assessing AI-generated videos' quality and integrity. These advanced tools play a critical role in identifying artifacts and ensuring produced content meets ethical standards and aligns with human values, utilizing comprehensive evaluation protocols and rich human feedback mechanisms. This approach enhances fidelity and authenticity, addressing potential biases and improving alignment with user expectations, fostering responsible use in filmmaking, advertising, and education [87, 26, 90, 42, 43]. As AI-generated media evolves, ongoing research and collaboration are crucial for enhancing detection methods and establishing robust defenses against misuse.

7 User Interaction and Engagement

7.1 Frameworks for User-Controlled Video Generation

Frameworks for user-controlled video generation are pivotal in facilitating content customization and enhancing interactivity in AI-generated media. Approaches like 'Make-Your-Video' and 'Interactive-Video' integrate textual context and structural guidance, allowing users to create vivid videos tailored to their imaginative scenarios [57, 31]. These frameworks employ advanced generative models and user feedback mechanisms, enabling users to specify motion dynamics, narrative flow, and visual aesthetics [113].

The Direct-a-Video framework offers users control over camera movements and object motions, overcoming traditional video generation limitations [65]. Meanwhile, MMVID integrates multiple modalities, allowing users to guide video content generation with visual and textual inputs [66]. The Mora framework's multi-agent system facilitates collaboration among generative components and incorporates human-in-the-loop mechanisms for data refinement, ensuring alignment with user preferences [64].

These frameworks represent significant advancements in AI, offering unprecedented levels of user interaction and customization. By leveraging cutting-edge techniques such as Latent Diffusion and Variational Autoencoders, they improve video generation accuracy and enable diverse applications in entertainment and education [28].

7.2 Stages of User Interaction in Video Generation

User interaction in video generation involves stages crucial for aligning outputs with user preferences. Initially, user input captures specific parameters guiding the video generation process, ensuring content reflects desired attributes like length, style, and narrative flow [66]. The integration of multimodal inputs enhances content relevance by allowing precise articulation of preferences [59].

During the generation phase, advanced models produce video content adhering to specified parameters, with frameworks like MMVID and Direct-a-Video enabling control over camera movements and object dynamics [65]. The feedback and refinement stage allows users to evaluate content and provide insights for adjustments, facilitating continuous improvement [64].

These interaction stages significantly enhance interactivity and personalization, enabling innovative applications across fields such as customized video generation [19, 31].

7.3 User Feedback Mechanisms and Trust

User feedback mechanisms are essential for aligning video generation models with user preferences and fostering trust in AI-generated content. These mechanisms enable the collection of user insights, allowing models to refine outputs based on real-world expectations [88]. Interactive interfaces that allow real-time user input, such as modifications or satisfaction ratings, empower users to engage actively with the generation process [8].

Building trust involves ensuring transparency and accountability, with robust evaluation metrics aligning with human perceptions of quality [60]. By prioritizing user engagement and transparency, AI systems can align more closely with human values and expectations, paving the way for personalized and trusted video generation applications [92, 31].

7.4 Generative Recommender Systems

Generative recommender systems enhance user interaction through advanced machine learning techniques that provide personalized content recommendations. These systems leverage generative models to align closely with individual user preferences, addressing traditional retrieval-based paradigm limitations [28, 86]. By integrating user feedback and historical interaction data, these systems adapt dynamically to changing preferences, ensuring recommendations remain relevant.

These systems effectively model intricate user-item interactions, capturing nuanced user preferences through AI-generated content (AIGC) and natural language instructions [86, 114]. The incorporation of multimodal data enhances recommendation relevance, addressing traditional systems' limitations and leveraging advancements in AIGC and large language models [115].

Generative recommender systems are integral to enhancing user interaction by providing personalized and contextually relevant recommendations. By utilizing advanced generative models and integrating user feedback mechanisms, these systems tailor content to meet individual preferences, enhancing engagement and satisfaction across diverse fields [8, 86].

8 Challenges and Future Directions

The evolving field of video generation faces numerous challenges that significantly impact its research and development trajectory. This section explores the ethical complexities, scalability issues, and future directions that are crucial as technology advances.

8.1 Ethical Considerations and Challenges

Key ethical concerns in video generation include data privacy, authenticity, and potential misuse of content. Advanced models like CustomCrafter struggle to accurately depict complex interactions, raising ethical implications [48]. Sparse input reliance complicates these issues, as nuanced interactions may be inadequately represented [78]. A lack of understanding of causal relationships and dynamic interactions further affects output quality, necessitating ethical scrutiny to prevent AI misrepresentation [40]. The inconsistency of human feedback can undermine model reliability, posing ethical challenges [80]. Additionally, expert feedback on synthetic medical images is resource-intensive, limiting scalability and accessibility [51]. Evaluation benchmarks, such as those for talking-head videos, must align with human perceptions to maintain ethical standards [93]. Frameworks like VBench++ improve ethical evaluations by reflecting human judgment, though current benchmarks require more granularity [20, 43]. Aligning reinforcement learning with human values presents both opportunities and challenges, emphasizing the need for unified evaluation methodologies and enhanced interpretability to address ethical concerns in AI-generated video technologies [22, 19].

AI-generated, for reference only.

The quality of training data, especially in underrepresented contexts, further underscores the ethical implications of data reliance [42].

8.2 Scalability and Computational Resources

Scalability and computational demands are significant hurdles in video generation. Detecting and tracing fake videos is challenging due to diverse generation techniques and the need for robust solutions [94]. Existing dynamics grades fail to capture real-world video dynamics, necessitating comprehensive evaluation frameworks [92]. Current benchmarks inadequately address adherence to physical laws, complicating assessments of physical commonsense in text-to-video models [116]. Real-time animation capabilities are hindered by low-dimensional latent spaces, leading to artifacts that challenge scalability [62]. The lack of benchmarking datasets and generalization of detection methods across media types further complicate scalability, highlighting the need for robust solutions against real-world distortions [111]. Training data biases and hallucinations in outputs also affect AI agents' reliability [7]. Models like VGPNN and RD-GAN face challenges in maintaining geometric consistency and stability, impacting video quality and duration [117, 118]. Existing benchmarks focus on image generation and lack effective methods for evaluating unsafe video generation, complicated by temporal information in videos [119]. Addressing these challenges requires innovative approaches to enhance the scalability and computational efficiency of video generation models.

8.3 Temporal Consistency and Motion Modeling

Temporal consistency and motion modeling are crucial for creating coherent and realistic video sequences. However, existing methods often struggle to integrate user-directed actions while maintaining consistent motion across frames [31]. Generating longer videos with multiple scenes requires detailed actions inferred from textual descriptions, adding complexity [37]. Processing numerous input and output images poses computational challenges, particularly as image numbers increase [50]. Optimizing generative frameworks to handle larger data volumes without sacrificing temporal coherence is essential. Future research should focus on methodologies that improve the generation of extended, multi-scene videos by effectively integrating user guidance and actions. Recent advancements in text-to-video synthesis suggest that structured guidance, such as frame-wise depth information, can enhance video quality [75, 31, 27, 76, 43]. Addressing these challenges will advance the field, ensuring high-quality, contextually relevant video content.

8.4 Ethical Considerations and Content Authenticity

Ensuring authenticity and addressing the potential misuse of AI-generated content are critical ethical considerations in video generation. Maintaining aesthetic and technical quality is challenging, as issues like flickering and inconsistent motion can undermine trust [71]. Models like DreamVideo struggle with multiple subjects and motions, necessitating robust approaches to maintain authenticity [120]. The misuse of video generation technologies raises ethical dilemmas, as they can be exploited to create misleading content [94]. The lack of proactive defenses in existing studies exacerbates trust issues across demographics, highlighting the need for inclusive solutions [111]. Future research should enhance models like DisenStudio for better resolution and detail handling, crucial for authenticity [121]. Integrating pluralistic approaches in RLHF is vital for developing humanistic language models that meet diverse societal needs [122]. Improving evaluation metrics aligned with human preferences is essential to ensure that generated content resonates with human values, thereby enhancing trust [123]. Frameworks like Collaborative Video Diffusion (CVD) that maintain consistency across camera angles represent significant advancements in ensuring authenticity [55]. Addressing ethical considerations and ensuring content authenticity requires a multifaceted strategy, including improving model capabilities through enhanced evaluation protocols, developing robust defenses against misuse, and aligning metrics with human values [94, 119, 43]. Focusing on these areas will advance the field ethically and in line with societal expectations.

8.5 Dataset Diversity and Evaluation Metrics

Dataset diversity and robust evaluation metrics are essential for advancing video generation. Diverse datasets enable models to generalize across contexts, enhancing the fidelity of generated content [124]. The Flickr8k dataset exemplifies the importance of varied training data in capturing real-world nuances [125]. Standardized evaluation metrics are urgently needed to objectively assess generative models' performance across domains. Current research reveals gaps in GAN performance evaluation in video generation, complicating model output comparisons [126]. Developing comprehensive evaluation metrics is essential for ensuring generated content meets quality and realism standards. Future research should enhance forgery detection and address diverse trustworthiness issues to develop practical evaluation metrics applicable in real-world scenarios [111]. By tackling these challenges, researchers can improve the reliability and applicability of video generation models, ensuring high-quality, contextually relevant content. Robust datasets and evaluation frameworks will play a pivotal role in advancing AI-driven content creation across various domains.

8.6 Integration with Advanced Models and Technologies

Integrating video generation with advanced models and technologies is a burgeoning area in AI, focusing on enhancing generative capabilities across domains. Future research should incorporate additional 3D cues and improve object tracking to enhance frame consistency, addressing challenges related to overlapping trajectories [65]. Expanding datasets and refining evaluation metrics to include a broader range of fluid dynamics phenomena is critical for ensuring robust and efficient video generation technologies [6]. Enhancing models' understanding of causal relationships and improving text-vision alignment metrics are vital for developing general models capable of processing multiple modalities [40]. Exploring training data with larger image numbers and investigating efficient model architectures will enhance performance in complex tasks [50]. Integrating emotional feedback into reinforcement learning frameworks offers a promising direction for advancing AI technologies, facilitating more nuanced interactions [28, 31]. The integration of advanced video generation technologies, such as customized synthesis using textual and structural guidance, long video generation techniques, and hybrid frameworks combining Variational Autoencoders with Generative Adversarial Networks, presents innovative opportunities across various fields, including entertainment, education, and marketing. This advancement enhances the quality and diversity of generated content while allowing precise control over video characteristics, pushing the boundaries of automated video production [28, 76, 31]. By focusing on these areas, researchers can significantly advance the field, ensuring that video generation technologies continue to evolve to meet diverse application demands.

9 Conclusion

This survey synthesizes key advancements and challenges in the multidisciplinary field of video generation, emphasizing the integration of video generation, human feedback, reinforcement learning, generative models, computer vision, and user interaction as crucial for advancing AI technologies. Controllable text-to-video generation methods, such as Control-A-Video, illustrate the transformative potential of merging diverse AI methodologies to enhance video realism and comprehension [71]. Frameworks like RLHF-Blender facilitate learning from varied human feedback, improving reward model training and enhancing AI adaptability [127]. Interactive frameworks, including Playable Environments, further advance video generation technologies, creating new opportunities for user engagement [128].

Significant advancements in generating human-object interaction videos, as evidenced by HOI-GAN, reveal the potential of relational adversaries in generative tasks, paving the way for broader applications [129]. The iVGAN model's superior performance in generating realistic videos underscores the need for ongoing research and innovation in this area [130]. In biomedical video generation, the survey highlights challenges and potential advancements through improved methodologies and evaluation techniques, recognizing machine learning's transformative impact on healthcare. The effectiveness of methods that produce realistic outputs aligned with user expectations, particularly in music video generation, reflects the broader applicability of AI-driven content creation [112].

The conclusion underscores DIGAN's state-of-the-art results in video generation and its potential for future research, particularly in enhancing the model's capacity for complex motion dynamics [131]. Key insights reveal significant performance gaps between closed-source and open-source models, emphasizing the necessity for continued research and innovation [123]. The integration of human feedback in generative tasks suggests avenues for future exploration of sophisticated feedback mechanisms [132]. Furthermore, incorporating RLHF significantly enhances image caption

AI-generated, for reference only.

quality, paving the way for improvements in generative AI models [125]. The advancements in high-resolution video synthesis and the effectiveness of proposed methods in generating realistic, diverse long-term human videos highlight the need for sustained research and innovation in this domain.

This survey highlights the importance of ongoing research and innovation in the multidisciplinary field of video generation, emphasizing the need to address existing challenges and fully harness AI technologies to create more realistic, interactive, and user-aligned video content. Advancements in character consistency, exemplified by StoryAgent, further illustrate the potential for significant developments in video generation technologies [101]. The introduction of OpenHumanVid serves as a valuable resource for enhancing human-centric video generation, demonstrating improvements in model performance and providing a benchmark for future research [38]. Additionally, the potential of GenDDS to enhance autonomous vehicle training by generating high-resolution, realistic driving videos under diverse conditions is noteworthy [4]. Experiments indicate that agents trained with the IBT reward model significantly outperform those trained solely with imitation learning, achieving high success rates in human interactions [133].

AI-generated, for reference only.

References

[1] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation, 2023.

[2] 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.

[3] Daoyuan Chen, Haibin Wang, Yilun Huang, Ce Ge, Yaliang Li, Bolin Ding, and Jingren Zhou. Data-juicer sandbox: A feedback-driven suite for multimodal data-model co-development, 2025.

[4] Yongjie Fu, Yunlong Li, and Xuan Di. Gendds: Generating diverse driving video scenarios with prompt-to-video generative model, 2024.

[5] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation, 2024.

[6] Ali Kashefi. A misleading gallery of fluid motion by generative artificial intelligence, 2024.

[7] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. Agent ai: Surveying the horizons of multimodal interaction, 2024.

[8] Jacob Sherson and Florent Vinchon. Facilitating human feedback for genai prompt optimization, 2024.

[9] Sandeep Manandhar and Auguste Genovesio. One style is all you need to generate a video, 2023.

[10] Naoya Fushishita, Antonio Tejero de Pablos, Yusuke Mukuta, and Tatsuya Harada. Long-term human video generation of multiple futures using poses, 2021.

[11] Chenxin Li, Hengyu Liu, Yifan Liu, Brandon Y. Feng, Wuyang Li, Xinyu Liu, Zhen Chen, Jing Shao, and Yixuan Yuan. Endora: Video generation models as endoscopy simulators, 2024.

[12] Fanda Fan, Chunjie Luo, Wanling Gao, and Jianfeng Zhan. Aigcbench: Comprehensive evaluation of image-to-video content generated by ai, 2024.

[13] Gaurav Shrivastava and Abhinav Shrivastava. Video prediction by modeling videos as continuous multi-dimensional processes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7236–7245, 2024.

[14] Corneliu Ilisescu, Halil Aytac Kanaci, Matteo Romagnoli, Neill DF Campbell, and Gabriel J Brostow. Responsive action-based video synthesis. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6569–6580, 2017.

[15] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025.

[16] Yannick Metz, David Lindner, Raphaël Baur, and Mennatallah El-Assady. Mapping out the space of human feedback for reinforcement learning: A conceptual framework, 2025.

[17] Cesar Roberto de Souza, Adrien Gaidon, Yohann Cabon, and Antonio Manuel Lopez. Procedural generation of videos to train deep action recognition networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4757–4767, 2017.

AI-generated, for reference only.

[18] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.

[19] Xiao Liu, Xinhao Xiang, Zizhong Li, Yongheng Wang, Zhuoheng Li, Zhuosheng Liu, Weidi Zhang, Weiqi Ye, and Jiawei Zhang. A survey of ai-generated video evaluation. *arXiv preprint arXiv:2410.19884*, 2024.

[20] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models, 2024.

[21] Mengyang Feng, Jinlin Liu, Kai Yu, Yuan Yao, Zheng Hui, Xiefan Guo, Xianhui Lin, Haolan Xue, Chen Shi, Xiaowen Li, Aojie Li, Xiaoyang Kang, Biwen Lei, Miaomiao Cui, Peiran Ren, and Xuansong Xie. Dreamoving: A human video generation framework based on diffusion models, 2023.

[22] Gabrielle Kaili-May Liu. Perspectives on the social impacts of reinforcement learning with human feedback, 2023.

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

[24] Pengyuan Zhou, Lin Wang, Zhi Liu, Yanbin Hao, Pan Hui, Sasu Tarkoma, and Jussi Kangasharju. A survey on generative ai and llm for video generation, understanding, and streaming, 2024.

[25] Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Golnari, David A. Clifton, Yuxiong He, Dacheng Tao, and Shuaiwen Leon Song. Renaissance: A survey into ai text-to-image generation in the era of large model, 2023.

[26] Abul Ehtesham, Saket Kumar, Aditi Singh, and Tala Talaei Khoei. Movie gen: Swot analysis of meta's generative ai foundation model for transforming media generation, advertising, and entertainment industries, 2024.

[27] Rui Sun, Yumin Zhang, Tejal Shah, Jiahao Sun, Shuoying Zhang, Wenqi Li, Haoran Duan, Bo Wei, and Rajiv Ranjan. From sora what we can see: A survey of text-to-video generation, 2024.

[28] Yitong Li, Martin Renqiang Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text, 2017.

[29] Nuha Aldausari, Arcot Sowmya, Nadine Marcus, and Gelareh Mohammadi. Video generative adversarial networks: a review. *ACM Computing Surveys (CSUR)*, 55(2):1–25, 2022.

[30] Amir Mazaheri and Mubarak Shah. Video generation from text employing latent path construction for temporal modeling, 2021.

[31] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Make-your-video: Customized video generation using textual and structural guidance, 2023.

[32] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation, 2024.

[33] Jiawei He, Andreas Lehrmann, Joseph Marino, Greg Mori, and Leonid Sigal. Probabilistic video generation using holistic attribute control, 2018.

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

[35] Andrew Marmon, Grant Schindler, José Lezama, Dan Kondratyuk, Bryan Seybold, and Irfan Essa. Camvig: Camera aware image-to-video generation with multimodal transformers, 2024.

[36] Hanxin Zhu, Tianyu He, Anni Tang, Junliang Guo, Zhibo Chen, and Jiang Bian. Compositional 3d-aware video generation with llm director, 2024.

[37] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[38] Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, and Siyu Zhu. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation, 2025.

[39] Remi Denton and Rob Fergus. Stochastic video generation with a learned prior, 2018.

[40] 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.

[41] Vladyslav Yushchenko, Nikita Araslanov, and Stefan Roth. Markov decision process for video generation, 2019.

[42] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation, 2024.

[43] Mingxiang Liao, Qixiang Ye, Wangmeng Zuo, Fang Wan, Tianyu Wang, Yuzhong Zhao, Jingdong Wang, Xinyu Zhang, et al. Evaluation of text-to-video generation models: A dynamics perspective. *Advances in Neural Information Processing Systems*, 37:109790–109816, 2024.

[44] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis, 2021.

[45] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International conference on machine learning*, pages 1174–1183. PMLR, 2018.

[46] Lijun Yu. Towards multi-task multi-modal models: A video generative perspective, 2024.

[47] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation, 2024.

[48] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities, 2024.

[49] 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.

[50] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, Hui Ding, Zhe Lin, and Hengshuang Zhao. Unireal: Universal image generation and editing via learning real-world dynamics, 2024.

[51] Shenghuan Sun, Gregory M. Goldgof, Atul Butte, and Ahmed M. Alaa. Aligning synthetic medical images with clinical knowledge using human feedback, 2023.

[52] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[53] Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. Demamba: Ai-generated video detection on million-scale genvideo benchmark, 2024.

[54] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Yuexian Zou, and Ying Shan. Image conductor: Precision control for interactive video synthesis. *arXiv preprint arXiv:2406.15339*, 2024.

[55] 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.

[56] Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, and Dahua Lin. Humanvid: Demystifying training data for camera-controllable human image animation, 2024.

[57] Yiyuan Zhang, Yuhao Kang, Zhixin Zhang, Xiaohan Ding, Sanyuan Zhao, and Xiangyu Yue. Interactivevideo: User-centric controllable video generation with synergistic multimodal instructions, 2024.

[58] Lijie Fan, Wenbing Huang, Chuang Gan, Junzhou Huang, and Boqing Gong. Controllable image-to-video translation: A case study on facial expression generation, 2018.

[59] Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation, 2024.

[60] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models, 2023.

[61] Isabela Albuquerque, João Monteiro, and Tiago H. Falk. Learning to navigate image manifolds induced by generative adversarial networks for unsupervised video generation, 2019.

[62] Shengqu Cai, Duygu Ceylan, Matheus Gadelha, Chun-Hao Paul Huang, Tuanfeng Yang Wang, and Gordon Wetzstein. Generative rendering: Controllable 4d-guided video generation with 2d diffusion models, 2023.

[63] Ximeng Sun, Huijuan Xu, and Kate Saenko. Twostreamvan: Improving motion modeling in video generation, 2020.

[64] Zhengqing Yuan, Yixin Liu, Yihan Cao, Weixiang Sun, Haolong Jia, Ruoxi Chen, Zhaoxu Li, Bin Lin, Li Yuan, Lifang He, Chi Wang, Yanfang Ye, and Lichao Sun. Mora: Enabling generalist video generation via a multi-agent framework, 2024.

[65] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion, 2024.

[66] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning, 2022.

[67] Aram Davtyan, Sepehr Sameni, Björn Ommer, and Paolo Favaro. Enabling visual composition and animation in unsupervised video generation, 2024.

[68] Wen Wang, Kecheng Zheng, Qiuyu Wang, Hao Chen, Zifan Shi, Ceyuan Yang, Yujun Shen, and Chunhua Shen. Gendef: Learning generative deformation field for video generation, 2023.

[69] Ming-Yu Liu, Xun Huang, Jiahui Yu, Ting-Chun Wang, and Arun Mallya. Generative adversarial networks for image and video synthesis: Algorithms and applications. *Proceedings of the IEEE*, 109(5):839–862, 2021.

[70] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

[71] 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.

[72] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models, 2024.

[73] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising, 2023.

[74] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022.

[75] Chengxuan Li, Di Huang, Zeyu Lu, Yang Xiao, Qingqi Pei, and Lei Bai. A survey on long video generation: Challenges, methods, and prospects, 2024.

[76] Hsin-Ping Huang, Yu-Chuan Su, and Ming-Hsuan Yang. Video generation beyond a single clip, 2023.

[77] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. *arXiv preprint arXiv:2104.15069*, 2021.

[78] Aram Davtyan and Paolo Favaro. Learn the force we can: Enabling sparse motion control in multi-object video generation, 2024.

[79] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: Controllable image-to-video generation with text descriptions, 2022.

[80] Dilip Arumugam, Jun Ki Lee, Sophie Saskin, and Michael L Littman. Deep reinforcement learning from policy-dependent human feedback. *arXiv preprint arXiv:1902.04257*, 2019.

[81] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19401–19411, 2024.

[82] Hiroki Furuta, Heiga Zen, Dale Schuurmans, Aleksandra Faust, Yutaka Matsuo, Percy Liang, and Sherry Yang. Improving dynamic object interactions in text-to-video generation with ai feedback, 2024.

[83] Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment, 2025.

[84] Wentao Lei, Jinting Wang, Fengji Ma, Guanjie Huang, and Li Liu. A comprehensive survey on human video generation: Challenges, methods, and insights, 2024.

[85] 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.

[86] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Generative recommendation: Towards next-generation recommender paradigm, 2024.

[87] Abhijay Ghildyal, Yuanhan Chen, Saman Zadtootaghaj, Nabajeet Barman, and Alan C. Bovik. Quality prediction of ai generated images and videos: Emerging trends and opportunities, 2024.

[88] Donald R. Honeycutt, Mahsan Nourani, and Eric D. Ragan. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy, 2020.

[89] Shachar Don-Yehiya, Ben Burtenshaw, Ramon Fernandez Astudillo, Cailean Osborne, Mimansa Jaiswal, Tzu-Sheng Kuo, Wenting Zhao, Idan Shenfeld, Andi Peng, Mikhail Yurochkin, Atoosa Kasirzadeh, Yangsibo Huang, Tatsunori Hashimoto, Yacine Jernite, Daniel Vila-Suero, Omri Abend, Jennifer Ding, Sara Hooker, Hannah Rose Kirk, and Leshem Choshen. The future of open human feedback, 2024.

[90] Katherine M. Collins, Najoung Kim, Yonatan Bitton, Verena Rieser, Shayegan Omidshafiei, Yushi Hu, Sherol Chen, Senjuti Dutta, Minsuk Chang, Kimin Lee, Youwei Liang, Georgina Evans, Sahil Singla, Gang Li, Adrian Weller, Junfeng He, Deepak Ramachandran, and Krishnamurthy Dj Dvijotham. Beyond thumbs up/down: Untangling challenges of fine-grained feedback for text-to-image generation, 2024.

[91] Amr Gomaa and Bilal Mahdy. Unveiling the role of expert guidance: A comparative analysis of user-centered imitation learning and traditional reinforcement learning, 2024.

[92] Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective, 2024.

[93] Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu. What comprises a good talking-head video generation?: A survey and benchmark, 2020.

[94] Yan Pang, Yang Zhang, and Tianhao Wang. Vgmshield: Mitigating misuse of video generative models, 2024.

[95] Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, László Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 4real: Towards photorealistic 4d scene generation via video diffusion models. *Advances in Neural Information Processing Systems*, 37:45256–45280, 2024.

[96] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5521–5531, 2022.

[97] Yuanhui Huang, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Di Zhang, Jie Zhou, and Jiwen Lu. Owl-1: Omni world model for consistent long video generation, 2024.

[98] 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.

[99] Taku Yamagata, Ryan McConville, and Raul Santos-Rodriguez. Reinforcement learning with feedback from multiple humans with diverse skills, 2021.

[100] Tsun-Hsuan Wang, Yen-Chi Cheng, Chieh Hubert Lin, Hwann-Tzong Chen, and Min Sun. Point-to-point video generation, 2019.

[101] Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. Storyagent: Customized storytelling video generation via multi-agent collaboration, 2024.

[102] Xue Song, Jingjing Chen, Bin Zhu, and Yu-Gang Jiang. Text-driven video prediction, 2022.

[103] Shohei Yamamoto, Antonio Tejero de Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Conditional video generation using action-appearance captions, 2018.

AI-generated, for reference only.

[104] Kunpeng Song, Tingbo Hou, Zecheng He, Haoyu Ma, Jialiang Wang, Animesh Sinha, Sam Tsai, Yaqiao Luo, Xiaoliang Dai, Li Chen, Xide Xia, Peizhao Zhang, Peter Vajda, Ahmed Elgammal, and Felix Juefei-Xu. Directorllm for human-centric video generation, 2024.

[105] César Roberto de Souza, Adrien Gaidon, Yohann Cabon, Naila Murray, and Antonio Manuel López. Generating human action videos by coupling 3d game engines and probabilistic graphical models, 2019.

[106] Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models, 2024.

[107] Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. Evaluation agent: Efficient and promptable evaluation framework for visual generative models, 2024.

[108] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models, 2024.

[109] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021.

[110] Yue Wu, Rongrong Gao, Jaesik Park, and Qifeng Chen. Future video synthesis with object motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5539–5548, 2020.

[111] Jingyi Deng, Chenhao Lin, Zhengyu Zhao, Shuai Liu, Qian Wang, and Chao Shen. A survey of defenses against ai-generated visual media: Detection, disruption, and authentication, 2024.

[112] Sarah Gross, Xingxing Wei, and Jun Zhu. Automatic realistic music video generation from segments of youtube videos, 2019.

[113] Chang Liu and Han Yu. Ai-empowered persuasive video generation: A survey, 2021.

[114] Dimitri von Rütte, Elisabetta Fedele, Jonathan Thomm, and Lukas Wolf. Fabric: Personalizing diffusion models with iterative feedback, 2023.

[115] 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, 2024.

[116] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation, 2024.

[117] Niv Haim, Ben Feinstein, Niv Granot, Assaf Shocher, Shai Bagon, Tali Dekel, and Michal Irani. Diverse generation from a single video made possible, 2021.

[118] Hongyuan Yu, Yan Huang, Lihong Pi, and Liang Wang. Recurrent deconvolutional generative adversarial networks with application to text guided video generation, 2020.

[119] Yan Pang, Aiping Xiong, Yang Zhang, and Tianhao Wang. Towards understanding unsafe video generation, 2024.

[120] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion, 2023.

[121] Hong Chen, Xin Wang, Yipeng Zhang, Yuwei Zhou, Zeyang Zhang, Siao Tang, and Wenwu Zhu. Disenstudio: Customized multi-subject text-to-video generation with disentangled spatial control, 2024.

[122] Kristian González Barman, Simon Lohse, and Henk de Regt. Reinforcement learning from human feedback: Whose culture, whose values, whose perspectives?, 2025.

[123] Ailing Zeng, Yuhang Yang, Weidong Chen, and Wei Liu. The dawn of video generation: Preliminary explorations with sora-like models, 2024.

[124] Zhao Wang, Aoxue Li, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. Customvideo: Customizing text-to-video generation with multiple subjects, 2024.

[125] Adarsh N L, Arun P V au2, and Aravindh N L. Enhancing image caption generation using reinforcement learning with human feedback, 2024.

[126] Ankan Dash, Junyi Ye, and Guiling Wang. A review of generative adversarial networks (gans) and its applications in a wide variety of disciplines – from medical to remote sensing, 2021.

[127] Yannick Metz, David Lindner, Raphaël Baur, Daniel Keim, and Mennatallah El-Assady. Rlhf-blender: A configurable interactive interface for learning from diverse human feedback, 2023.

[128] Willi Menapace, Stéphane Lathuilière, Aliaksandr Siarohin, Christian Theobalt, Sergey Tulyakov, Vladislav Golyanik, and Elisa Ricci. Playable environments: Video manipulation in space and time, 2022.

[129] Megha Nawhal, Mengyao Zhai, Andreas Lehrmann, Leonid Sigal, and Greg Mori. Generating videos of zero-shot compositions of actions and objects, 2020.

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

[131] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571*, 2022.

[132] Hyun-Cheol Park and Sung Ho Kang. Domain adaptation based on human feedback for enhancing generative model denoising abilities, 2023.

[133] Josh Abramson, Arun Ahuja, Federico Carnevale, Petko Georgiev, Alex Goldin, Alden Hung, Jessica Landon, Jirka Lhotka, Timothy Lillicrap, Alistair Muldal, George Powell, Adam Santoro, Guy Scully, Sanjana Srivastava, Tamara von Glehn, Greg Wayne, Nathaniel Wong, Chen Yan, and Rui Zhu. Improving multimodal interactive agents with reinforcement learning from human feedback, 2022.

AI-generated, for reference only.

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.

AI-generated, for reference only.