A Review on Background, Technology, Comparison, and Future Tendency of Video Generation

Zhiyu Han*

Department of software, Tianjin Normal University, Tianjin, China

*Corresponding author: hanzhiyu@stu.tjnu.edu.cn

Abstract. Video generation techniques incorporate recent advances in deep learning and generative modeling, and are widely used in film and television, education, advertising, virtual reality, and other fields. The background lies in the growing need to generate high-resolution, dynamically consistent, and semantically accurate videos to meet diverse scene requirements. Existing techniques, including Generative Adversarial Networks (GANs), Variational Auto-Encoders (VAEs), Transformer and Diffusion Models, have achieved significant improvements in video quality and generation efficiency. This paper systematically reviews the development history of video generation technology, from model principles, application scenarios to technical advantages and shortcomings, and analyzes the performance of the current mainstream models in detail. Combined with the experimental results, this paper summarizes the future trends of multimodal fusion, resolution improvement, generation efficiency optimization and 3D video generation. In the future, video generation technology will focus on deep alignment of multimodal fusion, real-time high-resolution generation, dynamic scene optimization and 3D modeling, which will promote its wide application in virtual reality, scientific research, and film and television production, and open up new paths for interactive content generation.

Keywords: Video Generation; Generative Adversarial Model; Variational Auto-Encoders; Transformer Model; Diffusion Model.

1. Introduction

The advent of generative large-scale models has catalyzed a surge in the demand for video generation technology. This nascent field, a beacon of the fusion of machine learning and deep learning's cutting-edge capabilities, promises transformative applications across a myriad of sectors, from cinema and entertainment to advertising, ephemeral content creation, animation, video prediction, education, healthcare, and scientific inquiry. Bolstered by the immense computational muscle of modern computers, video generation technology has evolved to produce video content of exceptional quality, meticulously tailored to specific criteria, vividly realistic in its physical depiction, and distinguished by its high resolution. It operates with agility across unimodal and multimodal data sources, including textual descriptions, static imagery, and pre-existing video material.

The state-of-the-art in video generation technology has ushered in a paradigm shift, with a marked improvement in both the quality and resolution of videos, an expanded array of content diversity and flexibility, and a significant enhancement in the continuity and logical coherence of narrative flow. The Sora model错误!未找到引用源。[1], released by OpenAI, exemplifies this progress by generating one-minute high-fidelity videos that are not only visually crisp but also meticulously aligned with their textual counterparts. The technology's versatility has reached a point where it can adeptly handle videos and images with a wide range of durations, aspect ratios, and resolutions, enabling the creation of everything from concise clips to extended high-definition productions based on textual prompts. This rich tapestry of diversity and adaptability has opened up a plethora of scenarios for the application of text-to-video technology. Moreover, the technology prioritizes the narrative's internal consistency and logical progression, ensuring that the resulting videos are not only visually fluid but also logically coherent with the textual script. This alignment of visual and logical elements contributes to a viewing experience that is both natural and authentic, greatly enhancing user satisfaction.

The era of video generation technology is currently experiencing a period of exponential growth, heralding the dawn of a new epoch focused on this innovative field. A cadre of researchers is fervently engaged in pushing the boundaries of video generation techniques through innovative research and optimization. As of now, prevalent methodologies encompass Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Transformer, and Diffusion Models, among others[2-5]. Tulyakov proposed a video generation method based on generative adversarial networks, called MoCoGAN [6], which achieves high efficiency and diversity in video generation by representing motion and content separately in the latent space. It is shown that MoCoGAN is able to capture more realistic dynamic changes and object motion in the generated video, and especially performs well in the generation of video scenes with complex motions. Saharia proposed a diffusion model-based video generation method called Imagen Video [7]. This method generates a video by utilizing a diffusion model under the condition of high-quality text that is able to capture the exact alignment of the text description with the video content. It is shown that Imagen Video can achieve very high spatial quality and temporal consistency in the generated videos, especially in the generation and detail rendering of dynamic scenes. Compared to other traditional methods, Imagen Video is an important breakthrough in the field of text-to-video generation, as it improves the speed of generation and is able to handle complex spatio-temporal dependencies.

This paper endeavors to provide a comprehensive synthesis of the advancements in video generation technology from its inception to the present. It meticulously reviews and analyzes the pertinent research literature, elucidating the fundamental principles and model applications of this technology. Furthermore, it juxtaposes experimental outcomes from recent scholarly works, delineating the strengths and weaknesses of various techniques. Ultimately, the paper addresses the challenges and hurdles faced in the deployment of video generation technology, concurrently exploring its future trajectory to offer insightful guidance for its continued advancement and proliferation.

2. Primary technology

2.1. Generative Adversarial Networks (GANs)

Generative Adversarial Network (GAN), as a pioneering technology in the field of video generation, has greatly contributed to the rapid development of the field with its unique generative and adversarial mechanisms, from the earliest simple frame modeling to the realization of high-quality, spatiotemporally consistent video generation in recent years. In the early stage, GAN mainly focuses on the generation of single-frame images, and gradually extends to video sequence modeling by introducing the time dimension, which makes the generation of dynamic scenes and object motion possible. In recent years, with the improvement of the model architecture and the significant improvement of training stability, GANs have become more and more realistic in the field of video generation, not only generating short videos with complex dynamic features, but also making breakthroughs in the modeling of long time sequences, inter-frame continuity, and so on. The following are the more popular models of GANs in the field of video generation.

Motion and Content GAN (MoCoGAN) is a pioneering work in video generation task. A separation of motion and content representations is proposed, where the generator generates a sequence of video frames by independently sampling the content representation and combining it with temporal noise, while evaluating the frame-level and video-level realism using a discriminator. This structured design significantly improves the temporal consistency and diversity of video generation. The ability to generate short videos of high quality with low computational resources is an important beginning for the task of video generation. This work lays the foundation for GAN-based video generation methods. However, it is limited by the GAN framework, and the resolution and duration of the generated videos are restricted and the training is unstable.

StyleGAN for Video Generation (StyleGAN-V) focuses on video generation, inheriting the ability of StyleGAN series to generate high-quality images [8], and introduces an explicit temporal modeling

module, focusing on the generation of video sequences. StyleGAN-V generates videos with significantly higher resolution and semantic consistency, and is capable of handling generation tasks up to 4K. The model adopts a dynamic generation mechanism for the temporal dimension, allowing smoother transitions between frames, and introduces a hierarchical style control mechanism for fine-grained generation of video content. The emergence of StyleGAN-V marks a breakthrough in the quality bottleneck of GAN in the field of video generation, and it becomes the benchmark model for current video generation tasks.

2.2. Variational Autoencoders (VAEs)

Specifically, the application of VAEs in video generation exploits their potential representation learning capabilities on high-dimensional data. By introducing variational inference methods, VAEs map the original video frames to a continuous low-dimensional latent space at the encoding stage and ensure that the distribution of the latent space is close to the standard prior distribution through regularization. In the decoding stage, VAEs can sample from the latent space and then generate continuous, time-dependent, high-quality video frames through the decoder. The following are the more popular models of VAEs in the field of video generation.

VideoGPT combines the discrete modeling capability of GPT with the VAE framework to enhance video generation and encoding by discretizing the video data representation and modeling it with GPT [9]. It proposes discretized coding strategies for video frames, enabling the generation task to efficiently capture inter-frame dependencies, and incorporates the capabilities of autoregressive modeling, providing a novel approach to the field of video coding and compression. The model demonstrates the powerful combination of sequence generation techniques and video generation, opening the path to high-quality video generation based on sequence modeling.

2.3. Transformer

Originally designed for natural language processing, the Transformer model efficiently captures global dependencies in sequential data through a multi-head self-attention mechanism. This feature is extended and applied in video generation tasks. Video generation involves both time series modeling and spatial feature processing, and Transformer provides a unified way to model the temporal dependencies between video frames as well as the spatial structure within frames. The following are some of Transformer's more popular models in the field of video generation.

CogVideo [10], a cross-modal generation model that transforms input text descriptions into high-quality video sequences, demonstrates the potential of Transformer for cross-modal generation. The model incorporates a cross-modal attention mechanism, realizes efficient alignment of text and video content, and supports the generation of complex dynamic videos, which are widely used in film and television production, advertisement generation, etc. CogVideo is an important milestone in cross-modal video generation, and its unique generative mechanism makes it a representative model for the task of generating video from text.

Latent Video Transformer (LVT) is efficiently optimized for Vision Transformer (ViT) to reduce computational complexity, and the computational overhead is reduced through structural optimization [11,12], enabling the Transformer architecture to efficiently generate videos in resource-constrained environments. The model is suitable for real-time video generation or mobile application scenarios, and is an important optimization for Transformer in the generation domain.

2.4. Diffusion Models

Diffusion Models have shown significant advantages in the field of video generation in recent years, becoming one of the important directions in generative modeling. The core mechanism is to gradually reduce high-quality samples from random noise through gradual denoising. This generative process is not only highly suitable for the demand of detail accuracy and dynamic continuity in video generation, but also well adapted to complex time series and multimodal input characteristics. The following are some of the more popular models of Diffusion Models in the field of video generation.

Video Diffusion Models (VDMs) apply diffusion models for the first time to the task of video generation to generate high-quality video sequences through a stepwise denoising strategy [13]. The model introduces a temporal feature encoder in the time dimension to ensure that the model can effectively model the dependencies between video frames, extends the traditional diffusion model from processing still images to sequence data, and combines convolutional operations to achieve the extraction of temporal and spatial joint features. The model maintains extremely high visual quality and temporal coherence when generating complex, high-dynamic-range video content, provides a powerful generative framework for high-resolution video generation, and lays the foundation for diffusion modeling in video generation.

Imagen Video is a high-quality text-to-video generation model proposed by Google, based on a multi-stage diffusion strategy to generate visually appealing video content. The model's multi-stage generation strategy optimizes the temporal coherence and spatial resolution of the video. It generates high-quality video clips through an efficient diffusion mechanism and demonstrates excellent text-to-video content alignment.

3. Technology comparison

3.1. Dataset

3.1.1 Kinetics

The Kinetics dataset is a large-scale benchmark widely used for video generation and action recognition tasks [14]. It contains hundreds of thousands of 10-second video clips categorized into 400, 600, or 700 action classes, ranging from everyday activities to sports actions. The dataset features high-resolution and high-frame-rate videos, capturing dynamic temporal and spatial characteristics essential for tasks like spatiotemporal modeling, frame-to-frame continuity, and action generation. Its diversity and clarity make it ideal for unconditional video generation using GANs, VAEs, and diffusion models, while the well-defined class labels facilitate supervised learning and benchmarking for action-related tasks.

3.1.2 MSR-VTT

The MSR-VTT dataset is a multi-modal video dataset comprising 10,000 video clips [15], each ranging from 10 to 30 seconds in length. Every video is paired with 20 detailed natural language descriptions, offering rich semantic annotations and diverse scene content. The dataset includes a wide range of video sources, such as real-world scenes, animations, and computer-generated content, making it well-suited for tasks like text-to-video generation (T2V), multi-modal video retrieval, and video understanding. As a foundational dataset for multi-modal research, MSR-VTT provides strong alignment between video and textual modalities, making it a critical resource for evaluating Transformer-based and diffusion-based models in conditional video generation and cross-modal learning.

3.2. Evaluation criteria

3.2.1 Fréchet Inception Distance (FID)

The Fréchet Inception Distance (FID) is a widely used metric to evaluate the quality of generated video frames by measuring the distributional similarity between the generated frames and real frames [16]. It utilizes a pre-trained Inception network to extract high-level features from both real and generated frames and then computes the Fréchet distance between their feature distributions. A lower FID score indicates that the generated frames are closer in quality and diversity to the real frames. FID is particularly useful for assessing the overall visual fidelity and diversity of generated videos in both unconditional and conditional video generation tasks.

3.2.2 Temporal Consistency

Temporal Consistency evaluates the coherence and smoothness of motion between consecutive frames in a generated video [17]. This metric measures the alignment of temporal dynamics, often by calculating differences in optical flow or motion fields between adjacent frames. A lower temporal inconsistency score indicates that the motion transitions in the video are natural and seamless. Temporal consistency is critical for video generation tasks, as even high-quality individual frames may fail to create a coherent viewing experience without smooth temporal transitions.

3.2.3 CLIP Score

The CLIP Score assesses the semantic consistency between the generated video and a given condition [18], such as a text prompt or an image. It leverages the CLIP model, which maps text and visual inputs into a shared embedding space, to calculate the cosine similarity between the generated video frames and the input condition. A higher CLIP Score signifies stronger alignment between the generated video content and the intended semantics of the input. This metric is particularly valuable for conditional video generation tasks, such as text-to-video or image-to-video generation, where ensuring semantic accuracy is essential.

3.3. Model Performance Comparison

A comprehensive comparison of four representative models from different video generation frameworks is performed using three key evaluation metrics: the Fréchet Inception Distance (FID), Temporal Consistency and CLIP Score (Multimodal Consistency). The models analyzed include StyleGAN-V, VideoGPT, CogVideo, and Imagen Video, representing GANs, VAEs, Transformers, and Diffusion Models, respectively. The evaluation is based on two benchmark datasets: Kinetics (for unconditional video generation) and MSR-VTT (for conditional video generation).

Imagen Video is the top performer in the comparison, significantly outperforming the other models in terms of generation quality, temporal consistency, and semantic consistency. Imagen Video demonstrates excellent performance on both datasets, proving its leadership in the field of video generation. Especially in the task of generating high-resolution and complex dynamic scenes, Imagen Video demonstrates its ability to generate videos that are close to real videos in terms of visual quality and dynamic smoothness, showing high detail reproduction and realism.

CogVideo, as a representative of the Transformer architecture, performs well in multimodal generation tasks, especially in text-to-video generation, where CogVideo's semantic consistency and temporal consistency are at a high level. CogVideo is ideal for multimodal conditional generation because it can generate semantically consistent videos based on input text descriptions, and CogVideo makes full use of Transformer's powerful sequence modeling capabilities to generate videos with high visual fidelity and excellent performance in complex semantic understanding.

StyleGAN-V performs better in short-sequence video generation, and the generated video frames are of high quality and have more detailed visual effects. However, there are some limitations in terms of temporal consistency and multimodal consistency. Although StyleGAN-V is capable of generating short dynamic sequences with high visual quality, the transition between frames is not as natural and smooth as other models when dealing with long sequences and complex scenes. Therefore, StyleGAN-V is more suitable for generating short video clips with high visual effect requirements, but not for generating long videos with complex scenes.

VideoGPT, as a representative of VAEs, improves the inter-frame coherence to a certain extent by its autoregressive modeling approach, although there is a certain gap in generation quality and temporal coherence compared with other models. Especially in the conditional generation task, VideoGPT is able to generate videos that match the input conditions with a certain degree of flexibility. Despite its relatively weak overall performance, due to its relatively lightweight architecture, VideoGPT is suitable for application scenarios that require low computational resources, especially for scenarios that require rapid prototyping or running with limited computational power.

In summary, Imagen Video is the best choice for unconditional and conditional video generation, especially in high-resolution and complex dynamic generation tasks, and generates videos with very high visual quality and detail reproduction. CogVideo is also strong in multimodal tasks, especially for text-to-video generation needs, and is capable of generating high-quality videos that conform to complex semantics. CogVideo is also a strong performer in multimodal tasks, especially for text-to-video generation, producing high-quality videos with complex semantics. For high-quality generation of short dynamic sequences, StyleGAN-V is a suitable choice for scenarios that require the generation of short videos with high fidelity. VideoGPT, on the other hand, is suitable for lightweight conditional generation tasks, and performs well especially in resource-constrained scenarios that require flexibility and computational efficiency. Diffusion models (especially Imagen Video) and Transformer models (CogVideo) are the most comprehensive in the current video generation field, and can satisfy the diversified needs of different application scenarios. From high-quality unconditional video generation to multimodal conditional generation, these models represent the latest advances in the field of video generation and provide powerful tools and foundations for future research and applications.

4. Future tendency

With the wide application of generative modeling in the field of video generation, technological advances have gradually met the users' needs for high quality, dynamic coherence and multimodal adaptability. Video generation technology has huge room for expansion in terms of quality, efficiency, flexibility and application areas. In the future, the development of video generation technology will advance along the following major trends.

4.1. Multimodal fusion

Multimodal fusion video generation has been initially realized, such as text-to-video and image-to-video tasks, relying on cross-modal attention mechanisms and conditional generation models, but there is still room for improvement in terms of semantic consistency and dynamic fluency. Multimodal generation will achieve higher semantic accuracy and dynamic coherence by optimizing cross-modal alignment, self-supervised learning and multi-stage generation strategies. The model will support dynamic combinations of multimodal inputs, and be applied to advertising, education, film and television to provide a more natural and intelligent interactive video generation experience.

4.2. Resolution Enhancement

Current video generation technology has supported high-resolution generation, mainly relying on GAN, diffusion models, etc. to optimize resolution and details through multi-stage generation, but there are still efficiency and consistency problems in long sequences and complex scenes. Future technologies will focus on real-time high-resolution generation, using latent space modeling and distributed training to improve efficiency; combined with dynamic modeling to optimize long time sequence generation. More applications include 8K video generation, virtual reality and scientific simulation scenarios, promoting high-quality generation in more fields.

4.3. Generation efficiency optimization

The current video generation efficiency optimization focuses on lightweight architectures (e.g., sparse attention, latent space modeling) and distributed computing frameworks, which have supported some of the real-time generation needs, but high-resolution and complex scene generation is still facing a resource bottleneck. Generation efficiency optimization will achieve real-time generation through more efficient lightweight design, potential space diffusion and parallel reasoning. Combined with large-scale pre-training and migration learning, future models will be more adaptable to resource-constrained scenarios, such as mobile and embedded devices.

4.4. 3D Video Generation

3D video generation is an important branch of video generation technology, aiming at generating dynamic content with depth information and spatial sense. With the continuous innovation of new technologies such as diffusion modeling and Transformer, 3D video generation will meet the needs of the fields of Virtual Reality (VR), Augmented Reality (AR), meta-universe, gaming, and film and television production. Compared with traditional video generation, 3D video generation focuses more on modeling spatial geometry, dynamic scenes, and perspective consistency, and faces higher technical challenges.

5. Conclusion

This paper systematically presents the technological history of the video generation field, detailing the methodologies of various techniques along with their respective advantages and disadvantages. It also offers insights into the future development trends of video generation technologies. Video generation technology is in a stage of rapid development, from GANs, VAEs, Transformer and Diffusion Models, all kinds of technologies continue to promote the improvement of generation quality, dynamic consistency and multimodal adaptability. Current technologies have been able to efficiently generate high-resolution, diverse, and semantically and logically coherent videos, and have initially realized the deep integration of multimodal inputs and generation.

Despite the significant progress, video generation techniques still face challenges such as long time sequence modeling, multimodal conditional alignment, real-time generation capability and high resolution generation efficiency. The future development of the technology will focus on the following aspects: optimizing the semantic consistency and dynamic smoothness of multimodal fusion; developing real-time high-resolution generation methods; improving efficiency through lightweight architecture and potential space generation; and promoting the application of 3D video generation in virtual reality, meta-universe, and scientific simulation. With technological breakthroughs and the expansion of application scenarios, video generation will become an important tool in the fields of film and television production, education and training, virtual reality and scientific exploration, bringing far-reaching impacts to the society and the industry. The future of this field is full of potential and deserves continuous attention and exploration.

References

[1] Liu Y, Zhang K, Li Y, Yan Z, Gao C, Chen R, Yuan Z, Huang Y, Sun H, Gao J, He L, & Sun L. Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv:2402.17177, 2024.

[2] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, & Bengio Y. Generative adversarial nets. Advances in Neural Information Processing Systems, 2014, 27, 2672-2680.

[3] Kingma D P, & Welling M. Auto-Encoding Variational Bayes. arXiv:1312.6114, 2013.

[4] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, & Polosukhin I. Attention is all you need. Advances in Neural Information Processing Systems, 2017, 30, 5998-6008.

[5] Yang L, Zhang Z, Song Y, Hong S, Xu R, Zhao Y, Zhang W, Cui B, & Yang M-H. Diffusion Models: A Comprehensive Survey of Methods and Applications. arXiv:2209.00796, 2022.

[6] Tulyakov S, Liu M-Y, Yang X, & Kautz J. MoCoGAN: Decomposing Motion and Content for Video Generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018. pp. 1526-1535.

[7] Ho J, Chan W, Saharia C, Whang J, Gao R, Gritsenko A, Kingma D P, Poole B, Norouzi M, Fleet D J, & Salimans T. Imagen Video: High Definition Video Generation with Diffusion Models. arXiv:2210.02303, 2022.

[8] Skorokhodov I, Tulyakov S, & Elhoseiny M. StyleGAN-V: A Continuous Video Generator with the Price, Image Quality and Perks of StyleGAN2. arXiv:2112.14683, 2021.

[9] Yan W, Zhang Y, Abbeel P, & Srinivas A. VideoGPT: Video Generation using VQ-VAE and Transformers. arXiv:2104.10157, 2021.

[10] Hong W, Ding M, Zheng W, Liu X, & Tang J. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. arXiv:2205.15868, 2022.

[11] Rakhimov R, Volkhonskiy D, Artemov A, Zorin D, & Burnaev E. Latent Video Transformer. arXiv:2006.10704, 2020.

[12] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, & Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929, 2020.

[13] Ho J, Salimans T, Gritsenko A, Chan W, Norouzi M, & Fleet D J. Video Diffusion Models. arXiv:2204.03458, 2022.

[14] Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleyman M, & Zisserman A. The Kinetics Human Action Video Dataset. arXiv:1705.06950, 2017.

[15] Xu J, Mei T, Yao T, & Rui Y. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas, NV, USA. IEEE.

[16] Heusel M, Ramsauer H, Unterthiner T, Nessler B, & Hochreiter S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. arXiv preprint arXiv:1706.08500, 2017.

[17] Lai W-S, Huang J-B, Wang O, Shechtman E, Yumer E, & Yang M-H. Learning Blind Video Temporal Consistency. arXiv:1808.00449, 2018.

[18] Hessel J, Holtzman A, Forbes M, Bras R L, & Choi Y. CLIPScore: A Reference-free Evaluation Metric for Image Captioning.arXiv:2104.08718, 2021.