A Detailed Review of Video Generation Using Textual Descriptions and Structural Guidance

Dr. Rachana P
Faculty, Department of Information Science and Engineering
Alva's Institute of Engineering and Technology
Mangalore, India
rachana@aiet.org.in

Disha
Student, Department of Information Science and Engineering
Alva's Institute of Engineering and Technology
Mangalore, India
dishagshetty6@gmail.com

Sujal Bandekar
Student, Department of Information Science and Engineering
Alva's Institute of Engineering and Technology
Mangalore, India
sujalbandekar786@gmail.com

Sanket
Student, Department of Information Science and Engineering
Alva's Institute of Engineering and Technology
Mangalore, India
biradarsanket380@gmail.com

Shankar C Akashkore
Student, Department of Information Science and Engineering Alva's Institute of Engineering and Technology
Mangalore, India shankarakashkore32@gmail.com

Abstract—Artificial intelligence based multimedia generation has experienced rapid growth in recent years. Among different multimedia applications, text guided video generation has become an important research area because of its capability to automatically create videos from textual descriptions.

Existing video generation systems often face challenges such as temporal inconsistency, motion instability, flickering artifacts, and poor structural preservation. Recent advancements in diffusion models have significantly improved video synthesis quality by introducing stable and realistic generation techniques.

The framework discussed in this paper combines textual descriptions with structural guidance to improve temporal coherence and controllability during video generation. Structural guidance mechanisms such as depth maps help preserve scene geometry and maintain object consistency across consecutive frames.

This review paper presents a comprehensive study of video generation using textual and structural guidance. The paper discusses diffusion models, latent diffusion architectures, temporal transformers, structural guidance mechanisms, methodology, applications, advantages, limitations, and future research directions.

The study also compares several existing frameworks including Make-A-Video, VideoFusion, Text-to-Video Zero, Imagen Video, and ControlVideo. The comparison highlights the importance of temporal learning and structural conditioning in improving realism, motion continuity, and controllability in modern video generation systems.

Index Terms—Video generation, diffusion models, temporal consistency, structural guidance, multimedia synthesis, deep learning

I. INTRODUCTION

Artificial intelligence has transformed the field of multimedia generation by enabling systems to automatically create images, audio, and videos with minimal human effort.

Video generation has become one of the most rapidly growing research areas because videos are widely used in entertainment, education, healthcare, gaming, advertising, and virtual reality systems.

Traditional video production requires professional editing skills, expensive software tools, and significant production time. Recent advancements in deep learning and generative models have simplified this process by introducing automatic video synthesis systems.

Text guided video generation focuses on converting textual descriptions into realistic video sequences. The generated videos must maintain smooth motion, temporal continuity, and structural consistency across consecutive frames.

Although several modern approaches have demonstrated promising results, many existing systems still suffer from flickering artifacts, unstable object movement, and weak motion control.

Diffusion models have recently emerged as powerful alternatives to traditional generative adversarial networks because of their stable training process and high quality generation capability [5], [6].

Modern diffusion based systems gradually remove noise from latent representations to generate realistic visual outputs. The integration of structural guidance such as depth maps further improves motion consistency and scene stability.

This review paper provides a detailed study of video generation using textual descriptions and structural guidance. The paper discusses the architecture, methodology, applications, advantages, limitations, and future developments of modern video generation systems.

II. BACKGROUND STUDY

Video generation technology has evolved significantly during the past decade due to rapid advancements in artificial intelligence and deep learning.

Early video synthesis methods mainly relied on generative adversarial networks. These models generated visually realistic frames but often suffered from unstable training and inconsistent temporal behavior.

Researchers later introduced transformer based architectures to improve sequence learning and motion understanding. Although transformers improved temporal modeling, these systems required very large computational resources and training datasets.

Diffusion models later became highly successful because they generate outputs through gradual denoising operations. These models progressively transform noisy latent representations into realistic images and videos.

Several modern text guided video generation systems such as Make A Video [2], VideoFusion [3], Text to Video Zero [4], and ControlVideo [8] demonstrated impressive video synthesis capability.

However, many of these systems still face limitations including weak temporal consistency, motion instability, poor object preservation, and difficulty in generating long video sequences.

Researchers introduced structural guidance methods such as depth maps, pose estimation, and optical flow to improve scene consistency and motion controllability during video synthesis.

III. NEED FOR STRUCTURAL GUIDANCE

Textual prompts mainly describe object appearance, environmental conditions, and scene context. However, text descriptions alone cannot accurately define object movement and spatial structure throughout a video sequence.

Structural guidance provides additional geometric information that helps maintain scene stability and realistic motion generation.

Depth maps are particularly useful because they represent object distance and scene geometry effectively. Structural guidance helps preserve object positions, movement trajectories, and spatial relationships between consecutive frames.

This significantly improves temporal consistency and reduces flickering artifacts.

The integration of depth based structural information into diffusion based video generation frameworks improves controllability and visual realism.

Structural guidance allows users to customize object movement and scene arrangement more effectively compared to systems that rely only on textual descriptions.

IV. OVERVIEW OF THE PROPOSED FRAMEWORK

The proposed framework combines textual descriptions with structural guidance to generate controllable and temporally coherent videos.

The framework utilizes latent diffusion models [5], temporal transformers [7], structural encoders [9], and attention masking mechanisms to improve video quality and motion stability.

The primary objective of the framework is to maintain smooth motion continuity while preserving scene structure and textual alignment.

The framework extends pre trained image diffusion models into video generation systems by introducing temporal modules and structural guidance techniques.

The major contributions of the framework include improved temporal consistency, better structural controllability, long video generation capability, and enhanced prompt alignment.

The integration of depth maps significantly improves scene realism and motion coherence.

V. ARCHITECTURE OF THE FRAMEWORK

The framework consists of several important modules including latent diffusion models, structural guidance encoders, temporal transformer modules, causal attention masking mechanisms, and decoders.

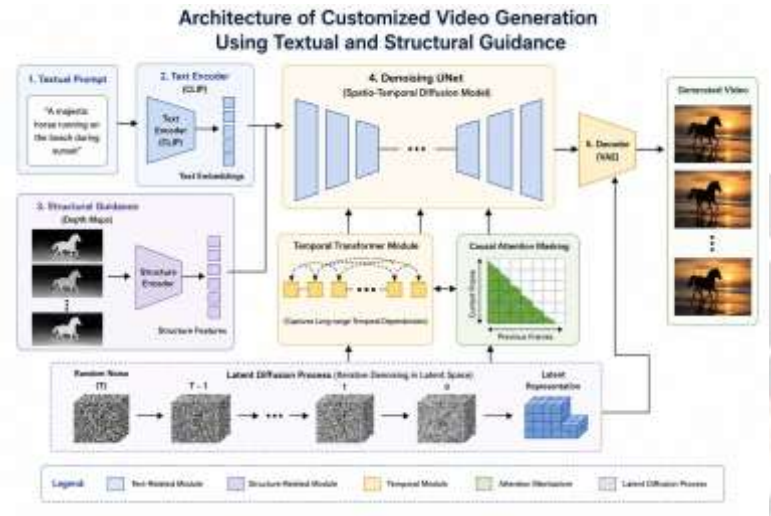
The latent diffusion model acts as the backbone architecture for video generation [5], [10].

Instead of operating directly in pixel space, the model performs computations in latent space, significantly reducing computational complexity while maintaining high visual quality.

Structural guidance encoders process frame wise depth maps and extract geometric features related to scene layout, object positions, and movement structure [1], [9].

These structural features guide the generation process and improve temporal stability.
Temporal transformer modules are introduced to model relationships between consecutive frames [7].
These modules apply temporal self attention mechanisms to maintain smooth motion continuity and reduce frame inconsistency.
The decoder reconstructs final video frames from denoised latent representations and generates high quality visual outputs.

Fig. 1. Architecture of Video Generation Framework



VI. METHODOLOGY

The proposed framework follows a systematic workflow for generating realistic and temporally consistent videos using textual descriptions and structural guidance. The methodology combines latent diffusion models, temporal transformers, and structural conditioning mechanisms to improve video quality and controllability.

A. Step 1: Text Prompt Encoding

The video generation process begins with textual input provided by the user. The textual description is processed using a text encoder to extract semantic feature representations [2], [7]. These semantic embeddings capture important contextual information such as object appearance, environmental conditions, scene composition, and motion description.

The encoded textual features guide the diffusion model during video synthesis and ensure alignment between generated frames and user prompts.

B. Step 2: Structural Guidance Extraction

Structural guidance information such as depth maps is extracted to preserve scene geometry and object consistency [1], [9]. Depth maps provide spatial information related to object distance, movement direction, and scene structure.

The structural encoder converts depth information into latent structural representations that guide the generation process and improve temporal stability across consecutive frames.

C. Step 3: Latent Space Initialization

Instead of directly operating in high resolution pixel space, the framework initializes noisy latent representations in compressed latent space [5], [10]. Latent diffusion significantly reduces computational complexity and memory requirements while preserving visual quality.

Random noise is progressively refined through iterative denoising operations during the generation process.

D. Step 4: Diffusion Based Video Generation

The latent diffusion model gradually removes noise from latent representations through multiple denoising stages [5], [6]. At each diffusion step, textual embeddings and structural guidance information influence the reconstruction process.

The diffusion mechanism generates realistic visual details, improves texture quality, and maintains semantic consistency with the textual prompt.

E. Step 5: Temporal Consistency Learning

Temporal transformer modules are introduced to model relationships between consecutive frames [1], [7]. Temporal self attention mechanisms capture motion continuity and maintain smooth object movement throughout the video sequence.

These modules reduce flickering artifacts and preserve temporal coherence during dynamic scene generation.

F. Step 6: Causal Attention Masking

Long video generation may become unstable when all frames interact simultaneously during attention computation. To solve this issue, the framework introduces causal attention masking mechanisms [1], [10].

Causal masking restricts future frame information from influencing previous frames and preserves sequential temporal order during generation.

G. Step 7: Video Frame Reconstruction

After iterative denoising and temporal refinement, the decoder reconstructs high quality video frames from latent representations. The generated frames are combined sequentially to produce the final video output.

The final generated video demonstrates improved realism, smooth motion continuity, structural consistency, and strong alignment with textual descriptions.

VII. DIFFUSION MODELS

Diffusion models have become one of the most important generative techniques in modern artificial intelligence based multimedia synthesis systems because of their stable training process and high quality generation capability [5], [6]. These models generate realistic images and videos by gradually transforming random noise into meaningful visual representations through iterative denoising operations.

The fundamental principle of diffusion models consists of two major processes known as the forward diffusion process and the reverse diffusion process.

A. Forward Diffusion Process

In the forward diffusion stage, controlled Gaussian noise is gradually added to training data over multiple time steps [5]. During this process, the original image or video frame progressively loses its visual information until it becomes nearly pure random noise.

Mathematically, the forward diffusion process can be represented as:

$$q(x_t|x_{t-1}) = \mathrm{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$ where:

- x_t represents the noisy sample at time step t
- β_t represents the variance schedule controlling noise addition
- N denotes Gaussian distribution
- I represents the identity matrix

The gradual addition of noise allows the model to learn robust feature distributions and improves generation stability.

B. Reverse Diffusion Process

The reverse diffusion process is responsible for generating realistic outputs from noisy latent representations [5], [6]. During this stage, the neural network learns to iteratively remove noise step by step until meaningful visual content is reconstructed.

The reverse denoising operation can be represented as: $p\theta(xt-1|xt)$

where:

- p_θ represents the learned denoising model
- x_t is the noisy latent representation
- x_{t-1} is the denoised output from the previous step

The model predicts the noise component present in the input sample and subtracts it progressively to reconstruct realistic images or video frames.

C. Latent Diffusion Models

Traditional diffusion models directly operate on high resolution pixel space, which requires very high computational resources. Latent Diffusion Models introduced by Rombach et al. [5] solve this problem by performing diffusion operations within compressed latent feature space instead of raw pixel space.

An encoder first compresses input images into lower dimensional latent representations. Diffusion operations are then applied within this latent space, significantly reducing computational complexity and memory consumption.

The decoder reconstructs high resolution outputs from the denoised latent representations after the diffusion process is completed.

D. Diffusion Models for Video Generation

In video synthesis systems, diffusion models are extended from image generation to temporal sequence generation [1], [6]. Instead of generating independent images, the model must maintain temporal consistency across consecutive frames.

Temporal transformer modules and spatio temporal attention mechanisms are integrated into diffusion architectures to preserve motion continuity and scene stability. Structural guidance information such as depth maps further improves object consistency and motion realism.

Modern video diffusion frameworks such as Make-A-Video, Imagen Video, and VideoFusion demonstrated that diffusion based architectures significantly improve video quality, motion coherence, and semantic alignment compared to earlier generative adversarial network based systems [2], [3], [6].

E. Advantages of Diffusion Models

Diffusion models provide several important advantages over traditional generative models:

- Stable training process
- High quality image and video synthesis
- Improved diversity of generated outputs
- Better semantic controllability
- Reduced mode collapse problems
- Improved temporal consistency in video generation

Because of these advantages, diffusion models have become the foundation of most modern text guided video generation systems.

TABLE I
COMPARISON OF EXISTING VIDEO GENERATION METHODS

Method	Temporal Stability	Structural Control	Long Video Support
Make-A-Video [2]	Medium	Low	No
VideoFusion [3]	Medium	Medium	Limited
Text-to-Video Zero [4]	Low	Low	No
ControlVideo [8]	Medium	Medium	Limited
Proposed Framework [1]	High	High	Yes

VIII. COMPARISON OF EXISTING METHODS

The comparison clearly shows that structural guidance and temporal transformers significantly improve motion continuity and video realism.

IX. APPLICATIONS

The proposed framework has several important applications in multimedia generation, entertainment, education, healthcare, gaming, and virtual reality industries. The ability to generate realistic videos from textual descriptions and structural guidance provides significant flexibility for automatic multimedia creation.

A. Entertainment and Film Industry

Artificial intelligence based video generation systems can simplify movie production, animation, and cinematic scene generation. Directors and content creators can automatically generate short video clips, storyboards, and visual effects using

simple textual prompts. Structural guidance mechanisms help maintain smooth object motion and scene consistency, improving cinematic realism.

B. Gaming and Virtual Reality

Video generation frameworks can improve gaming experiences through automatic generation of dynamic environments, animated characters, and realistic background scenes. In virtual reality systems, artificial intelligence based video synthesis can create immersive environments and interactive visual experiences with reduced development effort.

C. Education and E-Learning

Educational institutions can use video generation systems to create animated tutorials, instructional demonstrations, and interactive learning materials automatically from textual descriptions. Complex scientific concepts and technical procedures can be visualized more effectively using generated educational videos.

D. Healthcare and Medical Visualization

Artificial intelligence based video generation can support healthcare training and medical education by generating medical simulations, anatomy visualizations, and surgical procedure demonstrations. These systems can help medical students and healthcare professionals understand complex medical procedures more effectively.

E. Advertising and Digital Marketing

Customized video generation systems can automatically create promotional videos, product advertisements, and personalized marketing content. Businesses can generate multiple advertising videos with different themes and visual styles using textual descriptions and structural controls.

F. Social Media Content Creation

Social media influencers and digital content creators can use text guided video generation systems to quickly produce short videos, animated clips, and visual storytelling content for online platforms. This reduces production cost and minimizes manual editing effort.

G. Video Editing and Re-Rendering

The framework can also be utilized for video editing and re-rendering applications. Existing videos can be modified while preserving original motion structure and scene layout. Structural guidance helps maintain object consistency during appearance modification and style transfer operations.

H. Scientific Simulation and Research

Researchers can utilize video generation systems for visualizing scientific simulations, environmental changes, and engineering processes. Automatically generated visual content can improve understanding of complex scientific data and dynamic system behavior.

X. ADVANTAGES

The proposed framework offers several important advantages over traditional video generation systems and earlier generative architectures. The integration of diffusion models, structural guidance, and temporal transformers significantly improves the quality, realism, and controllability of generated videos [1], [5], [6].

A. Improved Temporal Consistency

One of the major advantages of the framework is improved temporal consistency across consecutive video frames. Temporal transformer modules and attention mechanisms help maintain smooth motion continuity and reduce abrupt frame transitions [1], [7]. This minimizes flickering artifacts and produces more realistic video sequences.

B. Better Structural Preservation

Structural guidance mechanisms such as depth maps preserve scene geometry, object positioning, and spatial relationships throughout the generated video [1], [9]. This improves scene stability and prevents distortion of moving objects during video synthesis.

C. Enhanced Video Realism

Diffusion based architectures generate visually realistic outputs with better texture quality, lighting consistency, and object appearance [5], [6]. Compared to traditional generative adversarial networks, diffusion models provide more stable training and higher quality video generation capability.

D. High User Controllability

The framework allows users to control video content using textual prompts and structural information [1], [8]. Users can customize object appearance, environmental conditions, scene layout, and motion style according to their requirements. This improves flexibility in multimedia generation applications.

E. Reduced Flickering Artifacts

Temporal attention mechanisms and structural conditioning reduce frame inconsistency and visual flickering [1], [7], [9]. Smooth transitions between frames improve viewing quality and overall motion realism.

F. Support for Long Video Generation

The framework introduces causal attention masking mechanisms to improve long duration video synthesis [1], [10]. This helps maintain temporal order and prevents instability during generation of extended video sequences.

G. Lower Computational Complexity

Latent diffusion models operate in compressed latent space rather than directly processing high resolution pixel data [5], [10]. This reduces memory consumption, computational overhead, and training complexity while maintaining high visual quality.

H. Scalability and Adaptability

The framework can be adapted for multiple multimedia applications including animation, gaming, education, healthcare visualization, virtual reality, and advertising systems. Its modular architecture allows integration with future artificial intelligence technologies [6], [7].

I. Improved Prompt Alignment

Cross attention mechanisms improve alignment between textual descriptions and generated visual outputs [1], [7]. The generated videos more accurately represent user provided prompts and scene descriptions.

J. Stable Training Process

Diffusion models generally provide more stable optimization compared to adversarial training approaches [5], [6]. This reduces training instability problems such as mode collapse and improves generation reliability.

XI. LIMITATIONS

Despite its strong performance, the framework still faces certain limitations.

Accurate depth estimation is necessary for effective structural guidance. Errors in depth maps may reduce motion quality and scene stability.

The framework also requires significant computational resources for training and high resolution video generation.

Rapid object movement may still produce motion artifacts in certain situations.

Ultra high resolution video synthesis and real time video generation remain challenging problems that require further research.

XII. LITERATURE REVIEW

Video generation using artificial intelligence has rapidly advanced due to developments in diffusion models and transformer architectures. Early video synthesis methods mainly relied on Generative Adversarial Networks (GANs), which generated realistic outputs but often suffered from unstable training and poor temporal consistency [?].

Ho et al. introduced Denoising Diffusion Probabilistic Models, which improved generation stability and visual quality through iterative denoising operations [11]. Rombach et al. later proposed Latent Diffusion Models that reduced computational complexity by performing diffusion operations in compressed latent space [5].

Singer et al. developed Make-A-Video, a framework capable of generating videos directly from textual descriptions without paired text video datasets [2]. Ho et al. proposed Imagen Video, which improved high definition video generation and temporal consistency using cascaded diffusion architectures
[6].

Khachatryan et al. introduced Text-to-Video Zero for zero shot video synthesis using pre trained diffusion models [4]. Zhang et al. proposed ControlVideo, which improved controllability and structural consistency during video generation [8].

Recent frameworks such as VideoFusion and CogVideo integrated temporal transformers and structural guidance mechanisms to improve motion continuity and scene realism [3], [7]. These studies demonstrate that diffusion based architectures significantly improve video quality, temporal consistency, and controllability in modern multimedia generation systems.

XIII. RESEARCH GAP

Although diffusion models and transformer architectures have significantly improved text guided video generation systems, several challenges still remain unresolved. Existing frameworks often suffer from temporal inconsistency, flickering artifacts, unstable object motion, and poor identity preservation during long video generation [1], [3], [6].

Many systems also require high computational resources and large scale datasets, limiting real time video synthesis capability [5], [12]. In addition, precise control over object movement, scene structure, and camera motion remains difficult despite the use of textual prompts and structural guidance methods [8], [15].

Current evaluation metrics mainly focus on visual quality and may not accurately measure motion realism and temporal coherence [13], [19]. Therefore, further research is required to improve controllability, efficiency, temporal stability, and real time high resolution video generation in artificial intelligence based multimedia systems.

XIV. FUTURE SCOPE

Future research in video generation may focus on real time video synthesis, ultra high resolution generation, and improved motion consistency. Integration of multimodal artificial intelligence, voice guidance, and interactive video editing can further improve user controllability and realism [6], [14].

Future frameworks are also expected to support personalized multimedia generation, virtual reality applications, and cinema quality video synthesis with lower computational complexity [19], [20]. Continuous advancements in diffusion models and transformer architectures will further enhance intelligent multimedia generation systems.

XV. CONCLUSION

This review paper presented a comprehensive study of video generation using textual descriptions and structural guidance. The paper analyzed the role of diffusion models, latent diffusion architectures, temporal transformers, structural guidance mechanisms, and causal attention masking techniques in improving modern video synthesis systems.

The study highlighted how structural guidance using depth maps improves scene consistency, motion stability, and object preservation across consecutive video frames [1], [9]. The integration of temporal transformer modules and attention mechanisms was shown to significantly improve temporal continuity and reduce flickering artifacts during video generation [1], [7].

The paper also compared several existing video generation frameworks including Make-A-Video, VideoFusion, Text-toVideo Zero, and ControlVideo, demonstrating that diffusion based architectures provide better realism, controllability, and motion coherence compared to earlier approaches [2], [3], [4],
[8].

In addition, the review discussed important applications of video generation systems in entertainment, education, healthcare, gaming, advertising, and virtual reality environments. The advantages, limitations, research gaps, and future scope of artificial intelligence based video generation systems were also examined in detail.

Overall, the proposed framework demonstrates that combining textual guidance with structural conditioning can significantly improve the quality, stability, and controllability of generated videos. Future advancements in diffusion models, multimodal learning, and real time video synthesis are expected to further enhance intelligent multimedia generation systems and enable more realistic and interactive video creation technologies.

REFERENCES

[1] J. Xing et al., “Make Your Video: Customized Video Generation Using Textual and Structural Guidance,” IEEE Transactions on Visualization and Computer Graphics, 2025.

[2] U. Singer et al., “Make A Video: Text Guided Video Generation Without Video Data,” International Conference on Learning Representations, 2023.

[3] Z. Luo et al., “VideoFusion: Diffusion Models for High Quality Video Generation,” Conference on Computer Vision and Pattern Recognition, 2023.

[4] L. Khachatryan et al., “Text to Video Zero: Zero Shot Video Generation Using Diffusion Models,” 2023.

[5] R. Rombach et al., “High Resolution Image Synthesis with Latent Diffusion Models,” Conference on Computer Vision and Pattern Recognition, 2022.

[6] J. Ho et al., “Imagen Video: High Definition Video Generation with Diffusion Models,” 2022.

[7] W. Hong et al., “CogVideo: Large Scale Text Guided Video Generation Using Transformers,” International Conference on Learning Representations, 2023.

[8] Y. Zhang et al., “ControlVideo: Training Free Controllable Video Generation,” 2023.

[9] P. Esser et al., “Structure Guided Video Synthesis Using Diffusion Models,” 2023.

[10] A. Blattmann et al., “High Resolution Video Synthesis Using Latent Diffusion Models,” 2023.

[11] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” Advances in Neural Information Processing Systems, 2020.

[12] A. Nichol and P. Dhariwal, “Improved Denoising Diffusion Probabilistic Models,” International Conference on Machine Learning, 2021.

[13] R. Yang et al., "Video Diffusion Models," Advances in Neural Information Processing Systems, 2022.
[14] S. Blattmann et al., "Align Your Latents: High Resolution Video Synthesis with Latent Diffusion Models," IEEE Conference on Computer Vision and Pattern Recognition, 2023.
[15] Y. Guo et al., "Animating Pictures with Text Driven Diffusion Models," International Conference on Learning Representations, 2023.
[16] M. Brooks et al., "Dreamix: Video Diffusion Models are General Video Editors," 2023.
[17] H. Tune-A-Video, "One Shot Tuning of Image Diffusion Models for Text to Video Generation," IEEE International Conference on Computer Vision, 2023.
[18] T. Ge et al., "Preserve Your Own Correlation: A Noise Prior for Video Diffusion Models," 2023.
[19] C. Wu et al., "NUWA XL: Diffusion over Diffusion for eXtremely Long Video Generation," 2023.
[20] L. Liu et al., "MotionDirector: Motion Customization of Text to Video Diffusion Models," European Conference on Computer Vision, 2024.