International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

AI Based Video Synthesis from Text

1st Mailaram Hepsibah Grace
Computer Science and Engineering
Rajiv Gandhi University of Knowledge Technologies
Basar, India
b201302@rgukt.ac.in

2nd Gangineni Anusha
Computer Science and Engineering
Rajiv Gandhi University of Knowledge Technologies
Basar, India
b200816@rgukt.ac.in

3rd Kotapati Shakeena
Computer Science and Engineering
Rajiv Gandhi University of Knowledge Technologies
Basar, India b200803@rgukt.ac.in

***Abstract*—Recent advancements in artificial intelligence have enabled the development of systems capable of generating multimedia content directly from textual descriptions. Text-to-video generation is an emerging research area that aims to automatically create video sequences based on natural language prompts. However, most existing approaches require large training datasets and high computational resources.**

This paper proposes a zero-shot text-to-video generation system that converts textual prompts into short video sequences using generative models. The proposed system integrates natural language processing techniques with generative video synthesis to generate visual frames corresponding to the input description. These frames are then combined to produce the final video output.

The system provides an interactive interface where users can input descriptive prompts, and the model generates corresponding video sequences. Experimental evaluation shows that the system can generate videos for various prompts with reasonable visual quality. The proposed approach reduces the dependency on large training datasets and demonstrates the potential of zero-shot generative models for automated multimedia content creation.

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

A. Background and Motivation

The rapid growth of artificial intelligence has significantly transformed the way digital content is created and consumed. With the increasing popularity of social media platforms, online learning systems, and digital entertainment services, there is a growing demand for automated multimedia content generation. Platforms such as YouTube, Instagram, and other video-sharing applications require continuous production of engaging visual content, which can be time-consuming and resource-intensive when performed manually.

Traditional video creation involves multiple stages, including script writing, scene design, recording, editing, and rendering. These processes require technical expertise, specialized software, and considerable effort. As a result, generating high-quality videos remains a challenging task for individuals and organizations with limited resources.

Recent advancements in generative artificial intelligence have introduced new possibilities for automating content creation. Text-to-image and text-to-video generation techniques allow machines to produce visual content directly from natural language descriptions. This enables users to generate multimedia content simply by providing textual prompts, reducing the complexity of traditional video production workflows.

However, most existing text-to-video generation systems rely on large-scale datasets containing paired text and video samples. Training such models requires significant computational resources and extensive data collection, which limits their accessibility. Additionally, generating coherent video sequences from textual descriptions involves challenges such as maintaining temporal consistency, understanding complex semantics, and producing realistic visual outputs.

Zero-shot learning has emerged as a promising approach to address these challenges. By leveraging pre-trained models, zero-shot systems can generate outputs for tasks without requiring additional task-specific training data. This approach reduces dependency on large datasets and enables models to generalize to unseen inputs effectively.

This project focuses on developing a zero-shot text-to-video generation system that can convert textual descriptions into video sequences using generative AI models. The aim is to simplify the video creation process by enabling users to generate videos directly from text prompts while minimizing computational requirements and improving accessibility for a wider range of users.

B. Challenges in Text to Video Generation

Text-to-video generation involves several challenges. First, it requires high computational power, as deep learning models such as diffusion and transformer-based architectures are resource-intensive and often depend on GPU acceleration.

Identify applicable funding agency here. If none, delete this.

Second, maintaining temporal consistency across video frames is difficult, as the generated frames must appear smooth and continuous. Any inconsistency between frames can reduce the overall quality of the generated video.

Third, understanding textual prompts accurately is challenging due to ambiguity, complex sentence structures, and variations in natural language. The model must correctly interpret objects, actions, and relationships described in the text.

Additionally, most existing systems rely on large datasets for training, which are expensive to collect and process. Although zero-shot approaches reduce this dependency, they may still face challenges in generating highly accurate outputs. Finally, generating realistic motion and smooth transitions between frames remains a complex problem. Advanced deep learning models are required to improve visual quality, maintain consistency, and produce meaningful video sequences.

C. Related Work and Existing Approaches

Previous research in text-to-video generation has explored a variety of machine learning and deep learning techniques for generating visual content from textual descriptions. Early approaches mainly focused on text-to-image generation using models such as Generative Adversarial Networks (GANs). These models were capable of producing static images from text prompts, but extending them to generate coherent video sequences remained a challenging task.

Traditional methods for video generation often relied on frame-by-frame synthesis techniques, where individual frames were generated independently and then combined to form a video. While these approaches were simple to implement, they often failed to maintain temporal consistency across frames, resulting in discontinuous or unrealistic motion in generated videos.

Recent advancements in deep learning have introduced diffusion-based models, which have significantly improved the quality of generated visual content. These models generate images by progressively removing noise and have been extended to video generation tasks. Diffusion-based frameworks have demonstrated strong performance in producing high-quality frames with better visual realism.

Transformer-based architectures have also played a crucial role in improving text-to-video generation. Models such as text encoders use self-attention mechanisms to capture contextual relationships within textual prompts. These models enable better understanding of complex descriptions involving multiple objects and actions.

In addition, recent approaches such as Make-A-Video and Phenaki have demonstrated the ability to generate video sequences directly from textual input. These models leverage large-scale pre-trained architectures and incorporate temporal modeling techniques to maintain consistency across frames.

However, most existing approaches rely heavily on large datasets containing paired text and video samples, which increases computational requirements and limits scalability. Furthermore, many models focus on improving visual quality but still face challenges in generating long-duration videos with smooth motion.

To address these limitations, zero-shot text-to-video generation approaches have been introduced. These methods utilize pre-trained models to generate videos without requiring additional training on task-specific datasets. Although zero-shot techniques improve flexibility and reduce data dependency, further improvements are needed to enhance output quality and temporal coherence.

D. Proposed Work and Contributions

This project presents a comprehensive approach for generating videos from textual descriptions using zero-shot generative AI models. The work is divided into two major phases.

In the first phase, the system utilizes pre-trained text-to-image and generative models to generate individual visual frames from textual prompts. The input text is processed using natural language processing techniques, and semantic features are extracted using transformer-based encoders. These features guide the generation of frames that represent the objects and scenes described in the prompt. This phase focuses on establishing a baseline system for generating visual content from text.

In the second phase, the system extends the basic framework by incorporating zero-shot text-to-video generation techniques. Instead of relying on large-scale paired text-video datasets, the proposed system leverages pre-trained generative models to produce sequences of frames without additional task-specific training. Frame sequencing and video composition techniques are applied to ensure temporal continuity and generate a coherent video output.

Furthermore, the system integrates a user-friendly interface that allows users to provide textual prompts and generate videos dynamically. The overall framework combines natural language processing, generative modeling, and multimedia processing to produce meaningful video sequences.

The key contributions of this work are as follows:

1) Development of a zero-shot text-to-video generation system using generative AI models.
2) Implementation of text processing and transformer-based encoding techniques for semantic understanding.
3) Generation of video frames from textual prompts and composition of frames into video sequences.
4) Integration of a user interface for interactive video generation.
5) Evaluation of the system using multiple textual prompts to analyze performance and output quality.

E. Organization of the Paper

The remainder of this paper is organized as follows. Section II presents the problem statement and objectives of the proposed system. Section III reviews related work and existing approaches in text-to-video generation. Section IV describes the system architecture and overall framework of the proposed model. Section V explains the methodology and algorithm used for generating videos from textual prompts. Section VI

provides implementation details and system components. Section VII presents experimental results and comparative analysis. Finally, Section VIII concludes the paper and discusses potential future work in improving text-to-video generation systems.

sectionDATASET AND INPUT PROCESSING

F. Input Data Collection

For this project, no fixed dataset is used. Instead, the system accepts user-provided textual prompts as input for generating video sequences. These prompts describe scenes, objects, or actions such as “a sunset over the ocean” or “a car moving on a road.”

Since the system follows a zero-shot approach, it utilizes pre-trained generative models that have already learned from large-scale datasets. Therefore, no additional dataset collection or manual annotation is required.

G. Input Processing

The input text undergoes preprocessing before being passed to the model. This includes cleaning, normalization, and tokenization. The processed text is converted into numerical representations using a transformer-based encoder.

These representations capture the semantic meaning of the prompt and guide the video generation process.

H. Input Characteristics

The input prompts used in the system have the following characteristics:

1) **Flexible Input:** Users can provide simple or complex descriptions.
2) **Unstructured Nature:** Inputs may vary in grammar, length, and style.
3) **Semantic Complexity:** Prompts may include multiple objects and actions.
4) **Real-Time Input:** Data is provided dynamically during runtime.

I. Output Visualization

The system generates multiple frames based on the input prompt, which are combined to form a video.

- Generated frames represent the visual interpretation of the text.
- Frames are arranged sequentially to maintain continuity.
- The final output is a video displayed through the interface.

II. Methodology and Model Implementation

A. Overall Framework

The primary objective of this project is to develop an automated system that generates video sequences from textual descriptions using zero-shot generative AI models. The methodology is structured into two major phases.

The first phase involves processing textual prompts using natural language processing techniques and extracting semantic representations using transformer-based encoders. The second phase focuses on generating visual frames using diffusion-based generative models and composing them into a continuous video sequence.

The complete workflow of the system consists of the following stages:

1) User input of textual prompt
2) Text preprocessing and normalization
3) Tokenization and encoding using transformer models
4) Frame generation using diffusion-based generative model
5) Sequential arrangement of frames
6) Video composition and output generation

B. Text Processing

Text preprocessing is an important step to ensure that the input prompt is suitable for model interpretation. The input text is cleaned and normalized by removing unnecessary characters and formatting inconsistencies.

The processed text is tokenized into smaller units and converted into numerical representations. Transformer-based encoders are used to extract semantic features, which capture relationships between objects, actions, and attributes described in the prompt.

Unlike traditional methods, minimal preprocessing is applied since transformer models are capable of handling raw text efficiently.

C. Frame Generation

The encoded text representation is passed to a diffusion-based generative model, which generates visual frames corresponding to the input prompt. The model gradually transforms noise into meaningful images by learning visual patterns from large-scale pre-trained data.

Multiple frames are generated to represent different stages or perspectives of the described scene.

D. Video Composition

The generated frames are arranged sequentially to create a coherent video sequence. Frame ordering ensures temporal consistency and smooth transitions between frames.

The final video output is generated by combining these frames and is displayed to the user through the interface.

E. Flow Diagram

F. Generative Models for Text-to-Video Generation

In this project, generative AI models are used to convert textual descriptions into visual frames and video sequences. Unlike traditional machine learning models, these approaches do not rely on handcrafted features but learn representations directly from large-scale data.

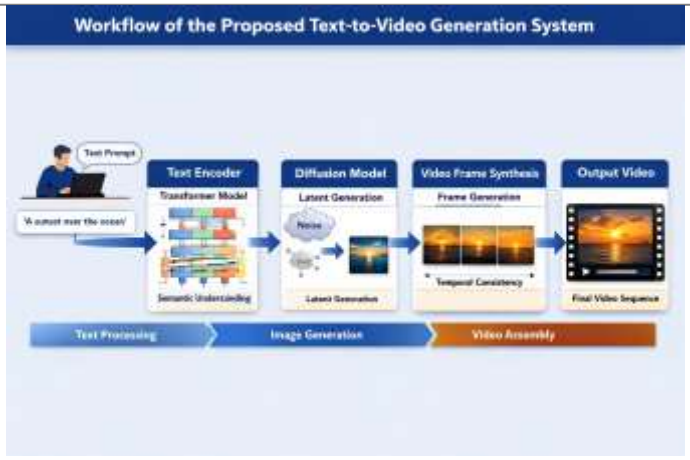

Fig. 1. Workflow of the Proposed Text-to-Video Generation System

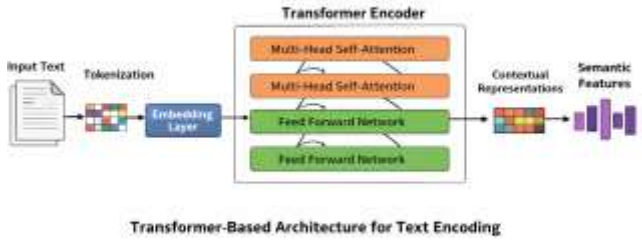

Fig. 2. Transformer-Based Architecture for Text Encoding

*1) **Transformer-Based Text Encoder**:* Transformer-based models are used to process and understand the input text prompt. These models utilize self-attention mechanisms to capture contextual relationships between words in a sentence.

In this project:

- The input text is tokenized and converted into embeddings.
- A pre-trained transformer encoder is used to extract semantic features.
- The encoded representation guides the visual generation process.

Transformer models are effective in understanding complex textual descriptions, including relationships between objects and actions.

*2) **Diffusion-Based Generative Model**:* Diffusion models are used for generating visual frames from textual embeddings. These models generate images by gradually removing noise from a random distribution.

In this project:

- The model takes encoded text features as input.
- It generates images through iterative denoising steps.
- Multiple frames are produced to represent different stages of the scene.

Diffusion models are capable of generating high-quality and realistic images compared to traditional generative methods.

*3) **Frame Sequencing and Video Generation**:* After generating individual frames, the system combines them to produce a video sequence.

- Frames are arranged in a logical order to maintain continuity.
- Basic temporal consistency is ensured between frames.
- The final output is a video generated from the sequence of images.

G. Transformer-Based and Generative Models

Transformer and generative models play a crucial role in the proposed text-to-video generation system. These models enable the system to understand textual input and generate corresponding visual outputs.

*1) **Text Encoder using Transformer Models**:* Transformer-based encoders are used to process and understand the input textual prompts. These models utilize self-attention mechanisms to capture relationships between words and extract meaningful contextual representations.

In this project:

- The input text prompt is tokenized and converted into embeddings.
- A pre-trained transformer encoder is used to extract semantic features.
- These features represent objects, actions, and relationships described in the text.

Transformer models are effective in handling complex and unstructured text, making them suitable for guiding visual generation tasks.

*2) **Diffusion-Based Image Generation Model**:* Diffusion models are used to generate visual frames from textual representations. These models work by starting with random noise and gradually refining it into meaningful images through multiple denoising steps.

In this project:

- The encoded text features are used as input to the generative model.
- The model produces high-quality image frames representing the input prompt.
- Multiple frames are generated to simulate motion and scene progression.

Diffusion models provide better image quality and realism compared to traditional generative approaches.

*3) **Zero-Shot Video Generation Approach**:* The system follows a zero-shot approach, meaning it does not require task-specific training on text-video datasets.

- Pre-trained models are directly used for video generation.
- The system generalizes to unseen prompts without additional training.

- This reduces dependency on large datasets and improves flexibility.

*4) **Frame Sequencing and Video Synthesis**:* After generating individual frames, the system combines them to produce a video.

- Frames are arranged in sequence to maintain continuity.
- Basic temporal consistency is preserved between frames.
- The final output is a video sequence generated from textual input.

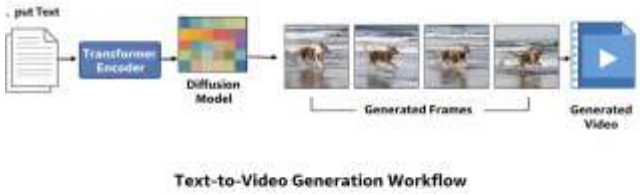

Fig. 3. Text-to-Video Generation Workflow

H. Drawbacks of Existing Systems

Existing text-to-video generation systems have shown significant progress, but several limitations still remain.

- **High Computational Cost:** Most existing systems rely on large-scale deep learning models such as diffusion and transformer architectures, which require powerful GPU resources and long processing times.
- **Dependency on Large Datasets:** Many approaches require extensive paired text-video datasets for training, making them difficult to scale and expensive to develop.
- **Poor Temporal Consistency:** Maintaining smooth transitions and consistency across video frames is challenging, often leading to visual artifacts and discontinuities.
- **Limited Understanding of Complex Prompts:** Existing models may struggle to interpret complex or ambiguous textual descriptions, affecting the quality of generated videos.
- **Limited Video Length and Quality:** Most systems generate short-duration videos with limited resolution and realism.

I. Advantages of Proposed System

The proposed zero-shot text-to-video generation system addresses several of these limitations by leveraging advanced generative AI techniques.

- **Zero-Shot Capability:** The system does not require task-specific training datasets, reducing dependency on large-scale annotated data.
- **Efficient Text Understanding:** Transformer-based encoders effectively capture semantic relationships in textual prompts.
- **High-Quality Frame Generation:** Diffusion-based models generate visually realistic frames from textual descriptions.
- **Flexible Input Handling:** The system can process a wide variety of user-provided prompts without restrictions.
- **Automated Video Creation:** The system simplifies video generation by automatically converting text into video sequences without manual intervention.

III. Experimental Results and Comparative Analysis

Fig. 4. Generated Video Frames from Text Prompts

A. Evaluation Metrics

Evaluating text-to-video generation systems is challenging, as it involves assessing both visual quality and semantic alignment between text and generated video. In this project, the performance of the system is evaluated using the following metrics:

*1) **Fre´chet Inception Distance (FID)**:* FID measures the similarity between generated images and real images by comparing their feature distributions. Lower FID values indicate better visual quality and realism of generated frames.

*2) **Structural Similarity Index (SSIM)**:* SSIM evaluates the similarity between generated frames in terms of structure, brightness, and contrast. It helps measure the consistency and quality of visual output.

*3) **CLIP Score**:* CLIP score measures how well the generated images align with the input textual prompt. It evaluates the semantic similarity between text and visual output. Higher scores indicate better alignment.

*4) **Visual Quality Assessment**:* In addition to quantitative metrics, qualitative evaluation is performed by visually inspecting the generated frames and videos. This includes:

- Clarity and realism of generated images
- Consistency across frames
- Accuracy in representing the input prompt

B. Results Analysis

The system was tested using multiple textual prompts representing different scenes and objects. The generated outputs demonstrate that the model is capable of producing visually meaningful frames from textual descriptions.

The results indicate that the system performs well for simple prompts involving clear objects and scenes. However, performance may vary for complex or ambiguous prompts, where maintaining temporal consistency becomes more challenging.

C. Comparative Analysis

The proposed zero-shot approach was compared with existing text-to-video methods based on qualitative evaluation.

- The proposed system reduces dependency on large datasets by using zero-shot learning.
- It provides flexible input handling compared to traditional trained models.
- However, some existing models may produce more temporally consistent videos due to extensive training.

IV. Future Scope

The proposed text-to-video generation system can be further improved and extended in several directions to enhance performance, scalability, and real-world applicability.

1) ***Improved Video Quality and Resolution****:* Future work can focus on generating higher resolution videos with improved visual quality and realism by using advanced diffusion models and better training techniques.

2) ***Enhanced Temporal Consistency****:* Maintaining smooth transitions between frames is a major challenge. Future improvements can include advanced temporal modeling techniques to generate more coherent and stable video sequences.

3) ***Longer Video Generation****:* The current system generates short video clips. It can be extended to produce longer-duration videos with more complex scenes and continuous motion.

4) ***Real-Time Video Generation****:* The system can be optimized for real-time applications by reducing computational complexity and improving processing speed using efficient model architectures and hardware acceleration.

5) ***Multimodal Input Support****:* Future systems can support additional inputs such as audio, images, or user sketches along with text to generate more detailed and customized videos.

6) ***Interactive User Control****:* Users can be given more control over the video generation process, such as specifying camera angles, object positions, motion speed, and scene transitions.

7) ***Integration with Applications****:* The system can be integrated into real-world applications such as content creation platforms, educational tools, animation systems, and social media applications.

8) ***Ethical and Responsible AI****:* Future work can focus on implementing safeguards to prevent misuse, such as generating harmful or misleading content, ensuring responsible deployment of generative AI systems.

V. Conclusion

VI. Conclusion

This project presents a zero-shot text-to-video generation system that converts textual descriptions into visual video sequences using advanced generative AI models. The proposed approach leverages transformer-based text encoders to understand semantic information from user input and diffusion-based models to generate high-quality visual frames.

Unlike traditional approaches that require large-scale paired text-video datasets, the proposed system utilizes a zero-shot framework, reducing dependency on task-specific training data. The system is capable of generating meaningful and visually coherent frames from a variety of textual prompts, demonstrating its flexibility and adaptability.

The experimental results show that the system performs effectively for simple and well-defined prompts, producing visually realistic outputs. However, challenges such as maintaining temporal consistency and handling complex descriptions still remain.

Overall, this work highlights the potential of generative AI in automating video creation and simplifying content generation workflows. The proposed system provides a foundation for future research in text-to-video generation, with opportunities to improve video quality, scalability, and real-time performance.

References

[1] A. Vaswani et al., “Attention Is All You Need,” NeurIPS, 2017.
[2] J. Ho et al., “Denoising Diffusion Probabilistic Models,” NeurIPS, 2020.
[3] A. Ramesh et al., “Hierarchical Text-Conditional Image Generation with CLIP Latents,” arXiv, 2022.
[4] C. Saharia et al., “Imagen: Photorealistic Text-to-Image Diffusion Models,” ICML, 2022.
[5] R. Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models,” CVPR, 2022.
[6] Meta AI, “Make-A-Video: Text-to-Video Generation without Text-Video Data,” 2022.
[7] Google Research, “Phenaki: Variable Length Video Generation from Open Domain Textual Descriptions,” 2022.
[8] Text2Video-Zero, “Text-to-Video Generation without Training,” arXiv, 2023.
[9] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision (CLIP),” ICML, 2021.
[10] Gradio, “Gradio: Build Machine Learning Web Apps Quickly,” Available: https://gradio.app