T2V Studio: A Diffusion-Based Framework for Text-to-Video Generation

Benakappa S M[1], Moulya R G[2], Poorvika[2], Pratheeksha D R[2], Rakshitha R[2]

[1]Associate Professor, Department of Computer Science and Engineering, JNNCE, Shivamogga, Karnataka, India
[2]UG Students, Department of Computer Science and Engineering, JNNCE, Shivamogga, Karnataka, India

---

Abstract - The design and implementation of T2V Studio, a diffusion-based framework for text-to-video generation, are presented in this paper. The proposed system combines motion-aware temporal conditioning with image-based diffusion models to produce short animated videos from natural language prompts. AnimateDiff motion adapters are incorporated to preserve temporal consistency across video frames, and Stable Diffusion 1.5 is used for frame-wise image synthesis. The framework's ability to produce coherent short video clips under constrained computational resources is demonstrated through its implementation and evaluation on Google Colab using an NVIDIA T4 GPU. The findings demonstrate that the system can generate visually consistent animations from text descriptions without the need for manual video editing or model training, which makes it appropriate for applications involving rapid content creation and prototyping.

***Key Words*:** Text-to-Video Generation, Diffusion Models, Stable Diffusion, AnimateDiff, Motion-Aware Video Generation, Generative AI.

1.INTRODUCTION

Traditional multimedia video content creation can be a huge manual process that requires technical expertise and computational power, especially in the design of animations, motion generation, and composition. These high barriers affect the scalability and applicability of video production through traditional tools and techniques. There have been developments in the field of deep learning that allow the creation of models that can generate multimedia content based on textual input.

In this paper, I describe the design and implementation process of the Text-to-Video Generation System that uses diffusion-based generative models like Stable Diffusion, as well as motion-based designs like AnimateDiff. This system allows the user to turn an input text message or phrase into an animated video composed of multiple short images that flow continuously from one frame to another. All these images or frames are designed to be temporally consistent.

The framework is implemented as an interactive application through a Gradio interface and thus enables the user to create videos based on simple text inputs. Google Colab is used to implement the framework on an NVIDIA T4 GPU and thus validate its usability despite limitations on computational power. This paper showcases the use of diffusion-based techniques for efficient text-to-video content creation.

2. RELATED WORK

The body of the paper consists of numbered sections that present the main findings. These sections should be organized to best present the material. The use of text-to-video (T2V) synthesis has become more feasible in recent years thanks to the developments in diffusion-based generative models. Owing to their semantic relevance, high-quality synthesis, and trainability, diffusion models have been widely used in visual content synthesis. Architectural improvements to diffusion-based T2V synthesis methods to address the challenges of flickering in the synthesized videos were proposed in the research conducted by Arkhipkin et al. [1]. Chauhan et al. [2] described methods related to text-image fusion and masked diffusion to address semantic alignment.

Make-Your-Video [3] proposed a two-stage diffusion framework for separating keyframe generation and motion synthesis. VideoGen [8] improved visual quality with reference-guided latent diffusion and temporal super-resolution for high-definition video generation. Other works have included LoRA-based fine-tuning [4], sketch-guided diffusion [6], multimodal style transfer [7], and LLM-guided diffusion for enhanced semantic control [9]. Together, these techniques demonstrate the increasing importance of language-guided generation and multimodal learning in T2V systems [10]. These developments are supported by fundamental research in machine learning and artificial intelligence. The foundation for contemporary data-driven models was established by the machine learning paradigms outlined by Mitchell [12] and the classical AI ideas presented by Russell and Norvig [11]. Scalable visual and linguistic representations were made possible by the revival of deep learning, which was fueled by hierarchical feature learning as described by LeCun et al. [13]. By using self-attention processes, transformer-based architectures—first presented by Vaswani et al. [14]—further transformed sequence modeling. Building on this, text-conditioned generative models have been directly impacted by Devlin et al.'s [15] demonstration of the efficacy of deep bidirectional transformers in language interpretation.

3. PROBLEM STATEMENT AND OBJECTIVES

3.1. Problem Statement

Historically, producing animated video content via traditional methods has required substantial manual input, specialized knowledge and skills, as well as a considerable amount of time. Traditional animation production operations consist of various steps leading from the start of the project (concept preparation phases) to completion of the finished product (post-production) with each step needing to be completed before proceeding to the next phase, which results in a situation where it is challenging or impractical to quickly create video content at scale. These limitations make it difficult for most users, including students, academics and amateur content creators, to create animated video content due to their lack of specialized knowledge, expertise and/or resources necessary to complete all the productions steps involved in animation production.

In the last few years, text-to-video Generation Systems through advances in Deep Learning have shown that it is feasible to produce (synthesize) an animated video using textual descriptions of what is included in an Animated Video. Currently, a majority of the "state-of-the-art" Text-To-Video generation systems rely heavily upon using highly complex and computational models that are trained on massive data sets and that require significant computing power for running. Therefore, these technologies are generally not adequate for deployment to devices that have limited processing capabilities and/or they cannot be used on a regular basis because they require knowledge of extensive software and hardware components.

Consequently, a working Text to Video generation System that is easily accessible is needed to allow users to download and generate videos in environments with low power consumption and that produces acceptable quality videos and that generates standardized (non-unique) Video for every single input text. The best solution would leverage available pre-trained generative machine learning models, circumvent the need for additional costly training, and provide users with a straightforward method to generate short videos (coherent or in sequences) based solely on user-supplied textual input.

3.2. Objectives

The primary objectives of this project are:

- To design and implement a diffusion based Text to Video generation framework that converts natural language text into short video content.
- To achieve temporal consistency in generated videos by incorporating motion-aware diffusion techniques.
- To ensure that the system can generate videos within a practical and acceptable inference time.
- To provide an interactive and user friendly interface that allows users to input text prompts and view generated videos.

4. PROPOSED SYSTEM

This section describes the overall architecture and working methodology of the proposed Text-to-Video (T2V) generation system. The system is designed to convert natural language descriptions into short, temporally consistent video sequences using diffusion-based generative models. By integrating Stable Diffusion with the AnimateDiff motion module, the system ensures high-quality visual output along with smooth motion transitions across frames

4.1. System Architecture

This system uses a combination of a Latent Diffusion Model (LDM) and a motion-aware module to convert text-based inputs into sequences of video frames. The system has four modules: text encoding, latent diffusion frame generation, temporal motion modelling, and video decoding.

The architecture has three primary components:

1. The Encoder (CLIP), which processes the input text prompt and produces latent vectors. These latent vectors contain semantics from the user's description and are used as inputs for the diffusion model; thus, utilizing CLIP-based Latent vectors results in high correlation between the text prompt and the video produced through the diffusion methodology.
2. U-Net with AnimateDiff Motion Module. The main network of the generation model is a modified version of the U-Net architecture, incorporating the AnimateDiff motion module. This includes additional temporal attention layers that help the model learn temporal dependencies from one frame to the next and create a coherent final output with reduced artefacts and flickering.
3. Variational Autoencoder (VAE) Decoder: Finally, the VAE decoder takes the denoised latent space representations produced during the diffusion process and converts them back into pixel-level information. This step produces frame images that can be combined into a video.

In the architecture diagram figure 1, the CLIP text encoder and VAE decoder are implicitly represented within the Stable Diffusion pipeline, as they are integral components of the pretrained diffusion model.

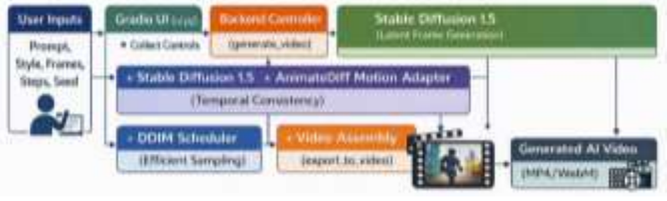

Fig. 1. *Architecture of the proposed Text-to-Video generation system*

4.2. Methodology

Utilizing a structured multi-step procedure, the methodology of the intended Text-to-Video generation system will consist of many procedures with each of the steps assisting with the production of video outputs that are temporally consistent and semantically related.

A natural language text prompt is submitted to a web-based application as the input. The text prompt will be transformed into a semantic embedding via the CLIP Text Encoder. The embedding serves as a basis for the diffusion phase by helping influence the denoising phases of the U-Net Architecture.

The diffusion will start with random Gaussian noise that exists within a Latent Space. Through a series of iterative denoising processes, the U-Net will take the random Gaussian noise and transform it into useful latent representations that are conditioned by the text embedding.

The AnimateDiff motion module is applied to the diffusion phase in order to supply temporal context to the outputs of the diffusion process and the means through which the data is shared among frames. A Denoising Diffusion Implicit Model (DDIM) scheduler will indicate the number of denoising steps. The use of the DDIM allows increased speed in inference due to the decreased number of sampling steps while preserving satisfactory output quality. The final denoised images will be fed to a VAE Decoder to produce the final frames and subsequently compiled into a video file.

4.3. Workflow of the Proposed System

The workflow of the proposed Text-to-Video (T2V) generation system outlines the complete process of turning a natural language text prompt into a short-animated video as seen in figure 2. This workflow aims to maintain semantic correctness, temporal consistency, and effective use of computing resources. Each stage of the pipeline performs a clear function and smoothly passes information to the next stage, resulting in a cohesive and visually meaningful video output.

The process starts with user interaction through a web interface. The user inputs a descriptive text prompt that outlines the desired scene, action, or visual concept. Along with the prompt, the user can set parameters such as the number of frames, inference steps, guidance scale, visual style, and seed value. These parameters let users influence both the creative and technical sides of the video generation process. After submitting the inputs, they are sent to the backend processing pipeline.

In the second stage, the CLIP-based text encoder processes the input text prompt. This encoder converts the natural language input into a high-dimensional semantic embedding that captures the prompt's contextual meaning. The generated embedding acts as a conditioning signal for the diffusion model, making sure that the visual content produced closely matches the user's description. This step is crucial for keeping semantic consistency between the text input and the generated video frames.

After text encoding, the diffusion-based generation process kicks off in the latent space. The system starts the latent representation with random Gaussian noise. The Stable Diffusion model then refines this noise through a denoising process guided by the text embedding. Working in latent space significantly cuts computational costs while maintaining high output quality. Each denoising step gradually changes the noise into meaningful latent representations that relate to the visual elements described by the input text.

To ensure smooth movement and consistency across frames, the AnimateDiff motion module is added to the diffusion pipeline. During the denoising process, AnimateDiff introduces temporal attention mechanisms that help the model learn the relationship between consecutive frames. Instead of creating each frame on its own, the motion module shares temporal information between frames, leading to consistent object appearances and realistic motion transitions. This step effectively reduces visual issues like flickering, jitter, and sudden changes between frames.

The denoising process is overseen by the DDIM (Denoising Diffusion Implicit Model) scheduler. The scheduler sets the number of inference steps and manages the noise removal process. By using DDIM sampling, the system generates video faster compared to traditional diffusion methods while keeping acceptable visual quality. Although the workflow includes a CPU execution path for completeness, the system is mainly designed for GPU execution. Diffusion-based Text-to-Video generation on a CPU is costly and not practical for regular use.

Once the latent video frames are fully denoised, they are sent to the Variational Autoencoder (VAE) decoder. The VAE decoder transforms the latent representations into pixel-space image frames. This decoding process changes abstract latent features into visual frames with color, texture, and spatial details. The decoder ensures each frame keeps the visual qualities learned during the diffusion process.

In the final stage, the generated image frames are compiled into a video file at a fixed frame rate. The frames are arranged properly to maintain motion continuity. The resulting video is temporarily stored and rendered directly in the Gradio interface, allowing users to preview the output instantly. Users can then download the generated video for further use.

Overall, the proposed workflow ensures effective coordination between natural language understanding, diffusion-based visual synthesis, motion modeling, and video rendering. By organizing these stages into a structured pipeline, the system successfully turns textual descriptions into short, cohesive, and temporally consistent video clips that are suitable for creative and practical uses.

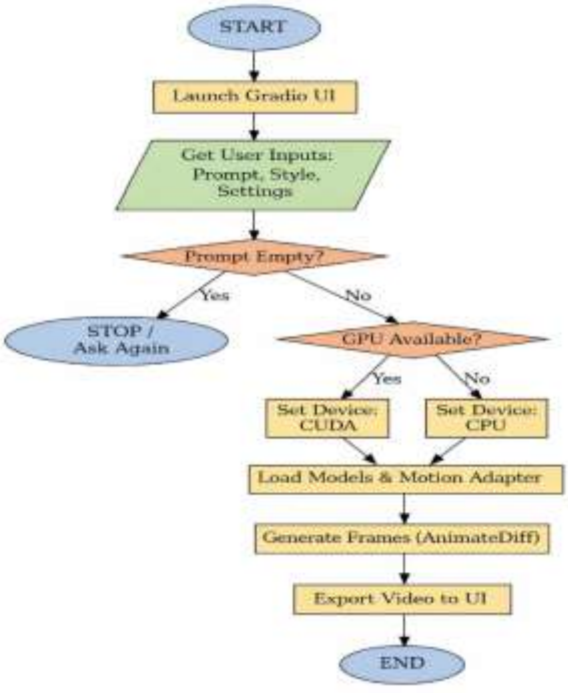

Fig. 2. *Workflow of the proposed Text-to-Video generation system.*

5. IMPLEMENTATION AND RESULTS

This section describes the practical implementation of the proposed Text-to-Video (T2V) generation system and presents the experimental results obtained during evaluation. The system was implemented using modern deep learning frameworks and tested under different configurations to analyze performance, output quality, and temporal consistency.

5.1. Implementation Details

The proposed system is implemented using Python as the primary programming language. The proposed system will be developed using the following technologies, Pytorch as a Deep Learning Framework, Diffusers Library to integrate Stable Diffusion 1.5 with AnimateDiff and is deployed onto a Google Colab Server with support for GPU (NVIDIA T4) acceleration during the diffusion inference process.

With the help of Gradio, created the web interface so that users can interact with the system; they are provided with prompts to type in; they can define visual styles; they can also adjust generation parameters (the number of frames generated, inference steps, guidance scale, and seed) which is illustrated in below figure 3. The back end (pipeline) of the application connects to the web interface via an events-driven architecture so users can generate and preview video in real time.

In order to ensure the system operates stably, several optimization techniques are used in the back end, including model caching, CPU offloading to the GPU, and VAE slicing; these processes significantly reduce the amount of GPU Memory available and help to mitigate out of memory issues during video generation.

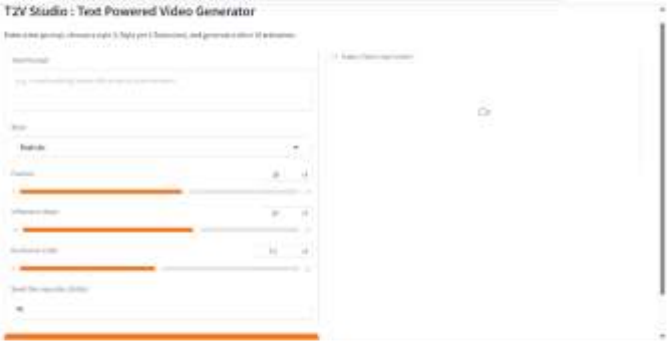
T2V Studio : Text Powered Video Generator

Fig. 3. *Gradio-based user interface of the Text-to-Video generation system.*

5.2. Experimental Setup

The experimental evaluation of the system was done by providing multiple text prompts as input for different images representing scenes, activities, and visual styles. These prompts include dynamic human movement, environmental, and stylized animated scenes. The experimental evaluation included both execution of the code via the graphics card (GPU) as well as the central processing unit (CPU) in order to assess any differences in the execution times of the two hardware methods.

The experiments were conducted using the following parameters:

- Total Frames: 8–24.
- Inference Steps: 16–32.
- Guidance Scale: 5.0–10.0.
- Seed Value: Fixed.

The DDIM Scheduler was used for all experimental evaluations in order to better balance the trade-off between output quality and the speed of each inference process. Qualitative evaluations of the generated videos were made based on the clarity of what was generated visually, the smoothness of movement, and the consistency of time.

5.3. Results and Performance Analysis

The proposed Text-to-Video (T2V) generation system was assessed to evaluate the quality of the videos produced, the consistency of movement across frames, the flexibility of styles, and the performance under limited computing resources. The evaluation focused on inference-only execution with pretrained diffusion models, without any additional training or fine-tuning.

5.3.1. Qualitative Results

The system generated short animated video clips that matched the input text prompts across various scenarios. Prompts about human movement, environmental scenes, and stylized concepts resulted in visually coherent frame sequences with smooth transitions between consecutive frames. The integration of the AnimateDiff motion adapter greatly improved the consistency of movement by reducing flickering and maintaining object continuity throughout the frames.

Figure 4 shows sample consecutive frames generated by the system for a specific text prompt. The frames demonstrate consistent object structure, smooth motion, and stable visual composition, indicating effective modeling of movement during the diffusion process. The use of latent diffusion allowed the system to maintain visual details while working efficiently with limited GPU memory.

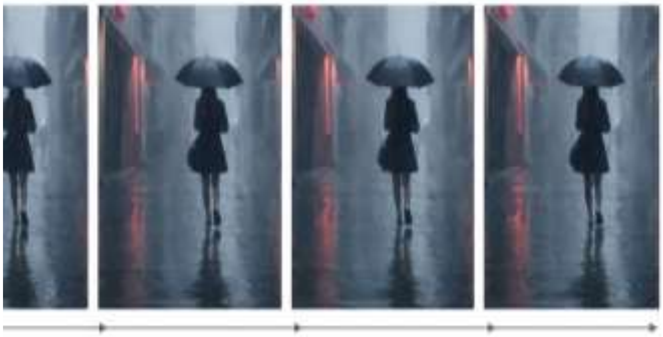

Fig. 4. *Sample consecutive frames generated by the proposed system*

5.3.2. Style-Based Variations

To further assess the system's flexibility, we ran experiments using the same text prompt but with different visual styles. Figure 5 illustrates the video frames generated by applying multiple styles to a single input prompt. While the semantic meaning of the prompt stays the same, the visual look of the generated frames varies depending on the chosen style.

These results show that the diffusion-based framework can maintain meaning while allowing stylistic changes through prompt conditioning. The ability to create diverse visual outputs from the same text highlights the system's creative potential without needing extra model training or specific style transfer methods.

Prompt- A girl walking in rain, cinematic lighting

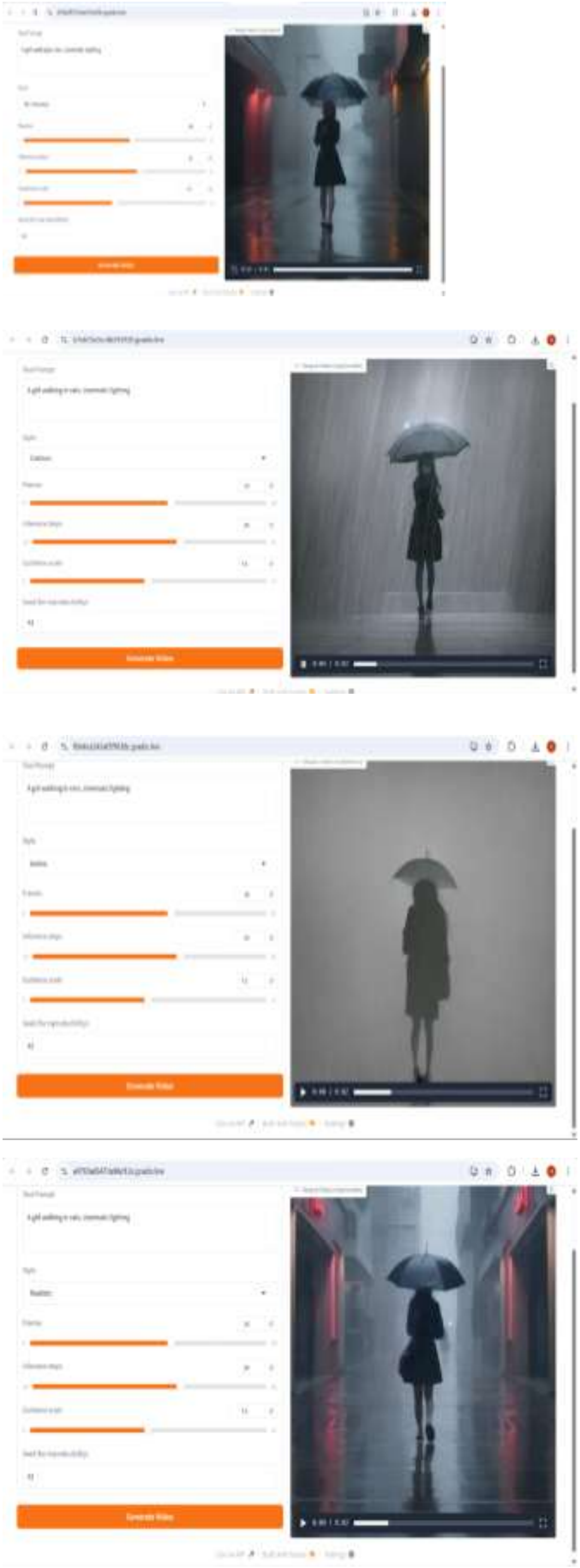

Fig. 5. *Generated video frames for the same text prompt under different visual styles*

5.3.3. Performance Analysis

We mainly evaluated performance on Google Colab using an NVIDIA T4 GPU. The system achieved practical inference times suitable for interactive use, especially when the number of frames and inference steps were kept within recommended limits. GPU execution provided significantly faster generation times compared to CPU execution, making GPU acceleration crucial for practical Text-to-Video generation.

While increasing these parameters improved visual quality and reduced noise, it also led to longer generation times. The use of the DDIM (Denoising Diffusion Implicit Model) scheduler helped balance this trade-off by lowering the number of sampling steps needed, enabling quicker inference while maintaining acceptable visual quality.

Although the workflow includes a CPU execution path for completeness, CPU execution proved to be expensive and unsuitable for regular use due to much higher inference times. Thus, the system is mainly designed for GPU deployment.

5.3.4. Impact of System Optimizations

Several optimization techniques helped ensure stable and efficient execution. Model caching reduced the need to repeatedly load pretrained weights, cutting down on initialization time. CPU offloading and VAE slicing minimized GPU memory use, preventing out-of-memory errors during video generation. These optimizations allowed the system to reliably produce videos with up to 24 frames on a resource-limited NVIDIA T4 GPU.

5.3.5. Discussion

The experimental results show that the proposed T2V Studio system achieves a balance between visual quality, movement consistency, and computing efficiency. While the system is limited to short video lengths and relies entirely on pretrained models, it effectively serves its purpose as an accessible and user-friendly framework for Text-to-Video generation. The results indicate that motion-aware diffusion techniques like AnimateDiff significantly improve temporal coherence, making the system suitable for fast content creation and prototyping applications.

6. APPLICATIONS

The suggested Text-to-Video generation system can be used in various areas that need quick and automated video content creation. In education, the system helps create short visual explanations for concepts and scenarios. In digital marketing and social media, it allows for the fast creation of promotional or storytelling videos based on text descriptions. The framework is also valuable for quickly prototyping animated ideas in creative fields, including concept visualization and pre-production planning. Additionally, the system aids content creators and designers by delivering visually consistent animated outputs without needing manual video editing skills.

7. CONCLUSION

This paper showed the design and implementation of T2V Studio, a diffusion-based Text-to-Video generation framework that can produce short animated videos from natural language prompts. By combining Stable Diffusion 1.5 with AnimateDiff motion adapters, the system achieves temporally consistent video generation without needing model training or fine-tuning. Using latent diffusion and DDIM sampling allows efficient inference even with limited computational resources. This makes the system suitable for platforms like Google Colab that have restricted GPU memory. Experimental results show that the system generates visually coherent video clips across different prompts and styles while keeping motion continuity and acceptable inference time. The interactive Gradio-based interface further improves usability, enabling users to easily control generation parameters and preview outputs. Overall, this framework offers an accessible and practical solution for quick Text-to-Video content creation and prototyping.

8. FUTURE SCOPE

Even though the system shows effective Text-to-Video generation, several improvements can be explored in the future. The framework could be expanded to support longer video durations and higher frame resolutions by using better memory optimization techniques. Future versions might incorporate explicit motion control or guidance for camera movement to boost motion diversity and realism. Adding audio generation and synchronization could facilitate multimodal video creation. Additionally, fine-tuning diffusion models on specific datasets may enhance visual quality for specialized uses such as education, animation, and marketing. Another potential improvement could be deploying it as a standalone web application or API service for wider access.

20. REFERENCE

[1] Arkhipkin, V., Shaheen, Z., Vasilev, V., Dakhova, E., Sobolev, K., Kuznetsov, A., & Dimitrov, D. (2024). ImproveYourVideos: Architectural Improvements for Text-to-Video Generation Pipeline. IEEE Access. DOI: 10.1109/ACCESS.2024.0429000

[2] Chauhan, S., Dudhane, A., & Murala, S. (2024). A Novel Scheme for Generating Context-Aware Images Using Text-Image Fusion with Masked Diffusion Models. IEEE Transactions on Multimedia.

[3] S. K. Alhabeeb and A. A. Al-Shargabi, "Text-to-Image Synthesis with Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction," IEEE Access, Accepted for publication. DOI: 10.1109/ACCESS.2024.3365043.

[4] W. Xiang, S. Xu, C. Lv, and S. Wang, "A Customizable Face Generation Method Based on Stable Diffusion Model," IEEE Access, vol. 12, pp. 195307–195318, 2024. DOI: 10.1109/ACCESS.2024.3520719.

[5] R. Rombach et al., “High-resolution image synthesis with latent diffusion models,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 10684–10695.

[6] Y. Peng, C. Zhao, H. Xie, T. Fukusato, and K. Miyata, "Sketch-Guided Latent Diffusion Model for High-Fidelity Face Image Synthesis," IEEE Access, Accepted for publication. DOI: 10.1109/ACCESS.2023.3346408.

[7] R. Togo, M. Kotera, T. Ogawa, and M. Haseyama, "Text-Guided Style Transfer-based Image Manipulation Using Multimodal Generative Models," IEEE Access, Accepted for publication. DOI: 10.1109/ACCESS.2021.3069876.

[8] Li, X., Chu, W., Wu, Y., Yuan, W., Liu, F., Zhang, Q., Li, F., Feng, H., Ding, E., & Wang, J. (2023). VideoGen: A Reference-Guided Latent Diffusion Approach for High-Definition Text-to-Video Generation. arXiv:2309.00398.

[9] Muhammad Waseem, Muhammad Usman Ghani Khan, and Syed Khaldon Khurshid, "LCGD: Enhancing Text-to-Video Generation via Contextual LLM Guidance and U-Net Denoising," IEEE Access, vol. 13, pp. 47073-47085, 2025, doi: 10.1109/ACCESS.2025.3550945.

[10] Kim, D., Joo, D., & Kim, J. (2020). TiVGAN: Text-to-Image-to-Video Generation with Step-by-Step Evolutionary Generator. IEEE Access, 8, 153113–153122.

[11] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach. Englewood Cliffs, NJ, USA: Prentice Hall, 2010.

[12] T. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997.

[13] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, 2015.

[14] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL), 2019, pp. 4171–4186.