iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84174

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

AI-Based Text-to-Video Generation Using Diffusion and Deep Learning Techniques

Vamshi Krishna Kurva[1], G. Narasimham[2]
[1]*Post Graduate Student, M. Tech, Department of Computer Science and Engineering, Jawaharlal Nehru, Technological University, Hyderabad, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Jawaharlal Nehru Technological, University, Hyderabad, India*

Abstract: ***Recent advances in Generative Artificial Intelligence have significantly improved automatic multimedia content creation. This paper presents an AI-based Text-to-Video Generation framework that converts natural language prompts into short animated video sequences using diffusion and deep learning techniques. The proposed framework employs the Stable Diffusion v1.5 model for generating high-quality image frames and AnimateDiff for introducing smooth temporal motion while preserving scene consistency.***
The DDIM scheduler is utilized to accelerate the denoising process and improve inference efficiency without compromising output quality. The generated frames are sequentially processed and assembled into MP4 videos using FFmpeg. The proposed system is implemented using Python, PyTorch, Hugging Face Diffusers, and Google Colab, providing a lightweight and cost-effective platform for video generation. Experimental evaluation demonstrates that the proposed framework generates visually realistic and semantically meaningful videos with improved motion continuity and reduced computational complexity compared to conventional frame-by-frame generation approaches. The proposed framework has potential applications in digital media, education, entertainment, advertising, animation, and content creation.
Keywords: Text-to-Video Generation, Stable Diffusion, AnimateDiff, Diffusion Models, Deep Learning, Generative AI, DDIM Scheduler, Motion Consistency.

I. INTRODUCTION

Artificial Intelligence (AI) has significantly transformed digital content creation by enabling machines to generate text, images, audio, and videos with minimal human intervention. Among these applications, text-to-video generation has emerged as a promising research area because it enables users to create visual content directly from natural language descriptions. The rapid growth of social media, online education, digital marketing, and entertainment platforms has increased the demand for automated video generation systems that reduce the time, cost, and expertise required for traditional video production.

Recent advances in deep learning and diffusion models have substantially improved the quality of AI-generated visual content. Stable Diffusion has demonstrated remarkable success in generating high-quality images from textual prompts, while motion-aware frameworks such as AnimateDiff extend these capabilities by introducing temporal consistency and smooth motion across consecutive frames. Despite these advancements, generating coherent videos remains challenging due to issues such as object inconsistency, flickering effects, scene discontinuity, and high computational requirements.

To address these limitations, this research proposes an AI-Based Text-to-Video Generation System that integrates Stable Diffusion v1.5, AnimateDiff, and the DDIM Scheduler to generate short video sequences from natural language prompts. Stable Diffusion is employed to synthesize high-quality visual frames, while AnimateDiff maintains motion continuity and temporal coherence throughout the generated sequence. The DDIM Scheduler accelerates the diffusion process, reducing inference time without compromising output quality. Finally, FFmpeg combines the generated frames into an MP4 video.

The proposed framework is implemented using Python, PyTorch, Hugging Face Diffusers, and Google Colab, providing an accessible cloud-based solution for AI-driven video generation. The system aims to minimize manual video editing, improve motion consistency, and generate visually meaningful videos suitable for applications such as digital content creation, education, storytelling, marketing, and entertainment. Experimental results demonstrate that the proposed framework effectively converts textual descriptions into coherent video clips while maintaining visual quality and computational efficiency.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

II. LITERATURE SURVEY

Generative Artificial Intelligence has significantly advanced multimedia content creation through diffusion models and deep learning techniques. Recent studies have demonstrated remarkable improvements in text-to-image and text-to-video generation by enhancing visual quality, semantic alignment, and motion consistency. However, challenges such as temporal inconsistency, computational complexity, and long-duration video generation remain active research problems.

Rombach et al. [1] introduced Latent Diffusion Models (LDMs) for high-quality image synthesis by performing diffusion in a compressed latent space instead of the pixel space. The proposed approach significantly reduced computational complexity while maintaining image quality and became the foundation of Stable Diffusion. However, the framework focuses only on image generation and does not address temporal consistency or motion generation required for video synthesis.

Guo et al. [2] proposed AnimateDiff, a framework that transforms pre-trained text-to-image diffusion models into text-to-video generators by introducing lightweight motion modules. The framework improves temporal consistency and smooth motion without retraining the entire diffusion model. However, generating long-duration and high-resolution videos still requires substantial GPU resources.

Ho et al. [3] presented Video Diffusion Models, extending diffusion models from image generation to video synthesis by jointly modeling spatial and temporal information. The framework generates visually coherent videos with realistic motion through iterative denoising. Nevertheless, the approach requires large-scale video datasets and high computational resources, limiting practical deployment.

Ho et al. [4] developed Imagen Video, a cascaded diffusion framework capable of generating high-definition videos from textual descriptions. Progressive super-resolution diffusion models improve video realism and semantic alignment with text prompts. Despite its impressive performance, the model has high computational cost and requires extensive training resources.

Singer et al. [5] introduced Make-A-Video, a text-to-video generation framework that eliminates the need for paired text-video datasets by transferring knowledge from text-image models. The system generates visually meaningful videos with good motion quality using temporal modeling. However, the generated videos are relatively short, and maintaining temporal consistency over longer sequences remains challenging.

Hong et al. [6] proposed CogVideo, a transformer-based large-scale text-to-video generation framework that combines language understanding with video synthesis. The model produces semantically rich videos that closely match textual descriptions. However, the framework requires extensive computational resources, large-scale pretraining, and exhibits relatively slow inference speed.

Overall, existing research demonstrates significant progress in diffusion-based text-to-video generation. However, many approaches suffer from high computational requirements, limited temporal consistency, and difficulty generating long coherent videos. To address these limitations, the proposed work integrates Stable Diffusion v1.5, AnimateDiff, and the DDIM Scheduler to generate temporally consistent short videos with improved computational efficiency in a cloud-based environment.

III. METHODOLOGY

A. Research Design

The proposed study follows an experimental research methodology to design and evaluate an AI-based text-to-video generation framework. The methodology focuses on converting a natural language prompt into a short video sequence by integrating text understanding, diffusion-based visual generation, motion adaptation, denoising, frame sequencing, and video encoding. Since the objective of the work is to build a practical generation system rather than train a new model from the beginning, the framework uses pre-trained deep learning models and evaluates their output quality under different prompt conditions.

B. Development Environment and Tools

The system was implemented using Python as the primary programming language. PyTorch and Hugging Face Diffusers were used for loading and executing the generative models. Stable Diffusion v1.5 was selected as the base text-to-image diffusion model because it can generate semantically meaningful visual frames from textual descriptions. AnimateDiff was integrated to extend the image generation capability into video generation by adding motion information across frames. A Motion Adapter module was used to improve motion continuity and reduce abrupt frame-to-frame changes. The DDIM scheduler was used during the denoising process to reduce inference time while maintaining acceptable visual quality. FFmpeg was used to combine the generated frames and export the final output in MP4 format. Experiments were executed in Google Colab using GPU support to reduce local hardware dependency.

C. Proposed Workflow

The methodology begins with the user entering a positive text prompt that describes the required scene, objects, action, and background. A negative prompt is also provided to avoid unwanted features such as blur, distortion, poor quality, deformation, and flickering. The input prompt is passed through the text encoder of the diffusion pipeline, where the semantic meaning of the prompt is converted into conditioning information for visual generation. After prompt conditioning, random latent noise is initialized and gradually refined through the diffusion denoising process.

Stable Diffusion v1.5 generates the visual representation of the prompt in latent space, while AnimateDiff and the Motion Adapter guide the generation process so that consecutive frames maintain similar object appearance, background continuity, and natural motion. The DDIM scheduler controls the iterative denoising steps and helps generate frames more efficiently than slower sampling strategies. The generated frames are arranged in sequence according to a fixed frame rate. Finally, FFmpeg encodes these frames into a playable MP4 video. The user can review the output and refine the prompt or parameters if the generated video is not satisfactory.

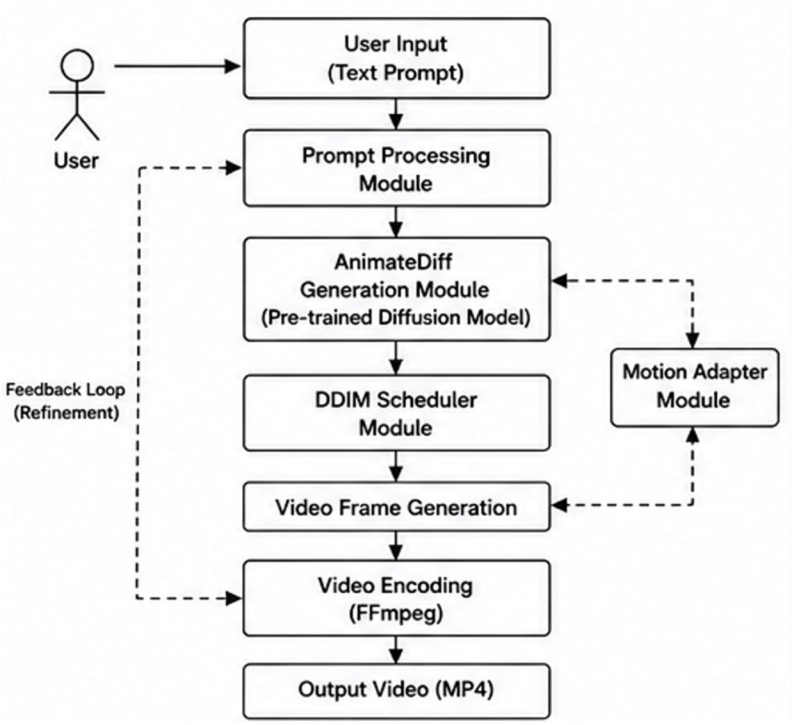


Fig. 1. System Architecture

D. Algorithmic Procedure

The overall procedure used in the proposed system is as follows:

Step 1: Accept the user text prompt and optional negative prompt.
Step 2: Preprocess the prompt and convert it into model-readable text embeddings.
Step 3: Load the Stable Diffusion v1.5 pipeline with AnimateDiff and the Motion Adapter.
Step 4: Initialize latent noise and apply DDIM-based iterative denoising.
Step 5: Generate temporally connected video frames based on the prompt.
Step 6: Arrange the frames sequentially and encode them using FFmpeg.
Step 7: Display the generated video and evaluate the output quality.

E. Evaluation Methodology

To evaluate the proposed framework, multiple prompts from different categories such as human activity, animals, urban scenes, and animated objects were tested. The generated videos were analyzed using both qualitative and quantitative criteria. Qualitative evaluation focused on whether the generated scene visually matched the input prompt, whether the object appearance remained stable, and whether motion appeared smooth across frames.

Quantitative evaluation was performed using Prompt Alignment Score, Temporal Consistency Score, Motion Smoothness Score, Frame Coherence Score, Video Generation Success Rate, and Average Video Generation Time. These metrics were selected because they directly measure the major requirements of a text-to-video generation system, namely semantic relevance, temporal stability, motion continuity, visual coherence, reliability, and computational efficiency.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed AI-Based Text-to-Video Generation framework was evaluated using multiple text prompts representing different scenarios such as human activities, animals, urban environments, and animated objects. The system integrates Stable Diffusion v1.5, AnimateDiff, the Motion Adapter, and the DDIM Scheduler to generate short videos from natural language descriptions. Experiments were conducted using Google Colab GPU, and the generated videos were analyzed based on semantic relevance, temporal consistency, motion smoothness, and visual coherence.

A. Generated Results

A Gradio-based user interface was developed to allow users to enter text prompts and generate videos interactively, as shown in Fig. 7. The interface accepts both positive and negative prompts and provides configurable inference parameters.

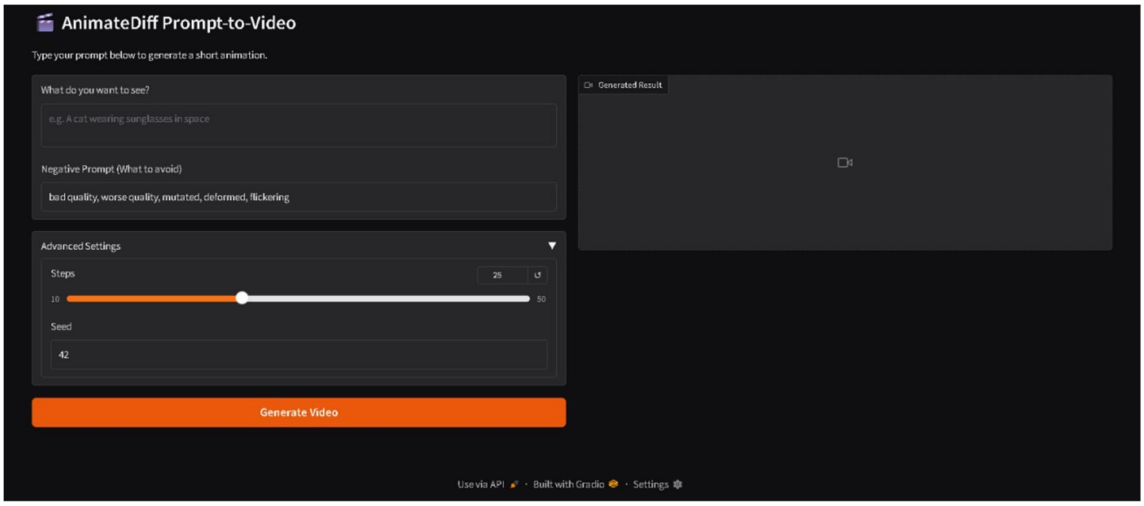


Fig. 2. User Interface of the Proposed System

The proposed framework successfully generated videos for various prompts such as "Cat Running on the Wall," "Man Walking on the Street," and "Robot Dancing with a Hat." The generated videos accurately reflected the input prompts while maintaining consistent object appearance, realistic backgrounds, and smooth motion across consecutive frames.

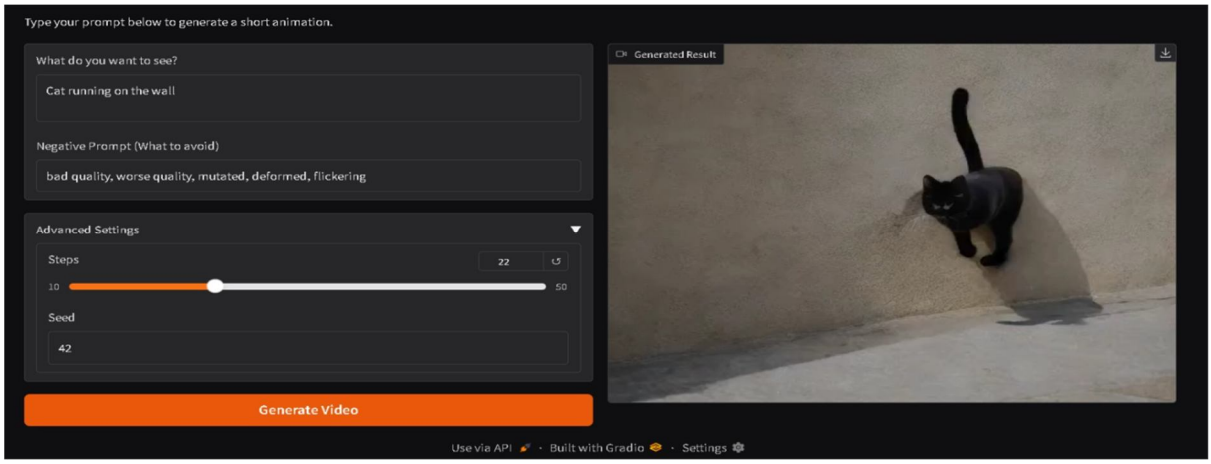


Fig. 3. Generated Output for "Cat Running on the Wall"

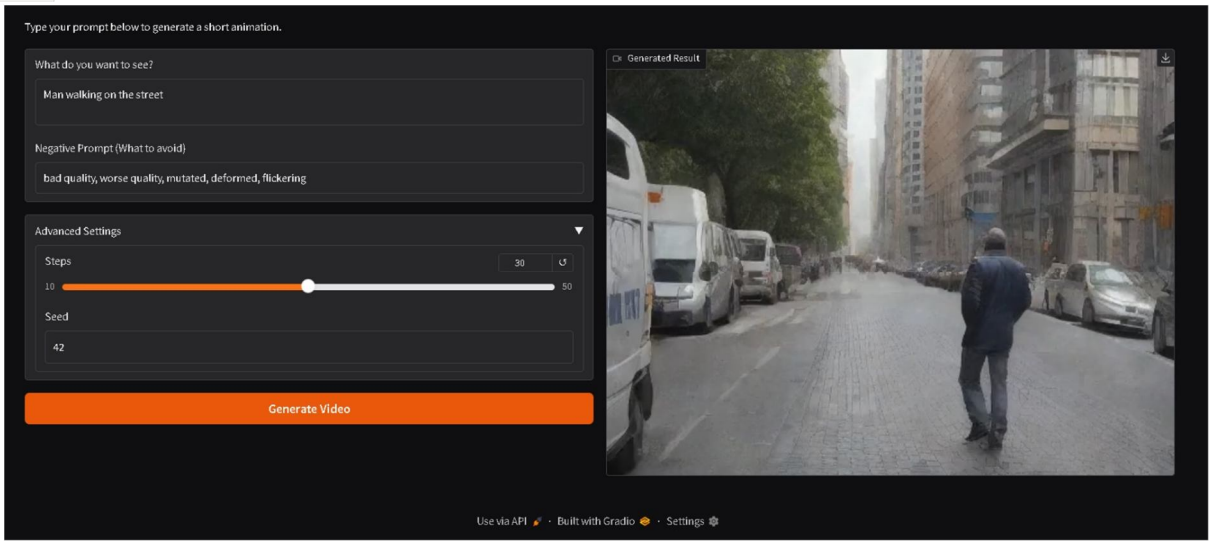


Fig. 4. Generated Output for "Man Walking on the Street"

B. Performance Evaluation

The performance of the proposed framework was evaluated using four quality metrics.

1) Prompt Alignment Score (PAS)

Prompt Alignment Score measures how accurately the generated video represents the semantic information contained in the input text.

$$PAS = \frac{N_c}{N_t} \times 100$$

where N_c represents the correctly generated semantic attributes and N_t denotes the total semantic attributes in the prompt.

2) Temporal Consistency Score (TCS)

Temporal Consistency measures visual continuity between consecutive frames.

$$TCS = \frac{1}{N-1} \sum_{i=1}^{N-1} S(F_i, F_{i+1})$$

Higher values indicate smoother transitions and reduced frame flickering.

3) Motion Smoothness Score (MSS)

Motion Smoothness evaluates the continuity of object movement across generated frames.

$$MSS = \frac{1}{N-1} \sum_{i=1}^{N-1} M_i$$

4) Frame Coherence Score (FCS)

Frame Coherence measures the consistency of objects, lighting, and background throughout the video.

$$FCS = \frac{1}{N} \sum_{i=1}^{N} C_i$$

5) Quantitative Results

Evaluation Metric	**Value**
Prompt Alignment Score	**92.6 %**
Temporal Consistency Score	**87.8 %**
Motion Smoothness Score	**87.9 %**
Frame Coherence Score	**90.8 %**
Video Generation Success Rate	**100 %**
Average Video Generation Time	**41.3 s**

Table 1. Quantitative Performance Evaluation

C. Discussion

The experimental results demonstrate that the proposed framework effectively converts natural language prompts into coherent video sequences. The integration of Stable Diffusion v1.5, AnimateDiff, and the Motion Adapter significantly improves temporal consistency, motion continuity, and semantic alignment compared with conventional frame-by-frame generation methods. The DDIM Scheduler further reduces inference time while maintaining high visual quality.

The proposed system achieved a Prompt Alignment Score of 92.6%, Frame Coherence Score of 90.8%, and Temporal Consistency Score of 87.8%, indicating strong semantic relevance and stable visual continuity. Although minor artifacts may appear during highly complex motion sequences, the framework provides an efficient and practical solution for automated text-to-video generation. The proposed approach is well suited for applications in multimedia content creation, education, digital storytelling, and AI-assisted video production.

V. CONCLUSION

This paper presented an AI-Based Text-to-Video Generation framework using Stable Diffusion v1.5, AnimateDiff, the Motion Adapter, and the DDIM Scheduler to generate short videos from natural language prompts. The proposed system successfully produced visually coherent videos with satisfactory semantic alignment, motion continuity, and temporal consistency. Experimental results demonstrated strong performance in prompt relevance, frame coherence, and motion smoothness while enabling efficient video generation on a cloud-based platform. Overall, the proposed framework provides an effective solution for automated text-to-video generation and has potential applications in education, digital content creation, storytelling, and multimedia production.

VI. FUTURE WORK

The proposed framework can be extended by supporting high-resolution and longer-duration video generation, improving character consistency and temporal coherence, integrating text-to-speech and background audio, and adopting advanced video diffusion models such as Stable Video Diffusion or Open-Sora. Future research may also focus on real-time video generation and deployment as a web or mobile application for broader accessibility.

REFERENCES

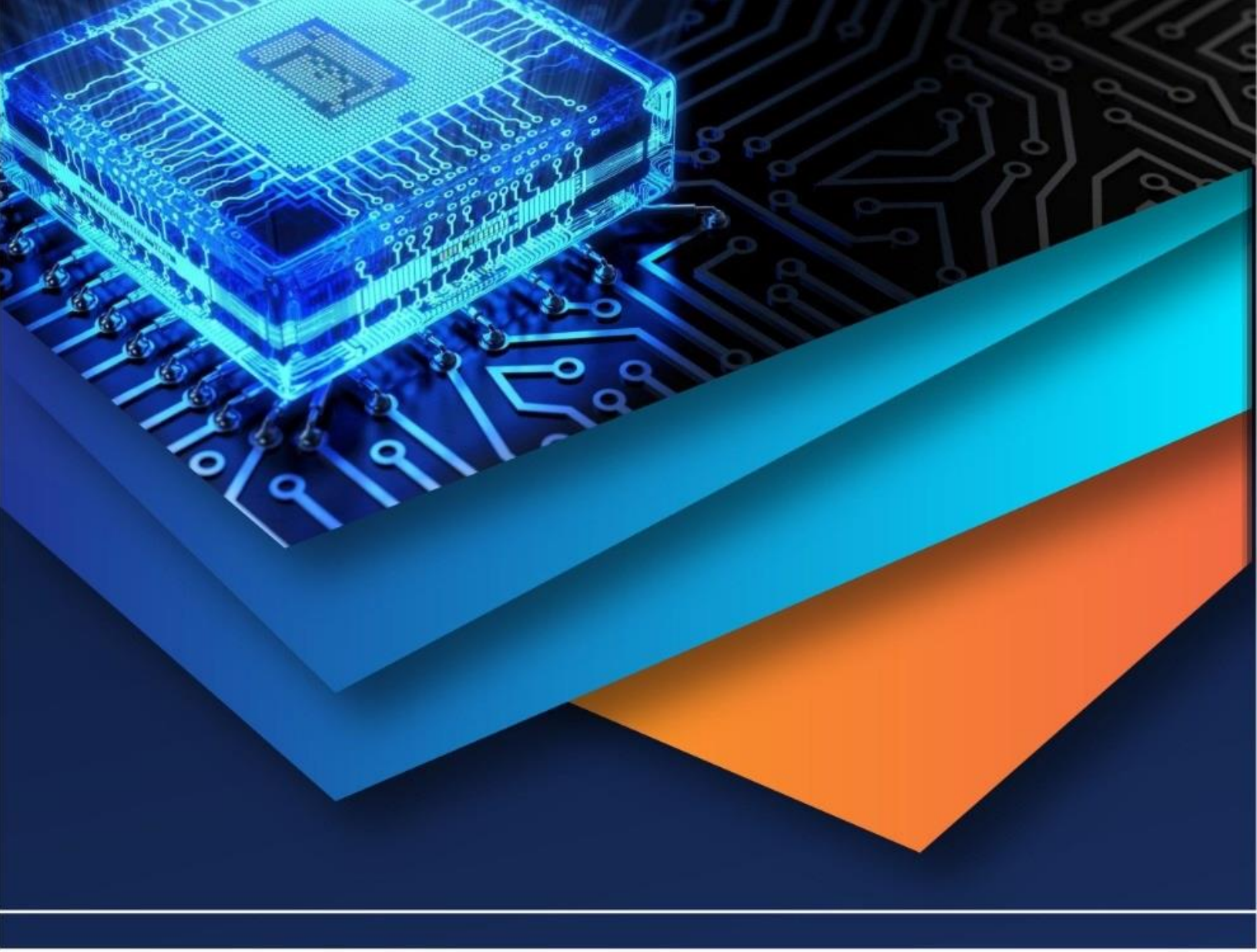
[1] T. Lee, S. Kwon, and T. Kim, “Grid Diffusion Models for Text-to-Video Generation,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 8734–8743.

[2] P. Esser, R. Rombach, and B. Ommer, “Diffusion Models for Video Generation,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2023.

[3] R. Rombach, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 10684–10695.

[4] J. An, S. Zhang, H. Yang, S. Gupta, J. Huang, J. Luo, and X. Yin, “Latent-Shift: Latent Diffusion with Temporal Shift for Efficient Text-to-Video Generation,” arXiv preprint arXiv:2304.08477, 2023.

[5] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, “Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021, pp. 1728–1738.
[6] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreiss, M. Aittala, T. Aila, S. Laine, and B. Catanzaro, “eDiffi: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers,” arXiv preprint arXiv:2211.01324, 2022.
[7] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, “Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models,” arXiv preprint arXiv:2304.08818, 2023.
[8] T. Brooks, A. Holynski, and A. A. Efros, “InstructPix2Pix: Learning to Follow Image Editing Instructions,” Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023.
[9] D. Ceylan, C.-H. P. Huang, and N. J. Mitra, “Pix2Video: Video Editing Using Image Diffusion,” arXiv preprint arXiv:2303.12688, 2023.
[10] H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang, C. Weng, and Y. Shan, “VideoCrafter1: Open Diffusion Models for High-Quality Video Generation,” arXiv preprint, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY