AI-BASED IMAGE TO VIDEO GENERATION SYSTEM

BY

EMENIKE DANIEL

REG. NO: GOU/U22/CSC/766

PROJECT REPORT PRESENTED TO THE DEPARTMENT OF COMPUTER SCIENCE AND MATHEMATICS, FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY, GODFREY OKOYE UNIVERSITY, ENUGU STATE

AS PARTIAL SATISFACTION OF THE REQUIREMENTS FOR GRANT OF THE DEGREE OF BACHELOR OF SCIENCE IN COMPUTER SCIENCE

SUPERVISOR:

ENGR DR. CHIBUEZE KINGSLEY

MAY, 2026

APPROVAL PAGE

This research, titled "AI BASED IMAGE TO VIDEO GENERATION SYSTEM," has been assessed and approved by the Department of Computer Science and Mathematics Committees in the Faculty of Computing and Information Technology, Godfrey Okoye University, Enugu, Nigeria.

Signature Date

Engr Dr chibueze Kingsley ....

Supervisor

External Examiner ....

External Examiner

Dr. Frank Okebanama ....

Sub-dean/HOD,

Department of Computer Science and Mathematics

Prof. I. A. Ajah ....

Dean, Faculty of Computing and Information Technology

CERTIFICATION PAGE

I certify that I completed the research included herein and that it was primarily based on my observations and investigation. Consequently, I will fully accept the University's decision if I am found to have violated academic integrity in any form.

_____

EMENIKE DANIEL

REG NO: GOU/U22/CSC/766

DEDICATION PAGE

This project is dedicated to my parents, Mr and Mrs Emenike, whose unwavering sacrifices, prayers, and encouragement made this academic journey possible, and to every Nigerian network professional working tirelessly to ensure that our network doesn't get congested and is secured

LIST OF TABLES

Table

LIST OF FIGURE

ACKNOWLEDGEMENTS

All glory and honor to god almighty for granting me the wisdom, health, and perseverance to finish this research. my faith in him has been my strength along the way.

the supervision by my supervisor, Engr Dr. Chibueze Kingsley, Dean, Faculty of Computing and Information Technology, Prof. I. A. Ajah and HOD, Dr. Frank Okebanama for their inception to its submission has been a great inspiration; his patience, expertise, and dedication are much appreciated. his careful feedback and academic guidance were priceless, and i am lucky to have had a dedicated supervisor.

i must express my gratitude to all the teaching staff of the Department of Computer Science and Mathematics, Godfrey Okoye University, for the good education they rendered to me and the passion for computing they gave me.

to my parents, i can't say how grateful 1 am. every page of this work reflects your prayers and sacrifices and your unconditional love. this accomplishment is mine as well. i am particularly grateful to my siblings. i can't thank you enough for your encouragement, understanding, and steadfast support during the toughest times of this research. you were there to support me through the times of doubt and made me feel confident to move forward. i am blessed to have you in my life.

thank you to all my friends and course mates for all the hardships, humour, and encouragement over the course of this programme. this trip was an unforgettable one.

Abstract

In this paper, the use of artificial intelligence in the process of generating video from images was investigated using deep learning algorithms pre-trained by machine learning algorithms. The study aimed at addressing the challenges associated with video generation systems, specifically high computation cost and inconsistent motion in the videos through the development of a hybrid 3D U-Net and diffusion models. The use of a dataset named Panda-70M that comprised video-text clips was done after the frames were extracted, resized, normalized, and filtered. In addition, transfer learning and temporal attention mechanism techniques were used to reduce computational cost and improve consistency of the generated video motion. The performance of the model was analyzed using SSIM (Identity) score of 0.6663, PSNR score of 14.98 dB, Motion Intensity score of 16.87, and Temporal Consistency of 0.0561.The system was able to obtain a mean video generation time of 118.27 seconds per sequence with a frame rate of 0.0676 FPS, indicating that the pretrained diffusion model was able to generate visual video sequences with motion consistency and structural preservation. React Native was utilized for the frontend interface, while Flask and FastAPI were used to implement the backend server handling the communication between the user interface and AI model components. Images along with text prompts can be uploaded by the user to generate motions, and the generated videos can be downloaded automatically by the user. In general, the experiment illustrated the applicability of the pretrained diffusion models for generating videos from images.

CHAPTER ONE
INTRODUCTION

1.1 Background of the study

Introduction

Artificial Intelligence (AI) is an area of Computer Science, which entails the development of systems that are able to execute tasks normally done by humans (Collins et al., 2021). For a long time now, there have been many changes in the functioning and application of AI where there are systems that not only perform specific tasks, but can learn, recognize visual elements and even generate new information and have developed explanation abilities. According to a research conducted by Lungu-Stan & Mocanu (2024), this change in AI has greatly impacted the multimedia sector where besides analyzing images, systems are now able to create real visual outcomes based on the training received from previous data (Akber et al., 2023).

The main impetus for such progress is Deep Learning (DL), which is one of the subcategories of AI that applies multilayered neural networks for pattern extraction from huge amounts of data. Nowadays, DL is getting more and more essential in such fields as image processing, movement prediction and content creation (Huang et al., 2024). Thanks to models like convolutional neural networks (CNN), transformers and diffusion architecture, machines can create animations and synthesis of movements in order to provide rich media output (Lungu-Stan & Mocanu, 2024).

The DL has been a very useful tool over the years when performing tasks such as image processing, where they extract patterns, objects, and descriptions or modify images to get new ones, while in natural language processing, they come up with powerful translators as well as programs that can read our intent from the messages. In addition, they create tools that can write articles on their own or articles on reinforcement learning, where the models have beaten the best human players in different games (Lungu-Stan & Mocanu, 2024). The achievements of deep learning have been surprising in many ways over the years, where we have seen DL and

Computer Vision (CV) used currently in the medical field for brain tumours detection, in Agriculture for pests and weed detection and classification, in transport for accident prediction and unauthorized parking detection, and many other areas of applications.

Another field closely associated with DL is Computer Vision (CV), and it is about making computers capable of interpreting and comprehending visual data (Cai et al., 2020). In its early stages, the concept of CV was concentrated on tasks like object detection and image categorization; however, today it has become capable of performing such operations as motion estimation, pose recognition, and even visual content generation. These abilities of the computer vision approach are critical in the context of animation systems as the comprehension of motions and structures is vital for realistic output creation (Akber et al., 2023).

The second concept that should be mentioned is Generative AI that presupposes the use of systems that can generate new content regarding images, sounds, and videos. Contrary to the previous generations of AI technologies that could only classify and predict some data, the new models will be capable of recognizing the main characteristics of the input data in order to generate realistic outputs (Lungu-Stan & Mocanu, 2024). Technologies such as GANs and diffusion models have been developed to improve the quality of generated media and allowed creating very realistic animations and visual effects as it was stated in the research conducted by Dhariwal and Nichol (2021).

The most intriguing application of generative AI is the generation of images into videos, which is the creation of a sequence of frames representing motion using a single static image. The idea will inevitably incorporate visual features from the initial image with previously learned motion patterns in order to produce content that is dynamic. Currently, this approach has gained

attention because it allows users to create animated outputs without the need for manual frame-by-frame design, which is the old standard practice (Akber et al., 2023).

The need for such systems arises from the limitations of static images. While images are effective for representation, they lack motion and temporal dynamics, which are essential for storytelling and user engagement. In contrast, video content is more expressive and interactive, making it more suitable for communication in modern digital environments. As a result, there has been increasing interest in automated systems that can generate videos from images efficiently (Lungu-Stan & Mocanu, 2024). Image-to-video generation systems have a wide range of applications, such as in the media and entertainment industry, where we can use them to create animations, special effects, and virtual characters (Ruiz et al., 2022). In education, they can help us to convert static teaching materials into dynamic visual explanations. Similarly, in marketing and advertising, they can be used to generate engaging promotional content, which studies have shown to enable generating interest as well as attract customers. These applications indicate the growing importance of automated video generation technologies, thereby stressing the significance of my study, which aims to design and implement an AI-based image-to-video generation system using pre-trained models.

1.2 Problem Statement

The traditional method of producing videos and video content is challenged by the fact that it requires significant time, financial resources, and specialised skills. And professional animators and video editors most of the time need to manually design and refine motion sequences, which makes the process labour-intensive and expensive.

Despite the global growth and advancement of AI, static images like photographs remain limited in conveying motion and interaction, as they do not effectively capture temporal changes, reducing their impact in many applications that require dynamic visual representation.

Some automated solutions that have been developed still face important limitations, such as high computational requirements, dependence on large datasets, and difficulties in handling diverse or unseen motion patterns.

Another critical issue is accessibility, as many existing AI-based video generation systems require powerful hardware or specialised environments, making them difficult for the common man to acquire or use in their daily designs or video content creation.

In summary, although there have been many developments already made in this field, the idea of creating videos from still images is quite a challenging one that requires much research, and this paper aims at developing a system that will be able to create videos from images using pretrained models in a computationally feasible manner.

1.3 Aim and Objective of the Study

The aim of this project, titled “AI-based Image-to-Video”, is to design and implement an AI-based image-to-video generation system using pre-trained models. The specific objectives are to:

1. Study existing AI-based image-to-video generation techniques
2. Collect and preprocess datasets for image-to-video generation
3. Implement a pre-trained AI model for motion generation
4. Test and evaluate the performance of the system
5. Develop a user-friendly web application interface for real-time validation

1.5 Significance of the Study

This study is important because it addresses the challenges associated with traditional video production by automating the process of generating video content from images, such that the system can significantly reduce the time and effort required for video and animation creation and also make it easy to accomplish without the need for the user to be a professional in it. This system, when designed and implemented, will benefit a lot of people, especially stakeholders in the government, media and entertainment sectors, and the education sector.

The system will also benefit people in the media and entertainment industry, where there is a constant demand for engaging visual content. And in the education sector, the system can be used to enhance learning by converting static materials into dynamic visual presentations, as it can improve understanding and engagement among students. The system also has practical applications in marketing and advertising, where dynamic content is more effective in capturing audience attention.

Overall, the study helps make AI-driven video generation more accessible, enabling users with limited technical skills or resources to create video content efficiently and effectively from anywhere in the globe as long as there is an available internet connection.

1.6 Scope of the Study

The purpose of this research is to build an application which will be capable of producing videos from the images by making use of already developed AI framework. The proposed solution will involve designing a graphical user interface to facilitate the process of conversion.

1.7 Limitations of the Study

And the study is limited to using existing pre-trained models rather than developing new models from scratch, as it will be cost-intensive to train or develop new models. And the study will only use a secondary dataset already available on online repositories like Kaggle.

1.7 Definition of Terms

1. Image-to-Video Generation

   This is the process of converting a single image into a sequence of moving frames using AI models.

2. Pre-trained Model

   They are models already trained on large datasets, for example, on object detection or image generation, and they are ready for reuse.

3. Generative AI

   They are AI systems that create new content from learned patterns.

4. Diffusion Model

   This is a model that refines the dataset noise to generate data.

5. Latent Space

   This is a compressed representation where features are encoded.

6. Temporal Consistency

   Temporal consistency refers to smooth and coherent motion across video frames.

7. Frame Synthesis

   This is the act of creating new frames between existing ones from video data.

8. Conditioning Input

   In machine learning, conditioning input is guidance provided to control the model output.

9. Image Embedding

   This involves numerically representing an image's features.

10. Fine-tuning

    This is the process of adjusting a pre-trained model for a specific task until a certain accuracy or confidence level is attained.

11. Inference

    This is a way machine learning engineers use a trained model to generate output.

12. Video Frame Rate

    It is the number of frames displayed per second in a video.

13. Motion Dynamics

    Motion dynamics explains how movement is represented in the generated video.

14. Resolution

    The level of detail in an image or video.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This section provides an overview of research on AI-based image and video generation, covering key areas such as AI video generation, image animation, diffusion models, talking face generation and generative models. It reviews and summarises the existing literature on the project topic and establishes a research gap.

2.2 Theoretical Background

2.2.1 Artificial Intelligence in Multimedia Generation

Artificial intelligence has become a core engine in modern multimedia generation, especially in the creation and transformation of both images and video content, as well as the optimisation of photos and videos. In recent years, generative AI systems, for example, those based on deep learning, have brought about increasing realistic and controllable synthesis of visual media, for instance, the conversion of static images to dynamic video sequences (Vondrick et al., 2023; Nichol & Dhariwal, 2021). In image generation, diffusion-based models are currently surpassing earlier approaches such as Generative Adversarial Networks (GANs) in a major way by leveraging a reverse denoising process in order to produce visually accurate images from latent noise while offering better sample diversity and training stability, which is a major benefit to the media and entertainment sector globally and in Nigeria (Song & Ermon, 2020; Dhariwal & Nichol, 2021). For video generation, deep learning architectures such as video GANs and diffusion-based spatiotemporal models have been modified and also adapted to capture motion clearly across frame and they do this by often learning joint spatial and temporal patterns from large video datasets they were trained on (Clark et al., 2022; Ho et al., 2022). These advances

demonstrate that emerging image-to-video systems, where an input image is combined with motion patterns learned from diffusion or hybrid GAN-diffusion frameworks, will be able to synthesise short realistic video clips which will be able to maintain structural consistency along with the original frame (Wang et al., 2023; Zhang et al., 2024). Overall, AI-driven multimedia generation now brings together GANs, diffusion models, and transformer-style architectures to build a powerful but flexible tool that will automatically create visual content that is suitable for applications ranging from digital media production to human–computer interaction.

2.2.2 Image-to-Video Generation Techniques

Image-to-video generation has evolved around three core families of techniques, such as motion transfer, frame synthesis, and video prediction, where each of these techniques plays a distinct role in turning a static input image into a coherent short video sequence. One of the families which is motion-transfer methods typically maps motion patterns from a source video or pose sequence onto a target image, often by extracting skeletal poses, optical flow, or dense trajectories from the driving clip, where it will now re-render the target under the specified motion fields using convolutional or diffusion-based generators (Chan et al., 2019; Li et al., 2021). Frame-synthesis approaches, a core family of image-to-video generation, on the other hand, focus on generating a sequence of new frames from the image key into the model input, either by moving smoothly from one representation to another in a learned video manifold or by iteratively denoising frames in a diffusion-based pipeline while preserving the original image's structure and appearance (Yang et al., 2021; Wang et al., 2023). Video prediction–style techniques extend this further by modelling how future frames are expected to evolve from the initial image, using temporal models such as recurrent networks, transformers, or spatiotemporal diffusion models to forecast plausible motion, whereas the scene is semantically kept consistent.

(Vondrick & Torralba, 2018; TextOCVP, 2025). Overall, these methods help control the output better, improve quality and will also help keep the video smooth over time, thereby forming the backbone of modern AI-based image-to-video systems.

2.2.3 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) provide a foundational architecture for data-driven image and video generation through a two-network setup, which consists of an adversarial setup comprising a generator and a discriminator trained in competition with one another (Goodfellow et al., 2014; Mirza & Osindero, 2014). The generator functions by learning how to map random noise or conditioned inputs into artificially created data that looks closely like real data patterns, while the discriminator is trained to classify whether each sample is real or generated, thereby providing a training signal that will progressively push the generator to produce outputs that are accurate with high precision (Goodfellow et al., 2014; Creswell et al., 2018). This step-by-step optimisation process, in which one part minimises, and the other maximises, is often stabilised using techniques such as gradient-penalty regularisation and spectral normalisation, in order to enable GAN-based systems to achieve the production of very clear and realistic images and videos, with smooth movement over time (Karras et al., 2019; Saito et al., 2017). In video generation, GANs have been adapted through architectures such as Video-GAN (VGAN) and its successors, which uses spatio-temporal generators to produce a series of images and multi-scale discriminators in order to assess both spatial quality and temporal consistency, thereby enabling applications such as video prediction, frame interpolation, and motion-aware video synthesis (Saito et al., 2017; Clark et al., 2019; Wang et al., 2020). A recent work that was reviewed combined GAN-style adversarial objectives with other generative paradigms, for example, diffusion, transformer-based models, in order to refine motion smoothness and visual fidelity,

and this demonstrates GANs' enduring relevance as a core component in an end-to-end AI-based process of turning an image into a video (Yang et al., 2021; Zhou et al., 2023).

2.2.4 Diffusion Models

Recently, diffusion models have emerged as a powerful paradigm in deep learning that is utilized for generating high-quality and real media, which is done by gradually denoising a latent representation from pure noise toward a structured output, where they are guided by a learned score function or noise predictor (Ho et al., 2020; Song & Ermon, 2020). In the image-to-video context, diffusion models extend the standard image diffusion framework by adding temporal dynamics, either through spatiotemporal UNets, latent-space transformers, or cascaded super-resolution stages that jointly refine spatial detail and motion coherence across frames of the video (Ho et al., 2022; Singer et al., 2022).

From the studies reviewed, Ho et al. (2022) and Singer et al. (2022) developed modern video-diffusion systems that often condition the denoising process on an input image (or its latent encoding) and a motion or text prompt, in order to enable the synthesis of short, fine-grained videos that will preserve the original image's structure while at the same time introducing plausible camera or object motion. Recent work has also shown that pairing diffusion-based image prior networks with lightweight motion adapters, such as those used in adapter-style image-to-video pipelines, will, in a major way, enable efficient, high-fidelity video generation without full retraining of the base model (Hu et al., 2023; Ho et al., 2020). These findings and developments place diffusion models as a central building block in present-day AI-based image-to-video systems, especially when the video needs to look real, move smoothly, and allow us, the users, to control the motion (Esser et al., 2023; Ho et al., 2022).

2.2.5 Talking Face Generation Systems

Talking face generation systems represent a specialised branch of AI-based image-to-video technology that creates realistic talking-face videos from a single image, and will also match speech to lip movement. Modern approaches typically combine audio-driven animation with neural rendering or generative models in order to map speech signals onto facial motion, especially lip movements, while preserving the identity and expression of the original face (Zhou et al., 2020; Nirkin et al., 2021). Many current systems employ GAN-based architectures, diffusion models, or lightweight neural renderers that generate a sequence of frames conditioned on an input audio waveform, which most of the time uses a separate lip-sync module trained with audio-visual alignment loss in order to enforce temporal coherence between speech and mouth motion (Zhou et al., 2020). These models enable the creation of lifelike virtual avatars, AI presenters, and digital humans that can be used for applications such as automated dubbing, personalized video messages, and interactive virtual assistants. By the integration of a robust audio-driven animation with identity-preserving face generation, the talking face systems exemplify how AI can bridge image-to-video synthesis with real-time multimodal communication in practical, user-facing environments.

2.3 Review of related works

D'monte et al. (2022) developed a system that converts images and videos into cartoonized outputs using machine learning and image-processing tools, with a controllable, adjustable framework for animation-oriented applications. The study employed a hybrid image-processing approach that combines deep learning and a GAN, and the system was deployed as a web-based prototype/system using a FastAPI backend and a React frontend, with an emphasis on image and video cartoonization and near-real-time processing.

Jiang and Wang (2024) analysed DL-based methods for animation scene generation and data mining, thereby reducing the manual design workload and providing data support for animation production and market understanding. The study utilised multimodal data, such as images and texts, to train a hybrid DL, and the results reveal a series of high-quality scenes across natural, urban, and indoor styles, with the DL method showing higher and more stable accuracy, recall, and F1.

Ling et al. (2023) utilised multiple public datasets containing 200 videos to train a hybrid DL/GAN model to automatically generate articulatory animations in different styles for English pronunciation learning, reducing the need to redraw hand-made animations for each style. The deep learning pipeline combined keypoint detection, motion transfer, dense motion estimation, and style transfer in order to build this system. Results indicate that the system can auto-generate articulatory animations from phonetic symbols and thereby reduce manual labour by over 90%.

Shobayo and Saatchi (2025) conducted a review of recent deep learning models for medical image analysis and interpretation, with an emphasis on improving diagnostic accuracy, automation, efficiency, and clinical adoption. The work initially identified over 500 studies for the review, but only 141 were included in the final review. Findings from the study identify factors such as data availability, interpretability, overfitting, image annotation burden, noisy images, data-sharing/privacy complexity, ethical concerns, trust, and computational/environmental costs as key challenges affecting deep learning models for medical image analysis and interpretation.

Wang et al. (2025) conducted an editorial review that outlined the current advances, challenges, and research trends in deep learning for image preprocessing, feature extraction, image processing, pattern recognition, transportation, hyperspectral imaging, biomedical imaging,

surveillance, and reconstruction. The study's findings present key challenges: interpretability, computational cost, data security, privacy-aware sharing, cross-scale modelling, and balancing performance and efficiency.

Kumar and Somasundara (2024) made use of a dataset of driving videos, facial photos and pre-trained bird datasets to train a deep learning model called FOMM-based animation model to generate realistic live video from a single source image using GAN-based deepfake techniques and a first-order motion model. The paper states that the finished video is realistic and suitable for several applications, and that extensive testing and validation were conducted, but no numerical results are reported.

Jiang et al. (2024) review several advanced strategies, including audio–identity fusion, landmark deformation with texture generation, emotional integration, secondary editing, diffusion-based generation, and advanced generative network-based methods, to provide a comprehensive survey of deep learning techniques for audio-driven facial animation. The study identifies several challenges in DL image-to-video systems, such as disentangling lip synchronisation from emotion, generalisation across speakers, dataset limitations, and the need for improved emotional expressiveness and personalisation.

A comprehensive survey conducted by Valente et al. (2023) on the advances in AI design and optimization solutions for image processing, with emphasis on deep learning and reinforcement learning developments point out findings that indicate many challenges such as large and complex datasets, computational demands, scalability, memory limits on a single machine, class imbalance, sparse or heterogeneous data, and difficulty keeping pace with rapidly evolving research.

Ripote and Wankhade (2025) reported consistent accuracy in their study, which developed a deep learning-based real-time image animation system for AR-assisted maintenance that automates facial expression synthesis and object motion generation, thereby improving efficiency and precision while reducing manual intervention. The study made use of an Image-sequence pairs dataset to train a hybrid system that uses Haar Cascade with AdaBoost, facial landmark detection, affine transformations, GANs, and VAEs.

Akber et al. (2023) conducted a comprehensive comparative review of DL-based motion style transfer approaches, enabling technologies, motion datasets, and contemporary challenges in realistic human motion generation. The key challenges include a lack of quantitative evaluation standards, a lack of a unified benchmark, difficulties with style retargeting, the need for paired/registered data, difficulties with few-shot style transfer, training overheads, and handling unseen motions/styles.

Lungu-Stan and Mocanu (2024) designed a system that facilitates and hastens the creative process in game/animation production by reducing repetitive manual work in 3D character animation and asset generation using deep learning. They utilised the CMU motion capture dataset to train AnimGPT and DenoiseAnimGPT, and then applied an LSTM-like comparator architecture for ablation. The results after evaluation show that AnimGPT achieved an MAE of 0.345 at 50 frames in 1b1 mode, while DenoiseAnimGPT achieved 0.2513 at 50 frames. And for "all" frames, AnimGPT 1b1 MAE was 0.381 vs LSTM 0.538123, and DenoiseAnimGPT 1b1 MAE was 0.2798 vs LSTM 0.506419. The generated results were described as imperceptible to the untrained human eye.

Alaaran and Soudani et al. (2025) designed a system that utilised a merged dataset built from three dedicated earthquake datasets: MEDIC, Disaster Scenes Database, and Ünlü et al.'s dataset,

to train an efficient lightweight CNN system for real-time post-earthquake damage-level prediction from UAV imagery, supporting rapid SAR decision-making on embedded devices. The results from the findings indicate a weighted average F1-score = 0.93 and accuracy = 0.93. Class-wise F1: 0.85.

Table 2.1Summary of Literature Review

Author (Year)	**Technique**	**Work Done**	**Findings**	**Limitations**
D'monte et al. (2022)	Hybrid ML + GAN (Image Processing)	Developed a system for cartoonizing images and videos using deep learning and GANs, deployed as a web-based system with FastAPI and React, supporting near real-time processing.	Achieved near real-time image and video cartoonization through web deployment.	Output resolution was poor.
Jiang and Wang (2024)	Hybrid Deep Learning (Multimodal)	Proposed DL-based animation scene generation using multimodal data (images and text), producing high-quality scenes with improved performance metrics.	Generated high-quality scenes with improved accuracy, recall, and F1-score.	No exact numerical results reported; multimodal fusion requires improvement.
Ling et al. (2023)	Hybrid DL + GAN	Developed a system for articulatory animation generation using	Reduced manual	Keypoint extraction

		keypoint detection, motion transfer, and style transfer, reducing manual effort by over 90%.	animation effort by over 90% through automation.	inaccuracies; occasional need for manual intervention.
Shobayo and Saatchi (2025)	Deep Learning (Review Study)	Reviewed 141 studies on medical image analysis, highlighting improvements in diagnostic accuracy and automation using DL.	Deep learning improved diagnostic accuracy, automation, and clinical efficiency.	Challenges include data availability, interpretability, overfitting, privacy, and computational cost.
Wang et al. (2025)	Deep Learning (Review Study)	Provided an overview of DL applications in image processing, feature extraction, and pattern recognition across multiple domains.	Deep learning enhanced image processing and pattern recognition performance.	Issues include interpretability, computational cost, data security, and scalability.
Kumar and Somasundara (2024)	GAN + First-Order Motion Model	Developed a model for generating realistic video from a single image using deepfake techniques and motion models.	Generated realistic videos from a single source image	High computational cost; complex training and

			effectively.	fine-tuning.
Jiang et al. (2024)	Deep Learning (Audio-Driven Models)	Reviewed techniques for audio-driven facial animation including diffusion models and generative networks.	Identified advanced techniques improving audio-driven facial animation generation.	Challenges in lip-sync, emotion modelling, generalization, and dataset limitations.
Valente et al. (2023)	Deep Learning + Reinforcement Learning	Surveyed AI-based optimization techniques for image processing, highlighting scalability and performance improvements.	AI optimization improved image-processing scalability, efficiency, and performance.	Large datasets, high computational demand, class imbalance, and evolving research challenges.
Ripote and Wankhade (2025)	Hybrid DL (GAN + VAE + Haar Cascade)	Developed a real-time animation system for AR-assisted maintenance using facial expression synthesis and motion generation.	Achieved consistent accuracy while reducing manual intervention requirements.	Not explicitly stated but involves system complexity and computational demand.

Akber et al. (2023)	Deep Learning (Motion Style Transfer)	Conducted a comparative review of motion style transfer techniques and datasets for human motion generation.	Provided comprehensive insights into realistic human motion generation techniques.	Lack of benchmarks, difficulty in style transfer, high training cost, and handling unseen styles.
Lungu-Stan and Mocanu (2024)	Deep Learning (AnimGPT Models)	Developed AnimGPT models for 3D animation using motion capture datasets, achieving strong MAE performance.	Produced realistic animations with low MAE and strong visual quality.	Performance degradation in long sequences; high computational requirements.

2.4 Research Gap

Having reviewed extensive literature by various authors, we find that existing image-to-video generation systems are based on diffusion models that exhibit significant limitations, including high computational requirements, temporal inconsistencies, limited motion control, and complex implementation processes. These challenges hinder their practical deployment, particularly in real-time and user-friendly applications. Therefore, my research proposes a 3D U-Net-based diffusion architecture integrated with transfer learning and temporal attention mechanisms to

improve video quality, reduce computational overhead, and enhance accessibility through a user-centred interface.

CHAPTER THREE
SYSTEM ANALYSIS AND DESIGN

3.1 Introduction

Image-To-Video Generation System Using AI Models has started to become one of the most trending areas in artificial intelligence, and it has also proved to be useful for media and entertainment, and with the development of generative AI, these systems have become the subject matter of research for many researchers. In this chapter, we have talked about the technique of analyzing and designing a system that will be used by us to explore the use of deep learning models in image-to-video generation.

3.2 Research Methodology

The research paper adopts a mixed-methodology approach that will involve the use of experimental research design and data collection for the deep learning models while Object-Oriented Analysis and Design (OOAD) will be adopted for system development in order to develop the mobile application to test a 3D UNet Diffusion model in a mobile application software for generating video from static images as well as viewing the results of the actions. The experimental design will involve the process of training the model and evaluation of the model using suitable metrics. Moreover, the process of data pre-processing will be done in order to improve generalization and reduce overfitting on the dataset followed by dividing of the dataset into training, validation, and testing datasets. On the other hand, the system development process involves the application of OOAD concepts and UML for developing an appropriate user interface and API for integrating the model.

3.3 System Analysis

Analysis of the system is an analysis of the proposed system where analysis of the system components and their interactions takes place. The next part of the research paper discusses

system analysis of the proposed system which includes analysis of the existing systems, problems in those systems and the ways to improve them. Analysis of the proposed system would be done through use case diagrams, activity diagrams and class diagrams.

3.3.1 Analysis of the Existing System

The current system is a traditional approach to video creation which involves a lot of manual work like filming, editing, and animation, which consumes lots of time and needs lots of resources as well as a professional or expert who has skills in video editing and creation. Another thing is that recent approaches to video generation use high-computational resources such as GPU. Furthermore, the current approaches are mostly developed as research prototypes and do not represent user-friendly applications due to their difficult interface. Most existing approaches focus only on text-to-video generation and are nothing but animation pipeline with little attention to single image-to-video conversion. Besides, the outputs of the current system, at least most of them, display inconsistent motion and unrealistic transition between different frames. In general, most solutions that have been suggested cannot be accessed through mobile devices and some of them are even not accessible for free usage. Therefore, there is a necessity to automate these approaches and develop an approach to the video creation based on an AI technology such as an image-to-video generation system which will be designed using pretrained 3D UNet Diffusion model.

3.3.2 Weaknesses of Existing Systems

Although different systems currently exist in the area of this study, these systems have fallen short in some areas, which are pointed out below:

i. The existing system requires a high computational cost and mostly limits accessibility for average users

ii. They have poor temporal consistency, which most times causes them to produce flickering frames and unnatural motion

iii. The user interface is not user-friendly, thereby reducing its usability.

iv. The existing systems are not optimised for real-time or fast processing, and they have limited user control over motion or output styles.

v. There is dependence on cloud services, which introduces challenges such as latency issues and data privacy concerns.

vi. There is a weak integration between AI models and usable applications.

3.3.3 Analysis of the Proposed System

The proposed system employs a DL technique, where it utilises a pretrained, model named 3D U-Net-based diffusion architecture, which was designed and trained to improve video quality and frame coherence. The new system integrates a transfer learning approach in order to reduce training cost and improve performance using pretrained knowledge. In the same vein, a temporal attention mechanism was incorporated to ensure smooth motion and to enhance frame-to-frame consistency of the output of the designed model. The proposed system will provide a mobile-based interface for accessibility by integrating the newly trained model API into the design mobile application for communication between system components and to optimise computational efficiency.

3.3.4 Advantages of the Proposed System

i. The proposed system will have the capability to produce realistic and stable video outputs.

ii. It will reduce the computational load by adopting pretrained models.

iii. The new system will enhance user accessibility by implementing and presenting a mobile interface and simplified user interaction.

iv. The new system supports faster or near-real-time processing, and it improves user experience through our intuitive system design.

v. The proposed system will bridge the gap between advanced AI models and practical usage because most of the applications or studies on it have been conceptual.

vi. The new system allows future scalability, such as text and audio integration, and will also have the ability to produce higher resolution outputs.

3.4 System Design Methodology

The suggested method of development is Object-Oriented Analysis & Design (OOAD). OOAD method is used for the development of AI based Image-to-Video generation system. Real world objects and their interactions are used in modelling the structure and behavior of the system in the case of OOAD methodology. OOAD method increases the modularity, scalability and code reusability properties which are needed in the development of complex, data intensive and integrated AI systems. In order to develop a well-defined and user-friendly system, UML is used for illustration and OOP concepts for programming.

Phases of OOAD include:

i. Requirement Analysis: During this stage, the system features were obtained from the end users, which included the static image and the prompt, and the system requirements were identified through the gaps present in the existing system along with the expectations of the AI-based Image to Video Generation System.

ii. Design of System Architecture: During this stage, UML diagrams like use case diagram, class diagram, activity diagram, and sequence diagram are drawn to capture the static and dynamic behavior of the system. There are interactions between the users and the generative AI and diffusion model in these diagrams.

iii. System Implementation: The architecture of the project is implemented in object-oriented programming languages and relevant frameworks. The backend is programmed using the Python (Flask) language, where the 3D U-Net-based diffusion models will be implemented, whereas the UI part is developed using React Native technology.

iv. Testing: All modules of the system undergo individual and combined testing. All functions undergo unit testing, while modules undergo integration testing. Model accuracy and system speed are validated based on performance measures such as RMSE, MAE, and response time.

v. Deployment: After the testing process, the system is deployed on the live server and is available for access by the users via web or mobile applications. Deployment also consists of deploying the models and the services of the databases to the secure cloud environment.

vi. Maintenance: In relation to the previous step, additional maintenance ensures the continual updating of the system with new data, updates on the model, and most current security. The accuracy of predictions may be maintained at the optimum level through the regular retraining of the models.

3.4.1 Unified Modelling Language (UML) diagrams

The UML diagrams are used in order to understand, design and organize the architecture and interaction of the components of the system. With the help of the UML diagrams, the developer and other relevant parties can understand the communication structure as well as the complete process flow in the system. The AI-based Image-to-Video Generation System consists of the following components:

A. Use Case Diagram for the Proposed System

Use case diagram depicts the functional requirements of the system as well as the interaction between the actors and the process within the system. The key actors in this system include

Users and AI model. There are 8 use cases in this system including sign in, log in, upload an image and see the generated video. Figure 3.1 depicts the use case diagram of the system.

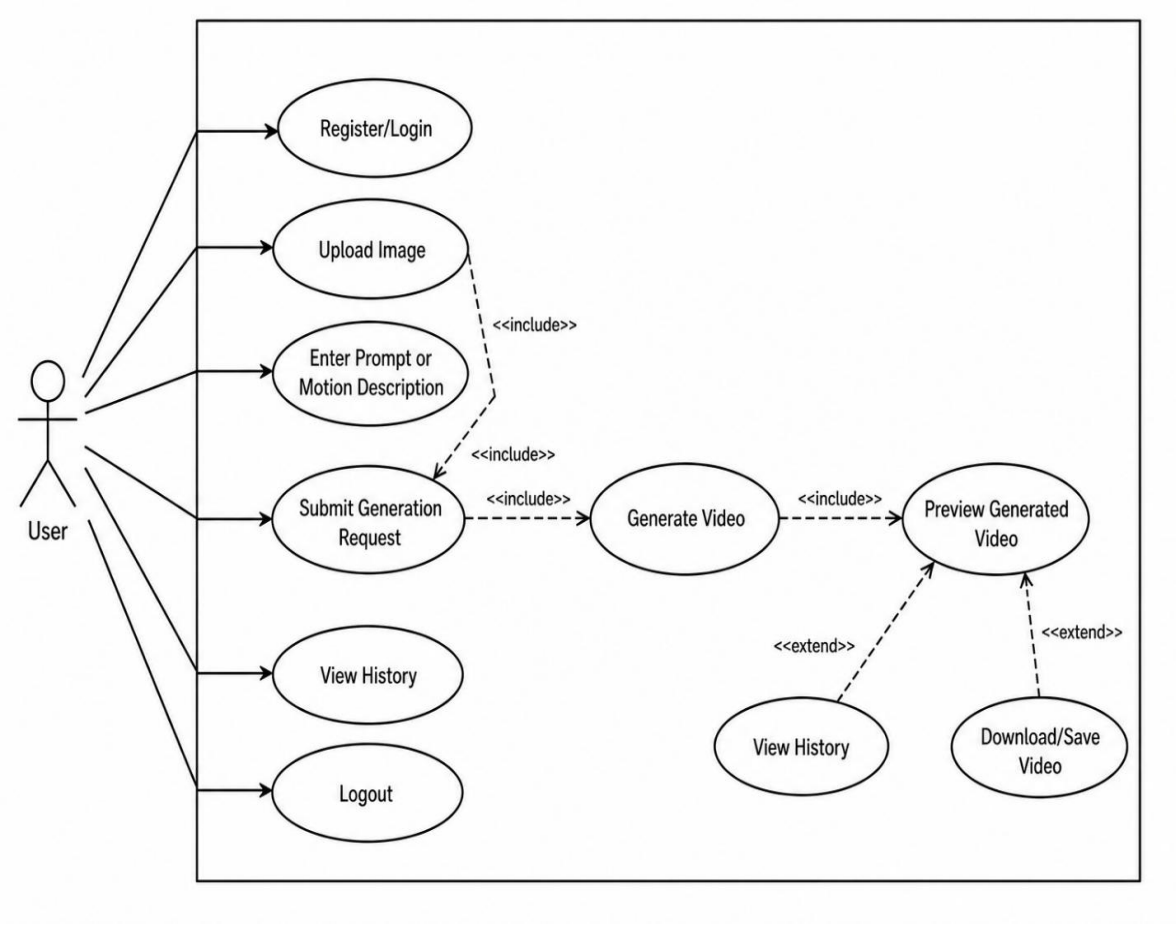


Figure 3.1 Use Case Diagram

3.5 System Architecture

The system architecture describes the structure and components of the proposed 3D U-Net-based diffusion model Image-to-Video generation system, integrated with mobile application software to create a user-centred interface that promotes a user-friendly experience. Figure 3.2 presents the AI-based Image-to-Video system architecture.

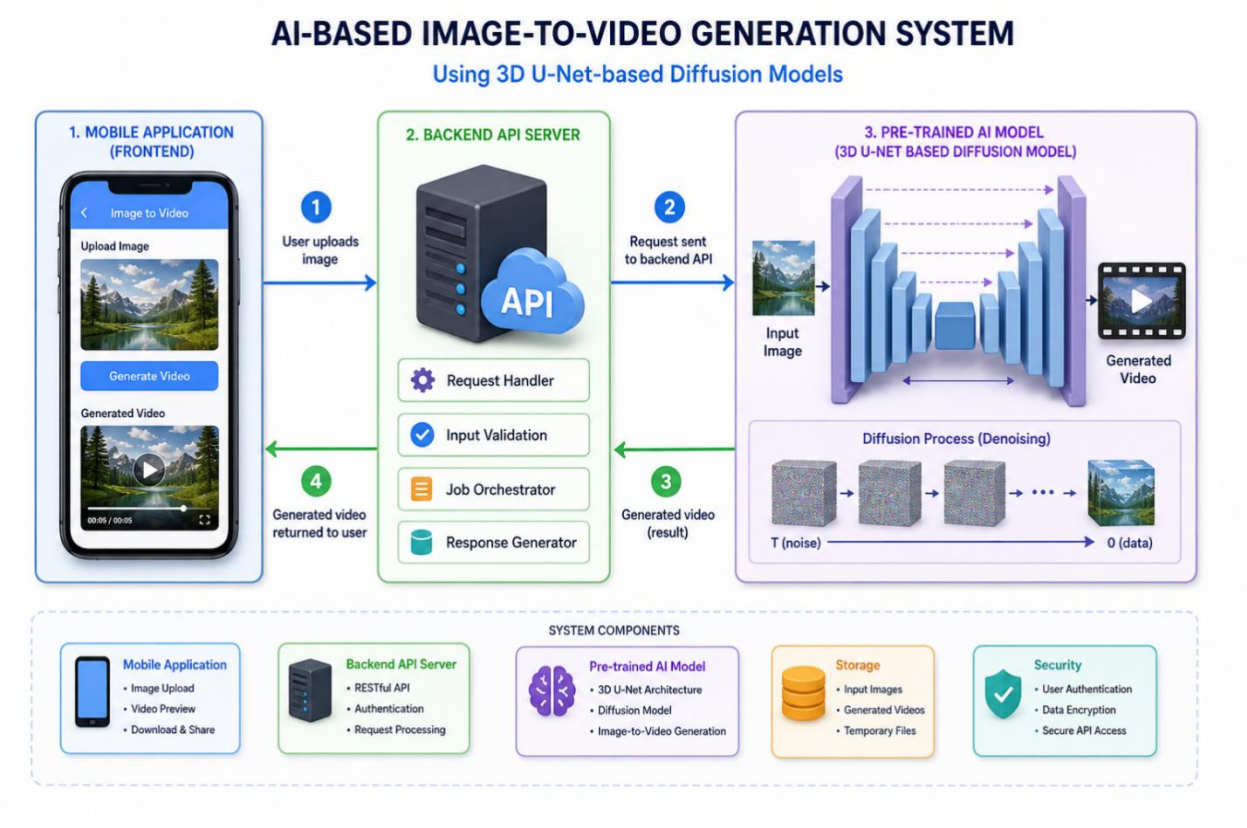


Figure 3.2 System Architecture

The interaction of the user from the mobile application front end is shown in Figure 3.2, where the user logs in to the mobile application in order to upload the image that will be used in creating the video. It may ask you to provide a motion/prompt description before clicking on generate video. The request together with the image and its metadata is sent using the HTTP/REST API to the backend server, which will process the request. It will do so by storing the input image temporarily before sending data to the AI model module, which will carry out the operation by encoding the input image in latent space before diffusion begins. 3D U-Net will generate a temporal frame, or rather the video sequence, which is then decoded to create a coherent video. The output frames are then compiled to create a video file with extensions such as MP4, and then sent to the backend server, which sends it back to the mobile application for

viewing purposes. Finally, the user can now decide the next action, which includes previewing the video, downloading or saving the video, sharing the video or viewing history.

3.6 Data Collection and Dataset

In this research, the secondary data obtained from Panda-70M, which is an open-source database consisting of 20,000 desirable video-text snippets, were used. The data consist of videos, images, and text snippets. This data set was utilized for training and evaluation of the model.

3.6.1 Sample Dataset

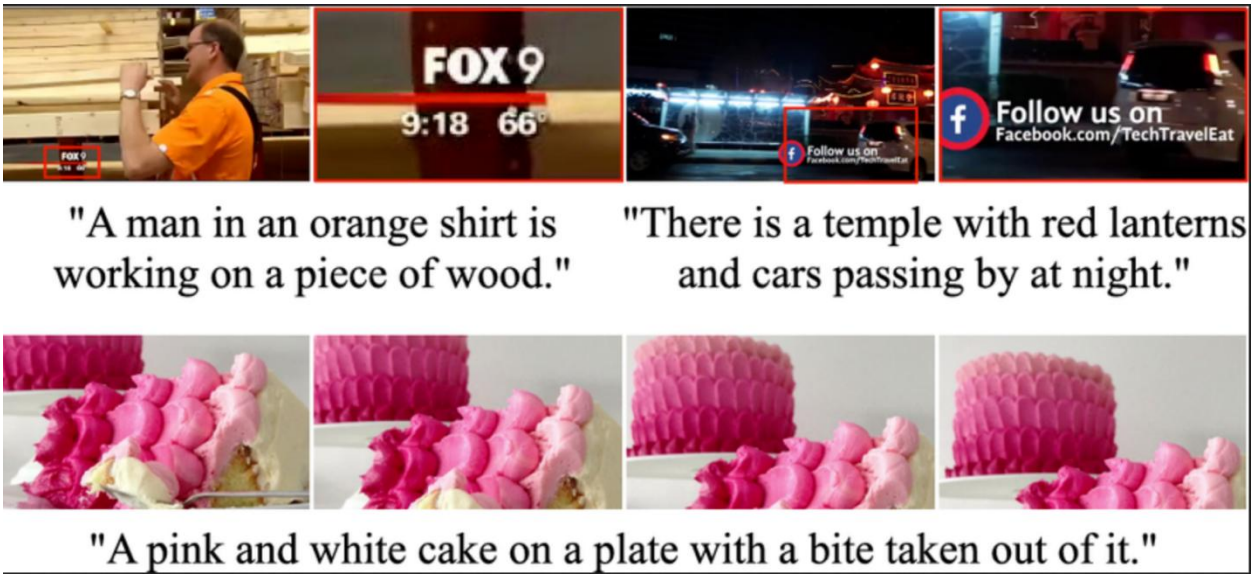


Figure 3.3: Samples of the collected dataset

Figure 3.3 shows a sample screenshot of the Panda 70M dataset containing millions of video clips and the structure in this metadata format for each segment

Segments:

1. 00:13 – 00:41 → "A woman in a white apron standing in a kitchen."

2. 00:55 – 01:36 → "A person is preparing food on a tray with ingredients."

3. 01:41 – 01:45 → "A woman in an apron preparing food."

3.6.2 Data Preprocessing

All the data collected from the Panda-70M (20,000 high-desirability video-text clips) dataset were prepared before training in steps that involved image resizing, frame extraction from video sequences, data normalisation, cleaning and filtering irrelevant data, and ensuring uniform format as well as improved model performance.

3.8 Model Design and Development

The suggested Image-to-Video generation model is a deep learning-based generative model that will transform a static image into video, optionally by using a text prompt/motion description prior to the generation of the video to ensure the generated output video fulfills the users' needs. The system architecture will be based on the combination of 3D U-Net architecture and a diffusion-based generative model. The 3D U-Net architecture is preferred due to the capability of extracting both spatial and temporal feature representation from image/video representations, whereas the diffusion model helps in generating high-quality frames by performing a gradual denoising procedure. In this way, the system will be able to learn the visual structure from the input image in order to generate frames in a way to obtain a smooth video output. Figure 3.4 illustrates the architectural design of the 3D U-Net architecture. Figure 3.5 illustrates the diffusion-based generative model

architecture

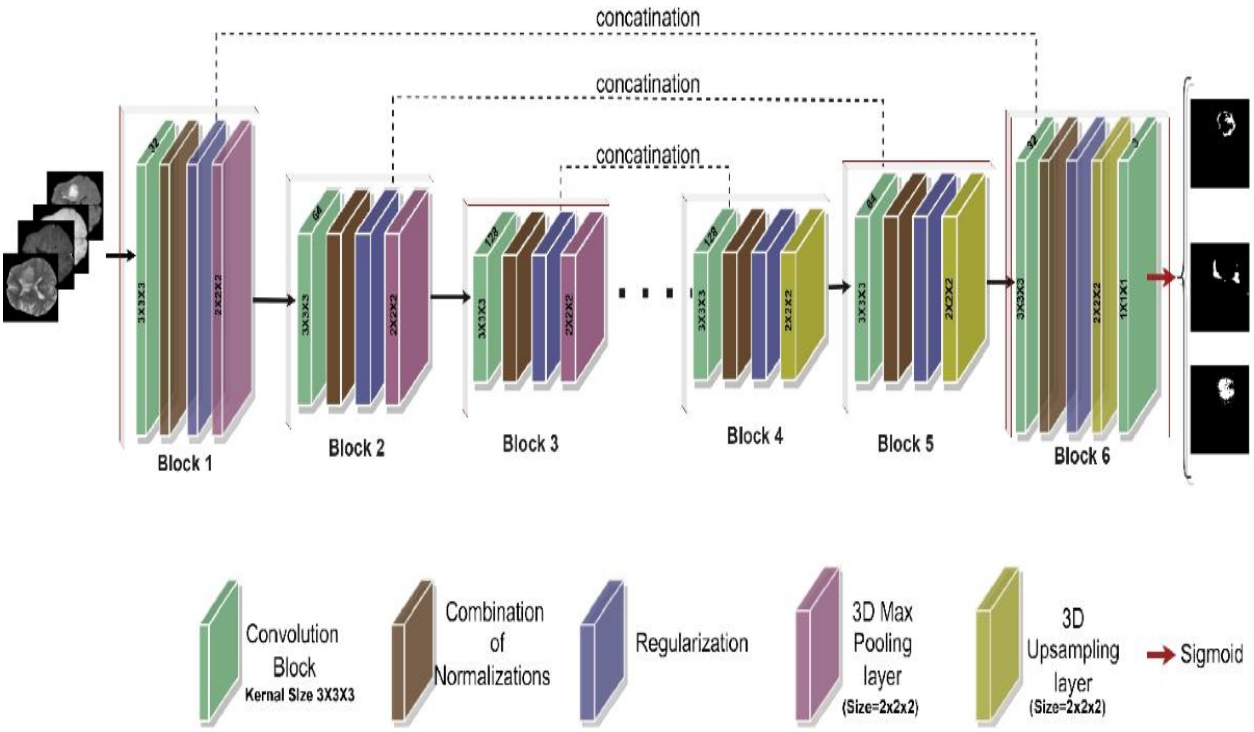


Figure 3.4 3D U-Net architecture

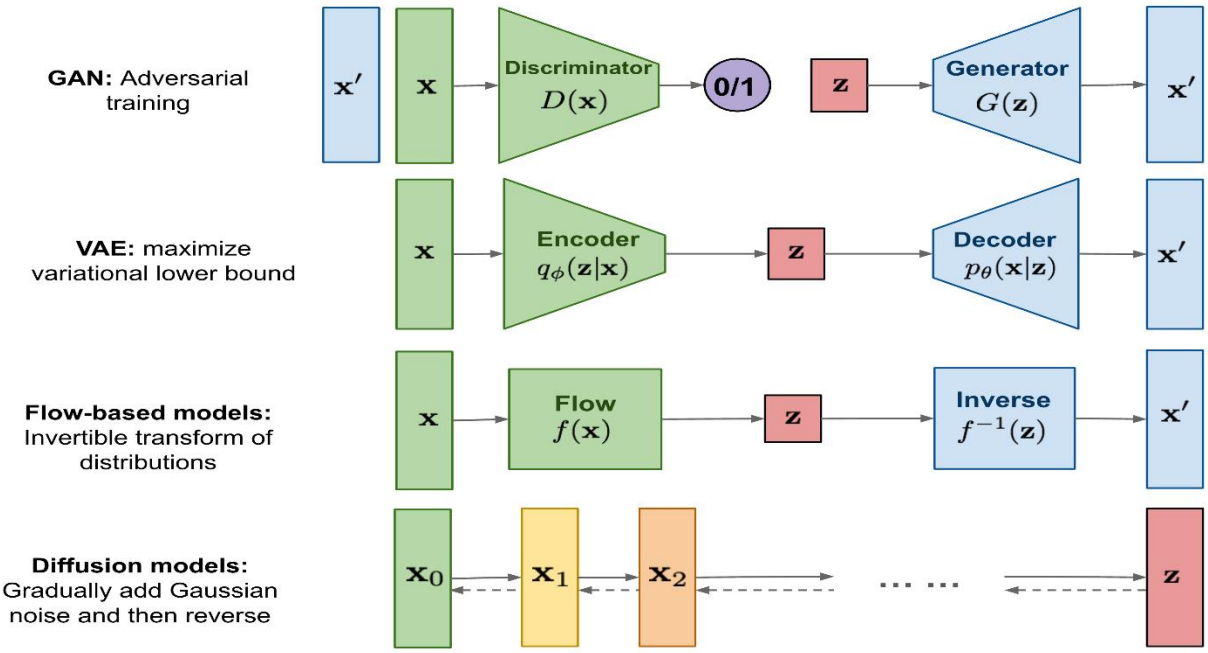


Figure 3.5 Diffusion-Based Generative Model Architecture

Figure 3.6 presents the proposed framework architectural diagram, which is a combined 3D U-Net + diffusion architecture illustrating how the model receives a still image serving as the base frame for the video generation, and in the process of the input, provisions are made for an optional text prompt, which is used to condition the input, thereby enabling semantic control over motion scene dynamics or context, unlike in the existing system.

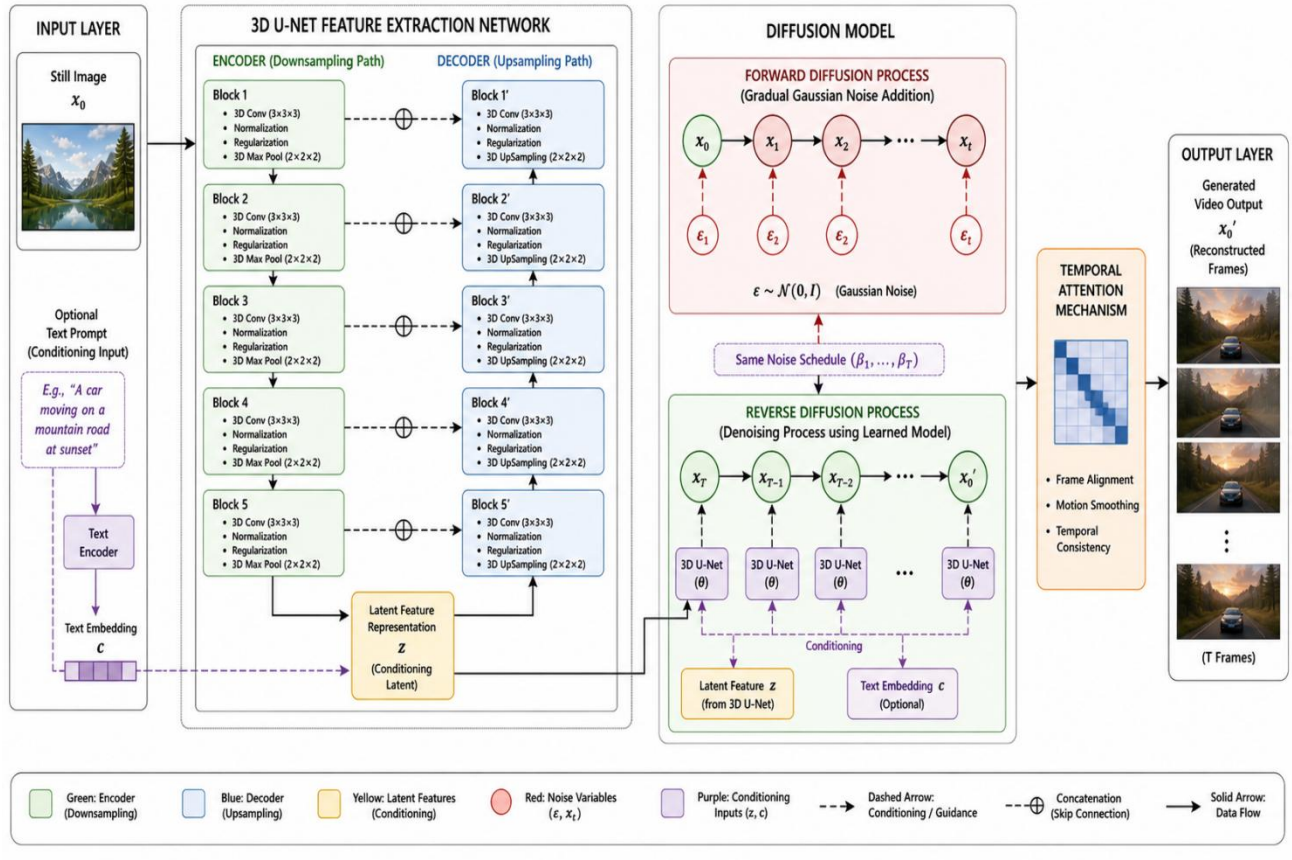


Figure 3.6 proposed model framework

The 3D U-Net feature extraction network is responsible for extracting spatial and temporal features. The encoder path progressively downsamples the input by extracting layered features using 3D convolutions, normalisation, and pooling operations. Additionally, each block of the encoder captures complex patterns over time, such as edges, textures, and structural patterns

relevant to motion synthesis. The decoder path is tasked with upsampling, reconstructing feature maps while targeting fine-grained details through skip connections, and skip connections are in charge of transferring feature information from encoder to decoder in a way that accurately preserves spatial consistency and improves reconstruction quality. The output of the 3D U-Net is a latent representation (z), which encodes the essential structure and motion cues of the input.

The diffusion model introduces a forward process, called the forward diffusion process, which allows the Gaussian noise to be gradually added to the input across multiple steps, and it is represented by equations 3.1 and 3.2

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \quad (3.1)$$

Equivalent sampling form:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, I) \quad (3.2)$$

Where:

$$\alpha_t = 1 - \beta_t$$
$$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$$

And the reverse diffusion process is applied such that the model can iteratively denoise the noisy signal in order to reconstruct a refined frame. The reverse diffusion process is represented by equations 3.3 and 3.4

$$p_\theta(x_{t-1}|x_t, z, c) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, z, c), \Sigma_\theta(x_t, t)) \quad (3.3)$$

A commonly used denoising mean form is:

$$\mu_\theta(x_t, t, z, c) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t, z, c)\right) \quad (3.4)$$

Where:

- x_0= original input image/frame
- x_t= noisy sample at timestep t
- ϵ= Gaussian noise
- β_t= noise schedule
- α_t= retained signal factor
- $\bar{\alpha}_t$= cumulative product of α_t
- z= latent feature representation from 3D U-Net
- c= optional text-conditioning embedding
- ϵ_θ= learned noise prediction network

$x_0^{'}$= reconstructed/generated output frames

Overall, this design for the proposed image-to-video generation system, using a 3D U-Net and a diffusion model, is effective, efficient, and robust because it combines feature extraction, generative modelling, transfer learning, and temporal consistency enhancement within a single pipeline.

3.9 System Requirements

System requirements refer to the necessary processes and specifications that should be present in a particular system. System requirements can therefore be divided into two categories, which include functional requirements and non-functional requirements.

3.4.1 Functional Requirements

i. The new system should allow users to register, log in, and log out securely, as well as manage the user sessions and restrict unauthorised access to system features.

ii. Users should be able to upload still images only in supported formats, such as JPG and PNG, and it should validate image size, resolution, and format before

iii. Users should be able to input motion descriptions or prompts to guide or personalise the video they want to generate.

iv. The system should preprocess input images, for example, resizing, normalisation, and augmentation if necessary, and it should prepare inputs in a format suitable for deep learning inference.

v. The system will extract spatial-temporal features using a 3D U-Net architecture, likewise having the capability to generate latent feature representation (z) for downstream processing.

vi. The system should implement forward diffusion, which involves noise addition, and then reverse diffusion, which means denoising, likewise having the ability to generate a sequence of frames from the latent representation.

vii. The system should allow users to preview generated videos before downloading, and it should also display generated outputs within the application interface.

viii. The system should allow users to download generated videos, and it should store user history and previously generated outputs.

ix. The system should return real-time, readable, and interpretable results to the user.

3.4.2 Nonfunctional Requirements

i. The system should have high levels of accuracy while maintaining low levels of error margin.

ii. The system should respond in real-time to user input and requests.

iii. The system should be scalable and be able to accommodate more than one user at once.

iv. User information/data should be safely stored and encrypted through normal means.

v. The system should provide a responsive interface that works on both mobile and desktop systems.

vi. The system should have high availability and low downtime.

3.7 Database Design

This design of the database is just a rough idea of how the database would appear after implementation. It shows the different tables that will be used to store information within the database. However, the major reason for designing a database is to show the model of the appearance of the database. The database of the proposed system is a relational database, and its design has been based on the use of the MySQL query language. It has entities and attributes for each entity.

Database schema for the proposed system

Table 3.1 User table for the proposed system

Attribute	**Data Type**	**Description**
Id	Int	The primary key of the table.
Username	Varchar	This is where the user's username is stored.
Email	Varchar	The field for storing the user's email address.
Password	Varchar	The field for storing the user's encrypted password.
Date_joined	Timestamp	The field shows the date the user registered on the system.

Table 3.2 Image upload table for the proposed system

Attribute	**Data Type**	**Description**
Id	Int	The primary key of the table.

User_id	Int	The foreign key that identifies the user who uploaded the image.
Image_name	Varchar	The field for storing the uploaded image's name.
Image_path	Varchar	The field for storing the uploaded image's storage location.
Image_format	Varchar	The field for storing the image format, such as JPG, PNG, or JPEG.

Table 3.3 Video Generation Table of the Proposed System

Attribute	**Data Type**	**Description**
Id	Int	The primary key of the table.
User_id	Int	The foreign key that identifies the user who generated the video.
Image_id	Int	The foreign key that links the generated video to the uploaded image.
Prompt	Text	The field for storing the user-entered text prompt or motion description.
Model_used	Varchar	The field for storing the model used, such as the 3D U-Net Diffusion model.
Generation_status	Varchar	The field for storing the status of the video generation process, such as pending, processing, completed, or failed.
Session	Timestamp	The session in which the video generation request was made.

Table 3.4 Generated Video Table for the Proposed System

Attribute	Data Type	Description
Id	Int	The primary key of the table.
Generation_id	Int	The foreign key that links the video to the video generation table.
Video_name	Varchar	The field for storing the name of the generated video.
Video_path	Varchar	The field for storing the location of the generated video file.
Video_format	Varchar	The field for storing the video format, such as MP4, AVI, or MOV.
Duration	Float	The field for storing the generated video's duration in seconds.
Resolution	Varchar	The field for storing the video resolution, such as 512×512 or 1024×576.
Date_generated	Timestamp	The field shows the date and time the video was generated.

Table 3.5 Model Parameter Table for the Proposed System

Attribute	Data Type	Description
Id	Int	The primary key of the table.
Generation_id	Int	The foreign key that links the parameters to a generation request.
Frame_count	Int	The field for storing the number of frames generated for the video.

Diffusion_steps	Int	The field for storing the number of diffusion steps used during generation.
Guidance_scale	Float	The field for storing the guidance value used to control prompt adherence.
Seed_value	Int	The field for storing the random seed used for reproducibility.
Created_at	Timestamp	The field shows when the model parameters were stored.

Table 3.6 History Table for the Proposed System

Attribute	**Data Type**	**Description**
Id	Int	The primary key of the table.
User_id	Int	The foreign key that identifies the user.
Generation_id	Int	The foreign key references the video generation request.
Activity	Varchar	The field for storing user activity, such as image uploads, video generation, previews, or downloads.
Session	Timestamp	The session in which the activity occurred.

Table 3.7 Download Table for the Proposed System

Attribute	**Data Type**	**Description**
Id	Int	The primary key of the table.
User_id	Int	The foreign key that identifies the user who downloaded the video.
Video_id	Int	The foreign key that identifies the downloaded video.

Download_status	Varchar	The field for storing whether the download was successful or failed.
Download_session	Timestamp	The field shows the date and time the video was downloaded.

CHAPTER FOUR

SYSTEM IMPLEMENTATION

4.1 Introduction

This is the phase where the software becomes alive. The next chapter looks at how the system has been implemented along with the user interface and the other tools utilized in the system. What this focuses on is how the implementation process was done for the system development.

4.2 Development Environment and Tools

Python was chosen for its flexibility and support for deep learning model frameworks. It is an open-source ecosystem that accelerates AI model development by supporting multiple ML frameworks, enabling rapid prototyping, testing, and evaluation of AI systems. Additionally, the Python programming language has a strong community support and extensive documentation available.

4.2.1 Programming Languages and Libraries

Programming Language Justification

Python supports deep learning and scientific computing efficiently because it has a simple syntax that improves readability and faster implementation. The Image-to-Video generation system that will employ a deep learning technique will leverage Python's large ecosystem for AI, ML and data processing to build and implement an effective and efficient system. Additionally, Python has compatible GPU acceleration libraries, and it is widely used in research and industry applications.

Libraries Used

In implementing my model, I used a lot of Python libraries, including TensorFlow for the 3D U-Net model training, Keras for simplifying neural network model development, PyTorch was used

in implementation of the flexible diffusion model, OpenCV for handling image preprocessing and transformations to the required format, and the NumPy library for supporting numerical computations and array operations. In addition to that, Matplotlib is used for visualisation and result plotting in accordance with the metrics considered. The diffusers library is used for diffusion-based video generation, while PIL handles image loading and basic processing, and Scikit-learn supports evaluation metrics and preprocessing of data utilised in training the 3D U-Net and Diffusion model.

4.2.2 Development Platform and Hardware Requirements

The development platform or integrated development environment employed for this study includes the Visual Studio Code (VSCode), which was used for coding and debugging, the Google Colab, which was extensively used for cloud-based GPU training in order for fast and efficient model training, and finally, the Git for version control, collaboration and deployment.

The hardware requirements for the system implementation are:

i. Minimum 8GB RAM required for smooth processing
ii. Recommended 16GB RAM for deep learning tasks
iii. Intel i5 or higher processor recommended
iv. NVIDIA GPU used for faster model training
v. CUDA support required for GPU acceleration
vi. SSD storage improves data loading speed
vii. High VRAM GPU preferred for diffusion models
viii. Stable internet required for cloud-based training

4.2 System Implementation

The new Image-to-Video generation system is designed using a modular architectural approach, with the frontend of the application software handling user interaction and input submission, while the backend is tasked with processing images and managing the generation pipeline. REST API connects the frontend and backend components of the software application system, such that the uploaded image is stored temporarily before processing that prepares the image for model compatibility. The prompt input will guide motion and scene generation for user-centred output. In the system engine, the 3D U-Net extracts spatial and temporal features, while the diffusion model generates sequential video frames, which are combined into the final video output, stored and made downloadable; error handling handles failed generation requests.

4.3 Model Training Process

The data includes image-video pairings in samples from the Panda-70M dataset. Preprocessing is done prior to training, where noise is eliminated, and the images are re-sized to have a consistent input size. Data is normalized for consistent training performance within the Google Colab setup, while augmentation is done to help the model generalize better.

Hence, the data is divided into training, validation, and test sets in a ratio of 70:15:15 respectively. The 3D U-Net and diffusion models are initialized for feature extraction and frame generation, and then transfer learning is applied for quick training. The loss function is set for reconstruction performance, and an optimizer is chosen.

For optimal results and high accuracy, the training was performed over 20 epochs with model checkpoints saved for best performance. The evaluation metrics were computed after training completion.

4.3.1 Result of Model Training and Validation

Table 4.1 presents the results of the Diffusion model, which generated realistic motion sequences with gradually decreasing identity preservation, frame quality, and temporal consistency across successive video frames.

Table 4.1: Evaluation Results of the Model based on Standard Metrics

Metric	**Value**
Average Motion Intensity	16.86849
Average Temporal Consistency	0.056117
Average Identity SSIM	0.666318
Average PSNR	14.97884
Generation Time (seconds)	118.2671
Effective FPS	0.067643

Table 4.1 shows SSIM values reduced steadily across generated video frames with the higher SSIM in early frames indicating stronger identity and structure preservation. Also, it demonstrates minimal temporal fluctuations despite increasing motion complexity. Overall table results confirm realistic motion generation with moderate quality degradation.

Subsequently, Figures 4.1, 4.2, 4.3, and 4.4 will present the graphical plot of the metrics used in evaluating the model, including motion intensity, average identity (SSIM), PSNR, and average temporal consistency.

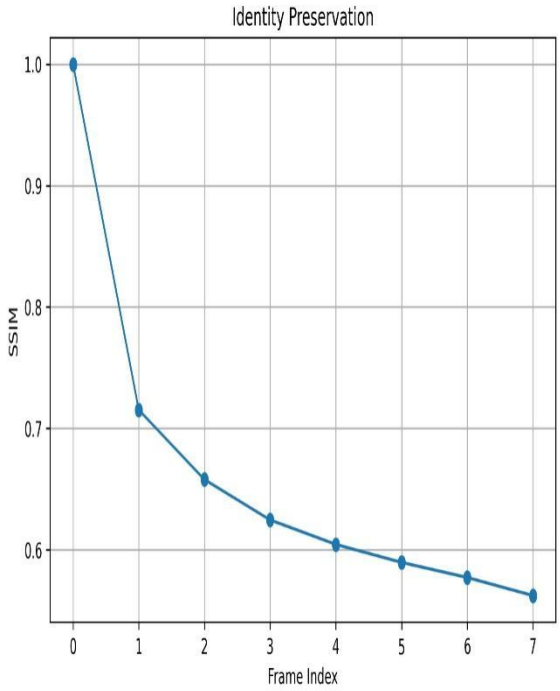


Figure 4.1: Identity Preservation Plot

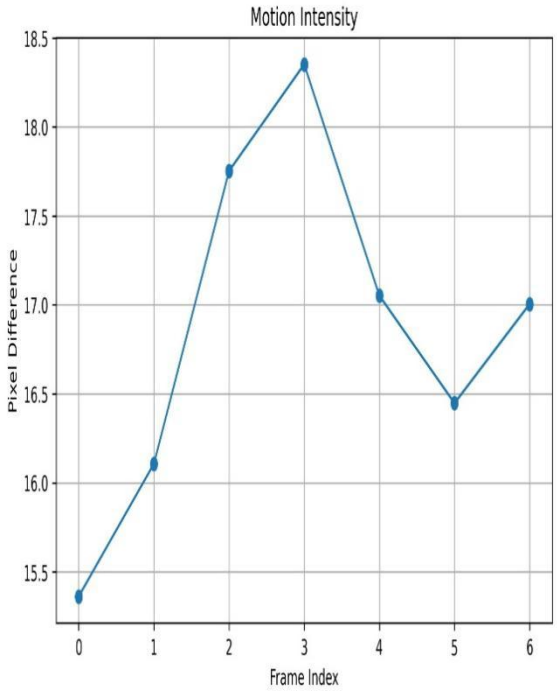
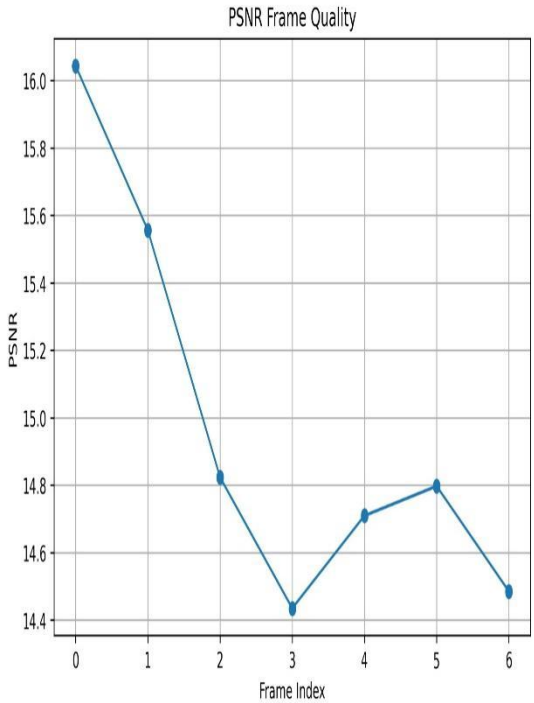


Figure 4.2 Motion Intensity Plot



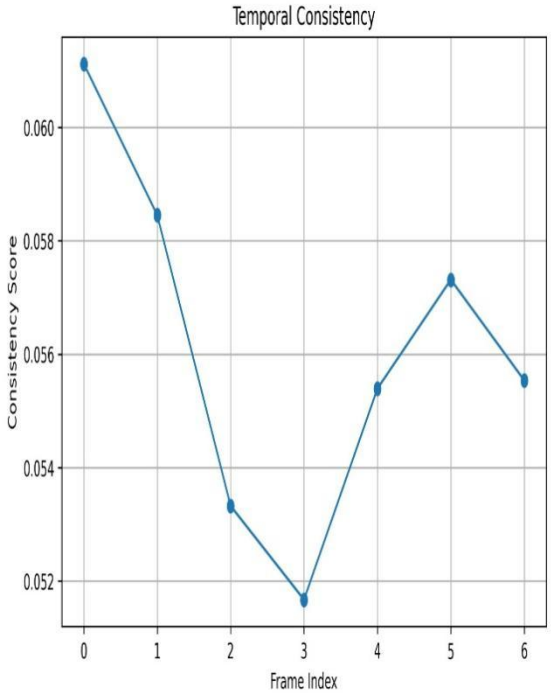


Figure 4.3: PSNR Frame Quality Plot Figure 4.4 Temporal Consistency Plot

Figure 4.1 shows that Identity preservation (SSIM), which dropped from 1.00 to 0.56 gradually and also weakened across frames, but remained visually recognisable throughout generation. Figure 4.2 indicates motion intensity peaked around frame index 3, with the pixel difference initially increasing, indicating stronger motion transitions between generated frames. However, the motion later stabilised, which indicates smoother temporal video progression after peak movement. Figure 3 shows that PSNR decreased from 16.04 to approximately 14.49, and the lower PSNR values postulate a gradual reduction in generated frame quality consistency, and frame three recorded the lowest PSNR, reflecting the highest visual distortion level. Finally, Figure 4 demonstrates how the temporal consistency slightly decreased across generated video frames, and from the plot, it can be seen that temporal consistency remained relatively stable despite fluctuations during middle frame generation. Overall, generated videos maintained acceptable continuity despite declining identity preservation metrics.

4.5 Discussion

The proposed model has successfully generated videos with smooth temporal frame transitions from a single image input with a high-quality visual input facilitated by the adoption of a diffusion-based generation framework. The 3D U-Net was able to capture the spatial and temporal relationships very effectively. Model achieved stable convergence after several training epochs with output videos preserving structural consistency from the input image, while the prompt integrated to the input helped improve motion realism, and on the other hand, noise artefacts were reduced with increased diffusion steps. Despite that, some outputs showed minor

blurring in complex scenes; however, model performance improved with higher-quality training data, and overfitting was minimised after we used data augmentation techniques.

The evaluation metrics indicated acceptable model performance for the system, with inference time varying with hardware capability. The system handled multiple requests with moderate efficiency, demonstrating the effectiveness of the hybrid model approach. Though limitations were observed in highly dynamic motion generation.

4.6 System Interface Description

A. User Interface: The User interface is designed for simplicity and ease of use, with the login and registration interface for user authentication.

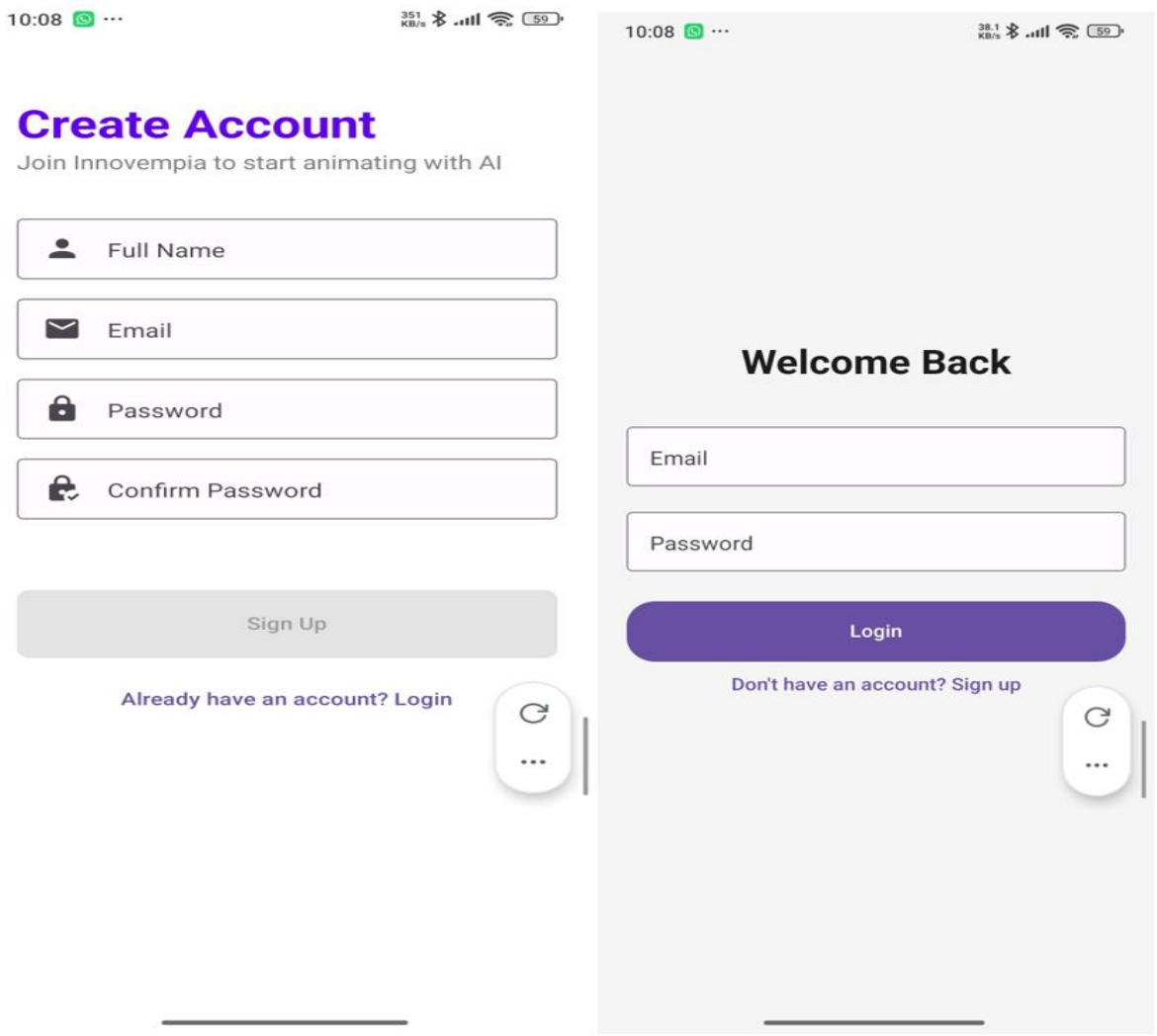


Figure 4.5 Account creation and Login Page

B. Image Upload Page: The image upload interface allows selecting an input image, and there is an input form for a prompt field that captures motion description, and a generate button, which will trigger the video generation process.

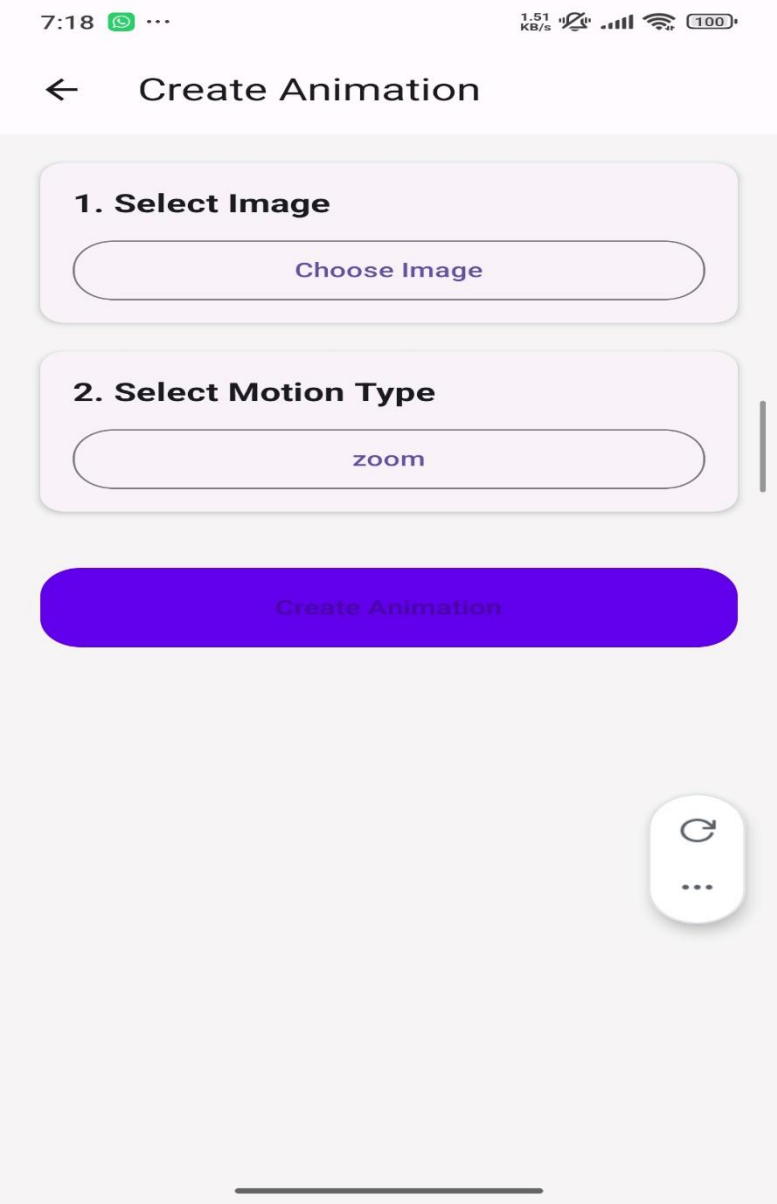


Figure 4.6 Image Upload Page

D. Output Page: The output section displays the generated video preview, with a download button that allows saving the generated video.

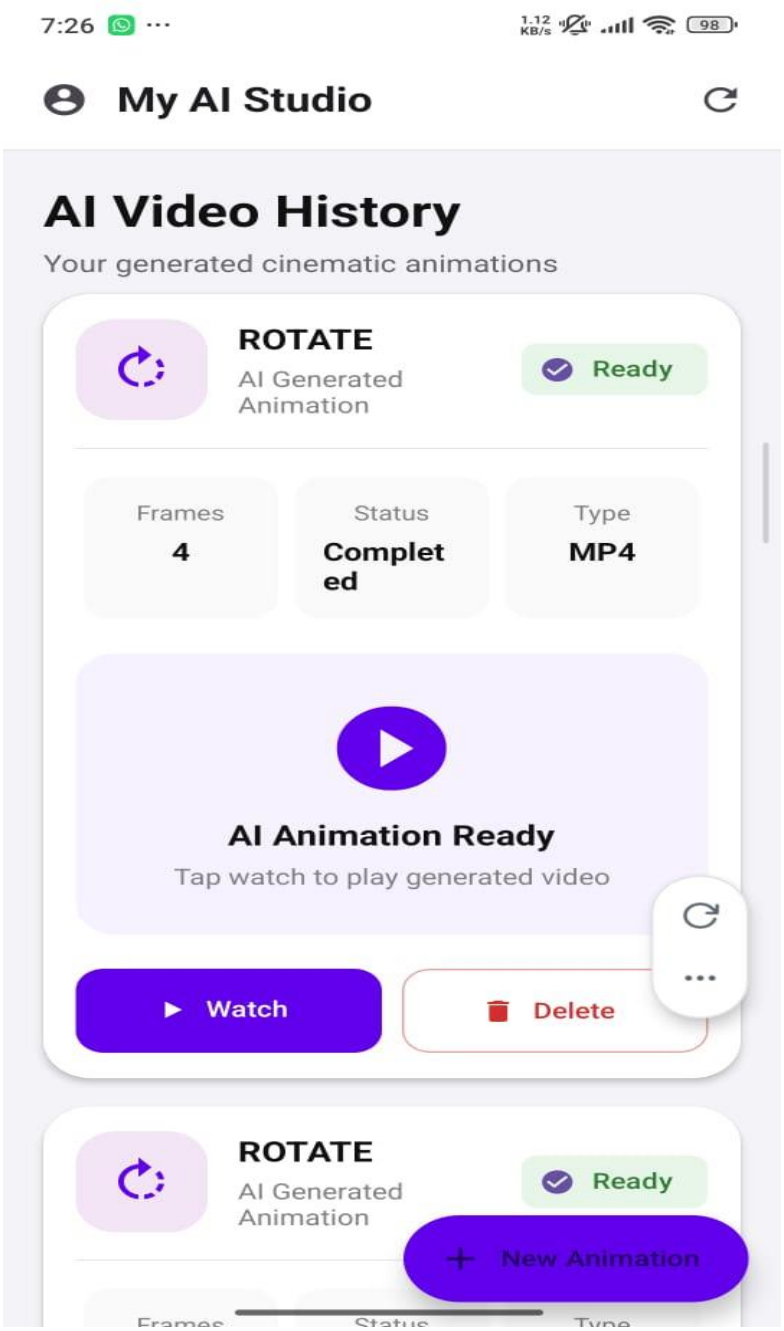


Figure 4.7 Result Page

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATION

5.1 Summary

The study developed an AI-based image-to-video generation system using pretrained diffusion models. The process started with the analysis of existing video generation systems in order to identify major technical limitations. Data was collected from the Panda-70M dataset and utilised in a transfer learning approach to train and evaluate the model. The dataset was preprocessed in a series of steps that included resizing, frame extraction, normalisation, and noise filtering procedures. Then, a hybrid 3D U-Net and diffusion architecture was implemented for video generation and transfer learning was integrated in order to reduce computational cost and training complexity.

Temporal attention mechanisms improved frame-to-frame smoothness and motion coherence, and for the implementation environment, programming languages and libraries, the study made use of Python, TensorFlow, PyTorch, OpenCV, and Diffusers. Model training was performed using Google Colab GPU acceleration environment, which enables us to divide the datasets in a ratio of 70:15:15 for the training, testing and validation. Evaluation metrics included SSIM, PSNR, temporal consistency, and motion intensity, with results demonstrating realistic motion generation with acceptable visual continuity across frames, and identity preservation gradually reduced across successive generated video frames. In the same vein, the motion intensity initially increased before stabilising during later frame generation stages

For system usability, the study developed and implemented a mobile application using React Native for the frontend and a Flask backend, enabling interactive user-system communication. The REST API connected the mobile interface with the deep learning backend engine, so that

when the user uploads an image, the system accepts it along with the optional motion prompts, processes it and generates frames that are compiled into downloadable MP4 video outputs automatically. Overall, the proposed system demonstrated practical usability for automated AI-driven video generation.

5.2 Conclusion

The study successfully implemented an AI-based image-to-video generation system using pretrained diffusion models, which improved generation quality and reduced training requirements significantly. The 3D U-Net captures spatial and temporal relationships between frames effectively so that the generated videos maintain acceptable continuity and realistic motion progression throughout generation. The model was evaluated considering appropriate metrics, which confirmed its stability and reliability in performance during testing stages. The adoption of transfer learning approaches enhanced model efficiency and reduced computational overhead effectively.

The mobile-based interface improved accessibility and usability for non-technical users through the integration of a prompt-guided generation, which improved motion realism and scene controllability in generated videos. And some generated outputs showed slight blurring within highly dynamic motion scenes. The proposed system bridged AI research concepts through the practical implementation of a real-time and deployment-ready system. The study further demonstrates diffusion models' effectiveness in multimedia content generation applications. Overall, the system achieved the research aim and stated project objectives successfully.

5.3 Recommendations

i. The study recommends that future studies utilise larger and more diverse video generation datasets so as to reduce bias, overfitting and improve generalizability of the system.

ii. Future systems should integrate more advanced temporal attention mechanisms

iii. Further studies improve the system so that it can support longer-duration video generation with improved stability.

iv. Application of optimisation techniques in order to reduce inference time and improve processing efficiency further.

v. Future research should integrate text-to-video and audio-to-video generation capabilities.

vi. Future studies should compare multiple pretrained diffusion models for performance benchmarking.

References

Akber, S. M. A., Kazmi, S. N., Mohsin, S. M., & Szczęsna, A. (2023). Motion style transfer tools, techniques and future challenges based on deep learning. Sensors, 23(5), 2597. https://doi.org/10.3390/s23052597

Alsaaran, N., & Soudani, A. (2025). Image-based classification using deep learning to predict post-earthquake damage levels using UAVs. Sensors, 25(17), 5406. https://doi.org/10.3390/s25175406

Cai, Y., Huang, L., Wang, Y., Cham, T. J., Cai, J., Yuan, J., Liu, J., Yang, X., Zhu, Y., & Shen, X. (2020). Progressive joint propagation for human motion prediction. In European Conference on Computer Vision (ECCV).

Chen, Y., Chu, W., Wu, Y., Yang, J., Mao, X., & Liu, W. (2025). COTA-motion: Controllable image-to-video synthesis with dense semantic trajectories. Neurocomputing, 657, 131671. https://doi.org/10.1016/j.neucom.2025.131671

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. IEEE Signal Processing Magazine, 35(1), 53–65. https://doi.org/10.1109/MSP.2017.2765202

D-ID. (2024). Speaking portrait: AI avatar presenters. https://www.d-id.com/speaking-portrait

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. In Advances in Neural Information Processing Systems, 34, 8780–8794.

D'monte, S., Varma, A., Mhatre, R., Vanmali, R., & Sharma, Y. (2022). Cartoonization of images using machine learning. International Research Journal of Engineering and Technology (IRJET), 9(5), 131–139.

Esser, P., Rombach, R., Blattmann, A., Müller, J., Treyer, M., Štys, D., & Ommer, B. (2023). Structure and content guided video synthesis with diffusion models

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems, 33, 6840-6851.

Ho, J., Chan, W., & Fleet, D. J. (2022). Video diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 895-904.

Hu, Y., Mittal, A., & Li, Y. (2023). Adaptive motion adapters for image-to-video diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,.

Jiang, D., Chang, J., You, L., Bian, S., Kosk, R., & Maguire, G. (2024). Audio-driven facial animation with deep learning: A survey. Information, 15, 675. https://doi.org/10.3390/info15110675

Jiang, J., & Wang, X. (2024). Animation scene generation based on deep learning of CAD data. Computer-Aided Design & Applications, 21(S19), 1–16. https://doi.org/10.14733/cadaps.2024.S19.1-16

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4401–4410).

Kumar, K. H., & Somasundara Rao, M. (2024). Image to live video transmission using GAN. International Journal of Creative Research Thoughts, 12(2).

Lungu-Stan, V.-C., & Mocanu, I. G. (2024). 3D character animation and asset generation using deep learning. Applied Sciences, 14, 7234. https://doi.org/10.3390/app14167234

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv. https://doi.org/10.48550/arXiv.1411.1784

Nichol, A., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In International Conference on Machine Learning (pp. 8162–8171).

Nirkin, Y., Hassner, T., & Wolf, L. (2021). FSGAN: Subject agnostic face swapping and reenactment. International Journal of Computer Vision, 129(11), 3157–3172. https://doi.org/10.1007/s11263-021-01511-7

Ripote, S. S., & Wankhade, N. R. (2025). Real-time image animation using AI machine learning and deep learning. Journal of Information Systems Engineering and Management, 10(40s).

Saito, W., Uchida, Y., Watari, Y., & Yokoo, T. (2017). Temporally coherent GANs for video generation. In Advances in Neural Information Processing Systems (pp. 6919–6

Vondrick, C., Shrivastava, A., & Gupta, A. (2023). Generative adversarial video generation with temporal coherence. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(8), 941–955. https://doi.org/10.1109/TPAMI.2022.3221234

Wang, A., Wu, H., & Iwahori, Y. (2025). Deep learning in image processing and pattern recognition. Electronics, 14, 1942. https://doi.org/10.3390/electronics14101942

Wang, T., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B. (2020). Video-to-video synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11), 3841–3855. https://doi.org/10.1109/TPAMI.2020.

Zhang, H., Li, Y., & Wang, O. (2024). Motion-aware diffusion for controllable image-to-video generation. ACM Transactions on Graphics, 43(4), Article 112. https://doi.org/10.1145/3647412

Zhou, K., Yang, Y., Cavallari, S., Torr, P. H. S., & Yang, Y. (2023). Learning hybrid generative models for video prediction and synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19876–19886).

Zhou, Y., Xu, Z., Assari, S. M., Takeuchi, M., & Sarkar, S. (2020). Talking face generation by audio-driven face animation. In