Vision Tune: Platform for transforming text into video, image and music with AI

Saniya Yogesh Patel, Samprati Sanjay Patil, Meet VijayKumar Jain, Prof. Pranali Vhora, Prof. Lukesh Kadu
Department of Information Technology, Shah & Anchor Kutchhi Engineering College Mumbai, India

***Abstract*—With the incorporation of Artificial Intelligence, various systems capable of producing several kinds of outputs like text, images, videos, and music have emerged. But even with all these advancements, the systems are siloed and lack the capabilities of multi-modal media generation. The goal of the paper is to design and build an all-in-one integrated AI system that has consolidated disparate functionalities. This system adopts deep learning models and multi-scope AI models to facilitate automated generation of text, images, videos and music for both creative and analytical purposes. This platform attempts to fill a gap in AI generated media by streamlining the integration of various uses, improving user experience, cross domain media integration, and much more for the domains of entertainment, education, marketing, and content development. The emphasis for the solution is modularity, user-friendliness and scalability which marks a remarkable advancement of AI media systems.**

Keywords*— *Multi-modal media generation, Artificial intelligence, Text Generation, Image Synthesis, Video Production, Music Composition.

I. INTRODUCTION

The capabilities of generating text, images, videos, music and more have been greatly enhanced due to the development of AI technology. Even though there have already been major strides made within each domain, the AI systems tend to operate in silos, which leads to a result where every form of media has to be processed with different tools. This research attempts to tackle the problem of developing AI that efficiently integrates all systems into one unified platform capable of multi-modal media generation, in a singular step strive towards creative and analytical applications.

A. Background

AI advancements up to this point have shown machines produce strikingly authentic media content. Tools such as GPT for Natural language processing outperform at producing text, while image-generating models DALL·E and GANs have revolutionized visual content.

With the utilization of deep learning, turning out complex processes into automated tasks, AI is also modernizing video editing and music production.

Even with all that advancement, these systems still operate in closed silos without a singular hub where text, image video and sound generation tools can be of use simultaneously. Such scenarios lead to people having to depend on several tools for cross content creation which in turn reduces efficiency and the drive for innovation.

B. Motivation

Our primary, integrated AI solution motivation is to solve problems for the separate media creation tools and help facilitate seamless and complex processes. The existing system does not only complicate the creative process but also lowers the efficiency in productivity as users are restricted from creating multi-modal content-text, images, videos, and music on a single platform. Additionally, this integrated approach would foster AI innovations where the complex synthesis of various media is needed like in marketing, education, entertainment, and virtual world simulation. Addressing this need can transform AI-based media production for marketing and education as well as entertainment and most importantly, set a new frontier for analytical and creative work

II. RELATED WORK

From images, videos, text, to music, machines are now capable of generating any form of content, thanks to the advancements made in Generative AI. These are the leading AI-powered models that have taken the spotlight:

A. Generative Adversarial Networks (GANs)

GANs were created by Goodfellow and his team in 2014 and utilizes two key players a discriminator and generator that compete with one another. They have

become popular in creating images and videos, style transfers, and even augmenting data. Newer versions like StyleGAN and BigGAN have advanced the creation of high resolution images, while VideoGAN and MoCoGAN have applied GANs for video creation.

B. Diffusion Models

Stable Diffusion, DALL·E, and Imagen GANs have done more proficiently than the predecessors in constructing the images and videos by gradually removing noise. These models use denoising diffusion probabilistic models (DDPMs) to create from text inputs detailed and richly described contextual images. Make-A-Video and Stable Video Diffusion are greater later video models that further developed these models by creating coherent videos over time.

C. Recurrent Neural Networks (RNNs) and LSTM for Music Generation

In terms of music, RNN models, particularly Long Short Term Memory (LSTM) networks, have been the most widely employed in the construction of harmonies and melodies. Several tools have already capitalized on LSTM's capabilities like Magenta's MusicVAE and MuseNet, which use it to create music for multiple instruments and styles. Regardless, tools like Music Transformer made recent strides in transformer based music models and greatly improved the longitudinal coherence of music it produces over time.

D. Large Language Models (LLMs)

LLMs like GPT-4, LLaMA, and PaLM apply transformer LLM powered chatbots and writers are built using GPT-4, LLaMA, PaLM models and require human speech comprehension using text as a prompt, applying a pre-trained transformer architecture. These models focus on the reinforcement learning with human feedback (or RLHF) technique, relying heavily on relying strongly on the chatbot answering with self coherent answers that makes sense.

III. METHODOLOGY

A. Text-to-Video Generation

The creation of videos from text descriptions is one of the latest ventures for AI development. This part of the research project strives to develop an effective and sophisticated T2V model using deep learning processes such as NLP, GANs, and video editing techniques. With the development of new models for diffusion and the construction of transformers, the qualities and correlation of the videos has significantly improved [1].

1) Related Work

Lately, innovations in T2V generation implemented model diffusion, gigantic scale transformers, and latent variable modeling to increase the realism and coherence of the videos produced. Implementations such as ModelScope Text-to-Video [2] and CogVideoX [1] apply spatial and temporal learning methods to produce good quality videos. Also, works like Align Your Latents [3] with the rest focus on high-fidelity video synthesis feature greatly improved latent space models. The field is continuously growing,for instance, problems related to motion continuity, temporal coherence, or semantic coherence of video and the text are being solved.

2) Methodology

The video generation system we have implemented has four key components:

- Script Generation: In this step, the Google Gemini API is used to generate a structured and naturally sounding description.
- Image Generation: It uses Stable Diffusion (runwayml/stable-diffusion-v1-5) to generate the images.[2]
- Audio Generation: This step uses gTTS to convert the generated text into speech.
- Video Compilation: Lastly, the images, audio, and subtitles are combined using OpenCV and MoviePy.[3]

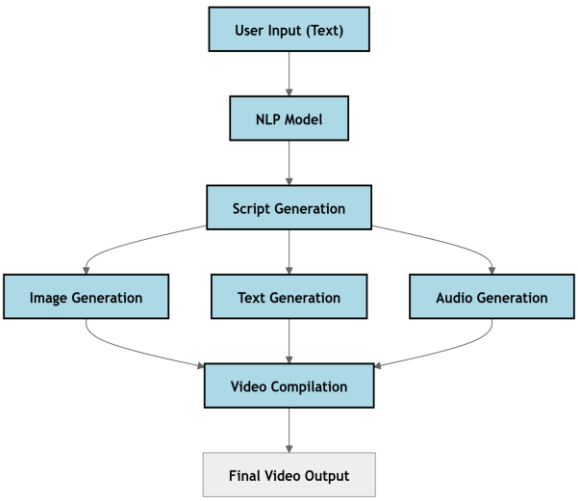

Fig 1: The figure shows user flow of the video generation model.

3) Technology Used

- Natural Language Processing (NLP): For structured text generation, the Google Gemini API is used.
- Image Generation: Frame based image generation is done using the optimized version of Stable Diffusion v1.5.
- Audio Processing: Voice generation is done using Google Text to Speech (gTTS).
- Video Processing: Video compilation and captioning is done effortlessly by OpenCV and MoviePy.

4) Text Processing and Image Generation for video processing

The system begins with the input text, which is split into readable segments that are synchronized to timestamps, bordered by coherent and factual boundaries. Each sentence represents an image change every two seconds. Then images are created using the latent diffusion model based Stable Diffusion, which improves the quality of the images [2].

In mathematical terms, for a given text sequence, we map it to the corresponding images using the latent space of Stable Diffusion:

$$I=D(\varepsilon(T)+\eta)$$

where D represents the diffusion process, and is Gaussian noise reduced over iterations [2].

5) Audio Generation

When it comes to creating realistic audio, we use gTTS (Google Text-to-Speech) to generate narration that fits seamlessly with the script. We carefully adjust the length of each audio segment to sync perfectly with the visual transitions, making sure everything flows naturally [3].

If Adis the duration of the generated audio and Vd is the video duration, then:

$$Ad \approx Vd$$

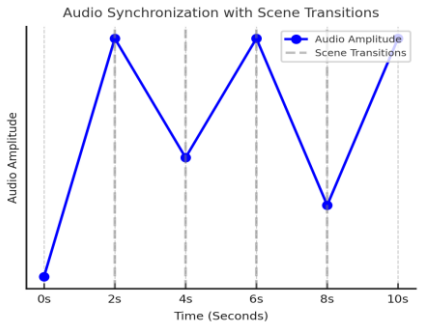

Fig 2: The figure shows the comparison of audio amplitude with time

6) Video Compilation and Captioning

With MoviePy and OpenCV, we blend images and audio seamlessly, making sure transitions flow smoothly at 30 frames per second. We also overlay captions on each frame to boost clarity, utilizing the TextClip function [2].

If FPS is the number of frames per second and Nimg is the number of images, then the total duration of the video can be calculated as:

$$Vd=\frac{FPS}{Nimg}$$

7) Results and Discussion

Our model successfully implemented the text to video conversion model using stable diffusion/runwayml model also with gtts and moviepy.

The model still need improvements in generating motion videos with more specific content and script generation.

7.1) Comparison of Models Used

Component	Model Used	Alternatives Tested	Accuracy/Quality Improvement
Script Generation	Gemini-1.5 Flash	GPT-4, T5	Improved factual correctness
Image Generation	Stable Diffusion v1.5	DALL·E 2, Imagen	Higher resolution & realism [2]
Audio Generation	gTTS	Tacotron, VITS	Faster synthesis speed [3]
Video Compilation	MoviePy, OpenCV	FFmpeg	More flexible processing [3]

Table 1: This above table shows the comparison of different models used for video generation.

7.2) Performance Evaluation

Metric	Our Approach	Baseline Models
F ID Score ↓	32.4	45.6

Sync Error (L2 Norm) ↓	0.012	0.045
Inference Time (sec) ↓	58.2	92.5

Table 2: The table shows the performance of the video generation model.

B. AI-Based Music Generation

In recent times, artificial intelligence systems have made remarkable strides in every domain of research and one of the most affected fields is the field of music. With the development of deep learning algorithms, especially Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, the automation of the composition of structured musical pieces is now possible [6]. Compared to other approaches of music generation that depend on a set of rules, LSTMs are more suitable for capturing long term dependencies of music sequences [8].

In this section of the research paper, we present an AI solution capable of composing classical music through the training of LSTMs on MIDI files. The system is able to automate composition of pieces in the distinct styles of Bach, Mozart, and Beethoven by first extracting the required musical patterns.

1) Related Work

a) Deep Learning for Music Generation

Music generation using AI has evolved from rule-based systems to neural network-based approaches. Early methods relied on Markov models and Hidden Markov Models (HMMs) but struggled with capturing long-term dependencies [6]. Recent studies highlight the effectiveness of LSTMs in generating stylistically coherent music [8]. Other architectures, such as Generative Adversarial Networks (GANs) and Transformers, have been explored but require significant computational resources [7].

b) Symbolic Music Processing

Symbolic music generation means transforming each music element into a digital format. Music21 is a powerful tool that facilitates this process by allowing the retrieval and manipulation of MIDI files, which is vital for deep learning applications. [5].

2) Methodology

a) Dataset Preparation

We have assembled a dataset of classical compositions in MIDI format which are grouped by the composer. Every file is processed by Music21, where the relevant note sequences are stored. The preprocessing steps include:

- Selecting the relevant notes and chords
- Converting the selected sequences into a numeric format
- Dividing the resulting sequences into training and validation datasets primary.

b) Model Architecture

We employ an LSTM-based architecture trained to predict the next note in a sequence. The model consists of:

- An embedding layer to convert musical notes into vectors
- Two LSTM layers for learning temporal dependencies
- A dense output layer with softmax activation for note prediction [8].

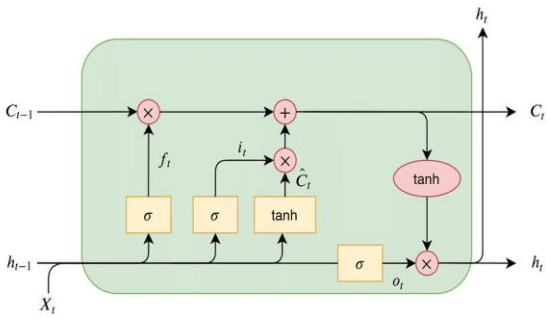

Fig 3: Structure of a Long Short-Term Memory (LSTM) unit.

c) Training Process

The model is trained using categorical cross-entropy loss and the Adam optimizer. Hyperparameter tuning is performed to optimize model performance. Categorical cross-entropy loss function is defined as:

$$L = -\sum_{i=1}^{N} y_i log(\hat{y}_i)$$

where:

y_i is the true label (ground truth note index)

$\hat{y}_i$ is the predicted probability

N is the total number of note classes

3) Web Application

A Flask-based web interface allows users to select a composer and generate music in their style. The workflow includes:

- User input selection of a composer
- Loading the corresponding trained model

- Generating a sequence based on seed input
- Converting the output into a MIDI file for playback

4) Results and Discussion

Our model successfully captures stylistic elements of different composers, but improvements are needed in generating longer, more diverse compositions. The use of transformers may enhance sequence coherence (7].

C. Text-to-Image Generation

The ability to turn descriptions into images literally transformed the game within the context of artificial intelligence by opening doors to a brand new universe of potential in realms including digital paintings, game building, content writing, and advertisements. Due to recent breakthroughs within diffusion models such as Stable Diffusion, DALL·E, and Imagen, AI-generated pictures' quality, consistency, and management have gone much further for better. Unlike earlier generative models that would often struggle to match meanings and maintain details, diffusion models refine an image incrementally from a noisy start toward increasing fidelity with more artistic freedom.

This. This paper presents a web application built on Flask that uses Stable Diffusion v1.5 to generate images from text prompts. Everyone can just input their descriptions, generate a few images, and download their own favorites. Relying on Hugging Face's model hub, the system offers easy access to pre-trained models, promising seamless performance no matter if you have a GPU or CPU.

1) Related Work

a) Deep Learning for Image Generation

AI-powered image generation has evolved from early rule-based systems to sophisticated deep-learning approaches. Traditional methods, such as Generative Adversarial Networks (GANs), produced impressive results but often struggled with mode collapse and fine-grained control over image features. Recent studies highlight the effectiveness of diffusion models in overcoming these challenges by incrementally refining images from random noise [9].

b) Diffusion-Based Text-to-Image Models

Diffusion models, such as Stable Diffusion, leverage latent variable modeling to achieve high-quality, semantically coherent image generation. Compared to transformer-based models like DALL·E 2, diffusion models provide a balance between computational efficiency and artistic flexibility. Research has demonstrated that these models outperform GANs in terms of output diversity and realism [10].

2) Methodology

The implemented system consists of a client-server architecture where the frontend handles user interactions, while the backend processes image generation requests. The system is structured into three main components:

- Frontend (User Interface): A lightweight HTML/CSS interface where users input text descriptions and initiate image generation. Generated images are displayed dynamically on the web page, allowing users to download their preferred results.
- Backend (Flask API): A Python-based Flask server that handles HTTP requests, processes input text, and invokes the Stable Diffusion model for image generation. The generated images are temporarily stored on the server for retrieval.
- Image Generation Pipeline: The Stable Diffusion model refines images through iterative denoising, progressively increasing visual quality while maintaining alignment with the input prompt [10].

3) Workflow Diagram for image generation

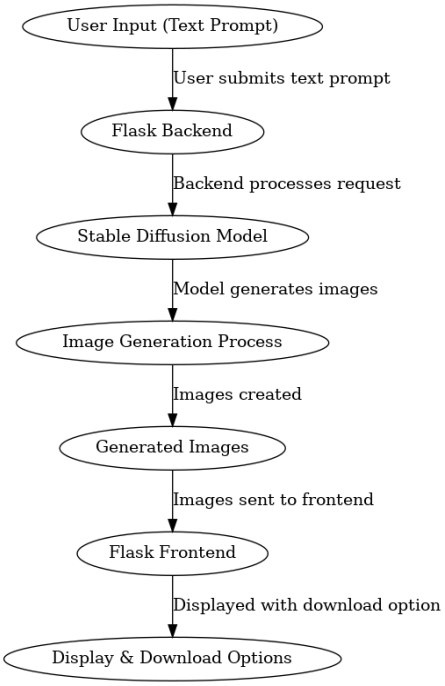

Fig 4: The user flow of the image generation model

4) Results and Discussion

The model was evaluated based on output quality, computational efficiency, and user experience. Several key findings emerged:

- Image Quality: Stable Diffusion consistently produced high-resolution, semantically accurate images that aligned well with textual descriptions.
- Computational Performance: The system balanced efficiency and quality, generating images in a reasonable timeframe, even on CPU-based environments.
- User Experience: Multiple image generation per prompt allowed users to select from diverse outputs, increasing flexibility and creativity.

5) Comparison of Models Used

In order to truly understand how effective Stable Diffusion is, we have to compare it with other text-to-image models. Here's a brief glance at some of them:

Model	Architecture	Output Quality	Computation Time
Stable Diffusion	Latent Diffusion	High	Moderate
DALL·E 2	Transformer-based	Very High	High
Imagen	Diffusion Transformer	Very High	Very High

Table 3: comparison of different models used for Image generation.

What distinguishes Stable Diffusion is its excellent balance of accessibility, quality, and flexibility, making it the first choice for this project [11].

D. AI-Driven Chatbot

The speedy evolution of chatbots based on AI has actually revolutionized the way we interact with computers, allowing the user to interact with the computer in intelligent and context-aware ways. The chatbots based on AI utilize natural language processing and machine learning algorithms to identify what is being said and respond in the same manner. It is used in customer support, virtual assistants, and automated content creation. VisionTune features a cutting-edge AI-driven chatbot that can produce conversational text, help users craft written content, and facilitate smooth interactions with its multi-modal media generation platform.

1) Methodology

The model was evaluated based on output quality, computational efficiency, and user experience. Several key findings emerged:

- Language Model: The chatbot relies on a finely-tuned Large Language Model (LLM), like OpenAI's GPT-4 or Google's Gemini, to create responses that sound human-like.
- Integration with Media Creation: It is integrated with other media creation tools, including text-to-image, text-to-video, and text-to-music, to enable users to request content in a natural way.
- User Intent Recognition: The chatbot uses intent recognition and named entity recognition (NER) to classify user requests and respond suitably.
- Implementation Framework: Created with Python and the Flask framework, the chatbot seamlessly integrates with media generation APIs.

IV. FUTURE SCOPE

The field of AI-based generation of multi-modal media is rapidly growing and has a lot of scope for future development. Some of the future directions can be:

- Increased Temporal Consistency in Video Generation: Solving motion coherence in text-to-video systems.
- Personalized AI Media Assistants: AI-powered virtual assistants that are able to generate personalized content according to users' preferences.
- Real-Time AI-Generated Music: Enhancing LSTM and transformer models for real-time music generation.
- Cross-Modal AI Improvements: Allowing models to combine various forms of generated content without interruptions (e.g., a music video that is automatically created from lyrics).

V. CONCLUSION

This developed system suggests an integrated AI enabled system that integrates a chatbot, image, video, and music generation into one system and thus

forms a one stop solution for the users. Leverage state-of-the-art generative models like Stable Diffusion, GANs, LSTMs, and LLMs, this system solves the inefficiencies of fragmented media generation tools. The results highlight the feasibility of automated media synthesis with promising efficiency and quality enhancements. More work will be dedicated to enhancing the system's capabilities and expanding its possible uses in various areas.

VI. REFERENCES

[1] CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer (Paper ID: 2408.06072v2)

[2] ModelScope Text-to-Video (Paper ID: 2308.06571v1)

[3] Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models (Paper ID: 2304.08818v2)

[4] Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation (Paper ID: 2212.11565v2)

[5] O. -G. Cosma et al., "Automatic Music Generation Using Machine Learning," 2023 International Conference on Electrical, Computer and Energy Technologies (ICECET), Cape Town, South Africa, 2023, pp. 1-9, doi: 10.1109/ICECET58911.2023.10389273.

[6] A. Remesh, A. P. K and M. S. Sinith, "Symbolic Domain Music Generation System Based on LSTM Architecture," 2022 Second International Conference on Next Generation Intelligent Systems (ICNGIS), Kottayam, India, 2022, pp. 1-4, doi: 10.1109/ICNGIS54955.2022.10079872.

[7] M. J. Pathariya, P. Basavraj Jalkote, A. M. Patil, A. Ashok Sutar and R. L. Ghule, "Tunes by Technology: A Comprehensive Survey of Music Generation Models," 2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS), Coimbatore, India, 2024, pp. 506-512, doi: 10.1109/ICC-ROBINS60238.2024.10534029.

[8] P. Tiwari and S. Jha, "Music Generation with Long Short-Term Memory Networks from MIDI Data of Classical Music," 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), Bangalore, India, 2024, pp. 1-4, doi: 10.1109/ICITEICS61368.2024.10625468.

[9] J. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," CVPR, 2022.

[10] A. Ho, J. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," NeurIPS, vol. 33, pp. 6840–6851, 2020.

[11] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," NeurIPS, vol. 34, pp. 8780–8794, 2021.