Journal of Computer Science

Research Article

Vision Tune: A Deep Learning Framework for Sentiment Driven Video, Image and Music Creation

Lukesh Rameshpant Kadu, Manoj Deshpande and Vijaykumar Pawar

Department of Computer Engineering, AC Patil College of Engineering, Kharghar, Navi Mumbai, India

Article history
Received: 07-09-2025
Revised: 21-12-2025
Accepted: 15-02-2026

Corresponding Author:
Lukesh Rameshpant Kadu
Department of Computer Engineering, AC Patil College of Engineering, Kharghar, Navi Mumbai, India
Email: lukesh.kadu@sakec.ac.in

Abstract: Artificial intelligence has enabled powerful generative models for text, images, video, and music, yet most tools still operate independently without a unified, multi-modal workflow. This article proposes an integrated AI framework, Vision Tune, that consolidates these isolated capabilities into a single, sentiment-aware platform for end-to-end media creation. The system leverages deep learning and multi-scope AI models to automatically generate written content, images, videos, and music for both creative and analytical applications, while emphasizing scalability, modular design, and user-centric interaction. By supporting cross-domain media synthesis and sentiment-driven customization, the framework targets real-world use cases in marketing, education, entertainment, and content production, where coordinated multi-modal outputs can enhance engagement and productivity. Beyond unification, the work highlights how the proposed architecture advances current AI media pipelines by reducing tool fragmentation, enabling cross-modal consistency, and providing a foundation for future extensions such as real- time generation, personalization, and human AI collaborative creation.

Keywords: Artificial Intelligence, Text generation, Image Synthesis, Video Production, Music Composition, Multi-Modal Media Generation

Introduction

The advancement of AI technology has significantly increased the capacity to generate text, photos, movies, music, and more. Despite the fact that significant progress has already been made in each area, AI systems sometimes work in silos, which means that several tools must be used to process different types of media. This project aims to address the challenge of creating Artificial Intelligence (AI) that effectively unifies various systems into a single platform that can generate multi-modal media in a single step toward analytical and creative applications. In this work, the term multi-scope AI models refer to an interconnected family of specialized models that collectively operate across several modalities and tasks rather than a single monolithic network. Concretely, Vision Tune coordinates large language models for script and dialogue generation, diffusion-based models for images and video frames, and sequence models for music, all exposed through a shared interface that supports cross-modal conditioning and orchestration. This design allows individual components to focus on text, visuals, or audio while still contributing to a coherent, sentiment-aligned media experience at the system level (Yang et al., 2025; Pathariya et al., 2024).

Background

The development of AI thus far has demonstrated that machines can create remarkably realistic media material. While natural language processing tools like GPT are better at creating text, DALL·E and GANs, which generate images, have transformed visual content. Artificial intelligence is also revolutionizing music creation and video editing through the use of deep learning, which transforms intricate procedures into automated jobs.

Despite all of that progress, these systems continue to function in isolated silos without a single center where tools for creating text, images, videos, and sounds can be used concurrently. People in these situations are forced to rely on many tools in order to create cross-content, which lowers productivity and inhibits creativity.

Motivation

Our main goal for developing an integrated AI solution is to address issues with the various media creation tools and assist in enabling smooth and intricate procedures. Because users are unable to create multi modal content text, images, videos, and music on a single platform, the current system not only makes the creative process more difficult but also reduces productivity.

DOI: 10.3844/jcssp.2026.1476.1483

Furthermore, where the intricate synthesis of several media is required, such as in marketing, education, entertainment, and virtual world simulation, this integrated method would promote AI advancements. By meeting this need, AI-based media production for marketing, education, and entertainment can be revolutionized. More significantly, it will open up new opportunities for creative and analytical work.

Contribution and Novelty

The primary contribution of Vision Tune lies in its integrated architecture that unifies text, image, video, and music generation into a single, sentiment-aware media creation pipeline, rather than proposing yet another standalone model for a single modality. The framework introduces a modular orchestration layer that manages data flow between heterogeneous generative models, enabling cross-modal consistency (e.g., aligning video scenes, background music, and narration to the same emotional profile) while remaining extensible to new back-end models. Compared with existing tools that address text to video, music generation, or image synthesis in isolation, Vision Tune focuses on end-to-end workflow integration, user centric interfaces, and cross- domain applications such as marketing campaigns, educational content, and immersive storytelling.

Related Work

Thanks to developments in generative artificial intelligence, machines can now produce any type of material, including writing, music, films, and photos. The top AI-powered models that have gained attention are as follows.

Generative Adversarial Networks (GANs)

The discriminator and generator are two essential components of GANs, which were developed by Good fellow and his colleagues in 2014. They are now widely used for style transfers, data augmentation, and the creation of pictures and films. While VideoGAN and MoCoGAN have used GANs to create videos, more recent versions such as Style GAN and Big GAN have improved the production of high-resolution photos.

Diffusion Models

When it comes to building images and movies by gradually eliminating noise, Stable Diffusion, DALL·E, and Imagen GANs have performed better than their predecessors. These

Methods generate richly detailed contextual images from text inputs by using denoising diffusion probabilistic models (DDPMs). More recent video models, such as Make A Video and Stable Video Diffusion, improved upon earlier models by gradually producing cohesive films.

Recurrent Neural Networks (RNNs) and LSTM for Music Generation

RNN models in particular, Long Short Term Memory (LSTM) networks have been the most frequently used in the music industry for creating melodies and harmonies. LSTM's capabilities have already been leveraged by a number of programs, such as Muse Net and Magenta's Music VAE, which enable it to produce music for a variety of instruments and styles. In any case, Music Transformer and similar tools have recently advanced transformer-based music models and significantly enhanced the longitudinal coherence of the music they generate over time.

Large Language Models (LLMs)

Transformers are applied by LLMs such as GPT-4, LLaMA, and PaLM. Using a pre-trained transformer architecture, LLM-powered chatbots and writers are constructed using GPT-4, LLaMA, and PaLM models and need to understand human speech when given text as a stimulus. These models, which emphasize the reinforcement learning with human feedback (RLHF) technique, mostly rely on the chatbot to provide self-comprehensible and logical responses.

Methods

Text to Video Generation

One of the newest endeavours in AI development is the production of films from text descriptions. Using deep learning methods including Natural Language Processing (NLP), GANs, and video editing, this section of the research project aims to create an advanced and efficient T2V model. As new diffusion and transformer construction models have been developed, the videos' quality and correlation have greatly increased (Yang et al., 2025).

In order to improve the realism and coherence of the movies created, recent advancements in T2V creation have used model diffusion, enormous scale transformers, and latent variable modelling. Spatial and temporal learning techniques are used by implementations like Model Scope Text-to-Video (Pathariya et al., 2024) and CogVideoX (Yang et al., 2025) to generate high-quality films. Additionally, efforts that concentrate on high-fidelity video synthesis, such as Align Your Latent (Wang et al., 2023), have significantly improved latent space models. The field is always expanding; for example, issues pertaining to motion continuity, temporal coherence, or semantic coherence between text and video are being resolved.

Methodology

- Four essential elements make up the video generating system we have put in place

DOI: 10.3844/jcssp.2026.1476.1483

- Script Generation: This stage creates an organized and organic-sounding description using the Google Gemini API
- Image Generation: The images are produced using Stable Diffusion (runwayml/stable-diffusion-v1-5). (Pathariya et al., 2024)
- Audio Generation: The generated text is turned into speech in this stage using gTTS
- Video Compilation: Finally, OpenCV and MoviePy are used to merge the pictures, sound, and subtitles

Deeper Architecture and Data Flow Details

The system follows a pipeline in which user input (prompt plus optional sentiment or style constraints) is first processed by an LLM-based controller that produces structured scripts, scene breakdowns, and semantic tags for downstream components (Fig. 1). These tags, which encode entities, actions, and emotional tone, are passed to a diffusion-based image/video generator to create key frames or clips, while a sequence model (e.g., LSTM or transformer-based music model) generates background music conditioned on the same high-level descriptors. A text-to- speech module then converts the script into narration, and a compositor component aligns narration, visuals, and music on a shared timeline using explicit timing metadata derived from the script segmentation and scene plan (Pathariya et al., 2024).

At the implementation level, each modality is encapsulated as a service with clearly defined input-output contracts (e.g., JSON-based specifications for prompts, timestamps, and sentiment Scores), which simplifies substitution of alternative models such as different diffusion backbones or LLM providers.

The orchestration layer maintains a central project state that records prompt history, generated assets, and alignment metadata, enabling iterative refinement (for example, regenerating only the music while preserving the same video and narration) without restarting the entire pipeline (Pathariya et al., 2024).

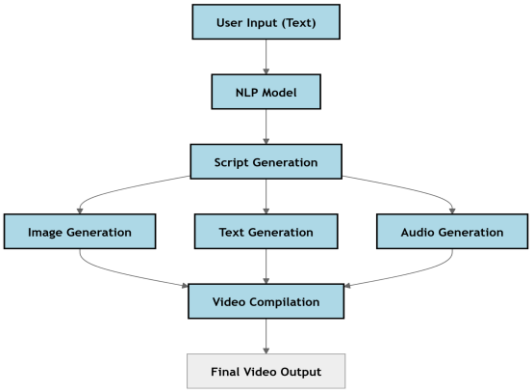

Fig. 1: User flow of the video generation model

Technology Used

- Natural Language Processing (NLP): Google Gemini
- API is utilized to generate structured text
- Image Generation: The optimized version of Stable
- Diffusion v1.5 is used to generate frame-based images
- Audio Processing: Google Text to Speech (gTTS) is used to generate voices
- Video Processing: OpenCV and MoviePy make it simple to compile and caption videos

Text Processing and Image Generation for Processing of Videos

The system starts with the input text, which is divided into readable sections that are times tamped and surrounded by factual and logical borders. Every two seconds, an image is represented by each sentence. The quality of the photographs is then enhanced by employing Stable Diffusion, which is based on the latent diffusion model (Pathariya et al., 2024) 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 (Pathariya et al., 2024).

Audio Generation

We utilize GTTS (Google Text-to-Speech) to produce narration that blends in perfectly with the script in order to produce realistic audio (Fig. 2). To ensure that everything flows organically, we meticulously modify the duration of each audio segment to precisely match the visual transitions. (Wang et al., 2023).

If Vd is the duration of the video and Ad is the duration of the created audio, then:

$$Ad \approx Vd$$

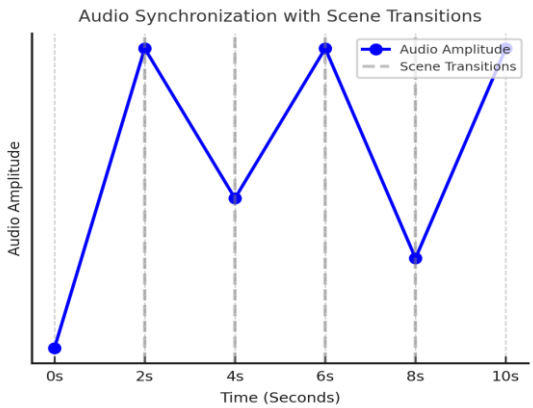

Fig. 2: Comparison of audio amplitude with time

Video Compilation and Captioning

We ensure smooth transitions between 30-40 frames per second by blending audio and pictures with MoviePy and OpenCV. To improve clarity, we additionally use the TextClip method to overlay captions on each frame (Pathariya et al., 2024).

If Nimg is the number of pictures and FPS is the number of frames per second, the video's overall time can be computed as follows:

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

Results and Discussion

We successfully developed a text-to-video conversion model utilizing the stable diffusion/ runwayml framework along with gtts and moviepy. However, the model requires enhancements to produce motion videos featuring more detailed content and script generation (Tables 1 and 2).

Comparison of Models Used

AI-Powered Music Creation

Recently, artificial intelligence technologies have made significant advancements across various research domains, and one area that has seen considerable impact is music. The evolution of deep learning techniques, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, has enabled the automated composition of well-structured musical works (Pawar and Raut, 2025). In contrast to other music generation methods that rely on specific rules, LSTMs are better equipped to understand long term dependencies in music sequences (Remesh et al., 2022).

This section of the research paper introduces an AI-driven solution that can create classical music by training LSTMs on MIDI files. The system is designed to automate the composition of works in the unique styles of Bach, Mozart, and Beethoven by initially identifying the necessary musical patterns.

Table 1: Comparison of different models used for video generation

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 (Pathariya et al., 2024)
Audio Generation	gTTS	Tacotron, VITS	Faster synthesis speed (Wang et al., 2023)
Video Compilation	MoviePy, OpenCV	FFmpeg	More flexible processing (Wang et al., 2023)

Table 2: Performance of the video generation model

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

Deep Learning for Music Generation

The use of AI for music creation has progressed from systems based on rules to those utilizing neural networks. Earlier techniques depended on Markov models and Hidden Markov Models (HMMs), yet they faced challenges in capturing long-term relationships (Pawar and Raut, 2025). Recent research underscores the success of LSTMs in producing music that maintains stylistic coherence (Remesh et al., 2022). Other architectures, such as Generative Adversarial Networks (GANs) and Transformers have been investigated, but they demand considerable computational power (Tiwari and Jha, 2024).

Symbolic Music Processing

Symbolic music generation involves converting every musical element into a digital format. Music21 serves as an effective tool for this process by enabling the access and modification of MIDI files, which is essential for deep learning applications (Dhariwal and Nichol, 2021).

Methodology

Dataset Preparation

We have put together a collection of MIDI-formatted classical music arranged by composer. Music21 processes each file, Storing the pertinent note sequences. Among the pre-processing actions are:

- Choosing the appropriate chords and notes
- Transforming the chosen sequences into a numerical representation
- Separating the final sequences into primary training and validation datasets

Model Architecture

Uses of LSTM-based architecture that has been taught to anticipate a sequence's next note. The model includes:

- A layer for embedding that turns musical notes into vectors
- Two LSTM layers for temporal dependency learning

DOI: 10.3844/jcssp.2026.1476.1483

- Softmax activation in a thick output layer for note prediction (Remesh et al., 2022)

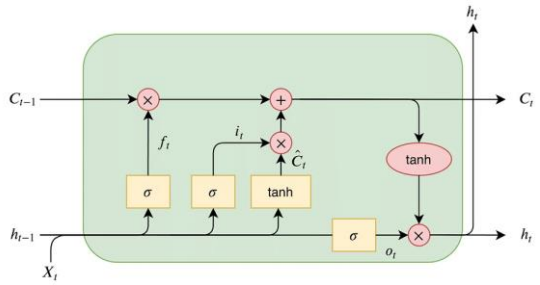

Fig. 3: Structure of a Long Short-Term Memory (LSTM) unit

Training Process

The Adam optimiser and categorical cross-entropy loss are used to train the model. To maximise model performance, hyper parameter adjustment is done. The definition of the categorical cross-entropy loss function is:

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

Where:
y_i is the true label (ground truth note index).
$\hat{y}_i$ is the predicted probability
□ is the total number of note classes

Web Application

Users can choose a composer and create music in their style using a web interface built using Flask. The process consists of:

- User-inputted composer selection
- The matching trained model is loaded
- Using seed input to generate a series
- Making a MIDI file out of the output so it can be played back

Our model does a good job of capturing the stylistic features of other composers, but it has to be improved in order to produce longer, more varied works. Sequence coherence may be improved by using transformers (Tiwari and Jha, 2024).

Text-to-Image Generation

In the area of artificial intelligence, the capacity to translate descriptions into graphics actually changed the game by creating new opportunities in areas like digital painting, game development, content creation, and advertising. Recent advances in diffusion models, including Stable Diffusion, DALL·E, and Imagen, have improved the quality, consistency, and management of AI-generated images significantly. In contrast to previous generative models, which frequently had trouble matching meanings and preserving details, diffusion models gradually improve an image from a noisy beginning towards greater fidelity while allowing for greater creative freedom.

This article describes a Flask-based web application that creates graphics from text prompts using Stable Diffusion v1.5. Anyone can simply create a few photographs, enter their descriptions, and download their own favourites. Regardless of whether you have a GPU or CPU, the solution promises smooth performance by providing simple access to pre-trained models through Hugging Face's model hub.

Deep Learning for Image Generation

From early rule-based systems to advanced deep-learning techniques, AI-powered image production has advanced. Though they frequently had trouble with mode collapse and fine-grained control over picture attributes, traditional techniques like Generative Adversarial Networks (GANs) delivered outstanding results. By gradually enhancing images from random noise, recent research demonstrates how well diffusion models work to overcome these obstacles (Pathariya et al., 2024).

Diffusion Based Text to Image Models

Latent variable modelling is used by diffusion models, including Stable Diffusion, to generate high-quality, semantically coherent images. Diffusion models offer a compromise between creative versatility and computational efficiency when compared to transformer-based models such as DALL·E 2. Studies have shown that these models perform better than GANs in terms of realism and output diversity (Pawar and Raut, 2025).

Methodology

A Client-server architecture makes up the implemented system, with the frontend managing user interactions and the backend handling requests for picture production. There are three primary parts to the system's structure:

- Frontend (User Interface): A simple HTML/CSS interface that allows users to create images and enter text descriptions. Users can download their desired results from the dynamically displayed generated photographs on the webpage
- Backend (Flask API): A Flask server with a Python foundation that manages HTTP requests, interprets text input, and uses the Stable Diffusion model to create images
- In order to facilitate retrieval, the created photos are briefly saved on the server
- Pipeline for Image Generation: By using repeated denoising to sharpen images, the Stable Diffusion

model gradually improves visual quality while staying in line with the input prompt (Pawar and Raut, 2025)

Workflow Diagram for Image Generation

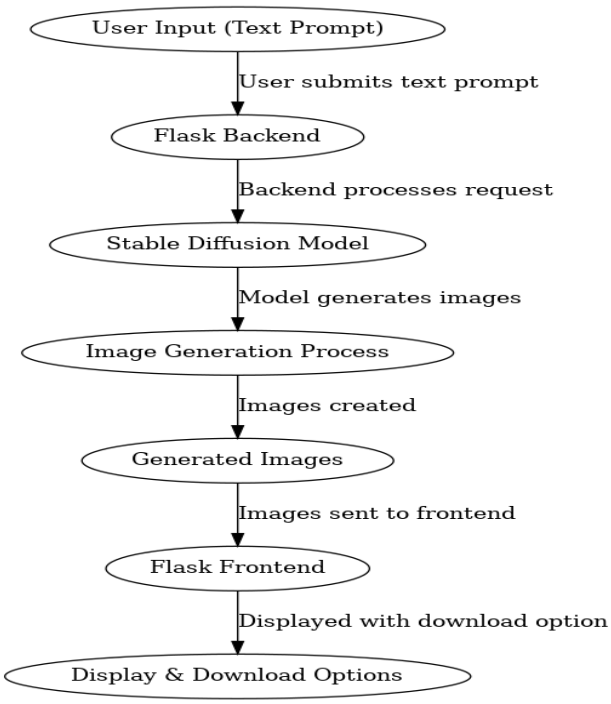

Fig. 4: The user flow of the image generation model

The model was assessed on the basis of user experience, computational efficiency, and output quality. A number of important conclusions were drawn:

- Image Quality: High-resolution, semantically correct images that complemented verbal descriptions were routinely created using Stable Diffusion
- Computational Performance: Even in CPU-based contexts, the system generated images in a fair amount of time by striking a compromise between efficiency and quality
- User Experience: By allowing users to choose from a variety of results, multiple image production per prompt enhanced freedom and creativity

Comparison of Models Used

We must contrast Stable Diffusion with other text-to-image models in order to fully comprehend its effectiveness (Table 3). Here is a quick look at a few of them: Stable Diffusion is the best option for this project because of its exceptional accessibility, quality, and flexibility balance (Li et al., 2020).

AI-Driven Chatbot

The rapid development of AI-based chatbots has truly transformed how humans engage with computers by enabling intelligent and context-aware user interaction.

Natural language processing and machine learning techniques are used by AI-based chatbots to understand what is being said and react accordingly. Automated content generation, virtual assistants, and customer service all use it. In addition to helping users create textual material, Vision Tune's state-of-the-art AI-powered chatbot can generate conversational text and enable seamless interactions with its multi-modal media generating platform.

Methodology

The model was evaluated based on output quality, computational efficiency, and user experience. Several key findings emerged: The model was assessed on the basis of user experience, computational efficiency, and output quality. A number of important conclusions were drawn:

- Language Model: To produce responses that sound human, the chatbot uses a precisely calibrated Large Language Model (LLM), such as Google's Gemini or OpenAI's GPT-4
- Integration with Media Creation: To allow users to request material in a natural way, it is integrated with other media creation tools, such as text-to-image, text-to-video, and text-to-music
- User Intent Recognition: To categorise user requests and provide appropriate responses, the chatbot employs named entity recognition (NER) and intent recognition
- Implementation Framework: The Flask framework and Python were used to create the chatbot, which easily connects to media generating APIs

Expanded Evaluation Strategy

To move beyond purely qualitative demonstrations, Vision Tune is evaluated along three dimensions: perceptual quality, cross-modal coherence, and system efficiency. Perceptual quality is assessed using established metrics such as FID for visual content and user preference studies for audio and multi-modal outputs, where participants rate clarity, aesthetic appeal, and emotional fit between modalities. Cross-modal coherence is measured through synchronization error between narration and scene changes, sentiment alignment scores computed by independent classifiers across text, audio, and frames, and expert judgments on narrative consistency. System efficiency is characterized by end-to-end Inference time, resource utilization, and scalability under concurrent user sessions, with comparisons against baseline workflows that chain separate tools for each modality.

Table 3: Comparison of models used for Image generation

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

Future Scope

Although Vision Tune demonstrates the feasibility of a unified, sentiment-driven media generation platform, several limitations remain. The current system depends heavily on pre- trained foundation models and third-party APIs, which can introduce latency, cost, and reproducibility concerns, and it does not yet guarantee fine-grained control over motion dynamics in video or long-term musical structure for extended compositions. Future work will focus on tighter temporal modelling for video, richer personalization based on user profiles and interaction history, and more robust human-in-the-loop tools for editing and steering multi-modal outputs, alongside a deeper exploration of ethical issues such as bias, content safety, and provenance tracking in generated media.

The field of creating multi-modal media using AI is expanding quickly and has a lot of room to grow. Among the potential future paths are:

- Improved Temporal Consistency in Video Creation: Addressing motion coherence in systems that convert text to video
- AI-powered virtual assistants that can provide customised material based on user preferences are known as personalised AI media assistants
- Improving transformer and LSTM models for real-time AI-generated music production
- Cross-Modal AI Improvements: Enabling models to seamlessly blend different types of generated content (e.g., an automatically generated music video based on lyrics)

Conclusion

This created system proposes an AI-enabled system that combines the creation of images, videos, music, and chatbots into a single system, providing users with a one-stop shop. Utilising cutting-edge generative models such as GANs, LSTMs, LLMs, and Stable Diffusion, this solution addresses the shortcomings of fragmented media generation tools. The outcomes demonstrate the viability of automated media synthesis with encouraging improvements in quality and efficiency. More effort will be put into improving the system's functionality and broadening its potential applications in many fields.

This AI-enabled multi-modal system offers a unified experience by integrating text, image, video, and music generation into a single ecosystem, eliminating the need for multiple tools. Through cross-modal creativity, it seamlessly combines text, visuals, and audio for instance, generating videos with custom background music and AI-powered narration while enhancing efficiency and productivity by automating creative workflows. Its personalization capabilities allow content to be tailored to user preferences, sentiment, and context, delivering highly customized outputs. Designed for scalability, the system supports diverse industries such as entertainment, marketing, education, and customer engagement. By fostering innovation, it empowers creators with advanced generative models that push the boundaries of design and storytelling, while remaining cost-effective by reducing reliance on multiple platforms and resource-intensive manual work.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Lukesh Rameshpant Kadu: Developing the main idea, research design, and approach. Contributions to conception and design, and/or acquisition of data, and/or analysis and interpretation of data.

Manoj Deshpande: Gathering data, conducting experiments/surveys, and performing analysis.

Vijaykumar Pawar: Preparing the manuscript, drafting, revising, and finalizing the paper.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Dhariwal, P., & Nichol, A. (2021). Diffusion Models Beat GANs on Image Synthesis. *Advances in Neural Information Processing Systems*, 8780–8794.

Li, Y., Pan, S., Zhang, K., Yin, Y., & Chen, Q. (2020). Financial sentiment learn- ing for stock return prediction. *IEEE Access*, *8*, 78899–78908.

Pathariya, M. J., Basavraj Jalkote, P., Patil, A. M., Ashok Sutar, A., & Ghule, R. L. (2024). Tunes by Technology: A Comprehensive Survey of Music Generation Models. *Proceedings of the 2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS)*, 506–512. https://doi.org/10.1109/iccrobins60238.2024.10534029

Pawar, V., & Raut, R. (2025). Comparative Study of Algorithms for Sentiment-Based Stock Market Prediction. *Proceedings of the 2025 4th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0*, 1–5. https://doi.org/10.1109/otcon65728.2025.11070531

Remesh, A., Paul, A., & Sinith, M. S. (2022). Symbolic Domain Music Generation System Based on LSTM Architecture. *Proceedings of the 2022 Second International Conference on Next Generation Intelligent Systems (ICNGIS)*, 1–4. https://doi.org/10.1109/icngis54955.2022.10079872

Tiwari, P., & Jha, S. (2024). Music Generation with Long Short-Term Memory Networks from MIDI Data of Classical Music. *Proceedings of the 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, 1–4. https://doi.org/10.1109/iciteics61368.2024.10625468

Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., & Zhang, S. (2023). ModelScope Text-to-Video Technical Report. *Computer Vision and Pattern Recognition*. https://doi.org/10.48550/arXiv.2308.06571

Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Zhang, Y., Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., & Tang, J. (2025). CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *ArXiv Preprint ArXiv:2408.06072*. https://doi.org/10.48550/arXiv.2408.06072