Volume: 13 Issue: VII Month of publication: July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73064

Realistic Video Synthesis from Audio using GAN

Sravanthi Karne[1], G. Narasimham[2]
[1]*Post-Graduate Student, Department of Information Technology, Data Science, Jawaharlal Nehru Technological University, Hyderabad, India*
[2]*Associate Professor, Department of Information Technology, Jawaharlal Nehru Technological University, Hyderabad, India*

Abstract: Realistic video generation from audio input is a challenging and emerging domain in the intersection of natural language processing, computer vision, and generative modeling. The ability to automatically generate coherent and visually compelling video content from raw audio has promising applications in media creation, virtual education, assistive technologies, and entertainment. Manual video creation remains time-consuming and skill-intensive, while automated solutions often lack semantic alignment and visual realism.

To address this gap, this project proposes an end-to-end intelligent pipeline that synthesizes realistic video content from audio input using Generative Adversarial Networks (GANs). The system begins by transcribing the user's audio using OpenAI's Whisper ASR model, followed by extracting meaningful textual descriptions via a language model (e.g., Groq LLaMA or OpenAI GPT). The script is used to generate key search terms for visual content retrieval, sourcing high-quality imagery from the Pexels API. Speech is generated using Edge TTS, and synchronized subtitles are created.The images are compiled into a dynamic video using MoviePy, and visual quality is further enhanced using Real-ESRGAN for super-resolution. The final output is a short, high-resolution, contextually accurate video with natural narration and relevant imagery. This work demonstrates the effectiveness of combining audio processing, NLP, GAN-based enhancement, and open content APIs to automate realistic video generation from scratch.

Keywords: Audio-to-Video Generation, Generative Adversarial Networks (GANs), Whisper ASR, Speech Recognition, Text-to-Video Synthesis, Image Super-Resolution, Real-ESRGAN.

I. INTRODUCTION

In an era dominated by digital media and short-form content, the demand for fast, automated, and high-quality video generation tools has never been higher. Creating explainer or educational videos traditionally involves multiple tools for scripting, narration, editing, and visual enhancement. These processes often require specialized software and expertise, creating a barrier for casual users, educators, and content creators with limited resources. Moreover, searching for quick video explanations online typically leads to long YouTube videos or irrelevant search results, offering little support for users seeking brief, focused insights.

To overcome these challenges, this project presents a fully automated system capable of transforming spoken audio into realistic, high-resolution videos with minimal user intervention. By combining state-of-the-art models and APIs, such as Whisper for transcription, Groq's language model for intelligent script generation, Edge TTS for speech synthesis, and Real-ESRGAN for visual enhancement, the system ensures a streamlined and coherent video creation process. The use of contextual visual assets retrieved from the Pexels API ensures that the resulting videos are not only informative but also visually engaging. This approach reduces the dependency on manual editing and democratizes multimedia content creation for a wider audience.

A. Objective

1) To build an automated system that takes audio input and turns it into a complete, high-quality video, helping reduce manual work and speed up the content creation process.

2) To make use of powerful AI tools—like Whisper for converting speech to text, Groq for generating meaningful scripts, Pexels API for finding suitable images, and ESRGAN for improving video quality—to create visually engaging and informative content.

3) To ensure the final video feels immersive by properly syncing audio with relevant visuals, offering a smooth and impactful viewing experience.

4) To provide an easy and innovative way for creators, educators, and media professionals to generate videos directly from audio, making the content creation process faster and more accessible.

II. LITERATURE SURVEY

1) TA2V(Text-AudiotoVideoGeneration–IEEETMM2024): This paper proposes a method to generate videos conditioned on both text and audio using 3D-VQGAN and diffusion models. It aligns with your approach where audio input is transcribed and processed to drive visual content generation. TA2V validates multimodal synthesis techniques, supporting your architecture's ASR → Script → Video flow.
2) IRE(ImprovedReal-ESRGAN–IEEEEarlyAccess2025): IRE enhances the Real-ESRGAN model for better visual upscaling, reducing artifacts and improving sharpness. It directly supports your project's stage of enhancing generated video resolution. Useful for producing clearer, high-resolution outputs from initially low-res visuals.
3) Audio PUGAN (ICASSP 2021): This paper presents a GAN model for high-fidelity audio generation using progressive upsampling. While your project focuses more on TTS than raw audio generation, PUGAN provides useful insights into GAN-based audio modeling. It informs the audio enhancement and synthesis parts of your pipeline.
4) SRGAN(Super-ResolutionGAN–Ledigetal.,2017): SRGAN is a foundational work in image super-resolution using GANs. It's the basis for Real-ESRGAN, which you use to enhance video quality. Understanding SRGAN helps contextualize how adversarial loss contributes to realism in upscaled frames.

III.METHODOLOGY OF THE PROPOSED SYSTEM

A. *Proposed System*

The proposed system aims to automate the process of creating short, high-quality videos using audio input as the starting point. It combines speech recognition, text generation, image retrieval, and video enhancement in a streamlined pipeline. When a user provides an audio file, the system first transcribes it into text using OpenAI's Whisper model. A meaningful script is then generated from the transcribed text, which is converted into natural-sounding speech using Edge TTS.

To match the script content visually, the system uses the Pexels API to fetch relevant background images. These images, along with the synthesized audio and captions, are combined into a complete video using MoviePy. To ensure the visual output is sharp and realistic, Real-ESRGAN is applied to enhance image resolution.

This system reduces manual effort and enables users to generate engaging video content with minimal input, making it useful for educators, content creators, and storytellers who want to produce multimedia content efficiently.

B. *System Architecture*

1) Audio Input: User provides a .wav file as input, serving as the base for content generation.
2) ASR with Whisper: Converts speech to text with word-level timestamps for synchronization.
3) Script Generation (Groq API): Refines raw transcript into a clean, structured script using LLaMA/GPT models.
4) Keyframe Extraction (Pexels API): Extracts keywords from the script to fetch relevant royalty-free visuals.
5) Video Assembly (MoviePy): Combines audio, visuals, and captions into a timeline using Python's MoviePy.
6) Visual Enhancement (Real-ESRGAN): Upscales low-resolution images/videos to improve visual quality.

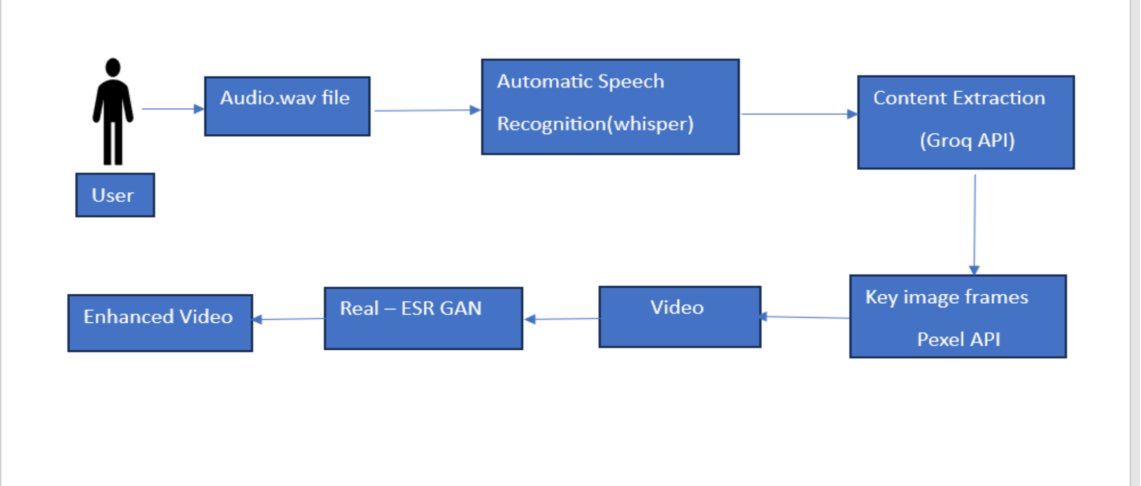

Audio.wav file
Automatic Speech Recognition(whisper)
Content Extraction (Groq API)
User
Key image frames Pexel API
Video
Real – ESR GAN
Enhanced Video

Fig: System Architecture

Final Output: Generates a polished video with aligned audio, visuals, and captions, ready for download or preview.
Algoritham

□ Purpose: Real-ESRGAN enhances low-quality images by increasing resolution and clarity without manual editing.
□ Input: Takes degraded images affected by blur, noise, compression, or downsampling.
□ Preprocessing: Converts images into normalized tensors and divides them into patches for training.

□ Generator

- Uses Residual-in-Residual Dense Blocks (RRDBs) to extract multi-level features.
- Initial layers detect basic patterns; upsampling layers increase resolution and detail.

□ Discriminator

- Applies Relativistic Discriminator logic to compare generated and real images.
- Uses convolutional layers to detect fine differences in textures and edges.

□ Workflow

- Image is formatted and passed through deep feature extractors.
- RRDBs progressively upscale and enhance the image.
- Noise/artifacts are removed, and final adjustments are made for sharpness and contrast.

□ Output: High-resolution, visually improved image suitable for display or video processing.

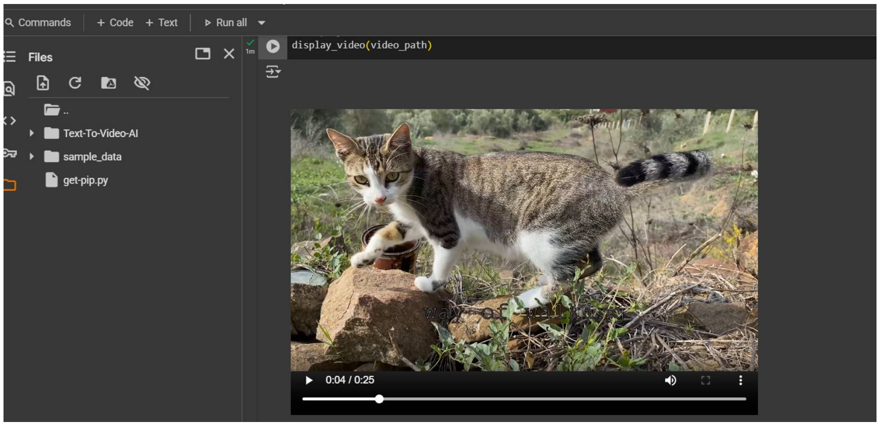
IV. IMPLEMENTATION AND RESULTS

A. Implementation steps

1) Audio Input Collection: The user begins by uploading an audio file (in .wav format). This serves as the input source for the entire video generation process.
2) Speech-to-Text Conversion (ASR): The uploaded audio is transcribed into text using the Whisper model. This model produces accurate transcripts along with timing information, which helps with later synchronization.
3) Script Generation using LLM: The raw transcript is passed to a language model (accessed through the Groq API), which condenses and rewrites it into an engaging and structured script suitable for visual storytelling.
4) Text-to-Speech Synthesis: The generated script is transformed back into spoken audio using Edge TTS. This new narration is clearer and optimized for synchronizing with video elements.
5) Visual Content Retrieval: Keywords and concepts are extracted from the script, and relevant images or short videos are fetched using the Pexels API. These visual assets represent the script content.
6) Video Creation using MoviePy: The narration audio and visuals are combined into a timeline using MoviePy. Subtitles can also be added for better accessibility. The result is a fully composed video.
7) Visual Upscaling with Real-ESRGAN: The assembled video is enhanced using Real-ESRGAN, which applies deep learning techniques to upscale the resolution and improve visual clarity.
8) Output Rendering: The final, high-quality video is exported and made available to the user for download or sharing.

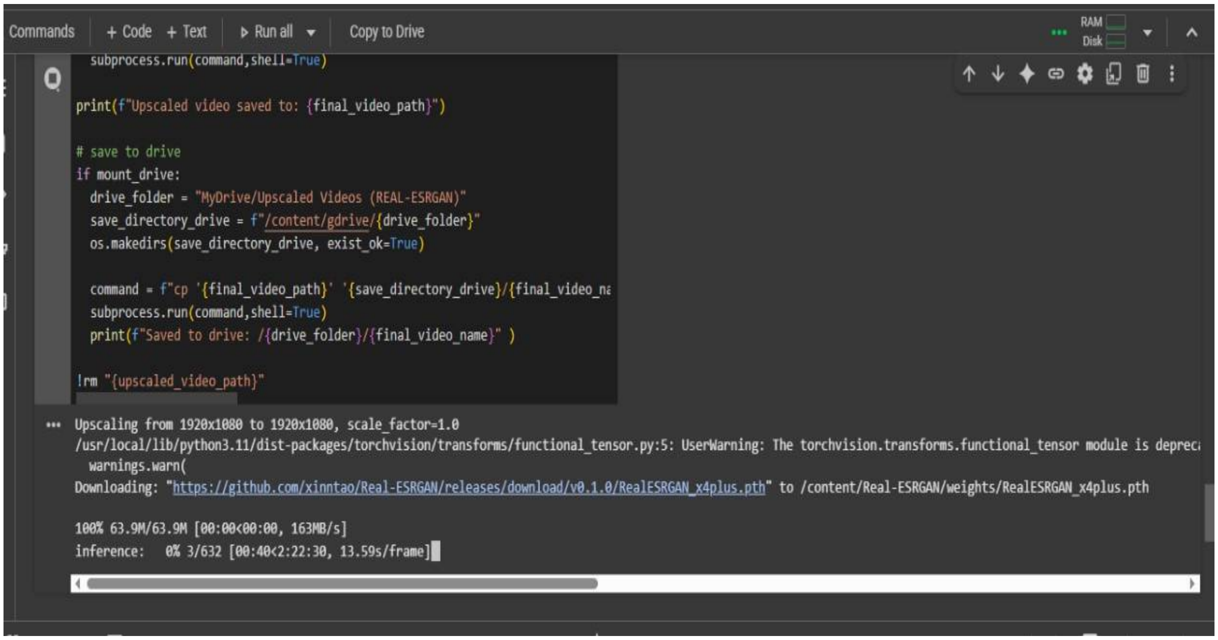
B. Results

The results section presents the practical outcomes obtained from implementing the *"Realistic Video Synthesis from Audio using GAN"* system. It highlights how each module—ranging from audio transcription and script generation to image retrieval, video composition, and enhancement—performs in an integrated pipeline. The system was tested using various audio inputs, and the final video outputs were evaluated in terms of synchronization, quality, and realism. These results demonstrate the capability of the proposed approach to automate the multimedia generation process efficiently, offering visually appealing and contextually accurate videos with minimal user input. This section provides both qualitative and quantitative insights that validate the effectiveness and relevance of the system.



```
display_video(video_path)
```

Fig: generated video



```
  subprocess.run(command,shell=True)

print(f"Upscaled video saved to: {final_video_path}")

# save to drive
if mount_drive:
  drive_folder = "MyDrive/Upscaled Videos (REAL-ESRGAN)"
  save_directory_drive = f"/content/gdrive/{drive_folder}"
  os.makedirs(save_directory_drive, exist_ok=True)

  command = f"cp '{final_video_path}' '{save_directory_drive}/{final_video_na
  subprocess.run(command,shell=True)
  print(f"Saved to drive: /{drive_folder}/{final_video_name}" )

!rm "{upscaled_video_path}"
```

```
Upscaling from 1920x1080 to 1920x1080, scale_factor=1.0
/usr/local/lib/python3.11/dist-packages/torchvision/transforms/functional_tensor.py:5: UserWarning: The torchvision.transforms.functional_tensor module is depreca
  warnings.warn(
Downloading: "https://github.com/xinntao/Real-ESRGAN/releases/download/v0.1.0/RealESRGAN_x4plus.pth" to /content/Real-ESRGAN/weights/RealESRGAN_x4plus.pth

100% 63.9M/63.9M [00:00<00:00, 163MB/s]
inference:   0% 3/632 [00:40<2:22:30, 13.59s/frame]
```

Fig: Upscaling Generated video using RealESRGAN

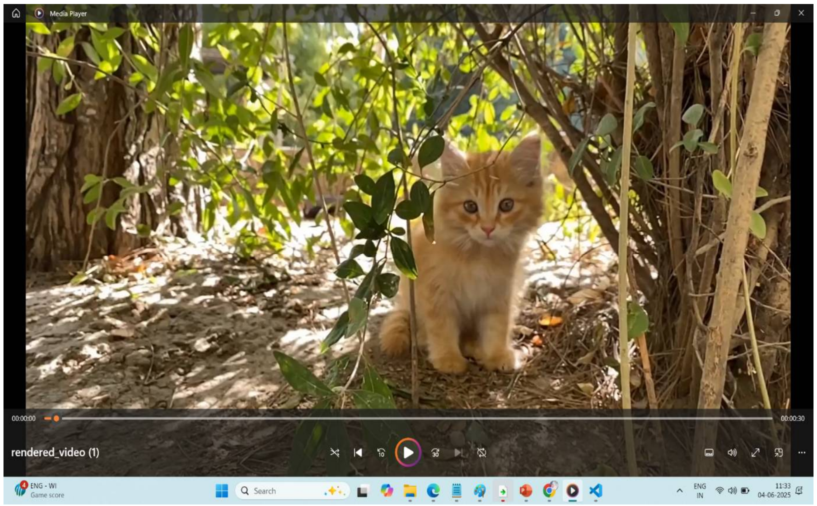


Fig: GAN genereted video

V. LIMITATIONS AND FUTURE SCOPE

While the system shows strong potential in automating video generation from audio input, it does face certain limitations, such as dependency on external APIs for image retrieval and challenges in handling noisy audio, which can reduce transcription accuracy. Moreover, longer processing times for high-resolution outputs may impact real-time usability. However, the project has promising future scope—such as expanding support for multiple languages, incorporating avatar-based animations, and enabling real-time video previews or script customization—which could significantly enhance its usability and appeal across education, media, and digital content platforms**.**

VI. CONCLUSION

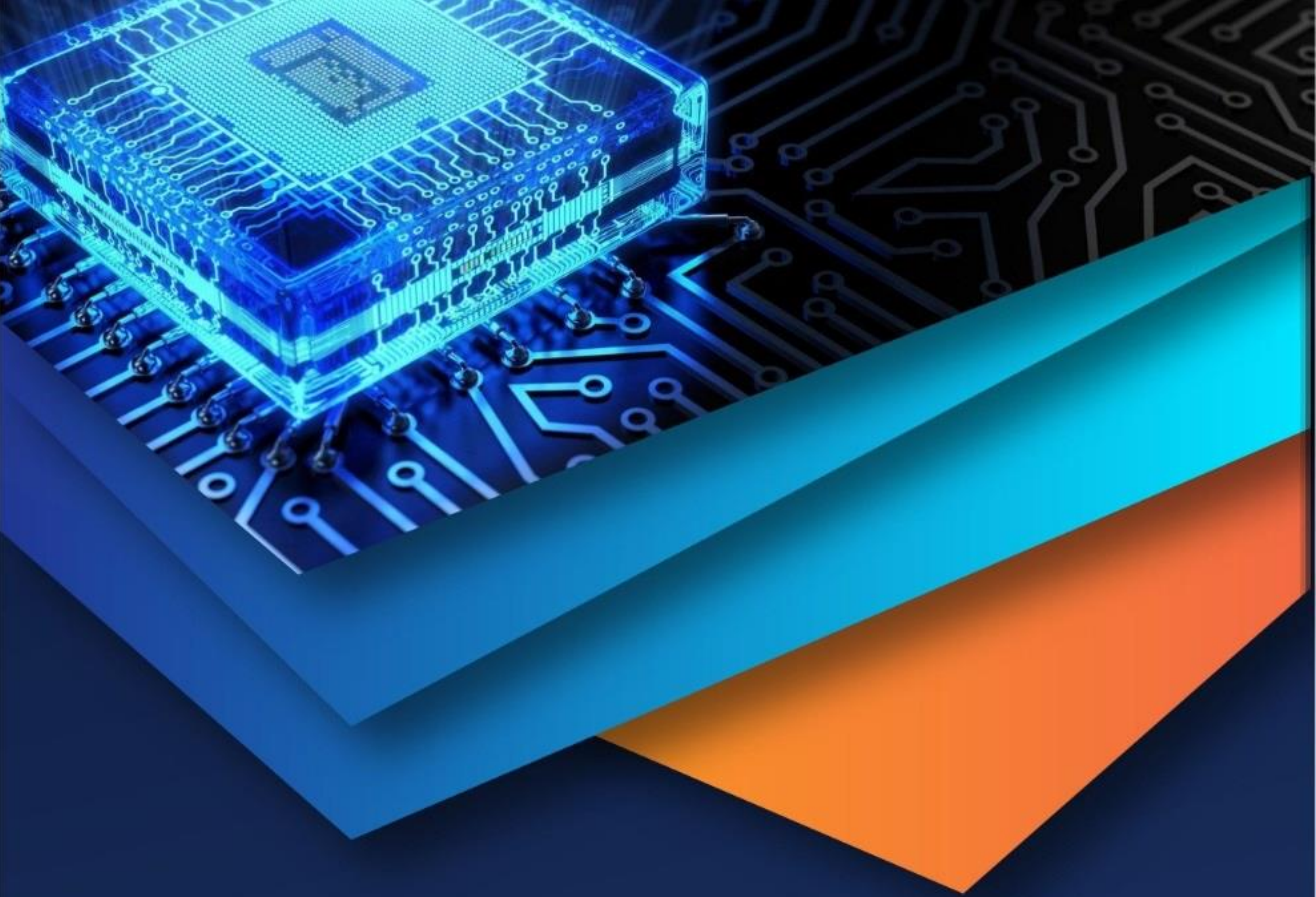
This project presents an automated end-to-end system for generating realistic and contextually accurate short videos from audio input. By integrating state-of-the-art tools such as Whisper for speech recognition, Groq/OpenAI for script generation, and Real-ESRGAN for visual enhancement, the system bridges voice-based input with coherent video synthesis. Each module—transcription, language modeling, image retrieval, TTS, and video assembly—contributes to a streamlined workflow requiring minimal user intervention. Designed for accessibility, the system simplifies content creation for users without technical expertise and is adaptable for applications in education, media, and infotainment. The results validate the feasibility of audio-to-video generation and highlight future potential in personalization, emotion-aware narration, and multilingual support.

VII. ACKNOWLEDGMENT

I sincerely thank Sri G. Narasimham sir for his continual support in helping me carry out this project. Also extend my heartfelt gratitude to the authors of the research papers reviewed in this study.

REFERENCES

[1] A. Radford et al., "Whisper: A large-scale, weakly-supervised model for speech recognition," arXiv, 2022. Available: https://arxiv.org/abs/2212.04356
[2] C. Ledig et al., "A GAN-based architecture for single-image super-resolution," in CVPR Proceedings, 2017, pp. 4681–4690.
[3] X. Wang et al., "Real-ESRGAN: A technique for improving low-quality images using adversarial training on synthetic examples," in ICCV Workshops, 2021.
[4] Y. Wu et al., "TA2V: Generating aligned video from audio and text using diffusion models," IEEE Transactions on Multimedia, 2024.
[5] Groq Inc., "Groq API for fast inference using LLMs (LLaMA/GPT)," 2024. [Online]. Available: https://groq.com
[6] □MoviePy Developers, "MoviePy: A Python library for editing video programmatically," 2023. [Online]. Available: https://zulko.github.io/moviepy/
[7] OpenAI, "Whisper: Open-source speech-to-text system," 2023. [Online]. Available: https://github.com/openai/whisper
[8] Pexels, "Pexels API: Access to royalty-free images and videos," 2023. [Online]. Available: https://www.pexels.com/api/



10.22214/IJRASET

INDEX COPERNICUS

45.98

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY