Since 1983

Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

Pathpulse.ai PREBULLS

THE X - FAIR HACKATHON

Problem Statement: Manual video creation slows scalability and wastes educator's time and effort.

DOMAIN: Open Innovation [AI&ML X Edutech]

Team Name: CamelCase

Team ID: xfair0078

AVINYA DeepQuantica MadHat CYBER BYTES WorqHat

TITLE / INTRODUCTION

Project Flash - Multimodal GenAI for Automated Educational Video Creation

Team Members:

Anannya Wakalkar - Front-End Developer

Gauri Gaikwad - Front-End Developer and UI/UX Designer

Abhinav Bombale - Backend & AI Pipeline Developer

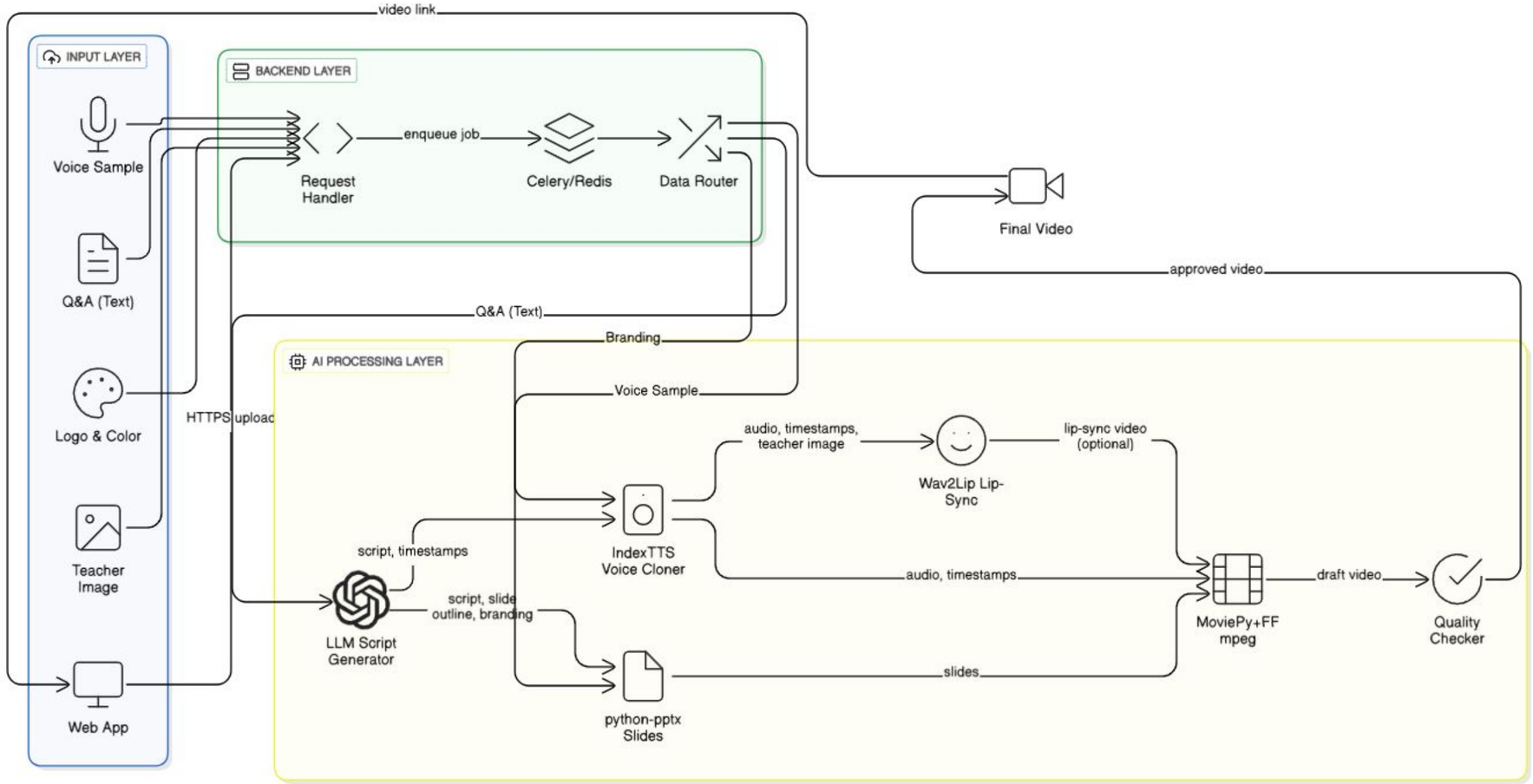
MIT World Peace University

Track : Hackathon [Open Innovation]

PROBLEM STATEMENT

- **Problem to be solved :** Manual question-solution video creation consumes excessive time and effort from educators.
- **Target Users**: Coaching institutes and online education platforms seeking to automate video-based solution creation at scale.
- **Importance or relevance**: A coaching institute with 50 teachers needs to record 1,000 question-solution videos weekly. Each 5-minute video takes ~20 minutes to prepare, resulting in over 300 teacher-hours lost per week — time that could be spent improving content quality or student interaction.

Proposed Solution



A multimodal generative AI system that automatically converts a question-and-answer pair into a fully synchronized, branded educational video — featuring a cloned teacher's voice and personalized, context-aware visuals.

FEASIBILITY AND VIABILITY

- **Feasibility:** The idea is technically possible with today's AI tools. LLMs can easily generate short and accurate scripts from Q&A pairs.Tools like python-pptx and MoviePy can automate video creation.The system can be easily scaled.Overall, both implementation and deployment are practical and realistic.

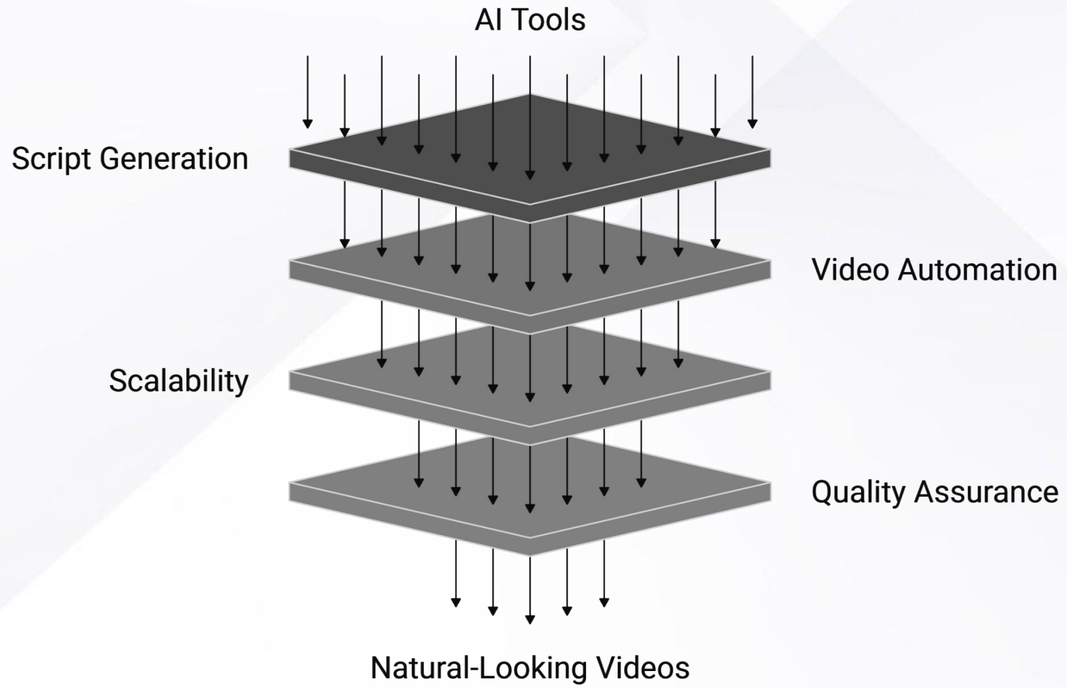
- **Potential challenges:**

Lip-sync videos can look unnatural without proper tuning.

Voice cloning may sound robotic if audio quality is poor.

- **Strategies for overcoming these challenges:**

Ensure high-quality audio and image uploads.



TECHNICAL APPROACH

- Front-End : HTML5 , CSS3, JS
- Frameworks: MERN (MongoDB, Express, React, Node.js) stack, FastAPI (backend API)
- Programming Language: Python (core pipeline and automation)
- Libraries: python-pptx, MoviePy, OpenCV, FFmpeg, Pydub, Pandas, Webview library
- AI Models: Ollama Gemma3 Model (for script generation), IndexTTS (for voice cloning), Wav2Lip (for lip-sync), SDXL-stable diffusion (for image generation)
- APIs: Index TTS API, SwarmUI FastAPI , WorqHatAPI

IMPACT & BENEFITS

Input Question

Teachers provide a question to be explained.

AI Generates Video

AI creates a video from the question.

Video Ready

Video is ready for use without editing.

- Teachers spend hours making videos that AI can create in minutes.

  Our multimodal GenAI turns simple Q&A into sleek, branded video solutions automatically.

  It's faster, smarter, and will revolutionize how coaching institutes deliver education.
- Turn any question and answer into a ready-to-use explainer video in minutes,
- **NO EDITING, NO RECORDING, NO HASSLE**

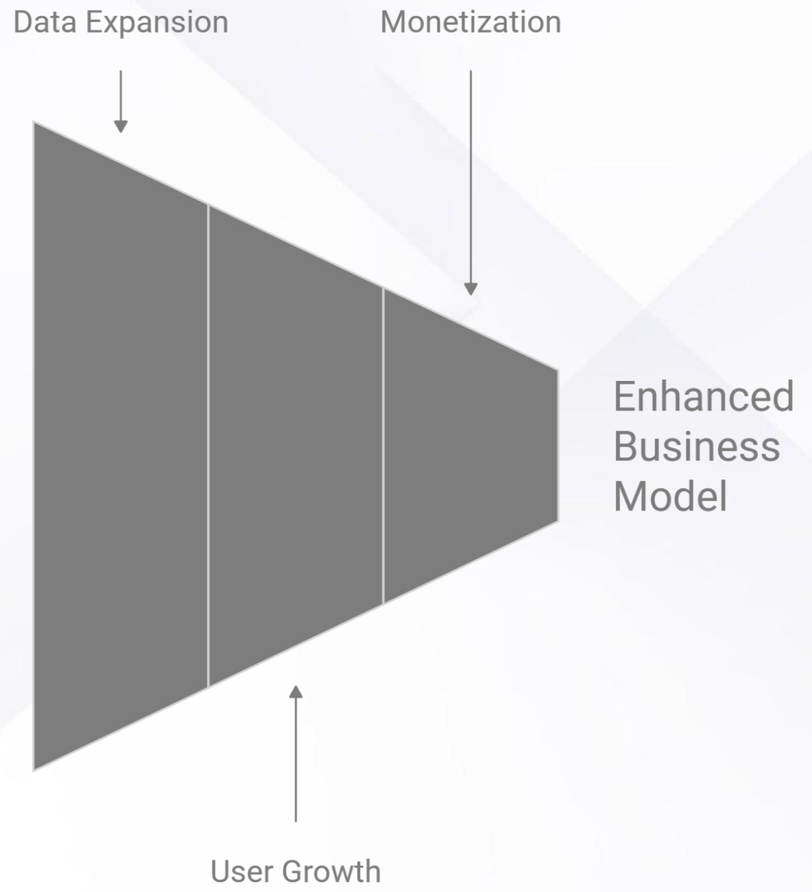
FUTURE SCOPE

Planned improvements / next phases:

- Faster rendering and lower GPU costs
- More human like voice generation
- Preview for teachers before final export

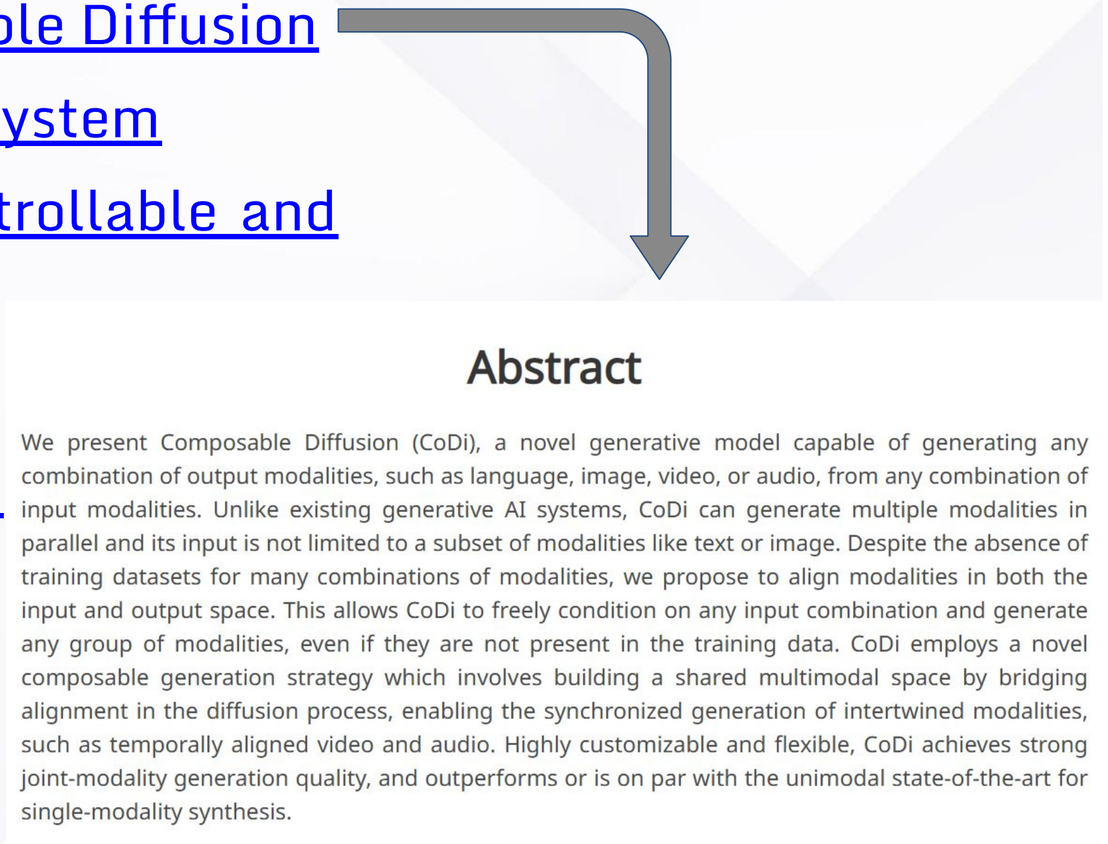
Long-term vision / scaling strategy :

- Data Scaling: Continuously expand the questions and solutions dataset sources to improve output accuracy and adaptability.
- User Scaling: Transition from smaller edtechs to enterprise clients and institutional adoption through APIs or SaaS models.
- Monetization Scaling: Introduce pay-per-use credits and white-label integrations.



REFERENCES

- Research Papers & Articles:
- [2305.11846] Any-to-Any Generation via Composable Diffusion
- [2206.03206] FlexLip: A Controllable Text-to-Lip System
- [2502.05512] IndexTTS: An Industrial-Level Controllable and Efficient Zero-Shot Text-To-Speech System
- Open Source Github Repositories:
- https://github.com/index-tts/index-tts/tree/v1.5.0?
- https://github.com/mcmonkeyprojects/SwarmUI
- Colab for making Wav2Lip high quality



Abstract

We present Composable Diffusion (CoDi), a novel generative model capable of generating any combination of output modalities, such as language, image, video, or audio, from any combination of input modalities. Unlike existing generative AI systems, CoDi can generate multiple modalities in parallel and its input is not limited to a subset of modalities like text or image. Despite the absence of training datasets for many combinations of modalities, we propose to align modalities in both the input and output space. This allows CoDi to freely condition on any input combination and generate any group of modalities, even if they are not present in the training data. CoDi employs a novel composable generation strategy which involves building a shared multimodal space by bridging alignment in the diffusion process, enabling the synchronized generation of intertwined modalities, such as temporally aligned video and audio. Highly customizable and flexible, CoDi achieves strong joint-modality generation quality, and outperforms or is on par with the unimodal state-of-the-art for single-modality synthesis.

Thank You