VIDEO GENERATION FROM TEXT PROMPT

Team: Ashish Kulkarni, Dhruv Chawla, Aryan V S, Aryan Rawther

Mentor: Dr. Jayashree R.

Introduction

Our project aims to address two main problems. The **first problem** is access to video-based educational content. Most students are visual learners. However, most of the content available is text-based. Only a small subset of topics is available in video format online.

The **second problem** is that whenever the government releases a new policy or a reform, its reach is limited by the population's literacy level.

Both these problems can be solved by automating the process of video generation.

The last few years have seen exponential advancements in the field of Generative AI. We have tools like ChatGPT for text generation, DALL-E 2 for image generation, etc. Tools like Imagen Video and Make-a-Video only generate videos of a few seconds in length, without multiple scenes or contextual understanding.

We propose the combination & enhancement of the coherence between existing models for generating videos with an associated script, animations, diagrams, and speech. The system will take a text prompt as input and generate a video that accurately represents the prompt.

Proposed Approach

To solve the limitations in the current models by creating a more sophisticated AI model that can generate engaging and meaningful videos with the following **functionality**:

1. **Contextual Understanding:** Establish a contextual connection between preceding and succeeding frames to obtain a video with a coherent flow and continuity.
2. **Script Generation:** Use large language models to generate a script for the video and natural language processing (NLP) to extract keywords from the prompt.
3. **Audio:** Use normal or AI-based text-to-speech software for video narration. Use voice cloning to give the video an authentic feel. Generate background music matching the tone of the video.
4. **Video:** Select relevant scenes using object detection, determine appropriate transitions using machine learning, and merge frames with editing techniques to create a seamless and engaging educational video. Evaluate using human evaluators and iteratively improve based on feedback.

The following innovative techniques are part of our approach:

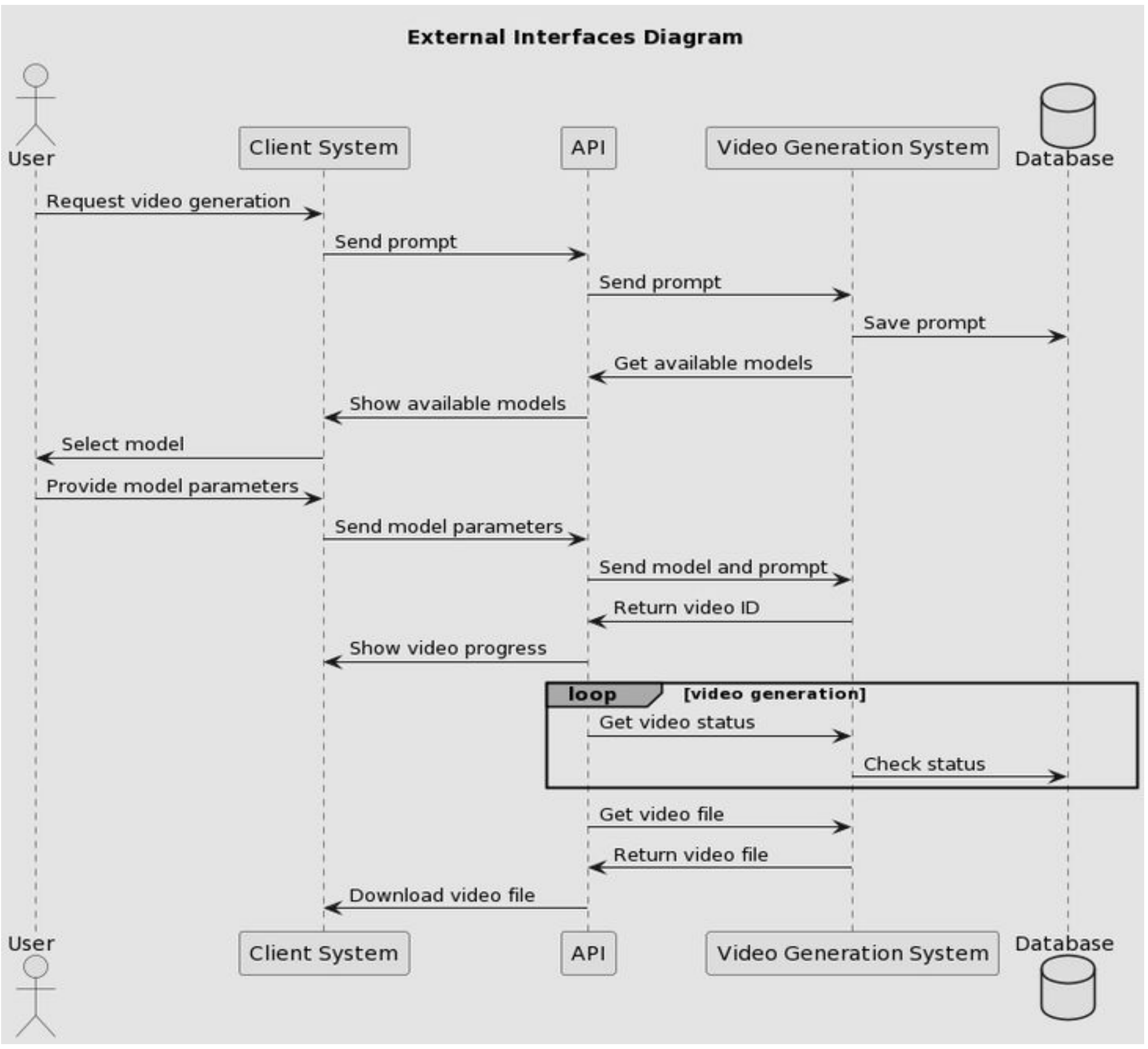
1. **Frame-by-frame Generation:** Divide the generated script into small intervals and generate audio and images for these small intervals to maintain audio-video synchronization.
2. **Video Frame Interpolation:** Generate frames in the middle of two adjacent images in consecutive frames to obtain a continuous video.
3. **Personalization:** Fine-tune the model on the content of a creator to generate videos in their style.

Dataset

YouTube is the most widely available database of educational and entertaining videos on the Internet. We have written code to scrape videos from YouTube and use OpenAI Whisper, which is an automatic speech recognition system, to extract the video transcript.

Architecture

The following diagram shows the flow of data and interactions between the user and the system for a web-based implementation of our model:



External Interfaces Diagram

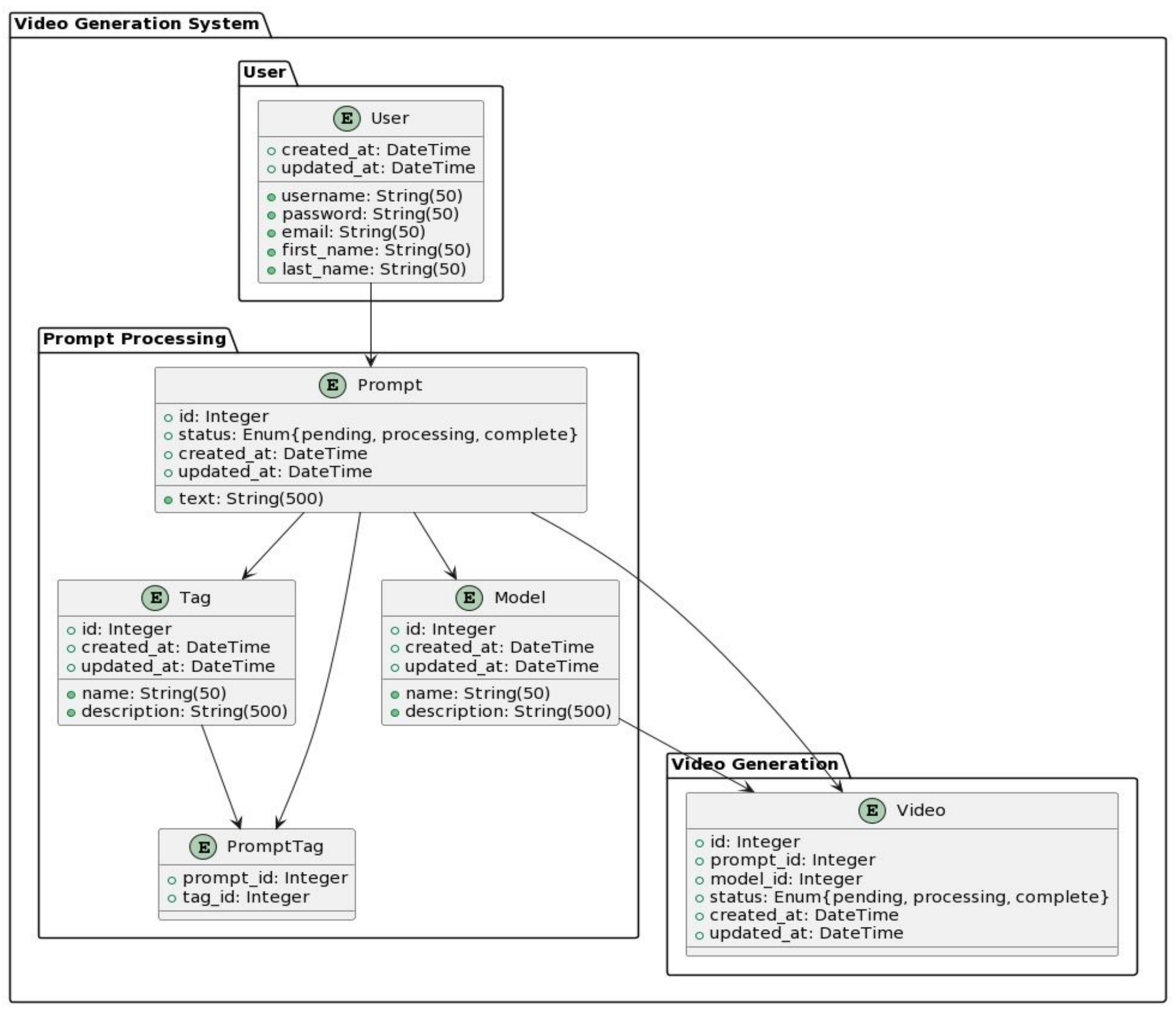
User
Client System
API
Video Generation System
Database

Request video generation
Send prompt
Send prompt
Save prompt
Get available models
Show available models
Select model
Provide model parameters
Send model parameters
Send model and prompt
Return video ID
Show video progress
loop [video generation]
Get video status
Check status
Get video file
Return video file
Download video file

User
Client System
API
Video Generation System
Database

The user can choose from a variety of different models based on their needs. After choosing the model, the user will provide the model parameters that include the prompt, tag, randomness, length etc. The prompt determines the content, the tag defines the video style and the randomness determines the level of creativity in the generated video. The backend will process this information and generate a video to return to the user.

The following class diagram shows the underlying database architecture for this system:



Video Generation System

User

E User
created_at: DateTime
updated_at: DateTime
username: String(50)
password: String(50)
email: String(50)
first_name: String(50)
last_name: String(50)

Prompt Processing

E Prompt
id: Integer
status: Enum{pending, processing, complete}
created_at: DateTime
updated_at: DateTime
text: String(500)

E Tag
id: Integer
created_at: DateTime
updated_at: DateTime
name: String(50)
description: String(500)

E Model
id: Integer
created_at: DateTime
updated_at: DateTime
name: String(50)
description: String(500)

E PromptTag
prompt_id: Integer
tag_id: Integer

Video Generation

E Video
id: Integer
prompt_id: Integer
model_id: Integer
status: Enum{pending, processing, complete}
created_at: DateTime
updated_at: DateTime

Conclusion

In conclusion, our proposed approach to automating video generation through the use of Generative AI has the potential to revolutionize the way we create educational content. Our innovative techniques such as frame-by-frame generation, video frame interpolation, and personalization further improve the quality and customization of generated videos. As we continue to refine and iterate on our approach, we believe that it will significantly impact the field of education and learning.

References

1. **LLaMA:** Open and Efficient Foundation Language Models [Feb. 2023]
2. **VALL-E:** Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers [Jan. 2023]
3. **Tune-A-Video:** One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation [Dec. 2022]
4. **VideoDubber:** Machine Translation with Speech-Aware Length Control for Video Dubbing [Nov. 2022]
5. **Imagen Video:** High Definition Video Generation with Diffusion Models [Oct. 2022]
6. **Make-A-Video:** Text-to-Video Generation without Text-Video Data [Sep. 2022]
7. **V2C:** Visual Voice Cloning [Nov. 2021]

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
PES UNIVERSITY
pes1ug20cs085@pesu.pes.edu | pes1ug20cs127@pesu.pes.edu