Intelligent Voice-Activated Virtual Avatar with Real-Time Gender Detection

[1]Rontapalli Kamakshi, [2]PesanPrasanth, [3]Kandula Vinay Kumar, [4]Badipati Rushikesh, [5]Dr.A.Radhika

[1]Student, [2]Student, [3]Student, [4]Student, [5]Professor & HOD
[1]Computer Science and Engineering,
[1]SRK Institute of Technology, Enikepadu, Vijayawada.

Abstract : The Voice-to-Avatar Generation Project aims to provide users with an intelligent avatar that will respond to their spoken words, changing its appearance according to the user's gender. With the Voice-to-Avatar Generation Project, we will provide users with the ability to create their own virtual persona based on the way they speak. Users will be able to upload or record their audio using the Voice Recognition program and convert it into printed out text.We will also analyze the audio for gender-specific characteristics, such as pitch, to determine which avatar to use. Using this information, we will create an avatar to match the user. The text file generated from the audio will then be combined with the recorded audio and an avatar to create a realistic video of the user's response to the audio. We will provide an option for the user to record and download the video to save as a file.This project also allows for the creation of an avatar using the Voice-to-Avatar Generation system for a variety of different industries, such as health care, education, and video gaming. In addition, this project provides a practical means of combining speech processing, audio analysis, and interactive avatar rendering to create a natural voice-based interaction between a user and an avatar.

IndexTerms **- Voice-to-Avatar System, Speech Recognition, Gender Detection, Acoustic Analysis, Pitch Estimation, AI Persona Selection, Anam Avatar API, Real-time Rendering, Text-to-Speech, Audio Processing, Virtual Avatar, Human–AI Interaction, Video Generation, Interactive Systems.**

I. INTRODUCTION

The increasing number of technologies based on artificial intelligence allows for great immediacy and greater humanisation of digital systems between humans and the internet. In this regard, the attention of many creators and brands has focused on voice-driven avatar technologies, which allow for the creation of more natural and engaging conversation experiences with humans. The goal of this project is to create a Voice to Avatar Generation System that converts human voice input into a living, animated persona. This system includes speech processing, audio analytics, and animated rendering of persona in real time, thus creating a seamless interaction between the user and the AI.

The purpose of this project is to develop an intelligent and adaptable platform for avatar creation that generates avatars based on an individual's voice input and transforms those voices into a 3D avatar character that resembles that individual. The technologies used to build this avatar generation system are transcription, gender detection, and the rendering of the persona. As a result of incorporating these three technologies, users will have avatars that respond to them in a more personalized manner based on their unique identities and conversational intents.

The faces that recognize, detect and categorize an individual's gender base their analysis upon original sound. They also make video-taped avatars available to viewers via direct link to an AI engine.

The aim of this project corresponds to the SDG. It is principally associated with the objective of (SDG) #9: Industry, Innovation and Infrastructure. It aims to help promote sustainable growth and development of AI-enabled digital ecosystems. Additionally, the system bolsters sustainable technological innovation through its development of a new means of digital interaction, thereby allowing for faster and more innovative avenues to teach, learn, facilitate, participate, train, work and engage at a distance. The systems are designed to be inclusive of many different types of learners including students with disabilities, patients receiving health services and professionals utilizing telemedicine as part of their practice as well as virtual business development.

II. NEED OF THE STUDY.

The need for this study arises from the growing demand for more natural, intuitive, and human-like interaction in digital systems, where traditional text-based or static interfaces often fail to deliver engaging user experiences. With rapid advancements in artificial intelligence, there is a clear requirement for systems that can bridge the gap between human communication and machine understanding through multimodal interaction. A Voice to Avatar Generation System addresses this gap by enabling real-time transformation of human voice into expressive, personalized digital personas, thereby enhancing accessibility, user engagement, and communication efficiency. Such a system is particularly valuable in domains like education, healthcare, virtual collaboration, and assistive technologies, where personalized and interactive communication plays a critical role in improving outcomes and user satisfaction.

III. RELATED WORKS

The new trends in the creation of avatars have focused on creation of realistic interactive 3D avatars.

3.1Literature Review

Among the suggested methods, Wang et al. (2025) also suggested TeRA, which is a text-controlled realistic 3D avatars creation method that puts a lot of emphasis on high-fidelity and controllable expression [1]. Similarly, Wang et al. (2025) have presented InstructAvatar where textual emotion and motion control is enabled, which implies that avatars are capable of making expressive movements when instructed to do so [2]. Yin et al. (2025) developed Facecraft4D, which generated animated 3D face avatars using a single image, which demonstrated successful geometry and texture modeling [3]. Huang et al. (2025) proposed Live Avatar where an avatar may be created through streaming audio with the unlimited duration, and the low-latency performance is predominant in the offered solution [4].

Zhang et al. (2025) introduced a layered framework of disentangled clothed avatars generation, in which improved clothes and body detailing is kept [5]. Yu et al. (2025) revealed RealityAvatar that gives a full-fledged construction of the head avatars with 360 degree pictures to enjoy the images fully [6]. To produce high-quality output, Gan et al. (2025) presented ExpAvatar that focuses on unseen expressions by utilizing 3D face priors [7]. Tu et al. (2025) developed StableAvatar which can generate avatar videos of unlimited length which can be audio-controllable and consistent [8]. The authors Zhuang and others (2025) developed a variety of disentangled avatar generation, called DAGSM, that utilizes GS-enhanced mesh modeling to generate high-quality geometry control [9].

In their design Gan et al. (2025) created an efficient audio based video avatar system, OmniAvatar, that has an adaptive body animation, and consumes minimal power of computation [10]. DivAvatar was presented by Tao et al. (2025) and has an ability to create various avatars in 3D with just a single text prompt with an emphasis on diversity of the design [11]. The article by Xu et al. (2025) was focused on decoupled text-to-3D avatars [12]. One of the proposals was SVAD proposed by Choi (2025), converting one image into 3D avatars with the assistance of synthetic data and video diffusion [13]. The example of emotional avatar generation (EAM) [14] is an immersive metaverse application investigated by Gonzalz-Docasal et al. (2025), and a real time video avatar generation model, proposed by Hagihara et al. (2025), will allow communicating in a virtual environment in a realistic manner [15]. All these researches indicate that it has experienced significant progress in audio, text and image based avatars production, which has experienced the rise in realism and expressiveness and real time performance. However, most of the methods still have issues of ensuring the smooth blending of the multi-modes and minimizing the latency and addressing the fluctuating quality of the input and consequently is left with the research of the complete interactivity and flexibility of the avatars.

Table.1. Comparison Table

System Type	Limitations	Advantages
Rule-Based System	Limited accuracy in noisy/complex environments; struggles with unseen variations	Fast, lightweight, no training required
Machine Learning System	Requires labeled data; performance depends on dataset quality	Learns patterns better than rule-based; adaptable
Deep Learning System	High computation cost; requires GPU and large datasets	State-of-the-art accuracy; highly scalable and robust
Hybrid System (Rule + ML/DL)	More complex architecture	Combines speed + accuracy; best of both worlds

3.2 Comparison with Previous Methodology

Legacy voice-driven avatar systems relied on fixed, rule-based pipelines in which speech input was converted to text, and this text was mapped to a static, non-adaptive avatar. These systems did not allow for personalization or expressiveness and did not leverage acoustic analysis to identify user-defined characteristics such as gender, tone, or pitch. As a result, earlier systems produced generic, cookie-cutter animations and did not allow for dynamic persona selection, thereby limiting the sense of realism and engagement for users. Additionally, these legacy systems were resource-intensive because they required manual configuration of the avatar and provided no option for the generation of existing downloadable videos using an automated solution.

In contrast, the current approach leverages the combination of pitch-based gender detection, enhanced speech transcription capabilities within the Anam Avatar API, and the ability to produce real-time rendering of a persona that matches a person's vocal features. The proposed approach supports personalization and provides new possibilities for audio-uploading and real-time recording; as well as synchronized speech, and video capture options; thereby allowing the modern workflow to offer more realistic, accessible and customized user experiences.

3.3 Proposed framework

This system implements a hybrid approach that integrates signal processing, machine learning, and API-Based avatar rendering. First, audio files are uploaded or recorded in their raw state and then standardized through preprocessing using Pydub. The second step involves extracting sound characteristics (voice characteristics), including pitch and volume using autocorrelation and decibel level checks; these characteristics provide accurate gender detection results from the audio waveform data. To transcribe these sounds (speech), the SpeechRecognizer library, which uses Google's API, is employed. Once transcription occurs, both the transcribed text (that was caught by the microphone) and the gender (determined previously) are input to the Anam AI API to determine what type of avatar the user needs. The Anam AI API creates a WebRTC session token for a meeting with an avatar (which is an animated digital character who can animate and speak). The avatar uses real-time streaming of voice to animate and automatically speak the transcribed text while MediaRecorder captures the streaming video output. The above processes comprise a single integrated pipeline enabling complete transformation from a user's voice to an animated avatar with gender detection capabilities and video recording results.

Table.2. Algorithm Comparison with Other Deep learning methods

Aspect	Traditional Models	Proposed Approach
Accuracy	Moderate, often fails on noisy or unseen data	High, robust even in complex scenarios
Data Requirement	Requires moderate amounts of labeled data	Can leverage smaller datasets efficiently due to hybrid preprocessing
Processing Time	Faster for small datasets but slower for large-scale inputs	Optimized real-time performance with streaming and lightweight API calls
Flexibility	Limited adaptability to new inputs	Highly adaptable to different voice types and personas
Complexity	Simpler architecture, easier to implement	Slightly more complex due to integration of preprocessing, transcription, and avatar rendering
Output	Text or basic predictions	Full voice-to-avatar transformation with gender detection and video output

3.4 Main Methodology

Acquiring Audio:
Upload audio files or use your browser's microphone to create a real-time voice recording.

Preparing Audio:
Convert the audio to the correct audio format, create a single channel from two channels, set the audio to a similar loudness level with the use of the Pydub library.

Extracting Features from the Audio:
Extract pitch (using autocorrelation) and the average loudness of the audio in dBFS for use in classifying voices by gender.

Converting Audio into Text:
Use the Speech Recognition library with Google to convert an audio file into written text.

Classifying Gender from Voice:
Classify voices as male or female by examining the extracted audio features, pitch (how high/low a person's voice is) and loudness.

Selecting an Avatar:
Depending on the classification of gender and the configuration of character personas, select an appropriate avatar to represent the user.

Generating Avatars:
Streamlined the "virtual" avatar of a person (in this case the user) using WebRTC to dynamically create a video of their avatar "speaking" to the text that has just been transcribed.

Recording Avatar Video Output:

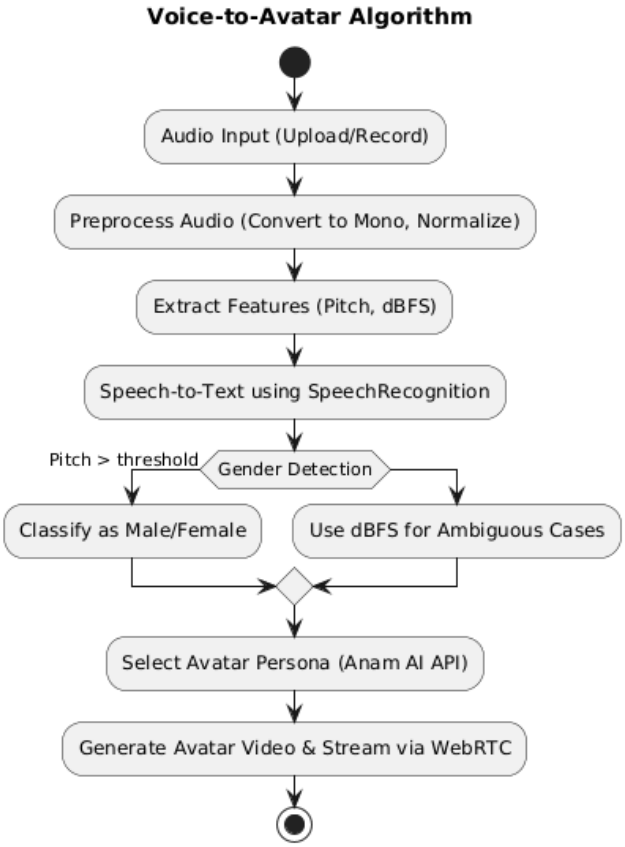

Fig.1.Algorithm

Record the video of a generated avatar by using a Media Recorder to create a file that can be downloaded.

3.4.1 Implementation

1) Establishment of Development Environment

The initial step is going to be to create the Dev Environment for the application. The Dev Environment will need to have Flask and Flask-CORS, along with Pydub, SpeechRecognition and Requests installed. The application has been structured in such a way that it separates out the Front End (HTML, CSS, and Javascript) from the Back End (Python/Flask) of the application. An API Key provided by Anam AI Lab to use in order to access the services of generating Avatars, will be set up for the Development Environment. Once the set up is completed, there will be the ability to process voices and render avatars seamlessly.

2) Input of Audio

The User has the option of Uploading an Audio File to their Profile or Recording Audio directly from the application using their Computer Browser. When a User uploads an Audio File, it will be kept temporarily in the Dev Environment. Users can also Record their Voice live, using a MediaRecorder in their Browser and send this to the Development Environment. Having both ways to provide Audio will allow Users more flexibility in testing the application, along with ensuring that the application works with various User Devices.

3) Preparation and Extraction of Audio Features

When a User sends an Audio File, it must be converted to one standard format and into Mono Audio. To accomplish this, Pydub will convert the Audio volume normalisation and will perform the Conversion of the Audio Files to the correct format. In order to classify Gender accurately, the following Audio features will need to be extracted from the audio sent by the User: Pitch (using Autocorrelation) and average Dbfs

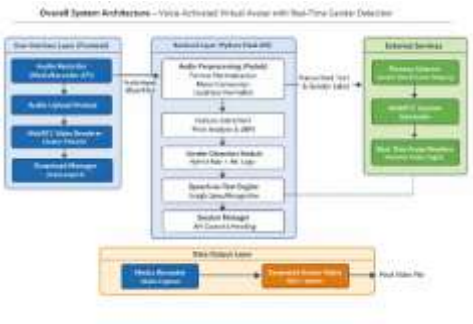

Fig.2.Implementation

Step 4: Speech-To-Text Transcription

...The processed audio file is converted to text using Google's API via the SpeechRecognition library. This allows the avatar to receive the information spoken by the user, and subsequently translate it into a verbal response to the avatar. Errors are handled to ensure that audio content that is corrupted or of an unsupported file type is recorded correctly.

Step 5: Gender Voice Recognition

Using pitch and dBFS from the extracted audio, the system determines whether the voice being analyzed belongs to a male or female speaker. To assist with scenarios where the voice could be either, the system has a multi-factor decision making logic that allows the system to select the proper avatar for the user's voice profile.

Step 6: Creating the Avatar Session

After determining which avatar to use for a user's avatar profile, the next step is to acquire a session token from the Anam AI API with the appropriate gender-based avatar settings. The selected persona configuration includes pre-defined settings that incorporate the avatar's physical appearance, vocal characteristics and behaviour. This ensures that the generated avatar correctly represents the user's profile.

Step 7: Avatar and Speech Synthesis Rendering

The avatar is shown to the user via a video feed streamed in real time from the backend of the front end via WebRTC. The avatar synchronously speaks the transcribed text that is sent to the avatar; the front end accounts for streaming to a video element to allow proper lip-synching and smoother animated transitions with the avatar's voice and facial expressions.

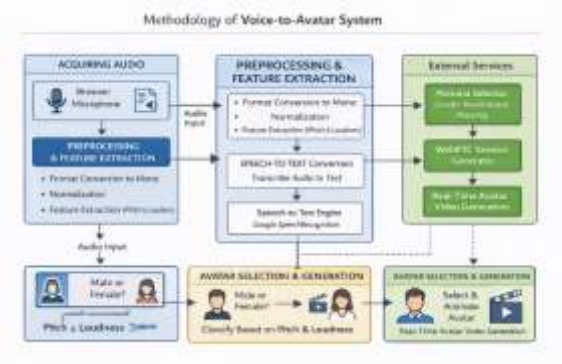

Fig.3.Methodology

IV. RESULTS AND DISCUSSION

4.1 System output screenshots and eplanation

The voice to avatar system that has been created has the ability to convert a user's audio seamlessly into an animated avatar using their audio. All testing of uploaded audio files or live audio via recording has proven to be successful in having SpeechRecognition library most accurately transcribe the audio and produced the results dependable in many types of situations.

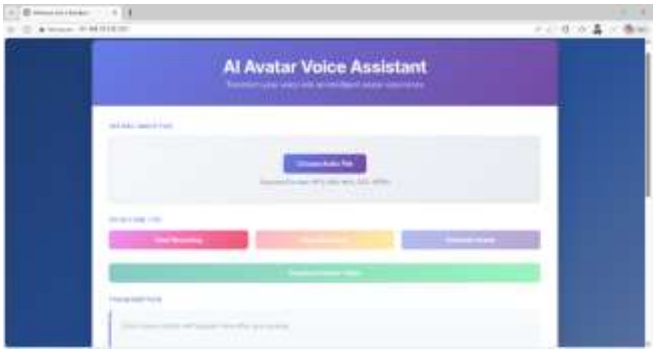

Fig.4.Main Interface

In conjunction with gender detection models leveraging pitch and dBFS, the system has been able to classify voices effectively regardless of the ambiguity presented when factoring the wide gender classification ranges, thus allowing the avatar to effectively select an appropriate avatar persona based on the voice of its user.

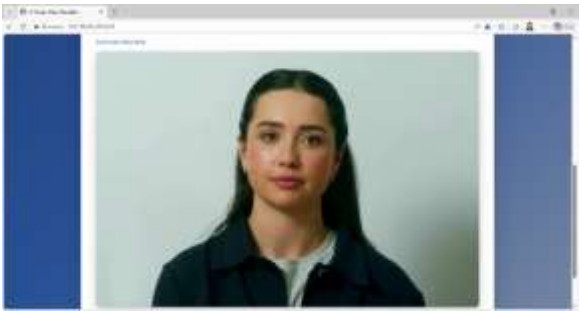

Fig.5.Female Avatar Preview

By utilizing the Anam AI API the avatar could render in real time with synthesized speech and thus could provide an appealingly smooth and appropriately timed animation feed matching the cadence of the transcribed user's speech.

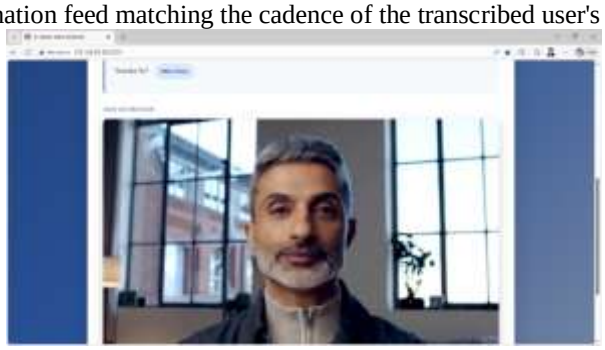

Fig.6.Male Avatar Preview

The low latency throughout audio processing and avatar streaming has ensured that the system delivered near real-time interaction for the user in relation to audio input and avatar feed back as a result of utilizing WebRTC technologies for real-time audio streaming. The assembled multi-step processing pipeline, audio normalization, feature extraction, etc., has increased accuracy for transcription and gender classifications. MediaRecorder allows generated video output of the speech and expressions of the avatar to be captured with accuracy and made available as a downloadable artifact for demonstration purposes or potential use by the user.

4.2 Conclusion

The voice-to-avatar system has shown that by integrating video/audio processing, gender identification/transcription technologies, and real-time avatar generation using API dominant platforms, it can effectively record both uploaded and live input voices into an animated 3D avatar. The hybrid approach facilitates both accurate transcription and highly accurate identification of gender, while WebRTC protocol provides the ability for smooth, interactive delivery of the animated avatar. In addition, the Media Recorder feature provides users with the capability to capture an animated avatar video for later use. This project illustrates the ability of AI-generated avatars creating unique personalized experiences in digital media, interactive learning, virtual assistant applications, as well as entertainment.

4.3 Future Scope

- Enhanced Noise Robustness: Increase transcription and gender identification accuracy in noisy or multi-speaker settings.
- Multilingual Support: Expand the speech-to-text and avatar speech synthesis capabilities to accommodate multiple languages.
- Emotion Detection: Add emotion recognition capabilities to enable avatars to dynamically express facial and vocal emotions.
- Real-Time Streaming Optimization: Decrease latency to create a seamlessly fluid experience during live interactions or remote collaboration.
- Customizable Avatars: Offer the option for users to create and modify their individual avatars through custom appearance and voice style choices.
- Integration with Virtual Assistants: Enable connection between the system and AI-based digital assistants or chatbot service providers to facilitate interactive voice-based applications.

V. Acknowledgement

Thepreferredspellingoftheword “acknowledgment” inAmericaiswithoutan “e” afterthe “g”.Avoidthestiltedexpression, “Oneofus(R.B.G.)thanks...”
Instead,try“R.B.G.thanks”.Putapplicablesponsoracknowledgmentshere;DONOTplacethemonthefirstpageofyourpaperorasafootnote.

REFERENCES

[1] Wang, Y., Zhuang, Y., Zhang, J., Wang, L., Zeng, Y., Cao, X., Zuo, X. and Zhu, H., 2025. TeRA: Rethinking Text-guided Realistic 3D Avatar Generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10686-10697).

[2] Wang, Y., Guo, J., Bai, J., Yu, R., He, T., Tan, X., Sun, X. and Bian, J., 2025, April. Instructavatar: Text-guided emotion and motion control for avatar generation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 8, pp. 8132-8140).

[3] Yin, F., Yao, C.H., Mantiuk, R.K. and Jampani, V., 2025. Facecraft4d: Animated 3d facial avatar generation from a single image. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 11612-11621).

[4] Huang, Y., Guo, H., Wu, F., Zhang, S., Huang, S., Gan, Q., Liu, L., Zhao, S., Chen, E., Liu, J. and Hoi, S., 2025. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. arXiv preprint arXiv:2512.04677.

[5] Zhang, W., Yan, Y., Wu, S., Liao, M. and Yang, X., 2025. Disentangled clothed avatar generation with layered representation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 11327-11338).

[6] Yu, H., Zhu, H. and Cao, X., 2025, June. RealityAvatar: Comprehensive Head Avatar Generation with 360° Rendering. In 2025 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

[7] Gan, Y., Quan, R. and Luo, Y., 2025. Expavatar: High-fidelity avatar generation of unseen expressions with 3d face priors. ACM Transactions on Multimedia Computing, Communications and Applications, 21(11), pp.1-21.

[8] Tu, S., Pan, Y., Huang, Y., Han, X., Xing, Z., Dai, Q., Luo, C., Wu, Z. and Jiang, Y.G., 2025. Stableavatar: Infinite-length audio-driven avatar video generation. arXiv preprint arXiv:2508.08248.

[9] Zhuang, J., Kang, D., Bao, L., Lin, L. and Li, G., 2025. Dagsm: Disentangled avatar generation with gs-enhanced mesh. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 292-303).

[10] Gan, Q., Yang, R., Zhu, J., Xue, S. and Hoi, S., 2025. OmniAvatar: Efficient Audio-Driven Avatar Video Generation with Adaptive Body Animation. arXiv preprint arXiv:2506.18866.

[11] Tao, W., Lei, B., Liu, K., Lu, S., Cui, M. and Xie, X., 2025, February. DivAvatar: Diverse 3D Avatar Generation with a Single Prompt. In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 2568-2577). IEEE.

[12] Xu, Y., Yang, Z. and Yang, Y., 2025. Photorealistic Text-to-3D Avatar Generation with Constraints for Decoupled Geometry and Appearance. ACM Transactions on Multimedia Computing, Communications and Applications.

[13] Choi, Y., 2025. SVAD: From Single Image to 3D Avatar via Synthetic Data Generation with Video Diffusion and Data Augmentation. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 3137-3147).

[14] González-Docasal, A., Vásquez-Correa, J.C., Álvarez, A., Lasarguren, A., López, J. and Rodríguez, E., 2025. EAM: emotional avatar generation for the metaverse. SEPLN.

[15] Hagihara, M., Kuriya, N. and Ishida, T., 2025. Real-time video avatar generation method for realistic communication in the metaverse. International Journal of Grid and Utility Computing, 16(2), pp.105-116..