

Meta AI Twin: A Smart Virtual Meeting Platform with 2D Avatar-Based User Presence and Cloud-Driven Language Services

CTM Praveen Kumar, Naziya Parveen, M Nausheen Fathima, Khalida Tabassum, P R Rachana

Department of Artificial Intelligence and Machine Learning Ballari Institute of Technology & Management,
Ballari, Karnataka, India

Emails: ctm.praveen@gmail.com, syedanaiziya506@gmail.com, nausheenf260@gmail.com,
khalidatabas903@gmail.com, prachana29@gmail.com

Abstract—Online meeting platforms are a crucial part of many sectors, such as education, business, and video conferencing. In this paper, an intelligent avatar is proposed that can join and participate in the meeting in real time in the absence of the attendee. The avatar in the meeting summarizes the discussion after the meeting is over and saves the notes in the history for later use. To make the user experience smooth, a number of predefined avatars are offered for selection while joining the meeting. The avatar is designed to automatically create a text summary of the discussion through speech-to-text conversion using Microsoft's speech services. The proposed system uses real-time video streaming through WebRTC. Unlike traditional online meeting platforms, where only the live video streams of the participants are shown, our avatar-based system proposes the use of lightweight avatar models as virtual representations of the users. This allows the users to stay engaged even when they are temporarily unavailable, as their avatars stay in the meeting as their representatives. From the user experience perspective, traditional systems lack personalization, with only the user's camera and microphone being used, whereas our system improves the user experience by offering predefined avatars and seamless browser-based performance.

Index Terms—online meeting platform, intelligent avatar, Microsoft services, personalization, summary, virtual meeting, WebRTC, speech-to-text

I. INTRODUCTION

Online meeting platforms have become an integral part of modern-day communication. These platforms have been able to support education, business, and remote collaboration. Although these platforms allow users to attend meetings from anywhere in the world, they often face issues while recreating the level of engagement that is experienced during physical meetings. An avatar is a virtual representation of a participant and offers a basic sense of presence. In addition, Microsoft Speech-to-Text and Language capabilities enable the automatic creation of meeting summaries to ensure that the user is always up to date, even when they are not physically present.

Background: Traditional meeting systems lack personalization, mainly depending on the user's camera and microphone. This results in user fatigue, disengagement, and inability to follow through on long conversations. Predefined two-dimensional avatar images, using web-based rendering and real-time communication technology, can solve these problems

by allowing users to be represented in a simple and efficient visual form. Motivation: The growing demand for remote communication emphasizes the need for solutions that make online meetings more natural and productive. Many users cannot attend all meetings in real-time, and even when they do, note-taking and engagement are difficult. This problem requires the creation of an avatar meeting system that not only allows users to be represented during conversations but also provides a clear summary of the conversations after the meeting. Contribution: This paper presents an intelligent avatar system for online meetings. The system allows users to choose from a predefined set of two-dimensional static avatar images that will represent them during meetings. The system uses Microsoft Speech-to-Text and Language services to provide automated summaries of conversations, which are stored as notes for future use. Through the combination of simple web-based avatar representation and WebRTC-based streaming, the proposed system improves personalization, engagement, and productivity in online collaboration.

II. LITERATURE SURVEY

The creation of realistic 3D avatars has recently received increased attention, with researchers aiming to improve virtual presence and human-like interaction in online spaces. Huang et al. [1] presented DreamWaltz, a system that can create complex 3D animatable avatars, allowing users to create highly detailed digital twins for virtual communication. Their system targets scene-level realism and animation control, adding to personalized digital interaction. Aliakbarian et al. [2] introduced HMD-NeMo, an online 3D avatar motion generation system that has the capability to generate avatar motion based on sparse observations. Using real-time motion capture and mapping, the system reduces the need for large datasets in guaranteeing realistic animation for virtual meetings. Similarly, Zhang et al. [3] presented MagicAvatar, a multimodal system for avatar creation and animation. Their system integrates speech, text, and image inputs to create responsive avatars that can be adjusted for different communication contexts, making it ideal for business and educational purposes. Huang et al. [4] proposed WildAvatar, a learning approach to derive 3D avatars from general internet

data "in the wild." It can generate avatars directly from general internet sources, improving overall generalization performance in different settings. Wu et al. [5] presented AnimPortrait3D, a text-driven avatar creation system that enables animation via morphable model alignment. The system enables control of avatars using natural language instructions while maintaining believable expressions. Other studies investigate the potential of avatars in enhancing presence, human interaction, and overall user experience in increasingly populated virtual worlds containing meetings and collaborative work activities. For instance, Yoshida et al. [6] conducted research to compare the use of AI-powered avatars and human-controlled avatars in the context of online video chats and demonstrated the benefits of their participation in virtual conferences. In addition, Zainudin et al. [7] proposed a framework aimed at integrating interactive avatars into systems via the utilization of edge learning and human activity recognition. Until recently, there has been an ongoing debate about the level of similarity between the avatar and the person to whose avatar it belongs, as well as the theoretical underpinnings of such a debate. For instance, Zhu et al. [8] showed that the more similar the avatar to a person, the more they become engaged in a virtual task in a collaborative team. On the other hand, Dobre et al. [9] investigated the avatars during mixed-reality meetings to identify the extent to which the new means of virtual communication affects communication, social presence, as well as emotions. Gasch et al. [10] provided a glimpse of how avatars in a virtual meeting contribute to influencing the participants when they were able to reduce technostress in a business meeting and guide the users accordingly. On the one hand, Menzel et al. [11] deployed the first actual live, mobile-friendly pipeline that enables users to develop and control avatars directly from their smartphones and hang out in a virtual space. This suggested the possibility of inexpensive live communication systems. On the other hand, Kohistani et al. [12] investigated the use of avatars and agents as surrogates in collaborative and social applications. Nan et al. [13] developed a pipeline to build an interactive digital avatar that could accommodate user-created responses to events through a speech and interaction module. Bidram et al. [14] focused their research on AI avatars in the metaverse and other virtual collaboration spaces, and their goal was to increase immersion levels among the participants. Kang et al. [15] created and introduced a real-time translation tool in the virtual environment and used gestures to input the information. In virtual environments, avatars don't just exist; they interact, and in this aspect, Meta's AI Twin takes the spotlight in the virtual world. Collectively, these studies reflect the rapid progress being made in the use of AI technology for avatar creation, focusing on realism, flexibility, and interactivity. They provide a foundation for projects such as Meta AI Twin, which seeks to provide secure, browser-supported 3D avatars that reflect user expressions and speech.

III. PROPOSED SYSTEM AND METHODOLOGY

A. Proposed System

The proposed system is an online meeting solution on the web that aims to make remote work or learning feel simple, reliable, and flexible. It allows you to create a new meeting and share a special code with others to join. Others can use the code to join only the specific meeting, which also prevents the need for extensive account setup for each meeting and increases the speed of joining. Once you are inside the meeting, you can raise your voice using the basic participation feature for live audio. There is also a chat box for text messages, which allows you to receive quick answers without disrupting the audio. This two-way communication system improves interaction and increases the efficiency and flexibility of meetings. The system also has an avatar feature for situations when you are busy, temporarily away, or simply do not want to communicate through live video. In this feature, a pre-set 2D static avatar is used instead of video, indicating your presence in the meeting without showing your face. This feature indicates to other participants that you are still in the meeting while reducing network and device resource consumption by not streaming video, but it still provides a basic visual identity. The system begins recording as soon as the meeting begins, recording the audio stream of all connected participants end to end. No manual input is required—the entire discussion is recorded. These audio streams are then processed by Microsoft Speech-to-Text, which converts the audio streams of all participants into a written text format. This audio recording contains the entire conversation and is the primary source for all that follows. Once it is transcribed, Microsoft Language services create a concise summary of the conversation. This summary points out the important points and topics of the conversation, weeding out the unnecessary information so you can understand the point of the conversation without having to read the entire text. The entire conversation and the final summary are stored in the meeting history, along with some basic information such as the meeting code and date. This history allows for easy access to previous meetings, viewing of critical discussions, and retrieval of summarized notes whenever required. The stored summaries are particularly useful for people who were not able to attend the meeting or arrived late. Another automated feature in meeting management is the idle session termination feature. This feature automatically terminates the session based on the number of active users in the meeting. The system automatically terminates the session if there are less than one active user, meaning all users have left the meeting. The platform is equipped with live audio, chatbox messaging, a pre-defined 2D avatar, automatic recording, cloud-based speech-to-text, smart summarization, and automatic termination. The platform provides a functional and friendly experience for online meetings. The design of the platform is intended to enhance usability, facilitate interaction, reduce manual documentation, and increase productivity in online meetings.

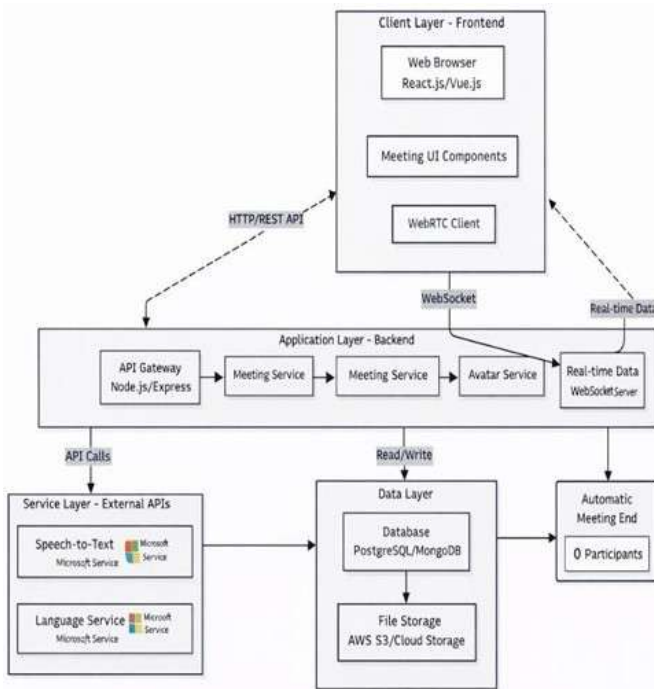


Fig. 1. System Architecture

B. Methodology

A. Creating Meetings and Joining In

The system allows users to create online meetings and join them using a special code. If a user wants to join a meeting, the system generates a special code for that user and sends it to the other participants in the meeting. The user then uses the code to join the meeting. Users get access to live audio communication and a chatbox for communication.

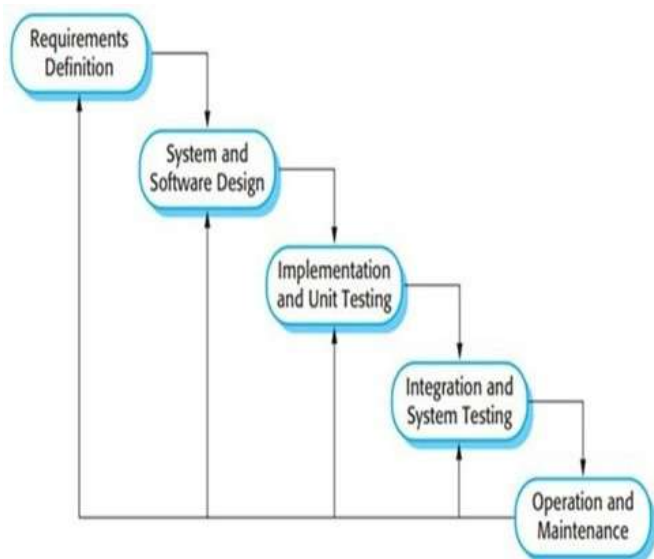


Fig. 2. Development model of the proposed online meeting platform

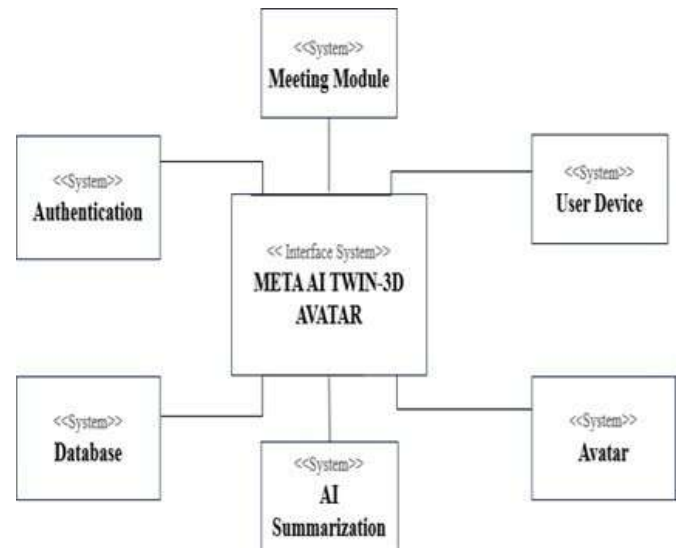


Fig. 3. Context diagram of the proposed online meeting system

B. Avatar Mode and Chatting

If a user has their hands full or doesn't want to participate in the meeting, they can turn on the avatar mode. The user will be represented by a special 2D static avatar in the meeting interface. Besides audio communication, users can use a chatbox for real-time communication.

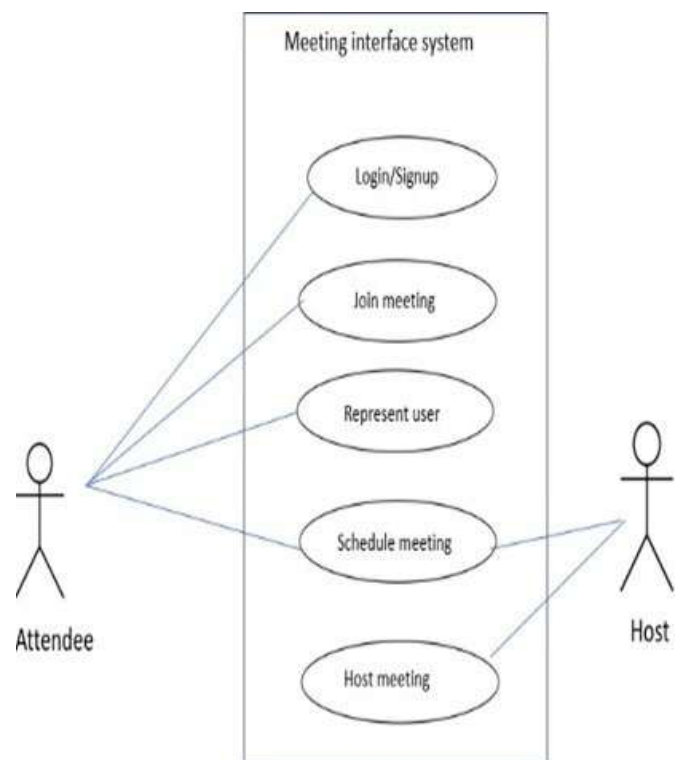


Fig. 4. Context diagram of the proposed online meeting system

C. Audio Stream Collection and Recording

As soon as the meeting starts, the recording process automatically begins. All audio streams from connected participants

are captured through the real-time communication layer and recorded automatically. The audio captured from all participants is then ready for transcription processing.

D. Speech-to-Text Conversion

All audio streams from various participants in the meeting are sent to Microsoft's Speech-to-Text service via a cloud interface during the meeting. The service automatically analyzes the audio streams, converting them into text in near real-time. The service processes continuous speech, recognizes words, and provides a structured text output that captures the entire discourse, including all statements made by various participants. The text output obtained from the service is the first input to the processing stages of the system. The automation of this process prevents the need to manually record statements during the meeting and also prevents the loss of important information. The cloud-based recognition service used prevents the performance and accuracy of recognition from being degraded by the device and network used.

E. Summary Generation Using Language Service

The Microsoft Language Services module automatically analyzes the text generated from the Speech-to-Text service and automatically generates a summary of the meeting. The module automatically analyzes the text to determine the key sentences, topics, and points that summarize the context of the meeting. The module automatically generates a summary of the meeting that captures the key information that was discussed. The irrelevant and less important information is eliminated, and the important points and ideas are retained. The summary generated is stored in the meeting history module along with the relevant meeting information.

F. Meeting History Storage

Once this summary is generated, both the transcript and the summarized meeting notes will be saved as a meeting history package. The package will contain a basic meeting code and date, as well as a newly formed summary. The user is able to access this as a reference to previous meetings and notable decisions.

G. Automatic Meeting Termination

The system also keeps track of the current number of participants. The system terminates the session when the participants are zero, meaning that there are no participants in the session. This termination makes the system efficient.

IV. EXPERIMENT RESULTS

The proposed system was implemented and tested in a real-time web environment to evaluate the functionality, reliability, and usability of the proposed system. A number of experimental meetings were conducted where the participants joined the meetings using the unique meeting codes provided by the proposed system. The experimental meetings were conducted to evaluate the functionality of the major components of the proposed system such as the functionality of the real-time audio communication, functionality of the avatar mode, functionality of the chatbox, automatic audio recording, speech-to-text conversion, summary generation, storage of the meeting history,

and automatic termination of the meeting. During the experimental meetings, the participants were able to successfully create and join the meetings using the unique meeting codes provided by the proposed system without any difficulties related to access and synchronization. All the participants were able to communicate using the real-time audio communication and the chatbox without any interruptions. The chatbox feature supported smooth text interaction alongside audio communication, enabling participants to exchange short messages, clarifications, and additional information during the meeting sessions. The avatar mode feature was also evaluated during the experiments. When users enabled the avatar mode, the predefined 2D static avatar images were displayed correctly in the meeting interface, allowing participants to remain visually represented without using video streaming. This confirmed that the platform can effectively maintain user presence while reducing bandwidth consumption and device resource usage, which is especially beneficial in low-network or limited-device environments. The automatic recording mechanism functioned as expected throughout the meeting sessions. Audio streams from all active participants were continuously captured from the beginning of the meeting until its completion. The recorded multi-user audio data was successfully transmitted to the Microsoft Speech-to-Text service, and accurate textual transcripts were generated for different speakers. The generated transcripts clearly reflected the flow of conversation and participant interactions, validating the reliability of the transcription module. Subsequently, the transcripts were processed using Microsoft Language services to generate concise summaries. The produced summaries effectively captured the main discussion points and important topics addressed during the meetings. These summaries, along with the corresponding transcripts, were stored correctly in the meeting history module and could be accessed by users after the meeting ended. This verified the correctness of the data storage and retrieval processes and demonstrated the usefulness of the history feature for reviewing previous discussions. The automatic meeting termination functionality was also tested as part of the experiments. When all participants left the meeting and the number of active users became less than one, the system automatically ended the meeting session without any manual intervention. This confirmed the proper operation of the participant monitoring mechanism and ensured efficient management of inactive sessions. Our results demonstrate that a seamless meeting system is achieved by our proposed system. The 2D avatar meeting, as well as the use of chat, transcription, and meeting summary within a meeting, rely on Microsoft's meeting cloud. Further ease is indicated by improving the meeting system, meeting records, and avatar meetings.

The predefined 2D static avatar images were displayed correctly when users enabled the avatar mode, confirming that the platform can effectively represent user presence without relying on video streaming. This avatar-based representation allows users to remain visually present in the meeting even when they are busy, have limited bandwidth, or prefer not to use a camera. Only the data packet from the system with static avatars enabled will be transmitted. This reduces the

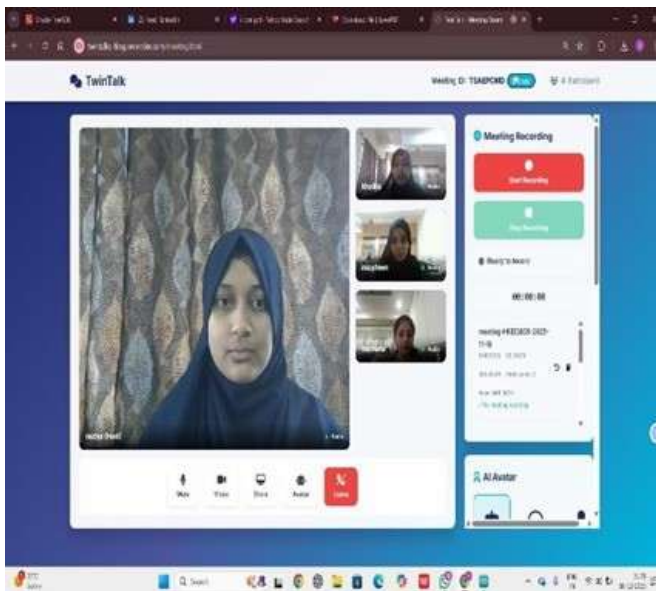


Fig. 5. Normal meeting interface where users are actively participating in a live meeting session

transmission load while still enabling the count of the number of participants in the meeting.

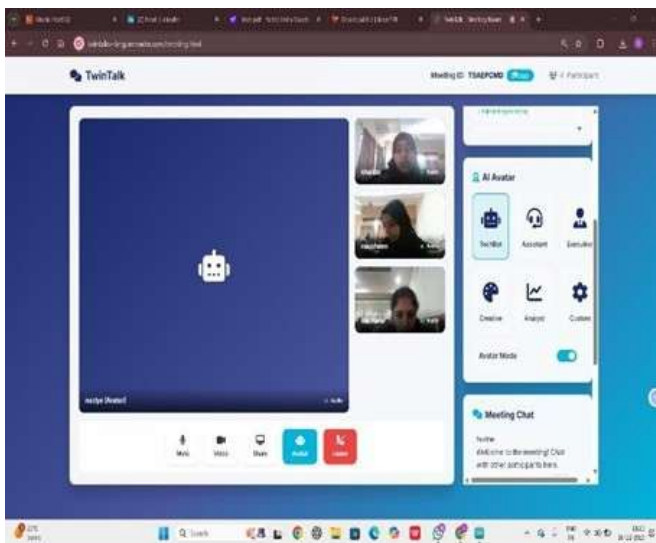


Fig. 6. Participants represented in the system with 2D avatar icons

The automatic recording module correctly captured audio streams from all participants throughout the meeting sessions. The recorded audio was successfully processed by the Microsoft Speech-to-Text service, producing accurate and readable transcripts for different speakers. The generated transcripts were then passed to Microsoft Language services, which produced concise summaries that reflected the main discussion points of each meeting. The meeting history module stored the generated transcripts and summaries correctly, allowing users to access previous meeting records after the session ended.

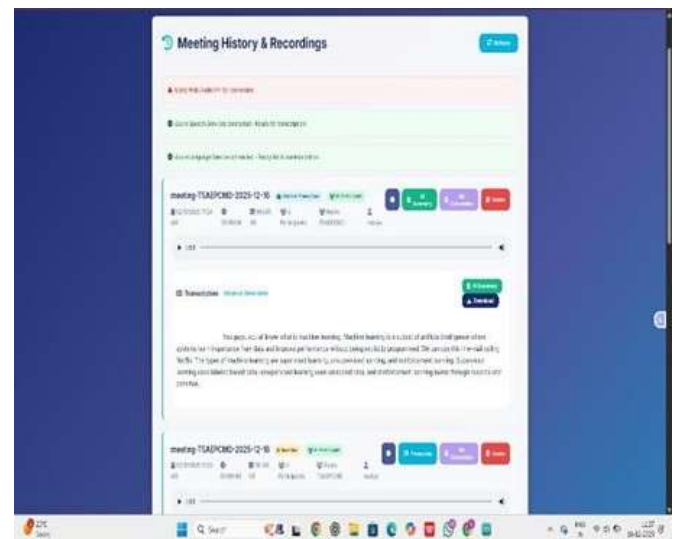


Fig. 7. Meeting history interface where generated summaries are stored for future reference

This verified the reliability of the data storage and retrieval mechanisms. The automatic meeting termination feature was also evaluated. When all participants left the meeting and the number of active users became less than one, the system automatically ended the meeting session. This confirmed the correct functioning of the session monitoring and termination logic.

Overall, the experimental results demonstrate that the proposed system is capable of supporting real-time online meetings with 2D avatar-based presence, chatbox communication, automatic transcription, and summarization using Microsoft cloud services. The results indicate that the platform improves usability and productivity by reducing manual documentation effort and by providing reliable access to meeting history for future reference.

V. CONCLUSION AND FUTURE WORK

This paper aims to design and demonstrate an online platform where members can join a group with a meeting code. Inside, the participant can have real-time conversations with a 2D avatar. The minutes of the meeting are logged with the help of Microsoft's Speech-to-Text, and people can join at the appointed time and even set their 2D avatars to unavailable, signaling the end of the meeting when everyone is offline. In other words, the end-product increases usability, productivity, and accessibility.

Thus, this technology is poised for future enhancements in various ways: multilingual chat, interpretation, and summaries; more lifelike and personalized 2D avatars of each participant that are more real; and AI-powered decision support tools to track decisions, action items, and questions using OpenAI and similar provider's language-model technology. This could be extended to audio and video summarization, especially when shared screens are utilized in presentations, and speech summarization under noisy environments. However, security

and authentication related issues would be the molders in how we will develop a truly smart and virtual meeting area.

REFERENCES

- [1] K. Huang, Y. Li, and S. Liu, "DreamWaltz: Scene-level Animatable Avatars from Internet Videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 1234-1245.
- [2] A. Aliakbarian, M. R. Osmani, and H. Escalante, "HMD-NeMo: Online 3D Avatar Motion Generation from Sparse Observations," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 2, pp. 1-13, 2024.
- [3] L. Zhang, J. Wu, and T. Chen, "MagicAvatar: Multimodal Avatar Generation and Animation for Interactive Communication," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 5, pp. 2678-2690, 2024.
- [4] K. Huang, X. Zhao, and L. Wang, "WildAvatar: Learning 3D Avatars In-the-Wild from Large-Scale Internet Data," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 543-559.
- [5] Y. Wu, Q. Li, and M. J. Black, "AnimPortrait3D: Text-Based Animatable Avatars via Morphable Model Alignment," *Computer Graphics Forum*, vol. 43, no. 7, pp. 1122-1135, 2024.
- [6] S. Yoshida, Y. Koyama, and Y. Ushiku, "Towards AI-Mediated Avatar-Based Telecommunication: Investigating Visual Impression of Switching Between Avatars," in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, 2024.
- [7] A. Zainudin, "Collaborative Edge Learning-Assisted Interactive Avatar Integration with Human Activity Recognition," *Engineering Reports*, vol. 7, no. 2, pp. 45-62, 2025.
- [8] R. Zhu, "Avatar Design in Metaverse: The Effect of Avatar-User Similarity on Task Engagement," *Computers in Human Behavior*, vol. 139, pp. 107-120, 2023.
- [9] G. C. Dobre, "Avatars in Mixed-Reality Meetings: Effects on Communication, Presence, and Mood Perception," *University of Augsburg Technical Report*, no. TR-2025-01, 2025.
- [10] C. Gasch, "A Study on User Acceptance, Preferences, and Technostress in Virtual Meetings with Avatars," *Applied Sciences*, vol. 15, no. 6, pp. 2345-2360, 2025.
- [11] T. Menzel, "Avatars for the Masses: Smartphone-Based Real-Time Dynamic Avatar Reconstruction for Virtual Interaction," *Virtual Reality Journal*, vol. 29, no. 3, pp. 123-145, 2025.
- [12] M. Kohistani, S. Momand, and A. F. Zhwak, "AI Driven Avatars in Virtual Reality: A Systematic Literature Review on Intelligent Agents for Enhancing Human-Computer Interaction," *Journal of Intelligent Systems*, vol. 34, no. 1, pp. 89-112, 2025.
- [13] E. Nan, "Real-Time Optimization and Implementation of a Digital Avatar with Speech and Interaction Modules," M.S. thesis, Dept. Comput. Sci., Northeastern Univ., Boston, MA, 2025.
- [14] R. Bidram, "AI Avatars in the Metaverse: Bridging Technology and Human Engagement," *Metaverse Journal*, vol. 12, no. 4, pp. 67-85, 2025.
- [15] J. Kang, "Real-Time Translation of Upper-Body Gestures to Virtual Avatars," *IEEE Transactions on Games*, vol. 17, no. 2, pp. 189-205, 2025.