IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, April 2026
Impact Factor: 8.2

Vaktar AI: A Conceptual Framework for AI-Based Talking Avatar Video Generation with Conversational Intelligence

Pritesh Patil, Osin Somkuwar, Shreeja Mahale, Kshitij Kumavat
Information Technology Department
Second Year, Information Technology
AISSMS Institute of Information Technology, Pune, India
pritesh.patil@aissmsioit.org, shreejamahale999@gmail.com,
osinsomkuwar26@gmail.com, kshitijkumavat.ks@gmail.com

Abstract: *Artificial Intelligence (AI) has significantly trans- formed digital content generation, particularly in the domain of interactive media[1], [5]. This paper presents the conceptu- alization of Vaktar AI, a system designed to generate personal- ized talking avatar videos and enable conversational interaction through video-based responses.*

The proposed system enables the generation of lip- synchronized avatar videos from static images using audio-driven facial animation techniques[3], [9]. In addition, it incorporates conversational AI capabilities that generate context-aware re- sponses, extending traditional text or audio-based systems into visual communication formats[1].

The integration of video generation and conversational intelli- gence represents a significant advancement over existing systems, which typically support these functionalities independently[4], [7]. By combining multimodal AI techniques, the system enhances user engagement and interaction[5].

This study focuses on the conceptual design, working princi- ples, system architecture, and advantages of the proposed frame- work, highlighting its potential to redefine human-computer interaction through visual and interactive communication..

Keywords: AI Avatars, Video Generation, Conversational AI, Lip Synchronization, Human-Computer Interaction, Digital Identity.

I. INTRODUCTION

The fast pace of innovation in the field of Artificial Intelligence (AI) and Deep Learning (DL) has resulted in considerable advancements in content generation tools[1], [5]. Video content is one of the most effective means of communi- cation used in education, marketing, and online engagement. Nonetheless, traditional video content generation remains highly resource-demanding, laborious, and time-consuming.

AI-powered avatar generation systems have been developed to address these challenges, enabling video content creation through digitized characters and audio-driven facial animation techniques[3], [9]. Although such technologies simplify the video creation process, most existing solutions lack personal- ization and interactive capabilities[4].

Conversational AI systems are capable of generating mean- ingful and context-aware responses[1]. However, these systems are typically limited to text or audio outputs and do not provide visual representations. As a result, a gap remains between visual communication and intelligent response generation.

This paper introduces Vaktar AI, a proposed system that integrates talking avatar video generation with conversational AI, thereby enabling interactive and visually enriched commu- nication.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, April 2026
Impact Factor: 8.2

II. RELATED WORK

The area of AI-generated media has grown rapidly with the emergence of avatar video generation systems and con- versational AI systems[4], [5]. Current systems like Synthesia and D-ID provide options for creating avatar videos; however, these systems mainly utilize preset avatars and scripted con- tent.

In deep learning, video generation has been made more realistic through the use of techniques such as lip synchroniza- tion and facial animation [3], [9]. In parallel, advancements in Natural Language Processing (NLP) have enabled conversa- tional AI systems to generate context-aware and human-like responses[1].

Nevertheless, current systems tend to approach video gener- ation and conversational AI separately. Video generation sys- tems primarily focus on visual output, whereas conversational AI systems prioritize text or audio-based responses[4], [7].

This paper aims to address this limitation by introducing a system where responses from conversational AI are presented in video form using personalized avatars, thereby combining visual generation and intelligent interaction into a unified framework.

III. LITERATURE REVIEW

The area of artificial intelligence (AI) media generation has developed immensely due to advances in deep learning, computer vision, and natural language processing [1], [5]. The following subsection highlights important research areas related to the proposed system – namely, AI-driven generators, lip sync models, and conversational AI systems.

The development of generative AI was initiated by the emergence of Generative Adversarial Networks (GANs). With GANs, researchers were able to develop highly realistic syn- thetic videos and images[6]. Although GAN-based technolo- gies have become the basis for current avatar and deepfake solutions by using dataset-driven methods to generate realistic content, they lack any interactivity[7].

Deep learning has further led to the development of lip synchronization models, allowing the generation of talking avatars whose lip movements coincide with generated audio streams[3], [9]. As with GAN-based generators, although the lip-syncing models can produce highly realistic audio- visual results, the models themselves usually process pre-synchronized inputs without any interaction capabilities.

Concurrently, conversational AI models have been devel- oped based on transformer networks that allow machines to comprehend context and generate human-like responses[1]. They have become common in chatbots and virtual assistants. Nevertheless, the responses provided by such technologies remain limited to textual or audio formats, without any visual representation.

Multimodal AI models have gained traction recently, with researchers focusing on incorporating various types of infor- mation, including text, images, and audio, to improve user interaction[5]. Such models have proven to be more effective and engaging than their unimodal counterparts. Nonetheless, most applications in existence do not yet consider video production and conversation as complementary processes.

An important problem that has been noted in the literature is the scarcity of models that fuse personalized talking avatar video generation with real-time conversational AI[4]. Most applications are concerned either with video production from scripts or responding to queries through conversational AI models in text or audio formats.

The proposed Vaktar AI model seeks to address this chal- lenge by offering a seamless platform for conveying conver- sational AI responses via generated talking avatar videos.

Most existing systems focus on generating responses in text or audio formats. However, there is limited research on extending these outputs into interactive visual representations, highlighting the need for systems like Vaktar AI.

IV. PROPOSED ARCHITECTURE

The proposed system, named Vaktar AI, represents an evolved form of traditional AI-based content generation sys- tems by introducing a visual layer to existing communication methods that primarily rely on text and audio [1], [5].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, April 2026 Impact Factor: 8.2

While current AI systems effectively generate textual and auditory outputs, they lack visual representation, which plays a crucial role in enhancing communication and user engagement.

To address this limitation, the proposed system generates talking avatar videos by integrating both visual and auditory modalities into a unified process. By utilizing avatar-based communication, the system combines the strengths of traditional audio-based AI systems with visual elements such as facial expressions and lip movements, thereby improving realism and expressiveness [3], [9].

In this context, the proposed system can be considered an advanced extension of conventional AI systems, where video generation capabilities are integrated with existing text and audio generation processes. This multimodal approach enhances interaction quality and provides a more immersive communication experience compared to traditional systems [5].

A. System Perspective

Conceptually speaking, the system acts like a pipeline for multimodal transformations where various inputs undergo processing to be transformed into a single video output [5]. In order to perform these transformations, the system undertakes various steps such as interpretation of input, production of responses, speech synthesis, and visualization [1], [2], [3].

In contrast to traditional systems where each of these steps is done individually, Vaktar AI performs all of them as part of an integrated workflow [4], [5].

B. Core Components

The system is organized according to two main components:

1) Talking Avatar Video Generation

This component centers on translating static information received from users into videos [3], [9]. The system takes an input in the form of a picture and text and produces video output wherein the avatar ”speaks” the input. The component involves the following steps:

a) Interpretation of input text provided by the user [1]

b) Conversion of text to speech [2]

c) Mapping of speech to mouth movements [3], [10]

d) Generation of output in the form of video that contains the animated character talking [9]

e) With this component, users will be able to produce video content without the need for actual record- ings and video editing, which makes the process of content production easier.

2) Conversational Video Interaction (AI Brain)

The second component turns the otherwise passive sys- tem into an interactive one [1], [8]. In this case, the AI communicates with the user through video messages instead of text responses. The interaction process in this case is outlined as follows:

a) Receipt of user inputs in the form of text/voice messages

b) Analysis of the inputs by the system using AI and generation of responses [1]

c) Conversion of responses into speech [2]

d) Displaying outputs as videos featuring animated avatars [3], [9]

TABLE I: LITERATURE REVIEW SUMMARY

Ref.	Year	System / Study Focus	Techniques Used	Key Strengths	Limitations	Research Gap Identified
[1]	2020	AI Content Generation Survey	Review of AI, ML & DL techniques	Provides comprehensive overview of AI media evolution	Lacks implementation of integrated systems	No unified system combining multiple AI modalities
[2]	20	Generative	GAN-based	Produces highly realistic	No interaction or	Focuses only on visual

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, April 2026 Impact Factor: 8.2

	17	Adversarial Networks (GANs)	image/video synthesis	synthetic visuals	conversational capability	generation
[3]	2022	Lip Synchronization Models	Deep learning-based lip sync	Accurate alignment of speech and lip movements	Works only on pre-defined or scripted inputs	No real-time conversational integration
[4]	2023	Conversational AI Systems	Transformer-based NLP models	Generates context-aware human-like responses	Output limited to text/audio formats	Lacks visual/video-based interaction
[5]	2024	AI Avatar Video Platforms	Multimodal avatar generation systems	Enhances engagement using avatars	Uses predefined avatars, limited personalization	No support for user-image-based avatars
[6]	2024	Multimodal AI Systems	Integration of text, audio, and vision	Improves interaction and system intelligence	Complex system design and limited real-time capability	Weak integration of video + conversational AI

e) By incorporating this component, the system shifts away from a traditional chatbot interface to a conversational one [8].

V. SYSTEM WORKFLOW

The system workflow of Vaktar AI is a set procedure through which the information fed by the user is translated into a dynamic talking avatar video format [5]. The design of this pipeline ensures efficient information processing, effective integration of various artificial intelligence modules, and consistent production of quality outputs.

It all starts from the first stage of information acquisition when a user enters an image together with his or her textual input. The former serves as the basis of an avatar created while the latter determines what should be conveyed semantically.

The third stage of the input processing stage occurs when the generated response to the user's request is created [1]. If there is no request, then a contextual response to the inputted information is produced. The main function of this stage is to ensure coherence of information. In the next stage, content transformation takes place when the generated text is turned into speech sounds [2].

Afterwards, the process of synchronization and animation mapping comes into play, where the created speech will be aligned with the lip movements of the avatar [3], [9], [10]. Through this process, there will be temporal synchronization of the audio and visual data, making it more realistic and engaging to the user.

Following this, the video generation process occurs, which is the process of animating the avatar through the synchroniza- tion of the previous stages and outputting an animated talking avatar video [3], [9]. This process involves the combination of the two forms of streams to generate one output. Lastly, the output delivery stage involves the delivering of the created video to the end user. The process aims to ensure that the output delivered is not only accessible but reusable [5].

VI. APPLICATIONS OF THE PROPOSED SYSTEM

The suggested AI model, which can create personalized videos, is useful in various fields because of its ability to make personalized videos [5]. The creation of personalized videos by integrating video creation and conversation allows users to have better communication than those who communicate using text or voice-based systems [1], [5].

• Education and E-Learning

The software can be deployed to develop interactive e-learning materials where ideas are presented through

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, April 2026
Impact Factor: 8.2

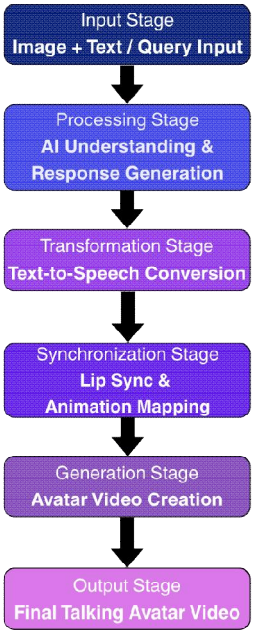

Fig. 1. Workflow of Vaktar

talking avatars [3], [9]. This helps to increase engage- ment, simplify difficult material, and improve knowledge retention.

• Content Creation and Social Media

Content producers can easily produce videos within min- utes without requiring any recording or editing technol- ogy [5].

• Corporate Communication & Business Meetings Businesses can take advantage of videos that feature talking avatars to send out news, conduct meetings, and deliver lectures without having to invest in video production services [5].

• Customer Support and Virtual Assistants

The software can be deployed in customer support to create videos instead of using the traditional chatbot systems [1], [8].

• Personal Branding & Digital Identity

The platform gives individuals an opportunity to design avatars that they can use in various online interactions [5].

• Healthcare & Awareness Campaigns

The platform enables users to use their avatars to com- municate about medical information and awareness cam- paigns [5].

VII. ADVANTAGES

These include the following with regard to usability, inte- gration of technologies, and effectiveness of communication:

• Improved Human-Computer Interaction

By using video output of AI rather than the conventional text, this would make it easier for the user to interact naturally with the machine [1], [8].

• Unique Personalization Using User-Supplied Pictures Using the pictures supplied by the user, it will be possible to create a personal avatar that will improve authenticity and enable digital identity establishment [5].

• Unification of Video Output and Conversation AI Unlike other current solutions which treat video genera- tion and conversation separately, in our case both aspects will be brought together [4], [5].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, April 2026 Impact Factor: 8.2

• Visual Communication for Improved Comprehension and Knowledge Retention
Through visual communication using avatar videos, knowledge and comprehension are improved [5].
• Lesser Dependency on Camera, Setup, and Editing Devices for Video Content Creation
The proposed system negates the need for camera, setup, and editing software, thereby reducing the difficulty involved in creating video content [5].
• Scalability of the System for Content Generation The system is capable of producing several videos simul- taneously, thus making the process scalable for purposes of teaching, advertising, and communications [5].
• Time and Cost Reduction
By automating the video content creation process, less time and money are required as compared to the manual process [5].
• Expandable Design
The conceptual model facilitates future improvements, which include emotion recognition, voice synthesis, and live rendering capabilities [8].
• Consistency in Video Quality
Automation guarantees a consistent level of video quality and prevents any variance that is observed in the manual creation process [5].
• A Connection between the Output of Audio AI and Visual Communication
The output from this system bridges the gap between the output of traditional audio AI and visual communication [5].

VIII. LIMITATIONS

Although the suggested Vaktar AI has an impressive feature set, there are certain challenges that need to be considered:
• Computationally Intensive Process
The generation of the avatar video includes several pro- cesses like voice synthesis, synchronization, and animation, and each of them requires powerful computational abilities, especially if the system is used by many people at once [2], [3], [9].
• Lag While Creating Videos
The process of creating the videos can be lagged because of their complexity, and this makes interaction difficult in the current version of the concept [3], [9].
• Dependency on the Quality of the Input Image
The quality of the resulting avatar depends significantly on the resolution and lighting of the picture the user provides [9].
• Scalability Issues
Scaling the system to work simultaneously with many users requires a strong infrastructure and optimizations; this can make the system complicated and costly [5].
• Potential Ethical and Misuse Concerns
The ability to generate realistic talking avatars raises concerns related to misuse, including the creation of misleading or deceptive content, similar to deepfake technologies [7].
• Technical Limitations to Audio-Visual Synchroniza- tion
Synchronizing speech and lip movement perfectly con- tinues to be difficult in a variety of languages [3], [10].
• Technical Limitations Relating to Infrastructure and Cost
For large-scale deployment, there may be a need for cloud-based GPUs, which would raise costs [5].

IX. CONCLUSION

In this paper, the conceptual framework of Vaktar AI will be discussed. This framework focuses on the integration of talking avatar video generation with conversational artificial intelligence [1], [3], [5]. In other words, the framework will discuss an advanced way of using artificial intelligence to generate videos rather than text or audio.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, April 2026 Impact Factor: 8.2

This concept entails the integration of the visual element into conversational artificial intelligence [5]. By generating visually appealing videos in real-time, this technique helps to bridge the gap between the static AI-based responses and human-like communication [8].

This is because AI technology helps to integrate video generation based on the user input which creates unique and visually appealing content [3], [9]. Such integration helps to develop a new form of digital communication where visual presentation of messages makes them more informative, ac- cessible, and engaging [5].

In conclusion, the use of talking avatars represents a novel way of integrating different elements of AI technologies into a single concept and a powerful tool for creating content [4], [5]. With the development of AI techniques, talking avatars and similar solutions are likely to play a key role in the future of communication [8].

REFERENCES

[1] N. Patwardhan, S. Marrone, and C. Sansone, "Transformers in the real world: A survey on NLP applications," Information, vol. 14, no. 4, p. 242, 2023.

[2] E. Casanova, K. Davis, E. Go¨lge, G. Go¨knar, I. Gulea, L. Hart, et al., "XTTS: A massively multilingual zero-shot text-to-speech model," arXiv preprint arXiv:2406.04904, 2024.

[3] Y. Xie, F. Ma, Y. Bin, Y. He, and F. Yu, "Audio-driven talking face video generation with joint uncertainty learning," in Proc. Int. Conf. Multimedia Retrieval (ICMR), pp. 1588–1597, 2025.

[4] M. Toshpulatov, W. Lee, and S. Lee, "Talking human face generation: A survey," Expert Systems with Applications, vol. 219, p. 119678, 2023.

[5] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, L. Liu, et al., "Multimodal image synthesis and editing: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., 2023.

[6] D. Saxena and J. Cao, "Generative adversarial networks (GANs): Challenges, solutions, and future directions," ACM Computing Surveys, vol. 54, no. 3, pp. 1–42, 2021.

[7] Y. Patel, S. Tanwar, R. Gupta, P. Bhattacharya, I. E. Davidson, R. Nyameko, et al., "Deepfake generation and detection: Case study and challenges," IEEE Access, vol. 11, pp. 143296–143323, 2023.

[8] T. Ki, S. Jang, J. Jo, J. Yoon, and S. J. Hwang, "Avatar Forcing: Real- Time Interactive Head Avatar Generation for Natural Conversation," arXiv preprint arXiv:2601.00664, 2026.

[9] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, et al., "SadTalker: Learning realistic 3D motion coefficients for stylized audio- driven single image talking face animation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 8652–8661, 2023.

[10] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talk- ing face generation with a high-resolution audio-visual dataset," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3661–3670, 2021.