International Journal
Information Technology Applications (ITA)

Volume 15, Number 2, August 2026

INTELLIGENT DEEPFAKE AVATARS FOR INTERACTIVE DIGITAL HERITAGE APPLICATIONS

Selma Rizvić, Aya Ali Al Zayat, Danijel Briški, Ján Lacko

Abstract:

Digital technologies have become an important means for the preservation and presentation of cultural heritage, enabling new forms of engagement through immersive and interactive experiences. Virtual reality (VR) heritage applications increasingly employ interactive digital storytelling to allow users to explore historical environments and narratives. In such environments, intelligent virtual characters can serve as guides or storytellers, supporting user engagement and facilitating interpretation of cultural heritage content. However, the visual fidelity and perceived realism of these avatars play a crucial role in shaping the overall user experience. This paper presents a method and supporting tool for the development of intelligent deepfake video avatars intended for use in digital heritage applications. The approach is based on video footage of a human actor recorded against a green screen and processed to produce a deepfake-based conversational avatar capable of interacting with users within immersive environments. The proposed approach was implemented in two case-study VR heritage applications. A user experience evaluation was conducted to assess usability and user perception of the avatar interaction. The results indicate that deepfake avatars achieve above-average usability scores and contribute positively to user engagement and immersion. The presented workflow provides a replicable method for integrating realistic conversational avatars into a wide range of digital heritage applications.

Keywords:

Digital heritage, deepfake avatars, intelligent virtual agents, virtual reality, interactive digital storytelling, user experience evaluation.

ACM Computing Classification System:

Computing methodologies - Artificial intelligence - Computer vision; Computing methodologies - Artificial intelligence - Natural language generation; Computing methodologies - Artificial intelligence - Intelligent agents; Human-centered computing - Interaction paradigms - Virtual reality; Applied computing - Arts and humanities - Media arts.

Introduction

Deepfake technology is one of the most prominent fields within artificial intelligence (AI), with a growing number of applications in both creative industries and content manipulation [23] [15]. In its broadest sense, deepfake refers to the creation of synthetic media where an individual's likeness is superimposed onto another person's using sophisticated image processing algorithms. This technology utilizes advanced deep learning methods to generate video or audio depicting fictitious events and statements [2]. Deep learning, a subfield of AI, emulates the neural processing of the human brain for tasks like speech recognition, object detection, language translation, and decision-making [18]. Among the most significant architectures in deep learning are Generative Adversarial Networks (GANs), introduced by Goodfellow et al. [6]. GANs, which typically leverage Convolutional Neural

Networks (CNNs) as their core components, enable the generation of highly realistic images and videos, including the synthesis of novel photorealistic faces or the high-fidelity replacement of existing ones [12].

The goal of this article is to present a tool for developing a deepfake video chatbot using footage of an actor recorded against a green screen. This approach enables the creation of interactive avatars capable of realistically simulating human communication, which presents numerous opportunities for application in education, entertainment, digital marketing, and simulations. This work will detail the end-to-end process of deepfake video creation, including video segmentation, voice-driven lip synchronization, and the generation of a final video composite. Particular focus is placed on creating a video chatbot - an avatar that leverages pre-recorded video for interactive communication. A case study will demonstrate the application of this process, enabling the dynamic synthesis of facial expressions and speech driven by text input generated from user queries.

The development of real-time deepfake chatbots marks the next stage in the evolution of human-machine interaction, blurring the line between digital assistants and human interlocutors. By integrating Large Language Model (LLM) based response generation, Text-to-Speech (TTS) synthesis, and real-time facial animation, it is possible to create highly realistic virtual entities capable of engaging in fluid, human-like conversation. These innovations are poised to revolutionize sectors such as customer support, digital marketing, and education by offering more interactive and personalized user experiences. Research in this domain facilitates the optimization of these systems for effective deployment and aids in identifying and addressing the challenges inherent in their implementation.

The primary objective of this research is to develop and evaluate a prototype real-time deepfake chatbot that integrates the aforementioned technologies: LLM-based response generation, speech synthesis, and real-time facial synthesis. The research is focused on understanding the technical capabilities and limitations of such a system and its potential for practical, real-world applications. Particular emphasis will be placed on system performance, the quality of user interaction, and the preservation of a transparent video background - a critical feature for seamless integration into diverse virtual environments.

1 State of the Art

1.1 Use of deepfakes in industries

Deepfakes have diversified significantly in recent years. This chapter details these various types along with their wide spectrum of applications, drawing from a comprehensive 2019 review by Kietzmann et al. that outlined the commercial uses of deepfake technologies [11]. The key classifications from their work are summarized in (Tab.1). Below we introduce key industries which are actively applying deepfake technology already.

1.1.1 Gaming and film industry

Deepfake technology offers a revolutionary alternative to traditional filming and animation in the film and gaming industries. Producers can generate realistic digital characters by manipulating existing actor footage or creating entirely synthetic ones [4]. This innovation enhances production efficiency. This technology can significantly lower production costs by providing an alternative to using stunt doubles or actor stand-ins. Instead of hours of filming, existing footage can be manipulated to meet new scene requirements. It also facilitates the real-time generation of realistic facial animations synchronized to dialogue, creating vivid characters without laborious manual animation [24]. The adoption of this technology is evidenced by film studios hiring deepfake artists [1].

AI-driven chatbots are achieving increasingly human-like interactions, capable of expressing nuanced reactions like confusion or surprise that enhance the illusion of reality [21]. This allows for the creation of non-player characters (NPCs) in games that can engage in spontaneous, natural conversations reflecting their unique personalities and emotions, providing players with a more immersive experience [9] [14].

Table 1. Types of deepfakes, their descriptions, and business applications.

Type	Description	Application
Photo Deepfakes	*Face and Body Swapping:* Altering, replacing, or blending a person's face or body with another's.	Virtual try-on for cosmetics, eyewear, hairstyles, and apparel.
Audio Deepfakes	*Voice Swapping:* Altering a voice or mimicking that of another person. *Text-to-Speech:* Modifying audio in a recording by supplying new text.	Creating diverse narrator voices for audiobooks (e.g. varying ages, genders, dialects). Correcting mispronunciations or script changes in post-production without re-recording.
Video Deepfakes	*Face Swapping:* Replacing a person's face in a video with another's. *Face Transformation:* Seamlessly morphing one face into another. *Full Body Transformation:* Transferring the bodily movements of one person onto another.	Placing an actor's face onto a stunt double's body for realistic action sequences. Allowing video game players to map their own faces onto game characters. Enabling public figures to conceal physical ailments during video appearances.
Audio-Visual Deepfakes	*Lip Synchronization:* Altering a speaker's mouth movement in a video to match a new audio track.	Dubbing advertisements and educational content into different languages while retaining the original speaker's voice.

1.1.2 Medicine

A notable medical application is the Revoice project, an initiative that uses deepfake audio technology to restore a personalized voice to patients who have lost the ability to speak due to conditions like throat cancer or neurological disorders. By training on prior recordings, the technology can synthesize a voice that retains the patient's unique tone, preserving their identity and facilitating communication. This project exemplifies the positive, assistive potential of deepfake technology [22].

Furthermore, chatbots are used in medical training to simulate patient interactions, allowing healthcare professionals to practice and refine their diagnostic and communication skills. These chatbots can be programmed with diverse scenarios and patient responses, helping trainees improve their ability to ask pertinent questions, interpret symptoms, and provide effective support in a controlled environment. This is particularly beneficial for training new medical professionals and for the continuing education of experienced practitioners [3].

1.1.3 Metaverse

In the metaverse, deepfake technology enables the creation of highly realistic digital avatars, fostering deeper user immersion in virtual environments. Major technology companies like Meta, Microsoft, and Nvidia are exploring deepfakes to enhance their respective metaverse platforms (Horizon Worlds, Mesh, and Omniverse).

While this promises more engaging virtual interactions, it also introduces significant risks, including impersonation and the spread of misinformation, which could undermine user trust and security [19].

The expanding application of deepfake technology across industries highlights its sophisticated and growing influence. It is being used to enhance user experiences, optimize processes, and create novel forms of interaction. While its full potential is still being explored, the role of deepfake technology is set to expand, presenting new opportunities alongside challenges that demand a responsible and ethical approach.

1.2 Models used for deepfake creation

1.2.1 Speech2Vid

The Speech2Vid model is designed to generate a talking-head video from a single static image of a person and a corresponding audio segment of speech. The model inputs a target face image and an audio clip and outputs a video where the facial movements are synchronized with the audio. A key feature noted by the authors is that the input image does not need to be from the training dataset, making Speech2Vid generalizable to novel faces and speech [10].

The Speech2Vid architecture comprises four main components: two separate encoders for audio and image data streams, a decoder that generates image frames conditioned on the audio signal, and a CNN-based deblurring module to enhance output frame quality. The video is generated frame-by-frame, processing the audio in 0.35-second sequences.

The encoder's architecture is a CNN based on AlexNet and VGG-M, with filter sizes adapted to the input dimensions. The identity encoder takes a static 112 x 112 x 3 image as input. To ensure the resulting identity vector captures features relevant to facial recognition, this encoder uses a VGG-M network pre-trained on the VGG Face dataset. The decoder receives the concatenated 256 dimensional feature vectors from the FC7 layers of both the audio and identity encoders. These vectors are progressively upscaled using transposed convolutions. The network also incorporates two skip connections, which merge encoder activations with decoder activations to preserve key identity features throughout the generation process [10].

1.2.2 Wav2Lip

Wav2Lip is an advanced deep learning model for synchronizing lip movements with an arbitrary audio track. Unlike previous models that only penalized reconstruction loss, Wav2Lip employs a sophisticated discriminator that evaluates the quality of the lip-sync itself. Furthermore, the model is speaker-agnostic, making it universally applicable to any face in any video, regardless of identity or language [16]. The key contributions of this work are [16]:

- The introduction of the Wav2Lip network, which achieves significantly higher accuracy and speed in lip synchronization compared to prior methods. It functions effectively on arbitrary speech and video without constraints.
- The development of a novel evaluation framework, including new metrics, to enable fair and robust assessment of lip-sync quality in unrestricted, "in-the-wild" videos.
- The creation of the ReSyncED dataset, a new benchmark designed to evaluate model performance on previously unseen subjects and scenarios.
- Wav2Lip is the first speaker-independent model to achieve lip-sync accuracy on par with real, synchronized video footage.

2 Design of the Application

The implementation process of the deepfake chatbot system represents a complex engineering endeavor that integrates generative artificial intelligence technologies, digital signal processing, and modern web services. The following sections detail the developmental path, from establishing the infrastructural environment to solving specific problems regarding the non-linear synchronization of audio-visual data.

2.1 Hardware Environment and Local Infrastructure Support

For the purposes of this study, the system is deployed on a dedicated workstation running the Windows operating system, optimized for intensive graphical computations. Local implementation ensures low latency in data transfer between the processing unit and video memory, which is critical for real-time synchronization of speech and imagery.

2.1.1 GPU Architecture Validation and CUDA Support

The core of the system's processing power is the NVIDIA graphics architecture. The first step of the automated implementation involves a system check for the presence of an NVIDIA GPU via the nvidia-smi diagnostic interface. This check is crucial because the system utilizes CUDA (Compute Unified Device Architecture) for the parallelization of deep learning algorithms. Without adequate NVIDIA drivers and GPU recognition, the system would be forced to utilize the Central Processing Unit (CPU), leading to unacceptably long video frame rendering times.

2.1.2 Workflow Architecture (Pipeline) and System Execution

Following the successful configuration of all components, the system enters an operational readiness phase controlled through an execution loop. This architecture allows the chatbot to remain active in a standby state, waiting for new input data. Local file access enables temporary data, such as audio recordings and generated PNG frames, to be processed directly on the workstation's SSD, dramatically accelerating the multiplexing phase of merging image and sound into the final WebM format.

2.2 Textual Processing Module: Llama and Groq Optimization

The heart of the intelligent interaction is Llama-3.3-70b-versatile, an exceptionally efficient model that provides performance comparable to massive systems but with the speed and cost-effectiveness of a much smaller model [7]. It is optimized for complex logical reasoning, code generation, and precise instruction following with a large context window. It represents the pinnacle of Meta's offerings for users seeking the best balance between high-end intelligence and practical real-time application.

To achieve real-time performance, integration via the Groq AI interface [8] has been implemented. Groq utilizes an innovative processor architecture that eliminates traditional memory bandwidth bottlenecks characteristic of standard CPU and GPU units. A key detail involves the query processing: with each call, a specific document (knowledge base) provided by the user is passed to the system along with the user's question. This document, combined with a defined system prompt, serves as the primary reference source, achieving high response accuracy based on factual data rather than solely on the model's internal knowledge.

2.3 Implementation of Neural Speech Synthesis via Edge-TTS Technology

The system converts generated text into speech using Edge-TTS (Microsoft Edge Text-to-Speech), a key component for achieving vocal realism. This module supports the project's autonomy,

enabling high-quality human voices without commercial API keys or external billing, fully aligning with the philosophy of a locally-hosted system.

The vocalization process uses an asynchronous programming paradigm, allowing communication with the voice generation service without blocking the main processing threads. This optimizes workstation resources and ensures a smooth transition from text to audio. The module sends text via an asynchronous client, using deep neural networks to model intonation, rhythm, and prosody, avoiding the robotic artifacts of older systems. The audio stream is encoded in real time and stored locally as an MP3 file.

2.4 Wav2Lip Model Optimization for WebM Format

A central part of the implementation involves modifying the Wav2Lip algorithm for lip synchronization. The standard model is designed to generate MP4 output, which is unfavorable for web applications requiring transparency.

2.4.1 Frame Analysis and Mel Spectrograms

The synchronization process begins by decomposing audio recordings into Mel spectrograms. Using the Short-Time Fourier Transform (STFT), the sound signal is transformed into the frequency domain on a non-linear scale. The Mel scale was chosen as it mimics human sound perception, emphasizing lower frequencies that carry the most information regarding lip formations during speech. In each batch, the model receives a visual frame of the face and the corresponding segment of the Mel spectrogram. The neural network then predicts the pixel transformations in the mouth region.

2.4.2 OpenCV Encoding Issues and Transition to PNG Sequences

Initial attempts at direct recording into WebM format using the OpenCV VideoWriter function failed due to incompatibility of FFmpeg tags for the VP9 codec within the library. A system error occurred indicating that the VP90 tag is not supported. The solution was found by changing the data storage paradigm: instead of direct video stream encoding, the system was modified to save each generated frame as a PNG image. This approach offers three advantages:

- Avoids quality loss caused by premature compression.
- Enables explicit preservation of the alpha channel (transparency) at the file level.
- Provides greater flexibility during final video composition.

2.5 Technical Challenges of Transparency and Alpha Channels

2.5.1 Data Loss in BGR Conversion

One of the greatest challenges was the inherent way OpenCV handles video. When loading frames using the cv2.VideoCapture function, data is converted to BGR format (3 channels). If the input video has an alpha channel (4th channel), it is automatically discarded. This means that even a transparent input video would become opaque as soon as it enters the model processing phase.

2.5.2 Chroma Key Implementation and FFmpeg Filtering

Following unsuccessful attempts to use the PyAV library (which caused memory allocation errors like malloc_consolidate), a chroma keying technique was applied in post-production. The logic is based on generating video on a uniform background of a specific hexadecimal value (e.g., pure green 0x00FF00).

In the final stage, FFmpeg is used with advanced color key filters. The processing command traverses all PNG frames, identifies the green color, and replaces it with transparency using the yuva420p pixel format. Similarity and blend parameters are precisely tuned to avoid "green artifacts" around hair and face edges, a common problem in automated background removal.

2.6 Summary of the Implementation Workflow

The final system operates as a closed loop: a user's textual query passes through the Llama model, audio is generated via Edge-TTS, audio is converted into Mel spectrograms, Wav2Lip modifies the PNG frames, and FFmpeg performs the final composition of the WebM video with an alpha channel. This process, although complex and composed of multiple heterogeneous modules, is optimized to provide results in near real-time, fulfilling the vision of an interactive deepfake assistant.

2.7 Integrated User Interface and Modular Tool

To consolidate all the previously mentioned modules into a functional system, a specialized tool with a graphical user interface (GUI) was developed. This application serves as a central orchestrator, allowing the user to manage the entire pipeline from a single environment. The tool is designed to accept several key inputs:

- Input Video: The source video of the persona that will be animated.
- User Inquiry: The textual question or prompt for the chatbot.
- Knowledge Base Path: A file path to the document used for RAG (Retrieval-Augmented Generation) to ensure factual accuracy.
- API Configuration: A dedicated field for the Groq API key to facilitate high-speed inference.

Furthermore, the interface, shown in (Fig.1), provides high levels of customization, allowing users to select the specific Llama model version, the target language, and the gender of the voice for the Edge-TTS synthesis, while also granting full control over native Wav2Lip parameters such as padding, rescale factor, smoothing, and batch size for optimal synchronization and performance.

The primary goal of this tool is to demonstrate the core functionality of a deepfake chatbot. However, the architecture is highly modular and designed for future expansion, including features such as direct voice input (STT - Speech to Text), real-time streaming, and integration into third-party environments or VR/AR platforms.

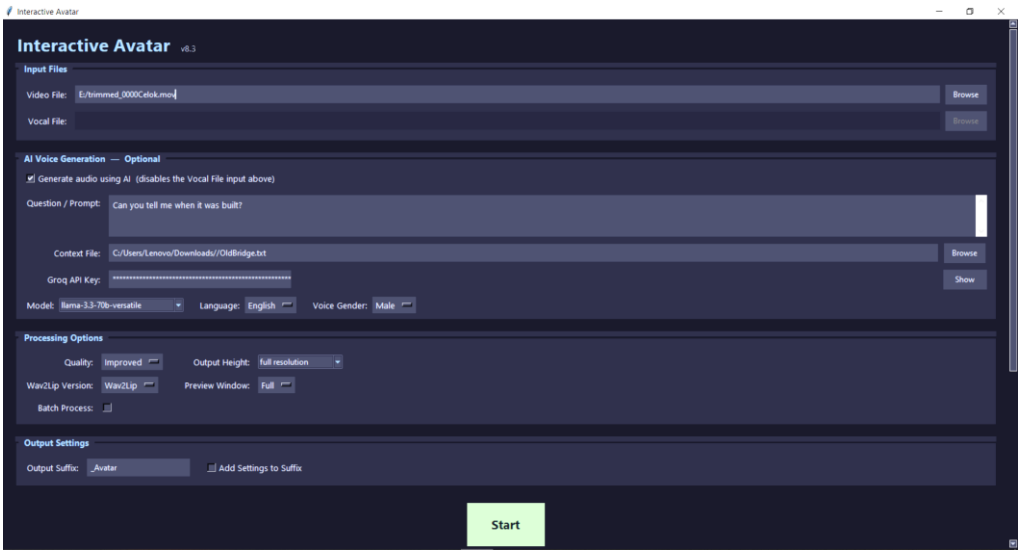

Fig.1. The integrated system interface showing input parameters and model selection.

Below is an example of the system's output. The following transformation demonstrates the process, starting from the source video of an actor filmed against a chroma key (green screen) background shown in (Fig.2), and resulting in the processed output presented in (Fig.3).

2.8 Practical Implementation in the Unity Environment

In addition to the standalone tool, a Unity-based project was developed to demonstrate the system's potential within interactive 3D environments. This implementation serves as a specialized use-case scenario, showcasing how the module can be integrated into broader gaming or simulation contexts. By embedding the deepfake chatbot into a virtual space, users can interact with digital personas or environmental elements that provide dynamic, AI-generated responses based on real-time game states or player progress. This architecture ensures that the chatbot is not merely a passive source of information but a reactive component that adapts to the player's investigative approach, providing personalized guidance and enhancing immersion.

Fig.2. Original input: actor on a green screen.

The visual layout and final design of this implementation are shown in (Fig.4), which presents the interaction interface within the 3D scene.

To connect all components within Unity, a dedicated C# script was developed to manage the communication between the Unity client and the backend server. The system captures the user's input from a specialized text field, encodes the string for URL compatibility, and sends an asynchronous request to the server.

The core logic handles the transmission of data and the subsequent retrieval of the generated video file. Once the server processes the request and returns the video content, the script saves the file locally to the device's persistent data path. This ensures reliable loading regardless of the hardware platform. Following the successful download, the system automatically configures the internal video player and audio components, transitioning from a loading state to the playback of the synchronized deepfake response. Upon completion of the video, the system triggers a final event to return the interface to its original state, completing a seamless interactive loop.

This Unity project demonstrates the flexibility of the deepfake chatbot, proving that it can be easily integrated into various interactive applications to provide personalized video communication.

Fig.3. Final output: synchronized deepfake with transparency and synthesized speech.

Fig.4. Conceptual design and final look of the Unity integration project.

2.9 Performance Analysis: WebM Generation and Alpha Channel Processing

The final testing phase focuses on the system's operational speed when generating video in WebM format with integrated transparency through chroma key removal. A critical aspect of this analysis involved testing the impact of video resolution on processing time, specifically comparing the original resolution (1280 x 720) with a lower resolution (640 x 360). For this analysis, the total processing time was decomposed into three key segments:

1. Model Inference: Frame generation using the modified Wav2Lip model.
2. Chroma Key Removal: Identifying and removing the green background from each frame.
3. Final Composition: Merging processed frames and audio into the final WebM container.

Measurements were conducted using a 10-second source video with audio lengths of 9, 6, and 3 seconds. The results, highlighting the significant performance gains achieved by reducing the resolution, are summarized in (Tab.2).

Table 2. Processing time comparison between Original (720p) and Lower (360p) resolution (in seconds).

	Audio: 9s		**Audio: 6s**		**Audio: 3s**	
Process Segment	**720p**	**360p**	**720p**	**360p**	**720p**	**360p**
Model Inference	38.62	38.62	28.26	28.26	20.25	20.25
Chroma Key Removal	22.01	22.01	13.78	13.78	6.83	6.83
WebM Video Creation	120.26	64.02	79.32	41.64	38.83	21.27
Total Process Time	180.89	124.65	121.36	83.68	65.91	48.35

The results demonstrate a substantial improvement in the "WebM Video Creation" phase when using the lower resolution. For instance, in the 9-second audio test, the composition time was nearly halved, reducing the total duration from 180.89 to 124.65 seconds. This trade-off is particularly beneficial for real-time applications where processing speed is prioritized over maximum visual fidelity.

3 Qualitative User Evaluation

3.1 Storytelling Applications for Evaluation

To confirm the proposed methods of using intelligent 2D avatars in the form of fake video in real deployment, two applications with local content related to monuments within the Slovak Republic were created.

The first of them was dedicated to partial construction development and historical events at Devín Castle, which is located near the capital of the Slovak Republic, Bratislava, above the confluence of the Danube and Morava rivers. The castle represented a key element in the defense of the territory and its administration during individual stages of historical development. The application processed the development of the castle in the Garay period and during the siege by the Ottoman troops. Within the application in virtual reality, the user is accompanied by an avatar with predefined scenarios in individual implemented scenes with the possibility of interaction within the scene. Subsequently, the user has the opportunity to interact with the avatar in direct communication.

The second application implemented was a virtual representation of the defunct narrow-gauge forest railway from the area of Topolčany town, which was implemented by the prominent family of Baron Stummer at the turn of the 19th and 20th centuries. Three scenes were implemented in this application. Communication with an avatar in the form of a moving image in the mansion - the seat of the Stummer family, the implementation of loading wood onto the forest railway in the Kulhan area and the arrival of the train for the first time to the sawmill and parquet factory in Topolčany as part of a ceremonial welcome. Both applications were implemented using the Unreal engine after which qualitative usability testing was carried out.

3.2 Qualitative User Evaluation of the VR Cultural Heritage Applications

Both applications integrate interactive scenes, task-based exploration, and a 2D fake-video avatar capable of conversational interaction in Slovak. The aim of the evaluation was to analyze:

- usability of the VR interaction
- clarity of educational content
- user engagement
- effectiveness of avatar-based communication
- perceived immersion and historical understanding

A qualitative evaluation using the Think-Aloud protocol [13] and a quantitative evaluation using the System Usability Scale (SUS) usability method [5] were conducted with 32 participants. Evaluated applications consist of the following scenes.

3.2.1 Devín Castle VR Application

Scene 1 - Introduction

Users enter an introductory environment presenting the historical context of the castle. A deepfake based avatar appears and introduces the historical development of the castle. The avatar delivers a short monologue in Slovak describing the importance of the location and the early development of the fortress.

Scene 2 - Building the Castle Well

Users participate in a simulation of constructing the castle well. User interactions include selecting tools, interacting with building elements and completing construction steps. After the interaction, users can speak with the avatar about construction techniques, castle infrastructure and historical periods of development.

Scene 3 - Defending the Castle Against the Ottomans

The final scene simulates a historical defensive situation. Users can operate defensive elements, observe battlefield events and respond to instructions. After completing the scene, the avatar provides additional explanations about military strategy and historical conflicts.

3.2.2 Stummer Forest Railway VR Application

Scene 1 - Mansion Interior

The first scene takes place in the interior of the Stummer mansion. It consists of a historical environment reconstruction where a portrait painting contains a 2D avatar. This avatar introduces the history of the Stummer family and their industrial activities.

Scene 2 - Forest Railway Operation

The second scene takes place in a forest environment with the railway. Users perform tasks containing carrying wood, assisting with loading operations and controlling the loading process onto the train.

Scene 3 - First Train Arrival to the Sawmill

The final scene reconstructs the first train arrival to the sawmill and parquet factory. The avatar explains industrial processes, wood processing and historical economic development of the region.

3.3 Methodology

3.3.1 Think-Aloud Protocol

Participants were instructed to verbalize their thoughts while interacting with the VR application, comment on confusion, expectations, and impressions. During the interview they had to describe their actions and interpretations. Sessions were audio recorded and observed by researchers. Each session lasted approximately 25-30 minutes. The evaluation followed these steps:

1. Introduction and explanation of the Think-Aloud method.
2. Brief familiarization with VR controls.
3. Interaction with Devín Castle VR.
4. Short break.
5. Interaction with Stummer Railway VR.
6. Semi-structured interview.

Observations and verbalizations were transcribed and coded. Qualitative data were analysed using thematic coding. Statements were categorized into:

- usability
- interaction
- educational value
- avatar communication
- immersion
- technical issues

Results

A total of 32 participants (Male - 19, Female - 13, Age range - 19-54 years, Average age - 27.6) took part in the study. Participants were recruited from university students, cultural heritage enthusiasts, and general public visitors. Their previous VR experience was as follows:

- No experience - 11 users.
- Occasional VR use - 14 users.
- Frequent VR users - 7 users.

A total of 514 user comments were recorded. From the point of view of overall user experience both applications were generally positively received as seen in (Tab.3).

Table 3. User comments of VR applications.

Category	Positive comments	Neutral comments	Negative comments
Devín Castle	142	21	18
Stummer Railway	138	26	19

Users highlighted a strong sense of presence, engaging historical storytelling and intuitive interactions. Avatar communication is a key feature. Users frequently commented on the deepfake based avatar interaction. Positive observations included:

- natural Slovak speech
- helpful explanations
- improved understanding of history

Example user statements:

- "It feels like a guide explaining the castle."
- "The avatar makes the information easier to understand."

We also identified the issues reported by some users related to uncertainty about when conversation could start and the desire for more spontaneous dialogue options.

Interaction and tasks were also positively evaluated by users. In the Devín Castle VR application users particularly enjoyed the well construction activity and defensive interaction. However, some participants needed time to understand task objectives. In the Stummer Railway VR application users appreciated manual work simulation and "physical" interaction with logs. Several users mentioned the loading interaction felt more game-like than educational.

Educational value of the applications was significant. Perceived learning (high - 21 users, moderate - 8 users, low - 3 users). Participants demonstrated improved understanding of castle defensive strategies, historical construction methods and early industrial forestry. Users reported strong immersion due to environmental realism, task participation, and narrative guidance.

Users perceived Devín Castle as more educational, while Stummer Railway felt more interactive and playful. The comparative analysis of applications is shown in (Tab.4).

The following usability problems were identified:

- Interaction discoverability: some users did not immediately understand which objects were interactive (observed in 10 participants).
- Avatar dialogue activation: users sometimes attempted to talk before the system allowed it (observed in 8 participants).
- Task instructions: certain tasks lacked clear visual guidance (observed in 6 participants).

Participants proposed several improvements based on more dialogue options with the avatar, clearer interaction indicators, additional historical scenes and a multiplayer educational mode.

Table 4. Comparative analysis of VR applications.

Criterion	Devín Castle	Stummer Railway
Historical narrative clarity	High	Medium
Interaction engagement	Medium	High
Educational content	High	Medium
Immersion	High	High

3.3.2 SUS Usability Evaluation

After completing interaction with each VR application, participants filled in the SUS questionnaire, which consists of 10 statements rated on a 5-point Likert scale:

- 1 - Strongly disagree
- 2 - Disagree
- 3 - Neutral
- 4 - Agree
- 5 - Strongly agree

Odd questions measure positive usability, even questions measure negative usability. The SUS statements (questions) are the following:

(1) I think that I would like to use this system frequently.
(2) I found the system unnecessarily complex.
(3) I thought the system was easy to use.
(4) I think that I would need the support of a technical person to use this system.

(5) I found the various functions in this system were well integrated.
(6) I thought there was too much inconsistency in this system.
(7) I would imagine that most people would learn to use this system very quickly.
(8) I found the system very cumbersome to use.
(9) I felt very confident using the system.
(10) I needed to learn a lot of things before I could get going with this system.

In the SUS scoring procedure, we computed for each participant the following values:

- Odd questions: score - 1
- Even questions: 5 - score
- Sum of adjusted values x 2.5

This produced a score from 0-100. After scoring we were able to interpret the usability with the SUS score (<68 - average usability, 69-78 - good usability, >80 - excellent usability).

Results

In (Tab.5) we summarise the mean responses per question by users for both applications. The responses obtained from the SUS questionnaire were processed using the standard SUS scoring procedure. For each item, the raw Likert scores were transformed into adjusted values to normalize positive and negative statements within the scale. In (Tab.6), we present the average adjusted values for odd and even questions, which were subsequently used to compute the overall SUS score for the evaluated VR applications.

Table 5. Mean response per question.

Question	Mean score (Devín Castle)	Mean score (Stummer Railway)
Q1	4.2	4.0
Q2	1.9	2.1
Q3	4.3	4.1
Q4	1.7	1.9
Q5	4.1	3.9
Q6	2.0	2.2
Q7	4.2	4.0
Q8	1.8	2.0
Q9	4.0	3.8
Q10	2.1	2.3

Table 6. Adjusted SUS calculation.

Question Type	Mean Adjusted Value (Devín Castle)	Mean Adjusted Value (Stummer Railway)
Odd questions	3.16	3.02
Even questions	3.08	2.90
Total adjusted sum	(3.16x5)+(3.08x5)=31.2	(3.02x5)+(2.90x5)=29.6
Final SUS score	31.2x2.5=78	29.6x2.5=74

From these results we can see that the Devín Castle application achieved the score: Good usability with high acceptance. Participants particularly appreciated the clear historical narrative, intuitive scene interactions and helpful avatar guidance. For Stummer Railway the achieved score was: Above-average usability. The slightly lower score was due to task complexity and interaction learning. Both applications achieved above-average usability. The Devín Castle VR application scored slightly higher, mainly due to clearer narrative structure and simpler task interactions, while

the Stummer Railway application received slightly lower scores due to more complex physical interactions and task discovery issues reported in the qualitative evaluation.

4 Discussion: Moral and Ethical Aspects

One of the most significant ethical challenges inherent in the development and application of deepfake technology is its capacity to generate visually and audibly convincing content capable of audience deception. Manipulated videos, which falsely depict individuals saying or doing things they never actually performed, possess the potential to be exploited in deliberate disinformation campaigns. Consequently, such content can precipitate a range of adverse consequences, including blackmail, public opinion manipulation, sabotage, and the destabilization of social and political structures. In the digital age, characterized by the rapid dissemination of news and the frequent bypassing of verification filters, deepfake technology further obfuscates the boundary between truth and falsehood, thereby complicating users' ability to discern reliable sources from manipulated ones. Even when the audience does not fully accept the authenticity of deepfake content, the technology's mere existence fosters an atmosphere of suspicion and skepticism. Research indicates that the prevalence of deepfake technology in the media landscape results in a general erosion of trust, impacting not only specific content but also institutional sources of information. This phenomenon reduces the public's willingness to accept the authenticity of any public statements or recordings. In such an environment, the sense of security and orientation within the digital space is severely compromised [20].

To date, only a handful of governments have adopted regulations specifically concerning artificial intelligence, while the majority have yet to do so, primarily due to concerns regarding the potential restriction of freedom of speech. Although most internet platforms, such as YouTube, have implemented certain legal mechanisms for content control, this process remains costly and time-consuming. The central challenge is that deepfake content often remains undetected by the naked eye. This difficulty has prompted both state institutions and private platforms to develop deepfake detection technologies and establish regulations governing their use. Proposed methods for combating misuse include the mandatory labeling of generated media, the development of forensic video analysis software, and the strengthening of the criminal-legal framework for cases of digital fraud [17]. Given the increasing ease of content creation and technological advancement, the detection of malicious content is an escalating challenge that demands continuous future work and investment.

Conclusion

In this paper, we presented a methodology for creating a functional deepfake chatbot, emphasizing the delivery of a convincing, real-time user experience. The model's current response time is within acceptable limits for a satisfactory level of interaction, and further convergence toward a true "real-time" response is anticipated with future technological developments, especially concerning the availability and capability of graphics cards for end-users.

During implementation, particular attention was devoted to optimizing resource consumption to ensure the fastest and most stable performance in response generation. In this context, the Wav2Lip model was utilized, which enables the synchronization of lip movements with the audio signal while imposing a minimal system load. By focusing exclusively on the lip region, this model reduces the complexity of image processing, which facilitates faster real-time response generation. This approach proved highly effective for chatbot interaction, where response speed is a critical factor in user experience. Furthermore, the functionality for removing the video background was successfully implemented, thereby providing an additional degree of flexibility and expanding the potential application of this solution across a wider spectrum of uses.

This work establishes a foundation for future research and optimization directed toward creating faster, more efficient, and more accessible real-time deepfake systems. Currently, the majority of the processing time remains dedicated to preserving visual transparency, which significantly impacts overall system performance. Consequently, it would be advantageous to consider implementing a predefined and optimized background. This strategy would reduce the number of operations the system must execute in real-time, thereby minimizing the need for additional algorithmic processing and shortening the response latency.

A crucial element for future implementation concerns the model's execution method. Given the considerable processing demands and the limited graphical resources on typical user devices, the optimal solution is to rely on a robust server-side infrastructure. Utilizing more powerful servers for model execution would ensure more stable and faster operation, reduce the computational load on the end-user device, and enable the broader adoption and usability of such solutions across diverse platforms and devices.

Finally, it is essential to stress that such systems are not yet fully prepared for the automation of complex or sensitive communication scenarios. To prevent the spread of irrelevant, inaccurate, or potentially harmful information, input queries must be carefully restricted, ensuring they are thematically limited and controlled. To mitigate misuse, these aspects must be addressed during the development phase through the implementation of robust security mechanisms, the limitation of output content, and the transparency of the algorithm's operation. This caution is especially critical when the technology is used in areas involving socially sensitive content. Ultimately, this technology necessitates continuous human oversight to guarantee its responsible, ethical, and secure application in real-world environments.

Acknowledgement

This work was supported by the project and the contributors involved in the development and evaluation of the presented VR cultural heritage applications. Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495.

References

[1] BBC (n.d.). Shamook: Star Wars effects company ILM hires Mandalorian deepfaker. Retrieved online, March 9, 2026, from https://www.bbc.com/news/entertainment-arts-57996094.

[2] Alanazi, S., Asif, S. (2024). Exploring deepfake technology: creation, consequences and countermeasures. Human-Intelligent Systems Integration. vol.6, no.1, pp.49-60.

[3] Battegazzorre, E., Bottino, A., Lamberti, F. (2020). Training medical communication skills with virtual patients: Literature review and directions for future research. In International Conference on Intelligent Technologies for Interactive Entertainment. Springer, pp.207-226.

[4] Bode, L., Lees, D., Golding, D. (2021). The digital face and deepfakes on screen. pp.849-854.

[5] Brooke, J. et al. (1996). SUS - A quick and dirty usability scale. Usability evaluation in industry. vol.189, no.194, pp.4-7.

[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM. vol.63, no.11, pp.139-144.

[7] Groq (2025). Llama-3.3-70B-Versatile. Retrieved online, February 27, 2026, from https://console.groq.com/docs/model/llama-3.3-70b-versatile.

[8] Groq Inc. (2026). Groq. Retrieved online, February 27, 2026, from https://groq.com/.

[9] Hardiman, J.P.W., Thio, D.C., Zakiyyah, A.Y. et al. (2024). AI-powered dialogues and quests generation in role-playing games using Google's Gemini and Sentence BERT framework. Procedia Computer Science. vol.245, pp.1111-1119.

[10] Jamaludin, A., Chung, J.S., Zisserman, A. (2019). You said that?: Synthesising talking faces from audio. International Journal of Computer Vision. vol.127, no.11, pp.1767-1779.

[11] Kietzmann, J., Lee, L.W., McCarthy, I.P., Kietzmann, T.C. (2020). Deepfakes: Trick or treat? Business Horizons. vol.63, no.2, pp.135-146.

[12] Kumar, M., Sharma, H.K. et al. (2023). A GAN-based model of deepfake detection in social media. Procedia Computer Science. vol.218, pp.2153-2162.

[13] Lewis, C., Rieman, J. (1993). Task-centered user interface design: A practical introduction.

[14] Liu, X., Xie, Z., Jiang, S. (2025). Personalized non-player characters: A framework for character-consistent dialogue generation. AI. vol.6, no.5, p.93.

[15] Masood, M., Nawaz, M., Malik, K.M., Javed, A., Irtaza, A., Malik, H. (2023). Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence. vol.53, no.4, pp.3974-4026.

[16] Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.V. (2020). A lip sync expert is all you need for speech to lip generation in the wild. In Proceedings of the 28th ACM International Conference on Multimedia. pp.484-492.

[17] Ramluckan, T. (2024). Deepfakes: The legal implications. In International Conference on Cyber Warfare and Security. vol.19, Academic Conferences International Limited, pp.282-288.

[18] Shiri, F.M., Perumal, T., Mustapha, N., Mohamed, R. (2023). A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU. arXiv preprint arXiv:2305.17473.

[19] Tariq, S., Abuadbba, A., Moore, K. (2023). Deepfake in the metaverse: Security implications for virtual gaming, meetings, and offices. In Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes. pp.16-19.

[20] Vaccari, C., Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media + Society. vol.6, no.1, p.13.

[21] Wang, X., Li, X., Yin, Z., Wu, Y., Liu, J. (2023). Emotional intelligence of large language models. Journal of Pacific Rim Psychology. vol.17, 18344909231213958.

[22] Whittaker, L., Kietzmann, T.C., Kietzmann, J., Dabirian, A. (2020). "All around me are synthetic faces": the mad world of AI-generated media. IT Professional. vol.22, no.5, pp.90-99.

[23] Zehra, Z. et al. (2023). Aesthetics of deepfake - sphere of art and entertainment industry. Facta Universitatis, Series: Visual Arts and Music. vol.1, pp.087-100.

[24] Zhang, C., Zhao, Y., Huang, Y., Zeng, M., Ni, S., Budagavi, M., Guo, X. (2021). FACIAL: Synthesizing dynamic talking face with implicit attribute learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision. pp.3867-3876.

Authors

Selma Rizvić, prof. Phd.
Faculty of Electrical Engineering
University of Sarajevo, Bosnia and Herzegovina
srizvic@etf.unsa.ba

Aya Ali Al Zayat, M.Sc.
Faculty of Electrical Engineering
University of Sarajevo, Bosnia and Herzegovina
aalialzaya1@etf.unsa.ba

Danijel Briški, M.Sc.
Faculty of Electrical Engineering
University of Sarajevo, Bosnia and Herzegovina
dbriski1@etf.unsa.ba

Ján Lacko, RNDr. PhD.
Faculty of Informatics
Pan-european university Bratislava, Slovakia
jan.lacko@paneurouni.com