A Platform for Interactive AI Character Experiences

RAFAEL WAMPFLER, ETH Zurich, Switzerland
CHEN YANG, ETH Zurich, Switzerland
DILLON ELSTE, ETH Zurich, Switzerland
NIKOLA KOVAČEVIĆ, ETH Zurich, Switzerland
PHILINE WITZIG, ETH Zurich, Switzerland
MARKUS GROSS, ETH Zurich, Switzerland

Fig. 1. Our system enables conversational and story-driven experiences with a digital character. Users can engage with the character on any topic through speech while the character mimics human behavior naturally.

From movie characters to modern science fiction — bringing characters into interactive, story-driven conversations has captured imaginations across generations. Achieving this vision is highly challenging and requires much more than just language modeling. It involves numerous complex AI challenges, such as conversational AI, maintaining character integrity, managing personality and emotions, handling knowledge and memory, synthesizing voice, generating animations, enabling real-world interactions, and integration with physical environments. Recent advancements in the development of foundation models, prompt engineering, and fine-tuning for downstream tasks have enabled researchers to address these individual challenges. However, combining these technologies for interactive characters remains an open problem. We present a system and platform for conveniently designing believable digital characters, enabling a conversational and story-driven experience while providing solutions to all of the technical challenges. As a proof-of-concept, we introduce *Digital Einstein*, which allows users to engage in conversations with a digital representation of Albert Einstein about his life, research, and persona. While *Digital Einstein* exemplifies our methods for a specific character, our system is flexible and generalizes to any story-driven or conversational character. By unifying these diverse AI components into a single, easy-to-adapt platform, our work paves the way for immersive character experiences, turning the dream of lifelike, story-based interactions into a reality.

CCS Concepts: • **Human-centered computing** → **Interaction techniques**; • **Computing methodologies** → **Discourse, dialogue and pragmatics**; **Procedural animation**; **Motion capture**; *Natural language generation*; *Speech recognition.*

Additional Key Words and Phrases: Conversational AI, Digital Characters, Embodied Conversational Agents, Large Language Models, Interactive Storytelling, Character Consistency, Personality Modeling, Memory Systems, Speech-Driven Animation, Speech Synthesis, Multimodal Interaction

ACM Reference Format:
Rafael Wampfler, Chen Yang, Dillon Elste, Nikola Kovačević, Philine Witzig, and Markus Gross. 2025. A Platform for Interactive AI Character Experiences. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25), August 10–14, 2025, Vancouver, BC, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3721238.3730762

Authors' Contact Information: Rafael Wampfler, ETH Zurich, Zurich, Switzerland, rafael.wampfler@inf.ethz.ch; Chen Yang, ETH Zurich, Zurich, Switzerland, chen.yang@inf.ethz.ch; Dillon Elste, ETH Zurich, Zurich, Switzerland, delste@inf.ethz.ch; Nikola Kovačević, ETH Zurich, Zurich, Switzerland, nikola.kovacevic@inf.ethz.ch; Philine Witzig, ETH Zurich, Zurich, Switzerland, philine.witzig@inf.ethz.ch; Markus Gross, ETH Zurich, Zurich, Switzerland, grossm@inf.ethz.ch.

SIGGRAPH Conference Papers '25, August 10–14, 2025, Vancouver, BC, Canada

1 Introduction

The vision of creating interactive, lifelike digital characters capable of engaging in meaningful, story-driven conversations has fascinated generations [Bates 1994; Cavazza et al. 2002; Lugrin et al. 2022]. From movie characters to digital representations of historical figures — such characters redefine how we experience storytelling and establish emotional connections to digital entities [Gong et al. 2023; Torre et al. 2019; Xu et al. 2024].

However, realizing this vision is a challenging task. It requires a seamless combination of conversational intelligence [Ramesh et al. 2017], character integrity [Schlenker 2008], personality and emotion [Bates 1994], knowledge and memory [Kope et al. 2013], voice synthesis [Torre et al. 2019], realistic animations [Wu et al. 2024], and integration into the physical environment [Li et al. 2022]. Due to this complexity, actors*actresses, as in Disney's "Turtle Talk with Crush", puppeteered characters in real-time [Casas and Mitchell 2019]. Even with major advancements in AI technology, conversational systems still struggle to provide interactive and story-driven experiences [Green and Jenkins 2014; Riedl et al. 2003]. Strategies often focus on isolated components, such as animation synthesis [Kim et al. 2024; Liu et al. 2022; Ng et al. 2024] or language modeling [Schmitt and Buschek 2021], rather than ensuring overall coherence across all components. Furthermore, requirements such as character consistency, customization, and real-time synchronization of the components often fall short [Lugrin et al. 2022].

In this work, we propose a modular system for creating believable conversational digital characters that support narrative experiences. By combining the power of large language models (LLMs) with multimodal sensing, expressive synthesis, and adaptive personality modeling, this system addresses the pertaining challenges, allowing for interactive, story-driven, and believable interactions. As a proof-of-concept, we present *Digital Einstein*, a digital representation of Albert Einstein, enabling users to have discussions on his scientific research, anecdotes from his life, and historical background. The system creates an immersive experience by integrating a story-driven AI character into a physical environment. While *Digital Einstein* only serves as an example application, the system architecture is highly modular. Individual components can be easily exchanged depending on the character and the specific target application. Thus, our work opens new possibilities for bringing interactive and believable digital characters to life.

Our system contributes several technical innovations that enable believable digital characters. It maintains character integrity using GPT-4o and a fine-tuned Llama 3 model, enhanced by synthetic conversation generation, embedding-based prompt steering, and a memory system for story consistency. Personality is dynamically adjustable, with emotional tone expressed through both speech and animation. Further, conversations are visually enriched with images generated by Midjourney. In addition, the character interprets its physical environment through a camera, enabling situational awareness. These components are integrated into a modular and extensible platform suited for diverse conversational and story-driven applications, anchored in a themed physical setup for immersive and emotionally engaging interactions.

2 Related Work

2.1 Conversational Digital Characters

Conversational digital characters have progressed from rule-based systems [Mateas and Stern 2003] to LLM-powered models [Qi et al. 2021], enabling dynamic, context-rich conversations. Personality modeling advances include dynamic personality infusion, where chatbot responses reflect predefined traits [Kovačević et al. 2024a,b]. Emotion-aware systems improve user engagement through speech emotion recognition [Hu et al. 2022] and text-based emotion detection [Kusal et al. 2024]. Further, modular architectures promote conversation consistency and scalability [Nguyen et al. 2022].

Conversational AI finds applications in education, healthcare, and storytelling. Narrative agents foster engagement through authored dialogue [Spierling 2005], while LLM-based storytelling supports coherent narratives [Li et al. 2024]. These systems simplify complex tasks and enhance accessibility [Qi et al. 2021]. However, sustaining meaningful interaction over multiple exchanges remains a core challenge. In particular, addressing memory limitations is key to maintaining multiturn coherence [Castillo-Bolado et al. 2024; Johri et al. 2025]. To overcome these memory limitations, RAG combines retrieval and generation to help chatbots maintain long-term context [Gao et al. 2024]. Dual-memory systems balance short- and long-term data for personalization [Zhang and Luo 2024] and flow, while selective memory improves user experience and retrieval efficiency [Sumida et al. 2024].

2.2 AI-Driven Narratives with Ethical Considerations

Early approaches, such as a semi-automatic artistic pipeline for recreating Einstein, demonstrated how small-scale productions could achieve realism with limited resources [Helzle and Goetz 2018]. Building on this, AI-driven conversational agents, as seen in the "Living Memories" concept, brought figures like Leonardo da Vinci to life [Pataranutaporn et al. 2023]. Larger-scale efforts, such as developing corpora for role-playing Chinese historical figures, highlighted the importance of contextual authenticity and low-resource data integration for nuanced depictions [Bai et al. 2024]. Meanwhile, ethical considerations have gained prominence. Research in "digital necromancy" examined the balance between preserving cultural heritage and addressing issues of authenticity and consent [Hutson and Ratican 2023]. Recent work shows how LLMs improve accessibility in digital humanities by generating concise portrayals of historical figures [Hasnain and Usman 2024], while ethical frameworks guide the reconstruction of narratives in education [Hutson et al. 2024].

2.3 Interactive Systems

Interactive systems have advanced conversational characters, enabling lifelike and engaging interactions. Recent frameworks align body movements with co-speech gestures, producing emotionally rich and context-aware responses [Kim et al. 2024]. Modular architectures further support such interactions by decoupling dialog management from embodiment, enabling customization and robust nonverbal communication [Santos et al. 2023]. Other end-to-end pipelines enhance virtual agents with real-time audio-video synchronization and anthropomorphic features [Rupprecht et al. 2024].

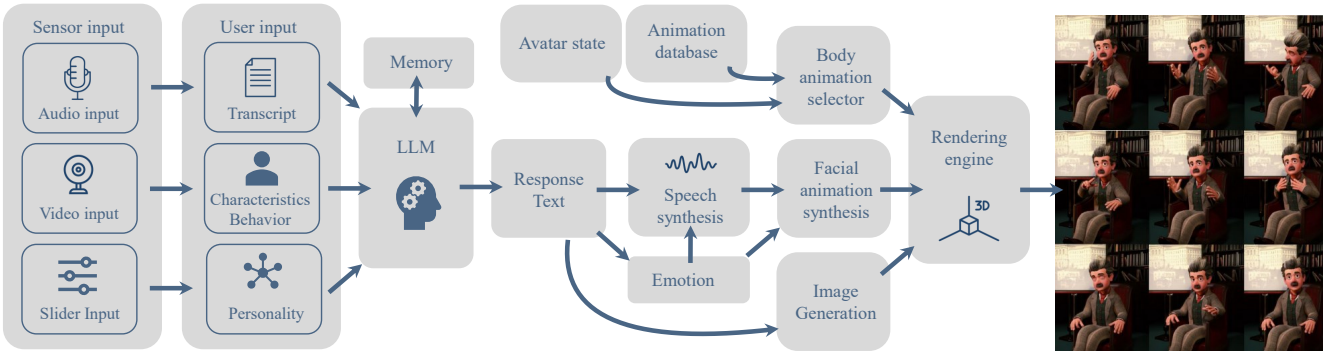

Fig. 2. System Overview: The pipeline processes sensor inputs, including transcribed speech and video-based user characteristics and behavior analysis, through an LLM-based chatbot supported by memory and an adjustable personality of the digital character. The chatbot's responses guide speech and facial animation synthesis based on emotions detected in the response, motion-captured body animation selected based on the avatar state, and image generation.

Immersive augmented and mixed reality systems also benefit from these advances. Systems combining speech recognition with real-time facial animation enhance character interactions in augmented reality [Casas and Mitchell 2023]. MoodFlow [Casas et al. 2024] extends this by using a prompt-embedded state machine to guide emotionally intelligent avatars in mixed reality. Platforms integrating vision and language models enable context-aware, real-time interactions [Maniatis et al. 2023], while hybrid systems support seamless user experiences through unobtrusive, spatially immersive interfaces [Encarnacao et al. 2000].

Unlike previous work, we address the AI challenges of story-driven and interactive conversations with believable characters through a unified, customizable framework.

3 System Design and Requirements

3.1 System Requirements

The development of our system was guided by several goals (i.e., believability, flexibility, realism) and constraints (i.e., low latency, resilience, technical complexity) that define our system requirements:

(1) *Modular and Scalable Design:* Supports easy upgrades and adaptations for diverse contexts.
(2) *User-Centric Approach:* Ensures intuitive and customizable interactions, allowing users to adjust the character's personality and tailor conversations to their preferences.
(3) *Real-Time Responsiveness:* Maintains low latency across all components to ensure seamless dialogue.
(4) *Robustness and Reliability:* Guarantees smooth operation in different settings, even under varying conditions.
(5) *Immersive Experience:* Combines a themed physical setup with spatial audio, realistic animations, and lifelike body movements synchronized with AI-generated responses to ensure natural and engaging interactions.

3.2 System Overview

Our system comprises several interconnected AI modules to address the complex AI challenges for interactive, story-driven characters. Figure 2 provides a high-level overview of the connection between our modules. A detailed interaction flow is shown in Figure 8.

The system's core component, implemented in Unity, features a digital character within a themed scene. The character transitions between four distinct states: *idle* (no interaction), *listening* (awaiting user input), *thinking* (processing input), and *speaking* (delivering responses). Inviting animations are played when a user is approaching, which is detected by the camera (our system queries the camera every 500ms). When the user sits down, the character transitions from *idle* to *speaking*, starting the conversation with a randomly selected welcome message, followed by awaiting user input. While *thinking*, the transcribed user input is processed by the cognitive module supported by an LLM-based chatbot. The system ensures character integrity by maintaining consistency in behavior and dialogue. Additionally, the character draws on its knowledge base and memory to provide contextually relevant responses. Users can further personalize their experience by adjusting the character's personality traits using physical sliders that dynamically affects response patterns. From the chatbot response, emotions are extracted to adjust speech and animations. Speech is synthesized using a fine-tuned Microsoft Azure neural voice model. The character's animations blend facial expressions, dynamically generated and synchronized from speech using Audio2Face [Karras et al. 2017], with motion-captured body movements. While the character is *speaking*, images are automatically generated using Midjourney based on the conversation context. Throughout the interaction, the character alternates between *listening*, *thinking*, and *speaking*. A fluid dialogue flow is maintained through coordinated state transitions based on user behavior and system responses. If the user remains silent for more than seven seconds, a signal is sent to the chatbot, prompting it to respond appropriately.

4 Design Challenges and Solutions

In this section, we discuss the key challenges encountered in developing a system for believable, conversational, and story-driven characters, along with our solutions.

4.1 Conversational Intelligent Chatbots

Creating believable digital characters requires sophisticated conversational abilities. The core challenge is to develop chatbots that

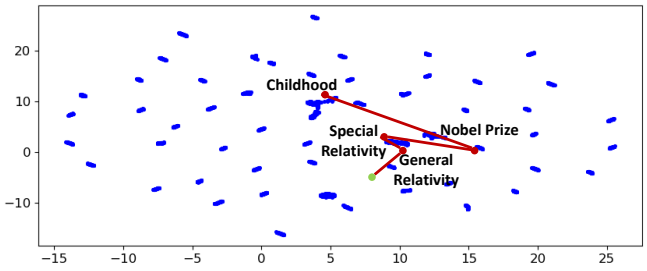

Fig. 3. Visualization of the embedding space from synthetic Einstein conversations. The clusters of blue points represent different topics. A user interaction trace is highlighted in red, with the starting point marked in green. The transition from the topic "Nobel Prize" to "childhood" is initiated by the user.

can process natural language, generate contextually appropriate responses, and maintain coherent, story-driven interactions that support user immersion. To address this challenge, we leverage state-of-the-art LLMs, specifically GPT-4o [OpenAI 2023] and Llama 3 8B [Touvron et al. 2023]. GPT-4o is used for real-time thematic consistency and high-quality engagement, while Llama 3 supports local deployment scenarios requiring privacy and cost efficiency, making it a valuable offline alternative. This dual-model setup enhances robustness, real-time responsiveness, and ensures continuity during cloud service outages, addressing requirements (3) and (4).

Character Stories and Topics. We compiled M topics related to the character's expertise and personal interests, such as hobbies and anecdotes. For the Einstein character, this set includes $M = 62$ topics in total, spanning his personal life, scientific theories, and musical interests. We used GPT-4o to generate N synthetic human-character conversations for each topic (see supplemental material). Each conversation turn t was embedded using Microsoft Azure's *text-embedding-3-large* model, producing a 3072-dimensional vector $\mathbf{e}_t$. To compute a representative vector for each conversation, we averaged its turn-level embeddings: $\mathbf{e}_{\text{conv}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t$, where T is the number of turns in the conversation. To obtain a topic-level representation, we further averaged all conversation embeddings within that topic: $\mathbf{e}_{\text{topic}}^{(j)} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{e}_{\text{conv}}^{(i,j)}, j = 1,\ldots,M$. Figure 3 shows a 2D UMAP projection where each point corresponds to $\mathbf{e}_{\text{conv}}$. Distinct clusters can be found per topic, and thematically related topics are positioned close to each other. The red trace illustrates a user's multi-topic conversation path through the embedding space.

Large Language Models. We fine-tuned Llama 3 8B to enhance its topic consistency and knowledge of character-specific topics on the raw text of the synthetic conversations. We used Axolotl [Axolotl Project 2025] and fine-tuned for three epochs with the Adam optimizer (learning rate of $2e^{-5}$, gradient accumulation of 8, cosine scheduling with 100 warmup steps) using a batch size of 1 on a cloud-based NVIDIA RTX 6000 GPU. For efficient operation, addressing requirement (3), Llama 3 8B runs locally with DeepSpeed [Rasley et al. 2020].

For GPT-4o (Microsoft Azure, version 2024-08-06), we use prompting to induce topic consistency. The system tracks the conversation flow using the embedding space. Each new turn is embedded and combined with past turns of the same session using an exponentially weighted average with a three-turn half-life ($h = 3$), giving higher weight to recent turns: $\mathbf{e}_{\text{agg}} = \sum_{k=0}^{\infty}\alpha_k\,\mathbf{e}_{t-k}$, where $\alpha_k = \frac{1}{Z}\cdot 2^{-k/h}$ and $Z = \frac{1}{1-2^{-1/h}}$. Here, $\mathbf{e}_{t-k}$ is the embedding of the turn k steps ago. The system then computes the cosine similarity between this aggregated embedding and all topic embeddings $\mathbf{e}_{\text{topic}}^{(j)}, j = 1,\ldots,M$. If the highest similarity exceeds a defined threshold, the corresponding topic j is selected as the *current topic*. To facilitate smooth topic transitions, the system performs a neighbor search between $\mathbf{e}_{\text{agg}}$ and all topic embeddings $\mathbf{e}_{\text{topic}}^{(j)}, j = 1,\ldots,M$. One of the three nearest neighbors is randomly selected as the *next topic*. The prompt is then modified to steer GPT-4o toward a smooth transition to the *next topic*.

For both models, the prompt first defines the role of the digital character and then six distinct contexts, followed by instructions and the conversation history (see Table 1 and supplemental material).

Table 1. Element descriptions and example responses for chatbot prompts.

Prompt Element	Description & Example Response
Location	Current location *Vancouver's vibrant energy at SIGGRAPH is a perfect reflection of the curious minds gathered here to shape the future of innovation.*
Scene	Descriptions of physical setup and Unity scene *This cozy setting feels perfect for deep, thoughtful conversations, as if we're sharing ideas over a cup of tea in a timeless library.*
User Description	Age, gender, number of people around, user appearance, user attention *You look ready for adventure in your blue shorts and white shirt, perfect for a curious mind like yours!*
Image Description	Metadata of displayed image *The swirling star remnants and slowing clocks perfectly capture how a black hole warps light and time, truly a cosmic wonder!*
Date	Day and time *Ah, nearly midnight, a wonderfully quiet time when the mind can wander freely and explore its most curious thoughts!*
Memory	5 Turns from past conversations
Instructions	Persona, tone, response guidelines, behavior, handling specific situations (e.g., silence), topic transitions.
Additional Data	Current conversation history

Knowledge and Memory. Our system implements a vector-based knowledge store to maintain conversation history. For each new user input, the system generates an embedding and performs a similarity search against stored conversations. The five most relevant conversation snippets are incorporated into the prompt's memory context. By focusing on topics and information most relevant to the current conversation context, this approach fulfills requirement (3).

4.2 Personality and Emotion

Beyond basic conversational abilities, a compelling digital character must master multiple dimensions of human-like interaction. Key among these is maintaining consistent personality traits while adapting to different conversational contexts, and demonstrating emotional intelligence through appropriate responses. To meet requirement (5), we integrated emotional intelligence into our system.

Fig. 4. Sliders for real-time personality adjustment across five traits: **V**ibrancy, **C**onscientiousness, **D**ecency, **A**rtificiality, and **N**euroticism.

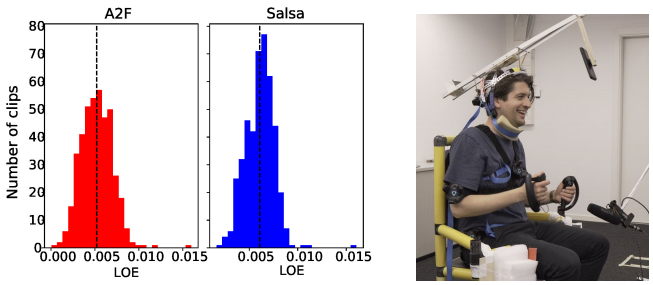

Fig. 5. Left: LOE distribution between retargeted motion-capture sequences and animations synthesized from the corresponding audio using Audio2Face (red, $\mu = 0.005$) and SALSA (blue, $\mu = 0.006$). Right: Motion-capture setup featuring eight positional trackers, two hand gesture trackers, and a full facial tracker mounted on a gimbal helmet.

An LLM can dynamically adapt the emotional tone of the responses based on the user's input [Chang 2024; Jin et al. 2024]. Thus, after response generation, we use GPT-4o-mini to determine one of 7 emotions matching the emotions supported by Audio2Face (i.e., amazement, anger, disgust, fear, joy, sadness, neutral). Furthermore, an intensity level ranging from 0.01 to 2.0 is regressed. These parameters are passed to Microsoft Azure Speech Synthesis, which adjusts its speaking style using the following mapping: amazement → excited, anger → angry, disgust → disgruntled, fear → fearful, joy → cheerful, sadness → sad, neutral → default style.

Chatbot personalities differ fundamentally from human personalities [Kovačević et al. 2024b]. We use a standalone method for dynamic personality infusion [Kovačević et al. 2024a], which rewrites the LLM responses to align them with predefined personality profiles using GPT-4o. Thereby, personality profiles are constructed from five key dimensions (vibrancy, conscientiousness, decency, artificiality, and neuroticism) on a 5-point intensity scale. To make personality control accessible to users, we built physical sliders using potentiometers, an Arduino, and a custom 3D printed case (see Figure 4) that directly adapts the prompt.

4.3 Speech Recognition and Synthesis

Realistic human-like interaction requires digital characters to process and generate oral communication effectively. Our system implements listening and speaking through Microsoft Azure's Speech Services. Speech recognition is integrated into Unity to minimize latency and maintain synchronized animations and state transitions, addressing requirement (3). The system incorporates a 1-second buffer for natural pauses in user speech before concluding the recognition process. The resulting transcribed text appears on the scene's whiteboard and is forwarded to the LLM for response generation. Our system employs Microsoft Azure's Custom Neural Voice model for speech synthesis, capable of conveying various emotional tones and intensities in the character's responses. The Custom Neural Voice model also supports voice customization, allowing fine-tuning to match specific thematic or stylistic requirements.

4.4 Animation Synthesis

In our system, facial animations are dynamically generated from the synthesized audio using cloud-based NVIDIA Audio2Face (A2F). It is a data-driven method that has been trained to align the audio signal with facial movements. Since audio primarily contains cues for lower face motion, researchers have incorporated static [Daněček et al. 2023; Peng et al. 2023; Wu et al. 2024] and dynamic [Witzig et al. 2024] emotion representations in data-driven animation models to generate lively upper face motion. A2F supports up to 10 predefined emotion labels. To ensure that the rendered facial expressions match the synthesized speech, we use the corresponding emotion label derived in Section 4.2 during animation synthesis.

To accommodate requirement (3), we developed a Python wrapper for A2F for incremental audio processing in windows of 0.5 seconds. Our wrapper streams the results directly from A2F, bypassing its default batch-oriented processing. If A2F is unavailable, we use SALSA LipSync Suite v2. While it has a higher absolute lip offset error (LOE) than A2F (see Figure 5, left), it ensures robustness against cloud service downtimes, addressing requirement (4). Furthermore, we animate the eyes procedurally: We define a look-at target at the user's head and generate saccadic eye movements.

While models like Audio2Gesture [Li et al. 2021] provide automated gesture synthesis, retargeting them to our stylized character leads to unnatural motion due to mismatched body proportions, requiring extensive manual adjustments. Instead, we animate the avatar procedurally using a curated library of motion-capture clips, categorized by the avatar state (*idle*, *speaking*, *listening*, and *thinking*). For each state, a dedicated set of clips is maintained, and a new one is randomly sampled based on the avatar's current state. If a clip ends and the state remains unchanged, a new clip is randomly selected from the same category. In Figure 5 (right), we show our custom motion capture setup for facial and body animations (see supplemental material for more details). Facial motion capture is used only for evaluation and is not part of our system.

4.5 Interaction With the Real World

For realistic interaction with the physical world, our digital character must perceive its surroundings and respond to users' non-verbal behaviors. We implemented this capability through a camera that serves as the character's "eyes", enabling it to detect user presence or absence, user characteristics, and user behavior.

Each frame is compared to a reference image of the empty armchair using the Structural Similarity Index (SSIM). If SSIM exceeds

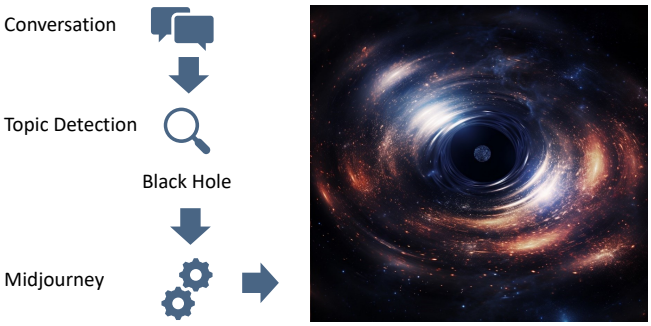

Fig. 6. The system identifies the topic from the conversation and uses it to generate a matching image (e.g., a black hole).

Fig. 7. Users engage with a character in an immersive environment that includes a screen (1), a table with a lamp, a hidden microphone, and personality sliders (2), decorative elements (3), an armchair (4), and a carpet (5).

a threshold δ, it indicates that someone is seated, triggering the conversation flow shown in Figure 8. For robustness, the conversation can also be started or stopped via designated keyboard keys, addressing requirement (4).

Furthermore, our system runs several OpenVINO [OpenVINO Toolkit 2025] models on the camera feed. First, faces are detected using the *face-detection-0200* model. If a face is detected but the SSIM is below δ, a person is likely to be close to the physical setup. The character tries to draw the user's attention by playing predefined motion-captured animations. The *age-gender-recognition-retail-00130* model classifies the user's approximate age and gender. The *head-pose-estimation-adas-0001* predicts the yaw angle of the user's head pose from a window of 15 seconds. If it exceeds 20 degrees, we classify the user as not attentive. The *face-reidentification-retail-0095* model allows for user re-identification. It returns a feature vector for the aligned face, which is then compared to a local database using cosine distance. If the distance is below 0.3, the user is recognized, enabling personalized memory retrieval. The database is updated continuously as new users are recognized.

Finally, to enrich user characterization while respecting privacy, an image of the user with the face blurred is sent to Microsoft Azure's GPT-4 Vision model. It returns descriptive details on clothing and accessories that is used as context by the chatbot.

4.6 Visual Storytelling Enhancements

Our system incorporates topic-relevant images by generating Midjourney prompts from up to five recent turns (see Figure 6). We compose the prompts with GPT-4o on Microsoft Azure and send the prompt to Midjourney through GoAPI [GoAPI 2025] to generate four image variations. The most suitable image is selected using the CLIP [Radford et al. 2021] score and then described by GPT-4 Vision on Microsoft Azure to provide context for the chatbot. An image is generated at most every two minutes and displayed for two turns or until the topic changes. As generation can take up to 32 seconds, we pre-generated five images per topic. Images are cached for reuse when similar topics arise, addressing requirement (3).

4.7 Physical Setup

The physical setup of the platform should balance thematic authenticity with functional requirements. Our system's environment evokes the aesthetics of the early 20th century while seamlessly integrating modern sensing, audio, and visual technologies (see Figures 7 and 9).

At its core is a 65-inch Samsung QM65R Public display within a custom-designed wooden frame mounted on a floor stand. The frame incorporates a hidden Logitech HD Pro Webcam C920, hidden behind a 3D-printed cover. Spatial audio experience is achieved through five speakers: two Visaton PL 8 RV speakers in the armchair, one Visaton WB 10 under the table, and two Visaton WB 10 speakers mounted behind the screen. A bass vibration module (Monacor BR-50) is integrated in the armchair. User interaction is facilitated through a microphone (Samson Go Mic Connect) hidden inside a book on the table. A unique feature of the setup is the inclusion of physical personality sliders (see Figure 4). The media box (APC AR109SH4), placed behind the screen, includes an Intel Core i9 computer with an NVIDIA RTX 3090 GPU and a Pioneer VSX-S510 AV receiver.

The design incorporates a curated selection of furniture and decor to maintain historical authenticity. An antique armchair purchased from an online auction platform was reupholstered using fireproof fabrics. The table, acquired from an online auction, was modified to house a speaker and microphone. A fireproof carpet was digitally scanned and adjusted to match early 20th-century styles. Additional decorative elements, such as a Mozart bust, antique books, a pocket watch, and postcards, were sourced from antique shops and auctions.

5 Evaluation and Application

5.1 Digital Einstein

We developed a stylized Albert Einstein avatar with an articulated body and a face rig driven by ARKit blendshapes. A stylized character mitigates the uncanny valley effect, reduces development time and cost, and enables real-time processing due to its lower level of visual detail. However, the modular design of our system supports realistic avatars as well. To synthesize Einstein's voice, we fine-tuned Microsoft Azure's Custom Neural Voice model using 1,618

recordings (average length of 4.02 seconds, SD=2.14, max=17.62) of an actor. We also captured motion data from the actor performing 7 *idle*, 12 *listening*, 1 *thinking*, and 39 *speaking* routines. Drawing from 62 core topics related to Einstein's life and research, we generated 71 synthetic human–Einstein conversations per topic. Conversations have 30 turns (SD=1.98, min=21, max=36) and comprises responses of 35 tokens (SD=13.85) on average. By embedding all 4,402 conversation transcripts, we can dynamically steer responses in GPT-4o.

5.2 Performance Analysis

Each system component achieves real-time performance to maintain a good user experience. Speech recognition introduces a one-second delay to account for natural pauses in spoken input. Personality rewriting runs in 1.03 seconds (SD = 0.11). GPT-4o and Llama 3 8B process the inputs in 1.16 seconds (SD = 0.46) and 1.6 seconds (SD = 0.51), respectively. Emotion prediction requires 0.61 seconds (SD = 0.07), and speech synthesis ends in 0.4 seconds (SD = 0.21). Synthesizing facial animations by Audio2Face begins with a 0.5-second buffer, whereas the asynchronous image generation takes 32 seconds on average. Finally, the user analysis through the webcam runs in 0.19 seconds (SD = 0.0029) but is excluded from the total, as Unity polls the webcam every 500 ms, reusing the latest result. The cumulative runtime is 4.7 seconds for GPT-4o and 5.14 seconds for Llama 3. The avatar's *thinking* state, accompanied by corresponding animations, effectively masks this latency.

5.3 User Evaluation

We conducted a user evaluation of our system by deploying the physical *Digital Einstein* setup at two large international events using GPT-4o due to its superior qualitative performance: GITEX GLOBAL 2024, a five-day tech event (374 sessions), and SIGGRAPH Asia Emerging Technologies 2024, a three-day scientific event (261 sessions). Furthermore, we collected 50 conversations with Llama 3 for comparison with the 4,402 synthetically generated conversations. Table 2 summarizes key statistics introduced by Toubia et al. [2021]: speed measures how quickly topics shift, volume represents the semantic range of the conversation, and circuitousness captures the directness of thematic progression. GPT-4o demonstrated higher engagement during the scientific event, with an average of 5.67 turns per session and longer responses (32.78 words on average) compared to the tech event (4.84 turns, 29.35 words). Llama 3 exhibited longer responses (33.10 words) but fewer turns (4.68). In particular, GPT-4o outperformed the synthetic conversations in both speed and volume metrics, indicating its ability to deliver concise yet engaging responses in real time. Llama 3 exhibited higher semantic speed (1.04), volume (0.70), and circuitousness (0.24), suggesting a more exploratory conversational trajectory, while GPT-4o, with its topic consistency mechanism, maintained greater coherence. These findings highlight our system's capacity to adapt to diverse contexts while maintaining conversational quality and thematic coherence.

Across all 62 topics, GPT-4o covered 49 and 42 topics at the tech and scientific events, respectively, compared to 22 topics for Llama 3. The topic frequencies are depicted in Figure 10. The most popular topics at the tech event included "GITEX" (241 occurrences), "Theory of Relativity" (64), and "Dubai" (61), while "SIGGRAPH Asia" (201), "Tokyo" (83), and "Theory of Relativity" (55) dominated at the scientific event. The most discussed topics in Llama 3 conversations were "Theory of Relativity" (10), "General Relativity" (10), and "Special Relativity" (5). This shows that event-specific discussions at the scientific and tech events naturally extended beyond the 62 Einstein-related topics, demonstrating adaptability to other topics.

Table 2. Quantitative analysis of conversation and engagement metrics for GPT-4o, Llama 3, and synthetic conversations. # Words indicates the average chatbot response length, with standard deviation in brackets. Metrics marked with ↑ indicate that higher values are beneficial for fostering more dynamic, contextually rich, and engaging interactions.

Statistic	GPT-4o (tech)	GPT-4o (sci)	Llama 3	Synthethic
# Sessions	374	261	50	4402
# Turns (avg)	4.84 (2.94)	5.67 (3.34)	4.68 (2.14)	30.00 (1.98)
# Words (avg)	29.35 (11.41)	32.78 (14.71)	33.10 (10.80)	25.23 (10.36)
# Topics	49	42	22	62
Speed (↑)	0.77 (0.05)	0.80 (0.05)	1.04 (0.05)	0.72 (0.05)
Volume (↑)	0.53 (0.03)	0.55 (0.04)	0.70 (0.03)	0.51 (0.04)
Circuitousness (↑)	0.02 (0.02)	0.02 (0.02)	0.24 (0.02)	0.02 (0.02)

5.4 Ethical and Privacy Considerations

Our system is designed with privacy and ethical considerations, balancing user experience and data protection. System components, such as the webcam service, can be disabled to prioritize privacy without compromising functionality, ensuring flexibility for diverse privacy norms. Critical processing tasks are conducted locally, which minimizes the transmission of sensitive data and enhances privacy protection. Furthermore, our system is adaptable to various regional privacy regulations. For example, all Microsoft Azure services utilized for speech recognition, synthesis, and language processing are hosted within a European region to guarantee GDPR compliance. As voice data contains biometric information, speech synthesis must also comply with regulations such as GDPR to protect user identity. This modular and region-aware architecture allows our system to balance rich interactive experiences with robust privacy safeguards, addressing the necessary trade-offs between enhancing system performance and preserving user privacy.

6 Conclusions and Future Work

We introduced a system that combines conversational AI with a carefully designed physical setup, demonstrating a significant step forward in the development of AI characters. Using the power of LLMs and combining it with multimodal sensing, expressive speech and facial animation synthesis, and adaptive personality modeling, we enable in-character behavior and story-consistent experiences. While we demonstrate *Digital Einstein* as an example application, our system is modular and readily extendable to other characters and their stories. This modularity allows seamless customization, enabling the creation of diverse characters and narratives tailored to various needs.

Despite these advancements, we acknowledge that certain limitations remain. Due to distributed computing, the system can occasionally experience latency in fast-paced conversations. Additionally, interrupting the interlocutor is not yet implemented, but is planned as future work. Furthermore, we will improve the animation synthesis model for more diverse and lively movements. We also intend to advance cognitive modeling to better simulate human-like understanding and reasoning. Finally, we will conduct structured user studies to assess the contribution of individual system components to user engagement and perceived immersion.

Acknowledgments

This work was supported by a Swiss National Science Foundation Grant under Grant No.: PZ00P2_216294. We thank Violaine Fayolle for modeling the *Digital Einstein* avatar and for her patience and dedication in continuously refining the rigs to meet our requirements. We also thank Patrick Karpiczenko for providing the speech recordings used to train the speech synthesis model.

References

Axolotl Project. 2025. Axolotl: Open fine-tuning framework for LLMs. https://github.com/axolotl-ai-cloud/axolotl. Accessed: May 2, 2025.

Ting Bai, Jiazheng Kang, and Jiayang Fan. 2024. BaiJia: A large-scale role-playing agent corpus of Chinese historical characters. arXiv:2412.20024 https://arxiv.org/abs/2412.20024

Joseph Bates. 1994. The role of emotion in believable agents. *Commun. ACM* 37, 7 (1994), 122–125.

Llogari Casas, Samantha Hannah, and Kenny Mitchell. 2024. MoodFlow: Orchestrating conversations with emotionally intelligent avatars in mixed reality. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE Computer Society, Los Alamitos, CA, USA, 86–89. https://doi.org/10.1109/VRW62533.2024.00021

Llogari Casas and Kenny Mitchell. 2019. Intermediated reality: A framework for communication through tele-puppetry. *Frontiers in Robotics and AI* 6 (2019), 60.

Llogari Casas and Kenny Mitchell. 2023. Intermediated reality with an AI 3D printed character. In *ACM SIGGRAPH 2023 Real-Time Live!* (Los Angeles, CA, USA) *(SIGGRAPH '23)*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/3588430.3597251

David Castillo-Bolado, Joseph Davidson, Finlay Gray, and Marek Rosa. 2024. Beyond prompts: Dynamic conversational benchmarking of large language models. arXiv:2409.20222 https://arxiv.org/abs/2409.20222

Marc Cavazza, Fred Charles, and Steven J. Mead. 2002. Interacting with virtual characters in interactive storytelling. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1* (Bologna, Italy) *(AAMAS '02)*. Association for Computing Machinery, New York, NY, USA, 318–325. https://doi.org/10.1145/544741.544819

Edward Y. Chang. 2024. Behavioral emotion analysis model for large language models. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 549–556. https://doi.org/10.1109/MIPR62202.2024.00094

Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. 2023. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers* (Sydney, NSW, Australia) *(SA '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3610548.3618183

Luis M. Encarnacao, Robert J. Barton, Oliver Bimber, and Dieter Schmalstieg. 2000. Walk-up VR: Virtual reality beyond projection screens. *IEEE Computer Graphics and Applications* 20, 6 (2000), 19–23. https://doi.org/10.1109/38.888003

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997 https://arxiv.org/abs/2312.10997

GoAPI. 2025. GoAPI: Generative AI API provider. https://goapi.ai. Accessed: May 2, 2025.

Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. 2023. Interactive story visualization with multiple characters. In *SIGGRAPH Asia 2023 Conference Papers* (Sydney, NSW, Australia) *(SA '23)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3610548.3618184

Melanie C. Green and Keenan M. Jenkins. 2014. Interactive narratives: Processes and outcomes in user-directed stories. *Journal of Communication* 64, 3 (2014), 479–500.

Muhammad Hasnain and Sardar Usman. 2024. Potential of large language models (LLMs) as supplementary tools for historical learning: Users' interaction and knowledge acquisition. *Foundation University Journal of Engineering and Applied Sciences (HEC Recognized Y Category, ISSN 2706-7351)* 4, 2 (2024), 60–66.

Volker Helzle and Kai Goetz. 2018. Digital Albert Einstein, a case study. In *ACM SIGGRAPH 2018 Talks* (Vancouver, British Columbia, Canada) *(SIGGRAPH '18)*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/3214745.3214782

Jiaxiong Hu, Yun Huang, Xiaozhu Hu, and Yingqing Xu. 2022. The acoustically emotion-aware conversational agent with speech emotion recognition and empathetic responses. *IEEE Transactions on Affective Computing* 14, 1 (2022), 17–30.

James Hutson, Paul Huffman, and Jeremiah Ratican. 2024. Digital resurrection of historical figures: A case study on Mary Sibley through customized ChatGPT. *Metaverse* 4, 2 (2024), 1–13. https://doi.org/10.54517/m.v4i2.2424

James Hutson and Jay Ratican. 2023. Life, death, and AI: Exploring digital necromancy in popular culture—Ethical considerations, technological limitations, and the pet cemetery conundrum. *Metaverse* 4, 1 (2023), 1–12.

Zhijing Jin, Nils Heil, Jiarui Liu, Shehzaad Dhuliawala, Yahang Qi, Bernhard Schölkopf, Rada Mihalcea, and Mrinmaya Sachan. 2024. Implicit personalization in language models: A systematic study. arXiv:2405.14808 https://arxiv.org/abs/2405.14808

Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Leandra A. Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M. Van Allen, David Kim, Roxana Daneshjou, and Pranav Rajpurkar. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine* 31, 1 (2025), 77–86. https://doi.org/10.1038/s41591-024-03328-5

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–12.

Sunwoo Kim, Minwook Chang, Yoonhee Kim, and Jehee Lee. 2024. Body gesture generation for multimodal conversational agents. In *SIGGRAPH Asia 2024 Conference Papers* (Tokyo, Japan) *(SA '24)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3680528.3687648

Andrew Kope, Caroline Rose, and Michael Katchabaw. 2013. Modeling autobiographical memory for believable agents. In *Proceedings of the Ninth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (Boston, MA, USA) *(AIIDE'13)*. AAAI Press, Palo Alto, California, USA, 23–29.

Nikola Kovačević, Tobias Boschung, Christian Holz, Markus Gross, and Rafael Wampfler. 2024a. Chatbots with attitude: Enhancing chatbot interactions through dynamic personality infusion. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3640794.3665543

Nikola Kovačević, Christian Holz, Markus Gross, and Rafael Wampfler. 2024b. The personality dimensions GPT-3 expresses during human-chatbot interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–36.

Sheetal D. Kusal, Shruti G. Patil, Jyoti Choudrie, and Ketan V. Kotecha. 2024. Understanding the performance of AI algorithms in text-based emotion detection for conversational agents. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 23, 8 (2024), 1–26. https://doi.org/10.1145/3643133

Changyang Li, Wanwan Li, Haikun Huang, and Lap-Fai Yu. 2022. Interactive augmented reality storytelling guided by scene semantics. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–15.

Danrui Li, Samuel S. Sohn, Sen Zhang, Che-Jui Chang, and Mubbasir Kapadia. 2024. From words to worlds: Transforming one-line prompts into multi-modal digital stories with LLM agents. In *Proceedings of the 17th ACM SIGGRAPH Conference on Motion, Interaction, and Games* (Arlington, VA, USA) *(MIG '24)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3677388.3696321

Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021. Audio2Gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 11273–11282. https://doi.org/10.1109/ICCV48922.2021.01110

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *Computer Vision – ECCV 2022* (Tel Aviv, Israel). Springer, Berlin, Heidelberg, 612–630. https://doi.org/10.1007/978-3-031-20071-7_36

Birgit Lugrin, Catherine Pelachaud, and David Traum (Eds.). 2022. *The handbook on socially interactive agents: 20 years of research on embodied conversational agents, intelligent virtual agents, and social robotics volume 2: Interactivity, platforms, application* (1 ed.). Vol. 48. Association for Computing Machinery, New York, NY, USA.

Apostolos Maniatis, Stavroula Bourou, Zacharias Anastasakis, and Kostantinos Psychogios. 2023. VOXReality: Immersive XR experiences combining language and vision AI models. In *Human Interaction and Emerging Technologies (IHIET-AI 2023): Artificial Intelligence and Future Applications*, Vol. 70. AHFE Open Access, New York, NY, USA, 139–148. https://doi.org/10.54941/ahfe1002938

Michael Mateas and Andrew Stern. 2003. Integrating plot, character and natural language processing in the interactive drama Façade. In *Proceedings of the 1st International Conference on Technologies for Interactive Digital Storytelling and Entertainment (TIDSE-03)*, Stefan Göbel, Norbert Braun, Ulrike Spierling, Johanna Dechau, and Holger Diener (Eds.). Fraunhofer IRB Verlag, Stuttgart, Germany, 139–151.

Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. 2024. From Audio to Photoreal Embodiment: Synthesizing Humans in Conversations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 1001–1010. https://doi.org/10.1109/CVPR52733.2024.00101

Ha Nguyen, Aansh Malik, and Michael Zink. 2022. Exploring realtime conversational virtual characters. *SMPTE Motion Imaging Journal* 131, 3 (2022), 25–34. https://doi.org/10.5594/JMI.2022.3153646

OpenAI. 2023. GPT-4 technical report. arXiv:2303.08774 https://arxiv.org/abs/2303.08774

OpenVINO Toolkit. 2025. OpenVINO: Open-source software toolkit for optimizing and deploying deep learning models. https://github.com/openvinotoolkit/openvino. Accessed: May 2, 2025.

Pat Pataranutaporn, Valdemar Danry, Lancelot Blanchard, Lavanay Thakral, Naoki Ohsugi, Pattie Maes, and Misha Sra. 2023. Living memories: AI-generated characters as digital mementos. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 889–901. https://doi.org/10.1145/3581641.3584065

Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. 2023. EmoTalk: Speech-driven emotional disentanglement for 3D face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Piscataway, NJ, USA, 20687–20697. https://doi.org/10.1109/ICCV51070.2023.01891

Peng Qi, Jing Huang, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. Conversational AI systems for social good: Opportunities and challenges. arXiv:2105.06457 https://arxiv.org/abs/2105.06457

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, Cambridge, MA, USA, 8748–8763.

Kiran Ramesh, Surya Ravishankaran, Abhishek Joshi, and K. Chandrasekaran. 2017. A survey of design techniques for conversational agents. In *Information, Communication and Computing Technology*, Saroj Kaushik, Daya Gupta, Latika Kharb, and Deepak Chahal (Eds.). Springer Singapore, Singapore, 336–350.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 3505–3506. https://doi.org/10.1145/3394486.3406703

Mark Riedl, C. J. Saretto, and R. Michael Young. 2003. Managing interaction between users and agents in a multi-agent storytelling environment. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '03)* (Melbourne, Australia) *(AAMAS '03)*. Association for Computing Machinery, New York, NY, USA, 741–748. https://doi.org/10.1145/860575.860694

Timothy Rupprecht, Sung-En Chang, Yushu Wu, Lei Lu, Enfu Nan, Chih-hsiang Li, Caiyue Lai, Zhimin Li, Zhijun Hu, Yumei He, David Kaeli, and Yanzhi Wang. 2024. Digital avatars: Framework development and their evaluation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (Jeju, Korea) *(IJCAI '24)*. International Joint Conferences on Artificial Intelligence Organization, California, USA, 1–4. https://doi.org/10.24963/ijcai.2024/1031

Carlos Pereira Santos, Phil de Groot, Jens Hagen, Agathe Boudry, and Igor Mayer. 2023. CUBE: Conversational user-interface-based embodiment: Developing a digital humans embodiment for conversational agents: Design, implementation, and integration challenges. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents* (Würzburg, Germany) *(IVA '23)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3570945.3607331

Barry R. Schlenker. 2008. Integrity and character: Implications of principled and expedient ethical ideologies. *Journal of Social and Clinical Psychology* 27, 10 (2008), 1078–1125.

Oliver Schmitt and Daniel Buschek. 2021. CharacterChat: Supporting the creation of fictional characters through conversation and progressive manifestation with a chatbot. In *Proceedings of the 13th Conference on Creativity and Cognition* (Virtual Event, Italy) *(C&C '21)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3450741.3465253

Ulrike Spierling. 2005. Beyond virtual tutors: Semi-autonomous characters as learning companions. In *ACM SIGGRAPH 2005 Educators Program* (Los Angeles, California) *(SIGGRAPH '05)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/1187358.1187365

Ryuichi Sumida, Koji Inoue, and Tatsuya Kawahara. 2024. Should RAG chatbots forget unimportant conversations? Exploring importance and forgetting with psychological insights. arXiv:2409.12524 https://arxiv.org/abs/2409.12524

Ilaria Torre, Emma Carrigan, Rachel McDonnell, Katarina Domijan, Killian McCabe, and Naomi Harte. 2019. The effect of multimodal emotional expression and agent appearance on trust in human-agent interaction. In *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games* (Newcastle upon Tyne, United Kingdom) *(MIG '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3359566.3360065

Olivier Toubia, Jonah Berger, and Jehoshua Eliashberg. 2021. How quantifying the shape of stories predicts their success. *Proceedings of the National Academy of Sciences* 118, 26 (2021), e2011695118.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. arXiv:2302.13971 https://arxiv.org/abs/2302.13971

Philine Witzig, Barbara Solenthaler, Markus Gross, and Rafael Wampfler. 2024. EmoSpaceTime: Decoupling emotion and content through contrastive learning for expressive 3D speech animation. In *Proceedings of the 17th ACM SIGGRAPH Conference on Motion, Interaction, and Games* (Arlington, VA, USA) *(MIG '24)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3677388.3696336

Sichun Wu, Kazi Injamamul Haque, and Zerrin Yumak. 2024. ProbTalk3D: Non-deterministic emotion controllable speech-driven 3D facial animation synthesis using VQ-VAE. In *Proceedings of the 17th ACM SIGGRAPH Conference on Motion, Interaction, and Games* (Arlington, VA, USA) *(MIG '24)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3677388.3696320

Zhenran Xu, Jifang Wang, Longyue Wang, Zhouyi Li, Senbao Shi, Baotian Hu, and Min Zhang. 2024. FilmAgent: Automating virtual film production through a multi-agent collaborative framework. In *SIGGRAPH Asia 2024 Technical Communications (SA '24)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3681758.3698014

XiuYu Zhang and Zening Luo. 2024. Advancing conversational psychotherapy: Integrating privacy, dual-memory, and domain expertise with large language models. arXiv:2412.02987 https://arxiv.org/abs/2412.02987

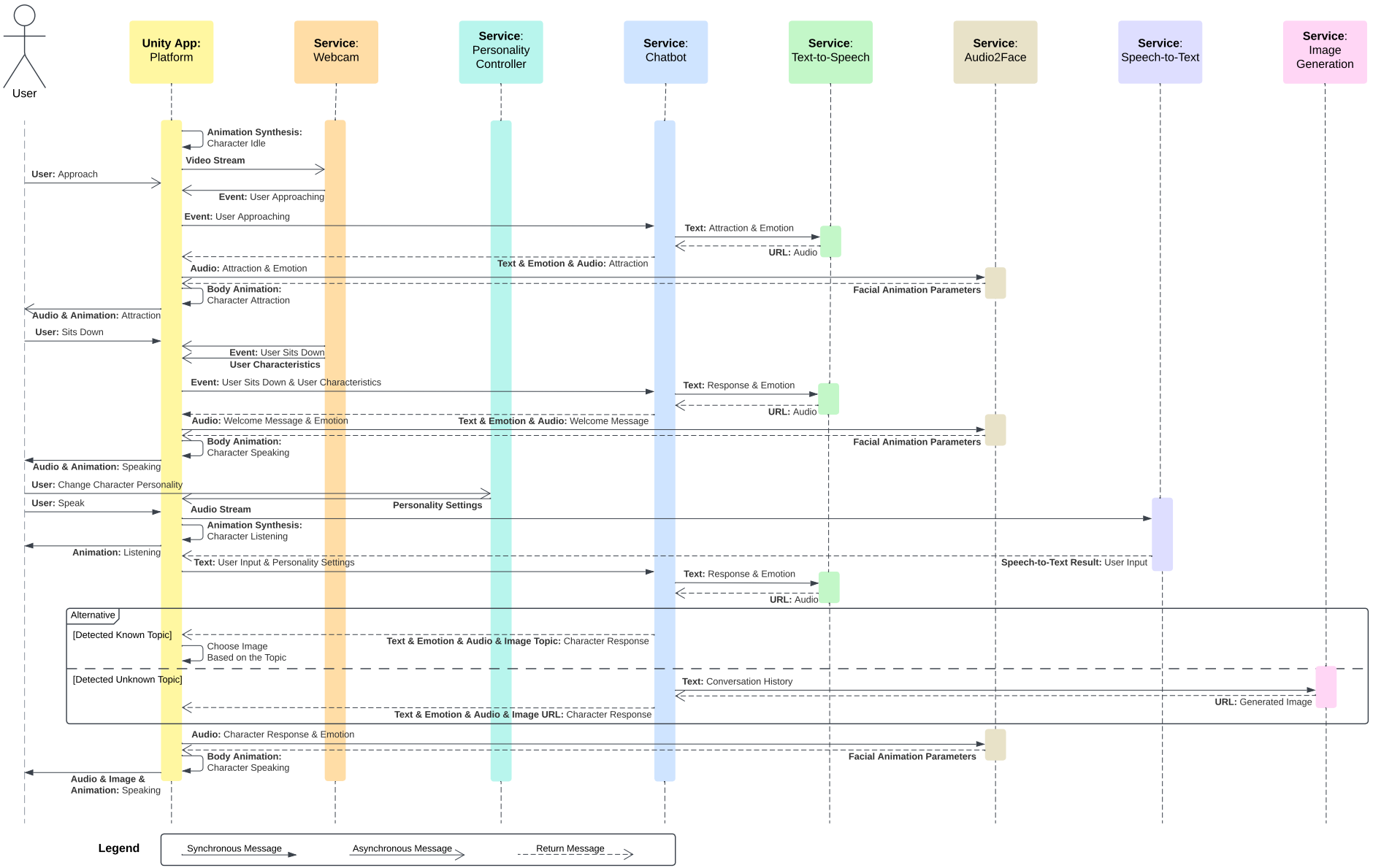

Fig. 8. System Interaction Flow and State Transitions: This sequence diagram illustrates the multi-modal interaction flow of our system, showing how user actions trigger transitions between character states (idle, speaking, listening, thinking) and the corresponding communication flow between the Unity app and AI service modules (webcam, Azure Speech Services, LLM chatbot, Audio2Face, and image generation). Attracting animations are predefined motion-captured clips designed to draw user attention when approaching.

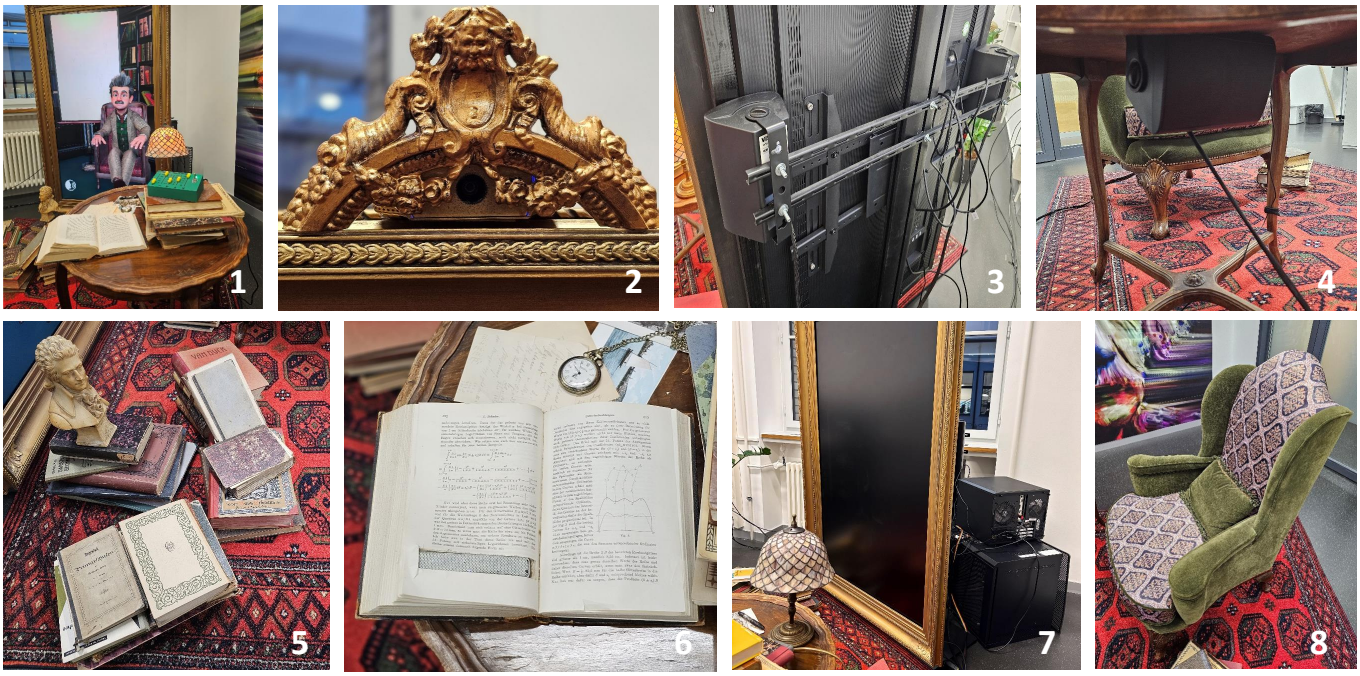

Fig. 9. A detailed view of our physical setup consisting of a screen and a table (1). The setup includes a webcam positioned at the top of the frame (2), speakers hidden behind the screen (3), and an additional speaker mounted below the table (4). The decor enhances the setup with books and a Mozart bust (5), while a microphone is discreetly placed within a book on the table (6). A media box is concealed behind the screen (7), and the arrangement is completed with a chair for seating (8).

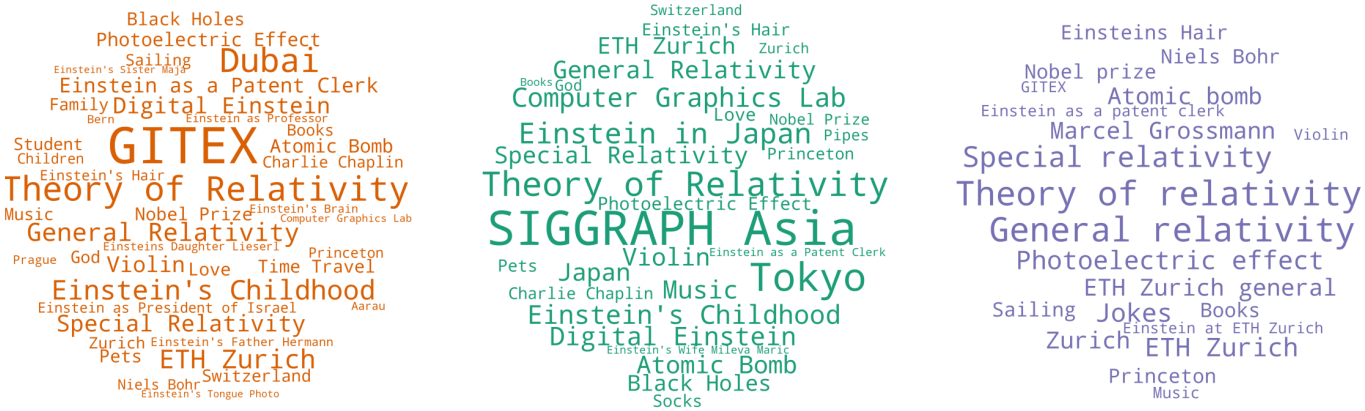

Fig. 10. Word clouds depicting the topics discussed during the tech event GITEX GLOBAL 2024 (left), the scientific event SIGGRAPH Asia Emerging Technologies 2024 (middle), and for Llama 3 (right). Word frequencies were square-root scaled to preserve the relative prominence of frequently mentioned terms while improving the readability of less frequent ones.