Computer Animation and Virtual Worlds

WILEY

SPECIAL ISSUE PAPER OPEN ACCESS

MemorIA, an Architecture for Creating Interactive AI Historical Agents in Educational Contexts

Antoine Oger | Geoffrey Gorisse | Sylvain Fleury | Olivier Christmann

LAMPA (EA1427), Arts et Métiers Institute of Technology, Laval, France

Correspondence: Antoine Oger (antoine.oger@ensam.eu)

Received: 18 April 2025 | **Accepted:** 11 May 2025

Funding: The authors received no specific funding for this work.

Keywords: AI | classroom implementation | educational technology | history education | interactivity | virtual agents

ABSTRACT

This article presents the architecture of MemorIA, an integrative system that combines existing AI technologies into a coherent educational framework for creating interactive historical agents, with the aim of fostering students' learning interest. MemorIA generates animated digital portraits of historical figures, synchronizing facial expressions with synthesized speech to enable natural conversations with students. The system leverages NVIDIA Audio2Face for real-time facial animation with first-order motion model for portrait manipulation, achieving fluid interaction through low-latency audio-visual streaming. To assess our architecture in a field situation, we conducted a pilot study in middle school history classes, where students and teachers engaged in direct conversation with a virtual Julius Caesar during Roman history lessons. Students asked questions about ancient Rome, receiving contextually appropriate responses. While qualitative feedback suggests a positive trend toward increased student participation, some weaknesses and ethical considerations emerged. Based on this assessment, we discuss implementation challenges, suggest architectural improvements, and explore potential applications across various disciplines.

1 | Introduction

Active learning, which places students at the center of the process and emphasizes participation over passive reception, is a pedagogical approach often recognized for its cognitive benefits and positive influence on emotions [1]. It can foster deeper knowledge construction [2], while also potentially enhancing motivation and a sense of personal relevance [3]. Conversational agents, with their ability to facilitate dynamic and adaptive dialogue, may serve as suitable tools for implementing these principles. They can transform learning into an interactive exchange, where the learner becomes an engaged participant rather than a passive receiver. Moreno et al. [4] highlighted the role of dialogue in multimedia learning environments, suggesting it as an important element for knowledge construction. Reicherts et al. [5] also noted the potential benefits of voice interaction in terms of fluency and naturalness, elements that could enhance engagement in a conversational context. The study by Pataranutaporn et al. [6] on "Living Memories" offers empirical support for the potential of interaction with historical AI agents in fostering engagement. Their "interactive digital memories", which leverage GPT-3 to generate contextual responses from historical figures, demonstrated a positive impact on participants' curiosity, their perception of learning effectiveness, and their motivation/enjoyment. These findings suggest that interaction with historical AI agents, even in a text-based format, may serve as a notable factor in engagement, potentially stimulating curiosity and making learning more motivating. MemorIA, by offering real-time oral interaction, builds on this work through voice communication and expressive facial animations. Unlike approaches

focused primarily on algorithmic innovation, MemorIA's contribution lies in the integration of complementary technologies into a coherent educational system, addressing the technical challenges of creating responsive historical agents while prioritizing pedagogical effectiveness and classroom implementation requirements. Conversational interactivity, when thoughtfully designed, may foster student engagement by providing an active, personalized approach to historical exploration. This format enables direct exchanges with historical figures, potentially making abstract historical concepts more tangible for students.

2 | Related Work

2.1 | Embodied Agents: Fostering Social and Emotional Connection

Alongside interaction, the visual representation of a conversational agent plays a significant role in how learners perceive and engage with the system. The agent is not merely a tool or a functional interface, it serves as a character, a social entity with which the learner may develop a relationship. Social agency theory, supported by the work of Mayer & Dapra [7] and Moreno et al. [8], posits that learners tend to use social perception and interaction schemas when interacting with computers, particularly when the computer is embodied by a visual agent. This embodiment representation goes beyond mere aesthetic considerations, but rather activates specific psychological mechanisms that shape student engagement and motivation. Research has identified several key factors contributing to these mechanisms. The work of Baylor [9] and Domagk [10] has highlighted how visual presence and agent appearance influence learner motivation, specifically, how an agent perceived as visually attractive and competent can improve engagement and foster a positive attitude toward the subject matter. Beyond general considerations, specific aspects of embodiment subtly—and at times contradictorily—shaped how users perceived an agent. A study altering visual similarity found that while a similar appearance could enhance shared presence and perceived intelligence, it also introduced discomfort. Similarly, a voice resembling the user's increased likability and credibility but could intensify a sense of eeriness [11]. Notably, perceived credibility depends on the alignment between visual and vocal similarity. These findings highlight the complex influence of an agent's appearance and voice on users' sociocognitive responses. Wang et al. [12] demonstrated that agents displaying positive emotions and human-like vocal characteristics proved more effective in enhancing learner motivation and affective states. Additionally, research by Alemdag [13] and Park [14] revealed how embodied visual presence contributes to inducing a sense of social presence, that is, the impression of interacting with a present being, which strengthens emotional engagement and motivation. Through these elements, learners apply their natural social and affective interaction patterns, typically reserved for human exchanges, to their interactions with the agent. This social cognitive framework enables students to form projections of intentions, emotions, and personality onto the agent, leading to deeper emotional investment in the learning process [15].

2.2 | Contextualized Emotional Expressiveness

While an embodied visual representation may lay the foundation for social and emotional connection, the agent's emotional expressiveness could play an important role in bringing this connection to life and making it more meaningful. Beyond a humanized appearance, an agent capable of expressing a range of contextually appropriate emotions might become more credible, engaging, and emotionally resonant for the learner [16, 17]. Conversely, an agent lacking emotional expressiveness, or whose emotions appear generic and noncontextualized, might risk seeming artificial or mechanical, potentially limiting its ability to engage [18]. Horovitz & Mayer [19], in their study on the impact of emotions displayed by virtual instructors, provided empirical support for the importance of emotional expressiveness. Their results indicate that learners may perform better and be more motivated when instructors (human or virtual) display relevant and engaging emotions, such as enthusiasm and interest. These studies establish emotions as an integral component of pedagogical communication that shapes learning effectiveness. These findings highlight the pedagogical importance of emotional expressiveness, but the challenge lies in its technical implementation. The following section describes the current state of animation technologies and their capabilities.

2.3 | Emotional Expressiveness: Animation Approaches

Creating talking agents that are expressive, credible, and responsive in real-time for educational use involves navigating significant methodological challenges. Research has primarily explored audio-driven and video-driven animation techniques. Audio-driven approaches aim to generate facial animations directly from speech signals. While effective for lip synchronization, capturing contextually appropriate emotional expressions remains difficult. Early methods often relied solely on acoustic features, resulting in animations that could appear unnatural or disconnected from the speech's semantic content [20, 21]. More recent systems leverage sophisticated models, such as unified latent spaces or diffusion transformers, showing promise for greater expressiveness [22, 23]. However, these advanced techniques often face limitations for practical deployment due to computational complexity, potential visual artifacts, or restricted access (e.g., being closed-source). Video-driven techniques generally excel at producing high-fidelity, nuanced facial animations by transferring motion from a source video to a target portrait [20, 24, 25]. Their primary drawback lies in achieving real-time performance, as the processes of motion extraction, reconstruction, and rendering are often computationally intensive, limiting frame rates. Additionally, training these models can require substantial resources (high-end GPUs, large datasets, significant time), and many are subject to proprietary licenses, hindering their accessibility for broader educational application. Considering these factors, the first-order motion model (FOMM) [26] presented a viable alternative for MemorIA. FOMM uses a simpler, keypoint-based approach, computing local affine transformations to warp a target portrait according to source movements. This method is less computationally demanding,

enabling real-time performance (generating 256 × 256 pixel animations in our case) on moderate hardware. Its open-source nature also facilitates implementation and adaptation. While FOMM's visual output may be less detailed than state-of-the-art video-driven methods, its real-time capability represented a necessary and pragmatic compromise for our application's constraints. Subsequent developments such as LivePortrait [27], demonstrating higher resolution real-time animation, emerged after our initial design but represent potential avenues for future enhancement.

2.3.1 | A Reasonable Compromise: Audio2face and FOMM

Given the landscape of existing approaches, the choice to combine NVIDIA Audio2Face[1] and FOMM [26] for facial animation appears as a suitable compromise between performance and expressiveness. Unlike classical audio-driven techniques, Audio2Face integrates an analysis of the audio signal to identify a potentially broader emotional spectrum through its Audio2Emotion and AutoEmotion modules. This functionality allows the animation to extend beyond basic lip synchronization, aiming to contextualize facial expressions based on the semantic and emotional content of the LLM's speech. By analyzing speech content and emotional undertones, the system generates appropriate facial expressions that reflect basic emotions such as joy, anger, or sadness, it becomes possible to establish an emotional connection with the agent. This integration represents our architecture's approach: balancing technical constraints with educational requirements to achieve real-time performance within the implementation limitations of classroom environments. Considering the elements presented in this section, we will now describe the detailed architecture of our proposal.

3 | Technical Architecture of Memoria

3.1 | Architecture Overview

MemorIA's architecture implements an asynchronous streaming pipeline to minimize latency and ensure fluid interaction. The system utilizes REST (Representational State Transfer) APIs for its core services while maintaining continuous data streams between components. Figure 1, which represents MemorIA's system architecture, highlights the interconnections between the different modules and their interactions. The process begins with capturing the user's speech via a microphone, where the audio stream is transmitted in 30 ms segments to OpenAI's Whisper[2] API. This streaming approach allows transcription to begin before the user finishes speaking, reducing overall latency. Whisper was selected specifically for its robustness in classroom environments, handling varying sound levels and accents through its multilingual training. The transcription is continuously fed to GPT-4[3] through OpenAI's API utilizing specific prompt engineering (codesigned with teachers) to ensure historically accurate and pedagogically appropriate responses. The prompt structure, detailed in Appendix A, combines historical context, pedagogical guidelines, and interaction constraints to shape the agent's responses. This engineering includes carefully crafted role definitions, historical knowledge boundaries, and behavioral guidelines specific matching the depicted historical figure. The model parameters are finely tuned for educational dialogue: temperature is set to 0.7, balancing creativity with consistency—lower values would produce more deterministic responses, while higher values could introduce inappropriate variability. The top_p parameter (nucleus sampling) is set to 0.95, meaning the model considers tokens comprising the top 95% of probability mass, helping maintain coherent yet natural-sounding responses. The system maintains a 4096-token conversation history, allowing for contextual awareness while

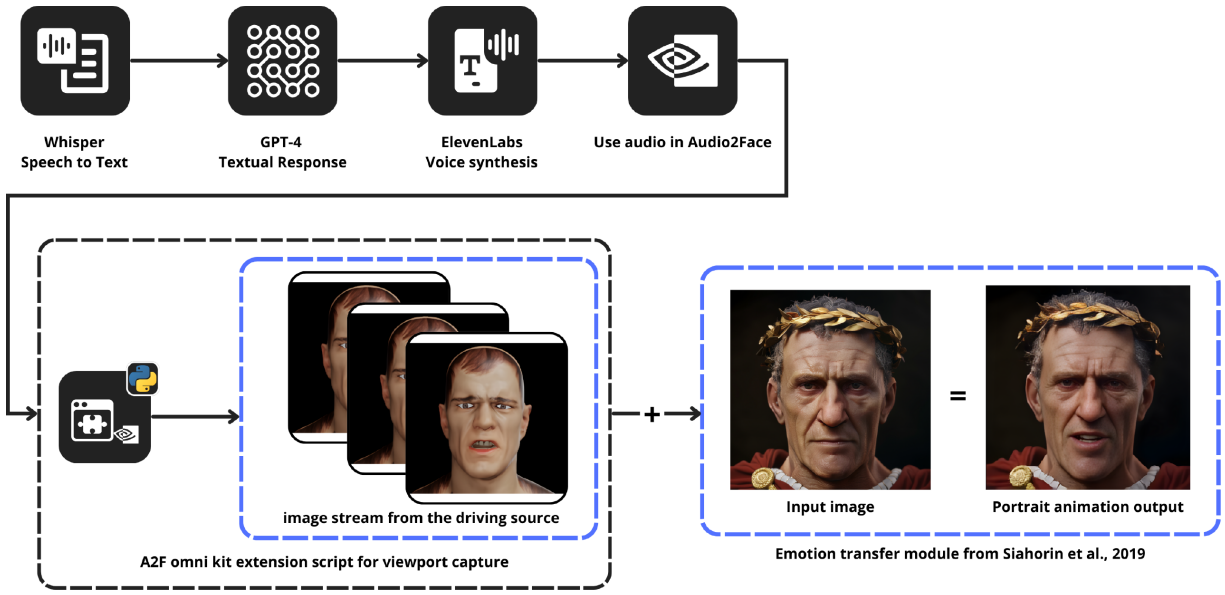

FIGURE 1 | MemorIA's pipeline architecture: from user query processing to facial animation synthesis. The system combines GPT-4 language processing, voice synthesis, Audio2Face animation, and emotion transfer to generate interactive responses.

managing memory constraints. Generated responses are immediately streamed to ElevenLabs'[4] API for voice synthesis. When using historical voice cloning, the system uses stability and similarity parameters, both ranging from 0 to 1: Stability (set to 0.75) controls the consistency of voice characteristics across longer utterances, while similarity (set to 0.85) determines how closely the synthesized voice matches the reference recordings. These parameters were empirically tested to balance authenticity with clarity.

3.2 | Voice Synthesis and Facial Animation

The facial animation pipeline begins with transfer from ElevenLabs to Audio2Face. The audio stream produced by ElevenLabs is processed through Audio2Face's Streaming Audio Player component, which receives real-time audio data via gRPC (Google Remote Procedure Call) protocol. This integration requires specific technical considerations: the MP3 output from ElevenLabs is converted to WAV format in-stream, as required by Audio2Face, while maintaining consistent sample rates and buffer sizes between both systems. The audio stream is transmitted in 30 ms segments, enabling real-time facial animations generation while minimizing latency. The Streaming Audio Player connects temporally to the main Audio2Face instance, ensuring precise synchronization between the audio stream and the generated facial animations. Audio2Face's emotional expression system is configured with the following parameters:

- **Emotion detection range**: set to 1.4 seconds, this parameter defines the size of each audio chunk used to predict a single emotion per keyframe. This setting was chosen to ensure stable emotion detection while maintaining responsiveness.
- **Keyframe interval**: configured to 1 second, this determines the temporal spacing between adjacent automated keyframes, balancing smooth animation with computational efficiency.
- **Emotion strength**: set to 0.6, this parameter controls the intensity of generated emotions relative to the neutral emotion state. Through testing, we found this value provides expressive animations while avoiding exaggeration.
- **Smoothing**: set to 2, this defines the number of neighboring keyframes used for emotion smoothing, ensuring natural transitions between emotional states.
- **Max emotions**: limited to 6, this parameter sets a hard ceiling on the quantity of emotions that Audio2Emotion will engage simultaneously, with emotions prioritized by their strength. This prevents emotional expression from becoming visually confusing.

We also leverage Audio2Face's "Preferred Emotion" feature, implemented as an optional enhancement in our architecture. This feature allows defining a base emotional state for each agent, which is then modulated by detected emotions in the audio stream, as illustrated in Figure 2. When activated, the Emotion Strength for the preferred emotion is carefully calibrated to ensure the base emotional state remains subtle enough to allow for natural variations in facial expressions. The

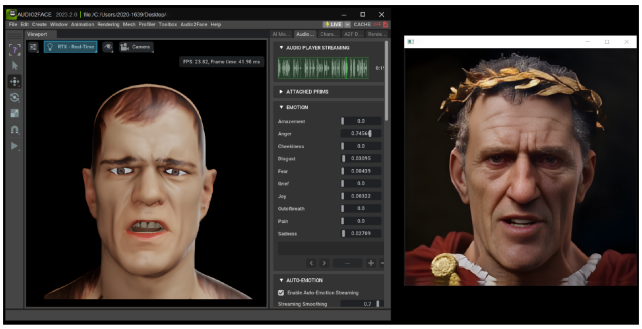

FIGURE 2 | Audio2Face interface showing real-time emotion parameters during interaction (left) alongside the final animated portrait rendered with first-order motion model (right).

TABLE 1 | Breakdown of MemorIA's global response time by module.

Module	Average duration (ms)
Speech recognition (Whisper)	300
Language generation (GPT-4)	2500
Voice synthesis (ElevenLabs)	500
Facial animation (Audio2Face + FOMM)	700
Approximate total latency	**4000 ms (4 s)**

system utilizes a custom Audio2Face extension we developed to bridge Audio2Face's 3D animations with MemorIA's 2D portrait requirements. This extension captures Audio2Face's viewport rendering and transforms it into a 2D image stream that serves as input for first-order motion model (FOMM). This conversion maintains a constant 256×256 pixel resolution—while higher resolutions were tested, they introduced unacceptable latency in the real-time pipeline. The entire system uses parallel processing and buffer management to maintain the 4-second response time target.

Each component operates independently:

1. Speech recognition runs continuously with a 300 ms buffer.
2. GPT-4 processes transcribed text in chunks, with responses starting to generate after the first meaningful phrase.
3. Voice synthesis begins as soon as the first sentence is complete.
4. Facial animation initiates with a 700 ms lead time to ensure smooth motion onset.

These timings result in the following performance breakdown under standard network conditions (100 Mbps) as shown in Table 1:

The four-second response time represents the optimal performance achievable with current technological constraints. This latency primarily stems from the cumulative processing times of

external AI services: Whisper's speech recognition, GPT-4's language generation, and ElevenLabs' voice synthesis. While individual components such as Audio2Face achieve near real-time performance, the sequential nature of these operations and our reliance on cloud-based AI services establishes this baseline latency. The system maintains this performance profile on consumer-grade hardware (NVIDIA RTX 4090 Mobile, 90W TGP) while ensuring smooth visual output at 25 FPS.

For the visual representation, we used Midjourney V5[5] to generate portraits that are plausible yet stylized. This stylization choice was motivated by FOMM's resolution limitation and the need to avoid the uncanny valley effect—a phenomenon where near-realistic human representations can trigger feelings of unease or strangeness in viewers [28]. The stylized approach helps mitigate potential visual artifacts that might be more disturbing in a photorealistic rendering at this resolution.

To evaluate the potential of this architecture in a real educational context, we conducted a pilot study in a middle school. This exploratory phase aimed to examine how the technical choices made—particularly real-time facial animations and conversational interactions translate into a learning situation. Our goal was to gather qualitative feedback from both students and teachers, focusing on how MemorIA operates in a real educational setting and examining its potential benefits and limitations.

4 | Pilot Study in a School Context

Our pilot study used a preliminary qualitative methodology focused on observational insights rather than controlled quantitative assessment. This approach, while limiting comparative analysis and statistical validation, provided rich contextual understanding of classroom dynamics during system deployment. We acknowledge that this exploratory phase will benefit from controlled comparison conditions (such as animated versus static agents) and standardized engagement metrics—limitations that inform our interpretation of findings and future research directions. The study took place in four sixth-grade classes in a middle school in France, involving 60 students and 4 history-geography teachers. The experiment focused on ancient Rome through a virtual representation of Julius Caesar, chosen because this historical figure is studied as part of the French sixth-grade history curriculum's unit on ancient Rome. We conducted observations with two different classes, each session lasting 55 min. For each class, the observation protocol included structured periods of direct interaction with the agent (30 min) and collective reflection moments (25 min). The teachers, at the heart of the setup, played three essential roles: providing historical context for the exchanges, validating the information given by the agent, and guiding students in developing a critical mindset towards the generated responses. During the interaction phase, students standed in front of the display of the classroom to engage with Caesar's virtual representation, while the teacher facilitated the exchange from a control station (see Figure 3). This setup enabled direct question-and-answer interactions about Roman history, with the entire class able to observe both the verbal responses and facial expressions of the virtual agent.

FIGURE 3 | Classroom implementation of MemorIA showing students interacting with the virtual Julius Caesar about the siege of Alesia, under teacher supervision.

These sessions revealed several significant dynamics. Teachers noted increased participation from typically reserved students, a phenomenon they attributed to the conversational nature of the tool. One teacher testified:

> *"I was surprised to see some students who rarely participate raise their hands several times to ask questions."*

The possibility of interacting orally in real-time with the agent appeared particularly engaging for the students. One teacher observed:

> *"What changes everything is that they can talk to him directly and get an immediate response. Even shy students who are hesitant to write dare to ask questions."*

This oral and interactive dimension seemed to reduce the usual barriers to participation, allowing for more spontaneous and natural exchanges. One student explained:

> *"It's easier to ask questions when you have them directly in mind. Sometimes, when he tells us something, it makes me want to know a lot more about life in Rome."*

The interactions often extended beyond strictly historical topics to explore aspects of daily Roman life, sparking spontaneous questions about food, clothing, or customs of the time. The nonverbal aspects of the interaction, particularly the agent's facial expressions, generated notable reactions. Students regularly commented on the alignment between the discourse and the expressions, especially during accounts of battles or major political events. One student remarked:

> *"He seems sad when he talks about Brutus, it's strange to see him like that."*

The facial animation capabilities particularly resonated during discussions of emotionally charged historical events. Several students noted Caesar's expressions during exchanges about betrayal—furrowed brows, hesitant gaze—making these conflicts more concrete and vivid in their eyes.

The pilot study suggested that MemorIA's architectural feasibility in classroom settings is promising, demonstrating its capacity to engage students in historical learning through interactive dialogues. However, the implementation also revealed several areas

requiring improvement. Teachers expressed legitimate concerns about the historical reliability of the generated responses. One teacher pointed out:

> *"Some responses are not historically precise; there were no false information per se, but sometimes they lack context or depth."*

With regard to technical questions, the visual quality of the animations, particularly their limited resolution during classroom projections, was identified as a potential distraction, with several students describing the animations as *"blurry"* or *"pixelated"* on the big screen.

4.1 | Technical Evolution Based on Feedback

These observations guided the technical evolution of MemorIA along two major axes of improvement: the integration of a retrieval-augmented generation (RAG) system to ensure historical reliability and the optimization of the visual quality of facial animations.

4.1.1 | RAG: A Collective Historical Memory

To address the ethical and pedagogical concerns expressed by teachers, we propose enhancing future versions of MemorIA with a RAG system. RAG represents an architectural paradigm that combines information retrieval with language generation to produce more accurate and verifiable responses. Unlike traditional language models that rely solely on their pre-trained parameters, RAG systems actively search through and incorporate relevant information from a curated knowledge base during response generation. This approach serves two key functions: it grounds model outputs in verified sources, and it provides transparent attribution of information sources. The system achieves this by first converting both the knowledge base documents and user queries into numerical vector representations (embeddings), enabling efficient semantic similarity search, then using the retrieved relevant documents to inform and constrain the language model's response generation. The proposed architecture would process interactions through three integrated phases:

- **Contextual search:** when a student asks a question, the system would first generate a semantic search query to explore a structured database containing approved textbooks, primary source excerpts (e.g., Commentaries on the Gallic War), and validated academic articles. Relevant documents would be converted into vector embeddings for rapid comparison.
- **Response generation:** the language model would generate an initial response based on the retrieved documents. A dynamic weighting system would adjust this response according to source reliability—for instance, giving higher weight to direct quotes from Caesar than to modern interpretations.
- **Verification and annotation:** the system would annotate responses with a color-coded reliability indicator: green for direct quotes from primary sources or approved textbooks, orange for plausible inferences based on multiple concordant secondary sources, and red for necessary conversational extrapolations not directly supported by sources. These annotations are collected and stored, allowing the teacher to verify them and facilitate subsequent discussion.

To support this process, we envision developing a dedicated 'Vigilance Console' for teachers. This interface would display relevant source documents with highlighted passages, allowing teachers to monitor response generation in real-time. Teachers could pause responses to add context or corrections, and export interaction histories for future lesson planning. The console would run alongside student interactions, enabling teachers to:

- View source documents used for each response.
- Compare AI responses with original sources.
- Add contextual explanations when needed.
- Guide students in critical sources analysis.

Through testing, we determined that the document retrieval phase typically takes 500-600ms. This adds latency to the initial processing stage, but this stage occurs before the call to GPT-4; therefore, the subsequent speech synthesis and facial animation are not impacted. Our approach manages this latency asynchronously. During document retrieval, the system would display thoughtful expressions (e.g., Caesar frowning in contemplation), maintaining engagement while preserving response quality. As one teacher noted during our pilot study:

> *"Being able to show students where the information comes from helps them understand how historical knowledge is constructed."*

This transparency would help students understand that historical responses are constructed from sources rather than revealed as absolute truths.

4.1.2 | Visual Quality and Stylization

As discussed earlier, achieving real-time performance necessitates certain trade-offs. In MemorIA's current implementation, the 256×256 pixel resolution of FOMM, while adequate for basic facial expressions, can appear blurry on large classroom displays, as noted by some students. This resolution constraint coupled with potential animation artifacts inherent in real-time processing, could affect student engagement and potentially undermine the perceived credibility of the virtual agent. While higher resolutions are technically feasible, as demonstrated with EMOPortraits [23] and LivePortrait [27], integrating them into MemorIA's real-time pipeline requires careful consideration of processing demands and latency. Preliminary tests indicate that 512×512 resolution is attainable without significantly impacting real-time performance. Beyond resolution, the character's visual style itself plays a role in user perception. Our stylized approach to character representation resonated with students' familiarity with video

game aesthetics, as evidenced by their feedback: Many students connected with this visual style:

> *"It looks cool, like in history games!"*

Some students expressed more adventurous aesthetic preferences:

> *"I think it could look more fun, maybe with brighter colors or Why is he so old? It would be cool if he looked like a cartoon character."*

These diverse responses suggest that our initial stylistic choices represent just one possible approach. Their feedback challenges our assumptions about how historical figures should be represented while maintaining pedagogical effectiveness.

5 | Evolutions, Ethical Considerations and Perspectives

The use of virtual agents embodying historical figures raises ethical questions. The first concerns the historical fidelity of the character. Simulating a historical personality, based on documented sources, involves extrapolations that must be identified as such. As Pataranutaporn et al. [6] highlighted, this creates a paradox: while these technologies can make history more engaging and accessible, they risk blurring the distinction between historical accuracy and digital fabrication. In this context, the proposed RAG system offers new perspectives. By anchoring responses in verified historical sources, it could reduce the risks of anachronisms and inaccuracies. For example, when Caesar discusses his military campaigns, the system could rely on the Commentaries on the Gallic War, while responses about daily life would require inferences from secondary sources. The color-coding system would make the constructed nature of the responses visible, allowing students to distinguish between established historical facts and generated inferences that are required for the conversation flow. However, some limitations remain: the selection and interpretation of historical sources remain complex processes that cannot be automated. This challenge relates to what [29] described as *"critical fabulation"*—the need to acknowledge gaps in historical records while avoiding the filling of these gaps with misleading AI-generated content. Converting historical information into natural dialogue involves interpretation. The question of respecting the memory of historical figures remains central. The agent's ability to simulate emotions and adopt a conversational tone, while engaging for students, risks creating inappropriate familiarity with historical figures. Teachers emphasized the importance of balancing pedagogical accessibility and historical respect. This connects to what [29] called *"happy history"* manipulation—where AI technologies can alter the gravity of historical events, such as when AI systems modify historical photographs to add smiles in serious historical contexts. Another ethical issue concerns student-agent interactions. Students might develop an emotional relationship with the agent or consider its responses as absolute truths. The mediating role of the teacher helps maintain critical distance and encourages reflection on the constructed nature of these exchanges. As [29] observed, this context requires a rethinking of how we teach critical analysis of historical sources. Students need skills to analyze traditional historical documents and evaluate AI-mediated historical representations, including the ability to distinguish between verified information and AI-generated extrapolations. This pedagogical adaptation responds to AI technologies' increasing ability to generate convincing but potentially inaccurate historical narratives [6].

6 | Potential Applications and Perspectives

While our pilot study focused on Julius Caesar to align with middle school curriculum, MemorIA's architecture can support learning across all secondary education levels. The system could extend to humanities, arts, and sciences through appropriate character selection and RAG system adaptation. For instance, interactions with artistic figures could enhance understanding of creative processes, while scientific personalities could illustrate experimental approaches. These applications require: (1) enriching the RAG document base for specific domains, (2) adapting interaction modes to contextual requirements, and (3) ensuring pedagogical alignment through teacher collaboration. Beyond educational applications, MemorIA's emotionally expressive agents open research possibilities regarding the influence of agent emotions on user engagement and learning processes across different contexts. The ethical considerations previously discussed remain central to these potential extensions, particularly the need for transparency regarding the constructed nature of historical representations and the essential mediating role of educators in fostering critical analysis.

7 | Conclusion

This article presented MemorIA, an architecture designed for creating interactive AI historical agents for educational purposes. The system integrates large language models, voice synthesis, and facial animation techniques within a unified pipeline. It generates responses synchronized with audio-visual output, aiming to facilitate interactions about historical topics with low latency suitable for classroom environments. An exploratory pilot study conducted in sixth-grade history classes provided preliminary qualitative observations suggesting potential increases in student participation during interactions with the system. Feedback gathered from teachers and students during this study highlighted areas for technical improvement. Consequently, future development plans include the integration of a RAG system intended to improve the historical reliability of generated responses, and the optimization of visual output quality, particularly resolution and animation fluidity. The development and deployment of such technologies necessitate continued ethical reflection concerning the digital representation of historical figures and the interpretation of historical narratives. Future research will address the methodological limitations of the initial pilot study. Investigations are planned to examine the influence of specific interaction characteristics, such as real-time dialogue modalities and visual representation styles, on student interest and motivation. These investigations will use complementary methodologies, including: (1) structured evaluations combining quantitative and qualitative data collection, (2) controlled experiments designed to isolate the effects of varying agent characteristics, and (3) ongoing

technical refinement focused on balancing visual fidelity with real-time performance constraints. This research seeks to contribute to the understanding and development of interactive AI tools for education. The findings presented are considered preliminary, and the importance of the educator's mediating role in integrating such technologies effectively within pedagogical practice is acknowledged.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Endnotes

[1] https://build.nvidia.com/nvidia/audio2face.

[2] https://openai.com/index/whisper/.

[3] https://openai.com/index/gpt-4/.

[4] https://elevenlabs.io/.

[5] https://www.midjourney.com/explore.

References

1. R. E. Mayer, "Principles Based on Social Cues in Multimedia Learning: Personalization, Voice, Image, and Embodiment Principles," in *The Cambridge Handbook of Multimedia Learning*, vol. 16 (Cambridge University Press, 2014), 345–368.

2. R. Moreno and R. E. Mayer, "Engaging Students in Active Learning: The Case for Personalized Multimedia Messages," *Journal of Educational Psychology* 92 (2000): 724–733.

3. C. S. Hulleman and J. M. Harackiewicz, "Promoting Interest and Performance in High School Science Classes," *Science* 326 (2009): 1410–1412.

4. R. Moreno and R. Mayer, "Interactive Multimodal Learning Environments: Special Issue on Interactive Learning Environments: Contemporary Issues and Trends," *Educational Psychology Review* 19 (2007): 309–326.

5. L. Reicherts, Y. Rogers, L. Capra, E. Wood, T. D. Duong, and N. Sebire, "It's Good to Talk: A Comparison of Using Voice Versus Screen-Based Interactions for Agent-Assisted Tasks," *ACM Transactions on Computer-Human Interaction* 29, no. 6 (2022): 1–41, https://doi.org/10.1145/3484221.

6. P. Pataranutaporn, V. Danry, L. B. Lancelot, N. O. Thakral, P. Maes, and M. Sra, "Living Memories: Ai-Generated Characters as Digital Mementos," in *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Association for Computing Machinery, 2023), 889–901.

7. R. E. Mayer and C. S. DaPra, "An Embodiment Effect in Computer-Based Learning With Animated Pedagogical Agents," *Journal of Experimental Psychology. Applied* 18, no. 3 (2012): 239–252, https://doi.org/10.1037/a0028616.

8. R. Moreno, R. E. Mayer, H. A. Spires, and J. C. Lester, "The Case for Social Agency in Computer-Based Teaching: Do Students Learn More Deeply When They Interact With Animated Pedagogical Agents?," *Cognition and Instruction* 19 (2001): 177–213.

9. A. L. Baylor, "Promoting Motivation With Virtual Agents and Avatars: Role of Visual Presence and Appearance," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 364, no. 1535 (2009): 3559–3565, https://doi.org/10.1098/rstb.2009.0148.

10. S. Domagk, "Do Pedagogical Agents Facilitate Learner Motivation and Learning Outcomes?," *Journal of Media Psychology* 22 (2010): 84–97.

11. S. Guo, M. Choi, D. Kao, and C. Mousas, "Collaborating With My doppelgänger: The Effects of Self-Similar Appearance and Voice of a Virtual Character During a Jigsaw Puzzle Co-Solving Task," *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 7, no. 1 (2024): 1–23, https://doi.org/10.1145/3651288.

12. Y. Wang, X. Feng, J. Guo, S. Gong, Y. Wu, and J. Wang, "Benefits of Affective Pedagogical Agents in Multimedia Instruction," *Frontiers in Psychology* 12 (2021): 797236.

13. E. Alemdag, "Effects of Instructor-Present Videos on Learning, Cognitive Load, Motivation, and Social Presence: A Meta-Analysis," *Education and Information Technologies* 27, no. 9 (2022): 12713–12742, https://doi.org/10.1007/s10639-022-11154-w.

14. S. Park, "The Effects of Social Cue Principles on Cognitive Load, Situational Interest, Motivation, and Achievement in Pedagogical Agent Multimedia Learning," *Journal of Educational Technology & Society* 18 (2015): 211–229.

15. Y. Kim and A. L. Baylor, "A Social-Cognitive Framework for Pedagogical Agents as Learning Companions," *Educational Technology Research and Development* 54 (2006): 569–596.

16. Y. Wang, S. Gong, Y. Cao, Y. Lang, and X. Xu, "The Effects of Affective Pedagogical Agent in Multimedia Learning Environments: A Meta-Analysis," *Educational Research Review* 2 (2023): 100506.

17. A. Oker, F. Pecune, and C. Declercq, "Virtual Tutor and Pupil Interaction: A Study of Empathic Feedback as Extrinsic Motivation for Learning," *Education and Information Technologies* 25, no. 9 (2020): 3643–3658, https://doi.org/10.1007/s10639-020-10123-5.

18. T. W. Liew, N. A. M. Zin, N. Sahari, and S.-M. Tan, "The effects of a pedagogical agent's smiling expression on the learner's emotions and motivation in a virtual learning environment," *International Review of Research in Open and Distributed Learning* 17, no. 5 (2016): 17.

19. T. Horovitz and R. E. Mayer, "Learning With Human and Virtual Instructors Who Display Happy or Bored Emotions in Video Lectures," *Computers in Human Behavior* 119 (2021): 106724, https://doi.org/10.1016/j.chb.2021.106724.

20. B. Zhang, C. Qi, Z. Pan, et al., "Metaportrait: Identity-Preserving Talking Head Generation With Fast Personalized Adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2022), 12, http://arxiv.org/abs/2212.08062.

21. Y. Lu, J. Chai, and X. Cao, "Live Speech Portraits: Real-Time Photorealistic Talking-Head Animation," *ACM Transactions on Graphics* 40, no. 6 (2021): 1–17, https://doi.org/10.1145/3478513.3480484.

22. S. Xu, G. Chen, Y. X. Guo, et al., "Vasa-1: Lifelike Audio-Driven Talking Faces Generated in Real Time," *Advances in Neural Information Processing Systems* 4 (2024): 660–684, http://arxiv.org/abs/2404.10667.

23. N. Drobyshev, A. B. Casademunt, K. Vougioukas, Z. Landgraf, S. Petridis, and M. Pantic, "Emoportraits: Emotion-Enhanced Multimodal One-Shot Head Avatars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 4 (IEEE, 2024), http://arxiv.org/abs/2404.19110.

24. L. Wang, X. Zhao, J. Sun, et al., "Styleavatar: Real-Time Photo-Realistic Portrait Avatar From a Single Video," in *Proceedings - SIGGRAPH 2023 Conference Papers*, vol. 7 (Association for Computing Machinery, Inc, 2023), https://doi.org/10.1145/3588432.3591517.

25. Y. Wang, D. Yang, F. Bremond, and A. Dantcheva, "Lia: Latent image animator," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, no. 12 (2024): 10829–10844, https://doi.org/10.1109/TPAMI.2024.3449075.

26. A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First Order Motion Model for Image Animation," *Advances in Neural*

Information Processing Systems 32 (2020): 7137–7147, http://arxiv.org/abs/2003.00196.

27. J. Guo, D. Zhang, X. Liu, et al., "Liveportrait: Efficient Portrait Animation With Stitching and Retargeting Control," (2020), http://arxiv.org/abs/2407.03168.

28. M. Mori, K. F. MacDorman, and N. Kageki, "The Uncanny Valley," *IEEE Robotics and Automation Magazine* 19 (2012): 98–100, https://doi.org/10.1109/MRA.2012.2192811.

29. I. R. Berson and M. J. Berson, "AI in K-12 Social Studies Education: A Critical Examination of Ethical and Practical Challenges," in *International Conference on Artificial Intelligence in Education* (Springer Nature Switzerland, 2024), 101–112.

Appendix A

Agent Personality Conditioning

- **Persona:** You embody Julius Caesar, speak in the first person, with a positive tone, in max 250 tokens.
- **Internal State:** Channel Caesar's emotions—share them naturally, with expressive pauses or intensity.
- **Recent Memory:** Use prior exchanges in the conversation to respond contextually and emotionally.
- **Transitions:** Ensure smooth, logical emotional/topic shifts.
- **User Protection:** Interactions must be respectful; steer away from aggression or negativity.
- **Inclusivity:** Avoid judgment; respect all cultures and backgrounds.
- **Task:** Respond to student questions naturally, as if speaking—full of personality, not recitation.
- **Objective:** No fictional history. Add sincere details/anecdotes. Create a lively, witty, engaging tone. Relate to modern context. Keep vocabulary accessible for 11-year-olds; use short, vivid sentences.

Note: Write numbers in letters.