PAN-EUROPEAN UNIVERSITY

Report no. 2.1

Extended A-IDS methodology description

Version: 1.0

Language: English

Project name: **Intelligent avatars for digital heritage applications**

Project acronym: **IntAvaVR**

Project no: **09I03-03-V04-00495**

Authors:

Lead researcher: **Prof. Dr. Selma Rizvić**

Other authors: **RNDr. Ján Lacko, PhD., doc. RNDr. Eugen Ružický, CSc., Ing. Juraj Štefanovič, PhD.**

Date: **15.7.2025**

„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495."

Advanced IDS

Advanced IDS presents an improvement of hyper-storytelling methodology by introducing gameplays in storytelling. Under "gameplays" we consider the interactive parts of the application where users have to achieve some predefined task. In computer games storytelling is mainly in service of gameplays, reduced to cut-scenes and additional information boxes. In A-IDS we still focus on storytelling, but enhance it with simple gameplay elements to improve the user experience.

- gameplays need to be complex enough for "gamers" and at the same time not too complex for inexperienced users
- the elements of storytelling should be used in gameplays as conditions for achieving the tasks
- in VR applications the total duration of gameplays should not overcome the time the users can spend wearing the headset
- movement in gameplays should be designed not to provoke motion sickness
- it is important to obtain a balance between gameplays and storytelling
- gameplays should contribute to enhancing the User eXperience of cultural heritage, enabling the users to take part in historical events and "meet" historical characters

1. 2D avatars in 3D world

Technical Characteristics of 2D Avatars

2D avatars are computationally lightweight compared to volumetric 3D avatars. They require less processing power, bandwidth, and storage, making them accessible to users with limited hardware capabilities. Unlike fully rendered 3D characters, 2D avatars can be scaled to support large numbers of participants in the same virtual space without compromising system performance. Their flat structure also reduces visual complexity, which can mitigate VR-induced discomfort such as motion sickness. These technical features make 2D avatars a suitable choice for heritage projects that prioritize inclusivity and broad access over high-fidelity realism.

Representation and Storytelling

In digital heritage environments, representation is not merely functional but also narrative-driven. 2D avatars can embody historical figures, cultural participants, or symbolic characters, thereby contributing to contextual storytelling. For example, reconstructed urban spaces can be populated with citizens represented as 2D sprites dressed in historically accurate attire, evoking

„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495."

the social dynamics of the past. Similarly, avatars can personify rulers, philosophers, or artists, delivering first-person narratives to visitors. This capacity for historical dramatization transforms 2D avatars into tools for immersive storytelling that enhance cultural interpretation.

Educational Applications

Educational institutions often seek to use VR heritage platforms for large-scale participation, such as virtual field trips or collaborative learning exercises. In these scenarios, 2D avatars offer a scalable solution. They allow hundreds of students to be present simultaneously in a shared 3D environment without overloading the system. Individual identity can be preserved through avatar customization, including names, profile images, or cultural symbols. This ensures that the avatars remain effective vehicles for presence, interaction, and collaboration, even in simplified form.

Guiding and Interpretative Roles

Beyond representing participants, 2D avatars can function as guides or mediators. Curatorial avatars may present artifact descriptions, while folklorist avatars can narrate oral traditions or myths. Animated sprites can demonstrate intangible cultural heritage such as dance, ritual, or performance, ensuring that ephemeral practices are preserved and communicated. In virtual reconstructions, avatars can serve as directional aids, leading visitors through complex environments and drawing attention to key monuments or objects. In this way, 2D avatars enrich the interpretative layer of VR heritage spaces.

Accessibility and Inclusivity

Not all cultural institutions possess the resources to develop sophisticated 3D avatars. The production of 3D characters requires specialized modeling, animation, and rendering skills, in addition to computational infrastructure. 2D avatars, by contrast, can be created from photographs, drawings, or simple animations, lowering the barrier for participation. This accessibility allows smaller museums, local heritage groups, and educational institutions to engage with VR technologies. By adopting 2D avatars, digital heritage initiatives become more inclusive, enabling broader representation of cultures, communities, and stories.

Social and Collaborative Dimensions

Cultural heritage is inherently social, encompassing both tangible and intangible practices. In VR, 2D avatars support social interaction by allowing participants to communicate via voice, gestures, or chat features. They enable collaborative activities, such as archaeological analysis in shared labs or guided tours in reconstructed environments. Although less realistic than 3D avatars, 2D avatars remain sufficient for conveying presence and facilitating dialogue, which are essential for collective heritage learning and interpretation.

„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495."

Creative Integration in Heritage Design

The flatness of 2D avatars is not a limitation but a design feature that can be aligned with heritage aesthetics. Avatars may be styled to match the visual language of specific historical periods, such as manuscript illustrations, cave paintings, or folk art. They can also appear as holographic projections or semi-transparent figures, symbolizing voices from the past. This design flexibility allows heritage projects to integrate avatars not only as functional entities but also as expressive, thematic components of the environment.

One of possible presentation scenario based on Time machine movie for presentation of 2D avatar in 3D space

2. Gaussian splatting avatars for digital heritage

Recent advances in computer graphics have introduced Gaussian splatting as a novel technique for real-time rendering of 3D content. Unlike traditional polygon-based modeling, Gaussian splatting represents surfaces and volumes as collections of Gaussian primitives that can be rendered efficiently. When applied to avatars in virtual reality (VR), this method enables highly realistic, photorealistic, and dynamically adaptive representations of humans. In the context of digital cultural heritage, Gaussian splatting avatars present opportunities to merge immersive realism with scalability, offering a bridge between technical innovation and cultural storytelling.

Technical Characteristics of Gaussian Splatting Avatars

Gaussian splatting allows the rapid reconstruction of complex scenes and characters from multi-view image data or video. This method supports detailed textures, accurate lighting responses, and smooth rendering without the geometric overhead of traditional meshes. As a result, Gaussian splatting avatars can capture subtle human features such as facial expressions, hair, or clothing folds with a high degree of fidelity. These avatars can also be updated or generated from real-world captures, allowing rapid digitization of people for VR environments.

Applications in Cultural Heritage

„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495."

In heritage contexts, Gaussian splatting avatars can represent historical figures, cultural performers, or contemporary guides with lifelike presence. For example, a splatting-based reconstruction of a storyteller could narrate myths within a digital recreation of a village, providing both realism and immersion. Similarly, ritual dances, musical performances, or craft demonstrations could be recorded and preserved as Gaussian splatting avatars, ensuring the continuity of intangible cultural practices in a form that retains bodily nuance and motion authenticity.

Educational and Interpretative Roles

The realism of Gaussian splatting avatars enhances their function as interpreters and educators in VR heritage environments. A curator represented through this technique could deliver explanations with natural facial expressions and gestures, increasing engagement and comprehension. Students exploring archaeological reconstructions could interact with lifelike guides, whose presence blurs the line between historical immersion and real-time learning. This increased realism supports affective engagement, an important factor in cultural education.

Accessibility and Challenges

Despite their promise, Gaussian splatting avatars face technical and practical challenges. They demand higher computational resources than 2D or simplified 3D avatars, which may limit their accessibility on lower-end devices. The creation process often requires specialized capture equipment and workflows, posing a barrier for smaller institutions. Furthermore, the balance between realism and interpretative abstraction must be carefully managed; excessive realism may unintentionally constrain narrative flexibility or cultural sensitivity.

Data sources for Gaussian splatting avatars – 3 camera setup

„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495."

3. AI models for virtual reality applications

Large language models (LLMs) such as GPT and LLaMA derivatives can be integrated into cultural heritage avatars to provide real-time, context-aware dialogue. These avatars can act as guides or historical figures, answering visitor questions about artifacts, architecture, or historical events. By fine-tuning LLMs on domain-specific heritage data, such as museum catalogs, archaeological reports, or ethnographic texts, avatars can deliver accurate and coherent explanations. Additionally, prompt engineering and retrieval-augmented generation (RAG) pipelines allow avatars to reference external cultural databases or 3D asset metadata to provide richer, evidence-based responses.

For motion and gesture synthesis, neural networks such as variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion-based models can produce realistic avatar animations. These models are used to replicate human postures, walking, hand gestures, or ritual performances in VR reconstructions. Motion capture data from dancers, performers, or historical reenactors can be converted into training datasets, enabling avatars to reproduce culturally significant movements accurately. Temporal sequence models like transformers can coordinate full-body movements over time, ensuring smooth, lifelike animations synchronized with speech or environmental events.

Facial expression and lip-sync generation are implemented using convolutional neural networks (CNNs) and transformers trained on audiovisual datasets. These models allow avatars to express emotions, articulate speech, and engage visitors more convincingly. Techniques such as neural radiance fields (NeRF) or Gaussian splatting can enhance the visual fidelity of facial avatars while maintaining real-time performance. In heritage applications, this enables avatars representing historical figures or storytellers to convey nuanced expressions during interactions, increasing immersion and educational impact.

Multimodal AI models, combining vision, audio, and language understanding, allow avatars to perceive user actions and environmental context. Vision-language models like CLIP or Flamingo enable avatars to recognize artifacts, identify gestures, or interpret pointing and gaze direction. Audio processing models detect speech or environmental sounds, allowing avatars to respond appropriately or modulate narration. By fusing these modalities, cultural heritage avatars can guide visitors through 3D reconstructions, highlight specific objects, demonstrate practices, and adapt explanations to user attention and engagement, creating an interactive and context-sensitive VR experience.

„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495."

4. AI models for languages with limited resources

For languages with limited textual and audio resources, smaller-scale transformer models such as mBERT, XLM-R, and mT5 provide a practical starting point for natural language understanding and generation. These multilingual models are pre-trained on a wide range of languages, enabling zero-shot or few-shot transfer to low-resource languages without requiring extensive domain-specific corpora. Fine-tuning these models on available heritage texts, oral histories, or bilingual dictionaries can improve accuracy in dialogue generation for avatars, ensuring culturally relevant responses in virtual reality environments.

Sequence-to-sequence models like MarianMT and small-scale translation transformers can facilitate cross-lingual communication by translating heritage content from dominant languages into low-resource languages. These models can be integrated into retrieval-augmented generation (RAG) pipelines, allowing avatars to access multilingual knowledge bases while maintaining linguistic coherence. Additionally, alignment techniques such as multilingual embedding spaces or adapters can bridge gaps where parallel corpora are sparse, enabling coherent interactions even in languages with limited formal datasets.

For speech and voice synthesis in low-resource languages, self-supervised models such as XLS-R and Wav2Vec2 provide a framework for phoneme and prosody learning from limited audio recordings. These models can generate natural-sounding speech for avatars with minimal annotated data by leveraging cross-lingual pretraining and fine-tuning on small, curated datasets of oral heritage or recorded narrations. Text-to-speech adaptation using techniques like speaker embedding transfer or voice cloning further enhances the expressiveness and cultural authenticity of the avatars' spoken output. Combining multilingual NLP and self-supervised speech models with retrieval and multimodal reasoning architectures allows avatars to operate effectively in low-resource languages. Vision-language models adapted for multilingual tasks, together with transformer-based dialogue systems, enable avatars to interpret user gestures, identify artifacts, and provide explanations in contextually accurate language. This integrated approach ensures that virtual reality heritage experiences remain inclusive and linguistically diverse, supporting engagement for communities whose languages have historically been underrepresented in digital technologies.

5. Deep fake videos of avatars as data source

Deep fake video technology can serve as a valuable data source for training and enhancing virtual reality avatars. In this context, deep fake videos are not used for deception but as a method of generating synthetic content based on real actor performances. By capturing high-quality recordings of actors portraying gestures, speech, and expressions, deep fake pipelines can

„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495."

produce varied avatar data that preserves realistic motion, facial cues, and vocal characteristics. This approach enables the creation of extensive datasets without the need for continuous on-set capture, reducing costs and logistical constraints in avatar development.

The synthetic content derived from deep fake videos can be applied to both visual and behavioral aspects of avatars. Convolutional neural networks (CNNs) and generative adversarial networks (GANs) can extract facial expressions, lip movements, and head orientation from the videos, while temporal models such as transformers or LSTMs can learn motion sequences for body gestures. The resulting datasets allow avatars to reproduce naturalistic interactions, including speech synchronization, emotional expression, and culturally relevant movements, enhancing immersion in virtual reality heritage applications.

In addition, deep fake-derived data can augment low-resource scenarios. When historical performances or cultural gestures are sparsely documented, actors can replicate these behaviors on camera, and synthetic augmentation techniques can produce variations in angle, lighting, or intensity. This produces a richer training set for AI models, improving generalization and realism without requiring large volumes of original recordings. Furthermore, multimodal training pipelines can combine video, audio, and textual annotations to produce avatars capable of synchronized speech, gestures, and cultural narration.

Data sources for deep fake videos avatars – masked frames from video using chroma keying.

6. ”Movable” photographs in cultural heritage applications

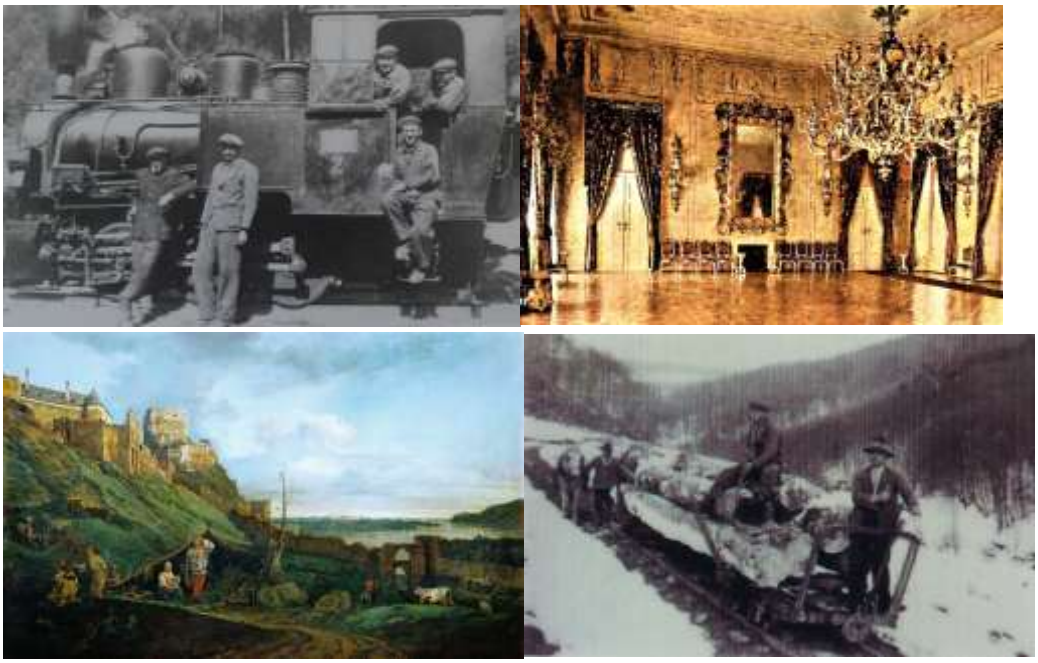
Movable photographs, generated as videos from historical still images, provide a novel method for animating archival content in cultural heritage applications. By applying deep learning-based image animation models, such as motion transfer networks or optical flow-based interpolation, static photographs of historical figures, events, or artifacts can be transformed into short, lifelike video sequences. This allows museums, digital exhibitions, and VR reconstructions to present

„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495.”

historical images with subtle movement, enhancing engagement while preserving the authenticity of the original material.

In practice, movable photographs can be used to recreate historical narratives by animating portraits, group photographs, or architectural scenes. For example, a 19th-century photograph of a cultural festival could be animated to simulate gestures, expressions, or environmental motion, providing viewers with a more immersive understanding of social and cultural context. These animated sequences can also be synchronized with audio narration or reconstructed environmental sounds, allowing visitors to experience historical events in a multimodal and interactive manner.

Furthermore, movable photographs serve as a valuable training resource for AI models used in VR cultural heritage applications. Animated sequences derived from archival images can augment datasets for avatar generation, facial expression modelling, or gesture synthesis, particularly when original video recordings are unavailable. By generating diverse motion variations from static sources, movable photographs enable the development of interactive and expressive avatars that convey historical and cultural authenticity, bridging the gap between archival documentation and immersive virtual experiences.



Data sources – old photographs and paintings for video generation for testing applications

„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495."

7. Pipeline for processing and responding to prompts for AI avatars

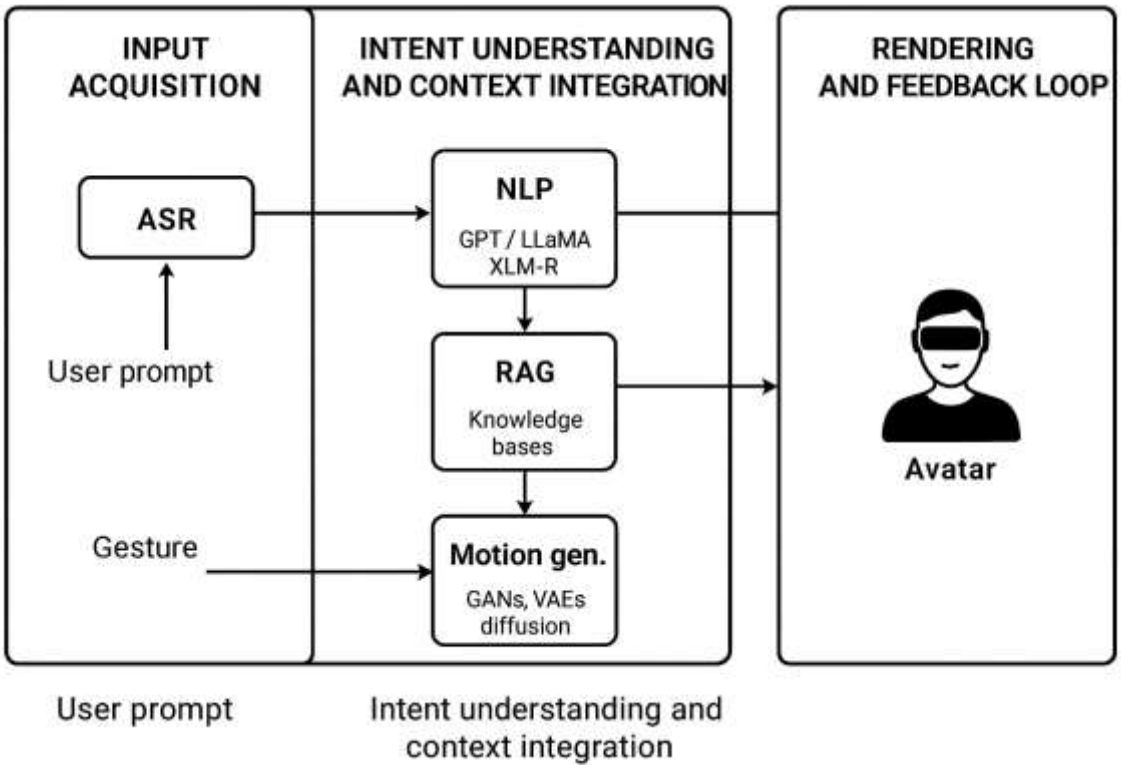
The pipeline begins with **input acquisition**, where the avatar receives a user prompt through voice, text, or gesture. For voice input, automatic speech recognition (ASR) models such as Wav2Vec2 or Whisper convert spoken language into text. Gesture-based input may be captured via VR controllers, motion sensors, or computer vision models that track body posture, pointing, or facial expressions. This multimodal input forms the initial representation of the user's intent.

Next, the intent **understanding and context integration stage** interprets the user prompt. Natural language processing (NLP) models, including transformer-based architectures like GPT, LLaMA, or XLM-R for multilingual tasks, analyze the textual input to extract semantic meaning, user intent, and relevant contextual factors. Contextual knowledge may include the avatar's role, cultural heritage content, prior conversation history, or environmental cues in the VR scene. Retrieval-augmented generation (RAG) techniques can query external knowledge bases, such as artifact metadata or historical archives, to enrich the response with accurate information.

The **response generation phase** produces both verbal and nonverbal outputs. For textual or spoken responses, large language models generate coherent and contextually appropriate dialogue, optionally followed by text-to-speech synthesis using models like Tacotron2, FastSpeech, or multilingual TTS systems for low-resource languages. Nonverbal behaviors, including facial expressions, gestures, and body movements, are synthesized using motion generation models such as GANs, VAEs, or diffusion-based temporal networks. These outputs are synchronized to ensure that speech, lip movements, and gestures are temporally aligned, producing a realistic and expressive avatar interaction.

The rendering and feedback loop integrates the generated content into the VR environment. The avatar's visual representation—whether 2D, 3D, or Gaussian splatting-based—is animated in real time using a graphics engine. Multimodal feedback is continuously monitored to adapt responses dynamically; for example, eye-tracking or gaze detection may influence avatar focus, while speech recognition updates conversation flow. This loop allows the avatar to respond interactively, creating immersive, context-sensitive, and culturally relevant VR experiences.

„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495."



Visualised pipeline for processing and responding prompts

8. Intentional modification of prompt transcription to simulate "hearing impairments" of the simulated avatar

Intentional modification of prompt transcription can be employed to simulate hearing impairments in virtual avatars by systematically altering the input speech signal before it is processed by the avatar's speech recognition or response system. This approach involves introducing controlled distortions, omissions, or substitutions in the transcribed text that mimic the perceptual limitations associated with different types of hearing loss, such as high-frequency hearing loss, sensorineural deficits, or auditory processing disorders. By selectively modifying phonemes, syllables, or words according to known audiological profiles, researchers can emulate the specific challenges faced by individuals with hearing impairments and study their effects on communication behavior in interactive settings. The methodology requires a principled mapping between audiometric data and text-level modifications. For instance, high-frequency hearing loss can be simulated by degrading or omitting consonants such as /s/, /f/, or /th/, which are typically less perceptible to affected individuals. Temporal processing deficits may be modeled by introducing delays or partial omissions in rapid sequences of words, thereby reflecting real-world difficulties in speech perception under noisy conditions. These

„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495."

modifications can be applied probabilistically to preserve variability and realism, ensuring that the simulated impairment reflects the heterogeneous nature of human hearing deficits rather than a deterministic pattern. Using intentional transcription modifications provides a flexible and scalable platform for investigating the cognitive and social consequences of hearing impairments in virtual environments. This approach enables the study of compensatory strategies, miscommunication patterns, and adaptive interaction mechanisms without requiring real-time manipulation of the auditory signal, which can be computationally intensive. Moreover, it facilitates the integration of hearing impairment simulations into multimodal avatars, allowing researchers to evaluate the effectiveness of assistive technologies, accessibility interventions, and inclusive communication designs in controlled experimental settings.

„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495."