PRESENCE

Funded by the European Union

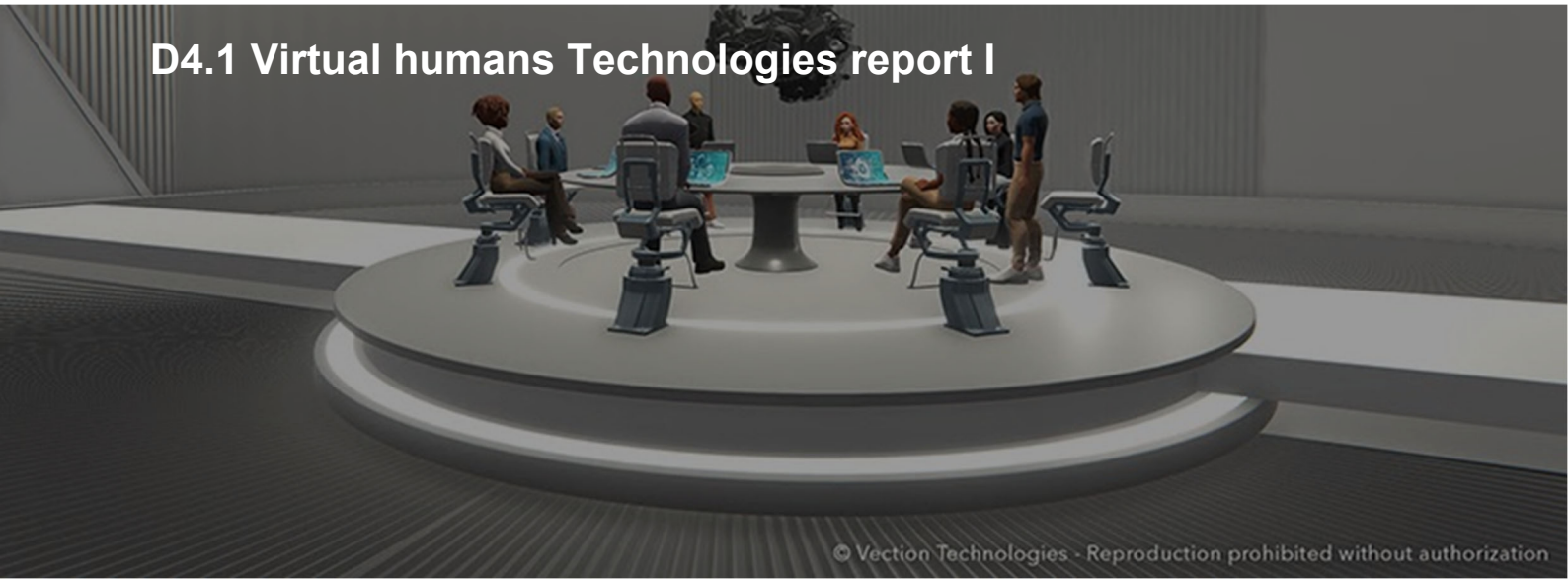
A toolset for hyper-realistic and XR-based human-human and human-machine interactions, PRESENCE

Grant Agreement nº 101135025

HE Call identifier: HORIZON-CL4-2023-HUMAN-01-CNECT
Topic: HORIZON-CL4-2023-HUMAN-01-21

Type of action: HORIZON Research and Innovation Actions

D4.1 Virtual humans Technologies report I



DISSEMINATION LEVEL

☒	**PU**	**Public**
	SEN	*Confidential, only for members of the consortium (including the Commission Services)*

Grant Agreement nº:	Project Acronym:	Project title:
101135025	PRESENCE	A toolset for hyper-realistic and XR-based human-human and human-machine interactions

Lead Beneficiary:	Document version:
UHAM	v0.1

Work package:
WP4 – Intelligent Virtual Humans

Deliverable title:
D4.1. Virtual humans technologies report I

Start date of the project:	Contractual delivery date:	Actual delivery date:
1st of January 2024	28th of February 2025	27st of February 2025

Editor(s):
Fariba Mostajeran, Ke Li, Frank Steinicke (UHAM)

LIST OF CONTRIBUTORS

PARTNER	CONTRIBUTOR
UHAM	Fariba Mostajeran, Ke Li, Frank Steinicke
DIDIMO	Silvia Bettencourt Ribeiro, Xenxo Alvarez, João Orvalho
JRS	Hannes Fassold
ARTANIM	Joan Llobera, Pierre Nagorny

LIST OF REVIEWERS

PARTNER	REVIEWER/S
CAPGEMINI	Denis Denisov, Jerome Pönisch

CHANGE HISTORY

VERSION	DATE	PARTNERS	DESCRIPTION/COMMENTS
V0.1	06 –01 – 2025	UHAM	First draft
V0.2	14 –01 – 2025	UHAM	First draft of T4.3 was added
V0.3	29 –01 – 2025	ARTANIM, UHAM	Description of T4.4 was added. Section 1-3, Introduction of WP4, T4.3 and T4.5 and Sections 5-7 were finalized.
V0.4	30– 01 – 2025	DIDIMO	Description of T4.1 was added.
V0.5	07– 02 – 2025	UHAM, ARTANIM, JRS	Ethical Considerations were added / version submitted to the EEA for review (12th of February).
V0.6	13– 02 – 2025	DIDIMO	Ethical Considerations about T4.1 were added.
V0.7	24– 02 – 2025	UHAM	Updated final image of the architecture diagrams
V1.0	27- 02 - 2025	i2CAT	Formatting, indexes and cross-references fixing, final version submitted to the EC

Executive summary

Virtual Humans Technologies Report I serves as a comprehensive overview of Work Package 4 (WP4) within the project framework. The report delineates the objectives, key performance indicators (KPIs), and essential tasks outlined in the Description of Action (DoA), serving to meet Milestone 3 concerning the technological pillars of smart avatars (SAs) and intelligent virtual agents (IVAs).

Key sections of the report include a detailed system architecture that illustrates the interconnections and requirements of WP4 tasks. A literature review highlights advancements in virtual human technologies to inform WP4's research and development strategies. The report also includes an evaluation of the initial developments and achievements within WP4, alongside an assessment of KPI status and any necessary mitigation plans for identified deviations. Additionally, it presents methods and findings from a preliminary assessment of the initial version of the Intelligent Virtual Human (IVH) SDK. Finally, the document outlines future experimental plans and anticipated publications within WP4, ensuring a roadmap for continued innovation in virtual human technologies

Table of contents

List of Tables

List of Figures

1. Introduction

1.1. Purpose, scope, and structure of the Document

The purpose of Deliverable 4.1 (D4.1): Virtual Humans Technologies report I, is to provide an overview of the goals and tasks of Work Package 4 (WP4). It will outline the required tasks and the roadmap to fulfill them as described in the DoA. This document aims at fulfilling Milestone 3: Technological pillars 1st delivery: WP4 middle delivery. It provides details about the first iteration of smart avatars (SAs) and intelligent virtual agents (IVAs) developed in WP4. The document is structured as follows:

- Section 2 provides WP4's objectives, key performance indicators (KPIs), and tasks according to the DoA. In addition, a detailed system architecture will illustrate the components of each task at a higher level, the hardware required, and the connection between the tasks.
- Section 3 reviews the current literature on virtual human technologies and advancements. This will provide a scientific basis for the research and development efforts in WP4 tasks.
- Section 4 elaborates on the i) developments and key achievements, ii) KPIs status, and if applicable, iii) deviations and mitigation plan of each task within WP4.
- Ethical considerations are addressed in Section 5 (a report of the project's Ethics External Advisor is annexed)
- Section 6 describes the methods and the results of a preliminary evaluation of the initial version of the IVH SDK.
- Section 7 provides an outlook and planned experiments and publications within WP4.

2. WP4 – Intelligent Virtual Humans

2.1. Objectives and KPIs

The following are the main objectives of WP4:

- Provide real-time photorealistic humanoid 3D models based on cost-efficient technology, which can represent users (avatars) or agents (IVA)
- Supporting natural and multimodal communication and interaction via speech, facial expressions, gaze, and full-body animations
- Moving from simple human-human communication to hybrid forms of interaction including multiple real and artificial users
- Create a set of technologies and libraries feeding WP1 and WP5 for the user-oriented tests and implementation of demonstrators

Based on these objectives, we aim to reach these Key Performance Indicators (KPIs) by the end of the project:

- KPI 4.1: Delivery of efficient 3D humanoid model generation pipeline

- KPI 4.2: Delivery of natural multi-user communication and collaboration capabilities for virtual humans with physics-based full-body animation
- KPI 4.3: Integration of multimodal interaction capabilities
- KPI 4.4: Deliver a set of APIs to integrate virtual humans with the two other pillars in the two demonstrators

2.2. Tasks

WP4 has five tasks that will be fulfilled by the end of the project. Table 1 provides a brief overview of these tasks, and Section 4 will elaborate on the status of each task in detail.

Table 1: WP4 Tasks

Task	Description

T4.1 Virtual Humanoid 3D Models (M04-M32), Lead: DIDIMO	• Development of
	  ○ a user-generated pipeline for the reconstruction of realistic humanoid 3D full-body models based on simple camera frames (as mobile cameras)
	  ○ AI-based algorithms to extract the most suitable poses for the creation of facial blendshapes from the RGB and depth video frames of the user
	  ○ neural rendering techniques to generate rigged, skinned, and fully animatable characters with high visual fidelity to generate hyper-realistic 3D humanoid models
	• Delivery of the avatars libraries and APIs that will be integrated into XR applications for user tests (WP1) and demonstrators (WP5).

T4.2 Motion Tracking and Action Classification (M04-M32), Lead: JRS	• Development of novel AI-based methods to predict and classify actions and intentions of users in the scope of the considered use cases. ○ Classify the actions of users and the smart avatars and predict likely subsequent actions, for example, to communicate via speech or gestures (like raising the hand), to walk towards a position or to interact with other users or IVAs. • In addition to employing deep learning based action recognition algorithms which have been pretrained on a fixed set of actions (e.g. from NTU-60 dataset), methods for few-shot action classification (i.e., training classifiers from few samples) or zero-shot action classification (without any training samples) will be developed.
T4.3 Speech and Facial Interaction (M04-M32), Lead: UHAM	• Development of methods for natural communication between humans and smart avatars as well as IVAs based on processing spoken language and facial expressions. • Analyse, process, and synthesize speech and facial expressions, adding AI-powered features to the humanoid 3D models (T4.1) • Use speech-to-text and text-to-speech synthesis to allow users to naturally communicate with humans and agents. • Train the models for the specific speakers based on the audio recordings

Task	Description
T4.4 Full Body Animation and Interaction (M04-M32), Lead: ARTANIM	• Development of a method for physics-based character animation in Unity3D • Development and test of method to scale ragdolls to learn from animated characters with different proportions • Development of methods to scale machine learning for physics-based character animation, testing, and evaluation of the methods • Development of physics-based animation methods for characters capable of dealing with proximal space, particularly the peri-personal space of VR users
T4.5 Multimodal Interaction (M04-M32), Lead: UHAM	• Integrate the results from T4.1-4 to provide intelligent virtual humans, which can represent real users and/or computer-generated agents for realistic multimodal communication and interaction between humans and agents. • Combine the individual contributions presented in T4.1-4 into a coherent representation, keeping in mind the requirements presented by the use-case scenarios.

2.3. Architecture

Figure 1 shows a schematic diagram of WP4 tasks and internal components. Each task and component will be described in detail in their respective part in Section 4.

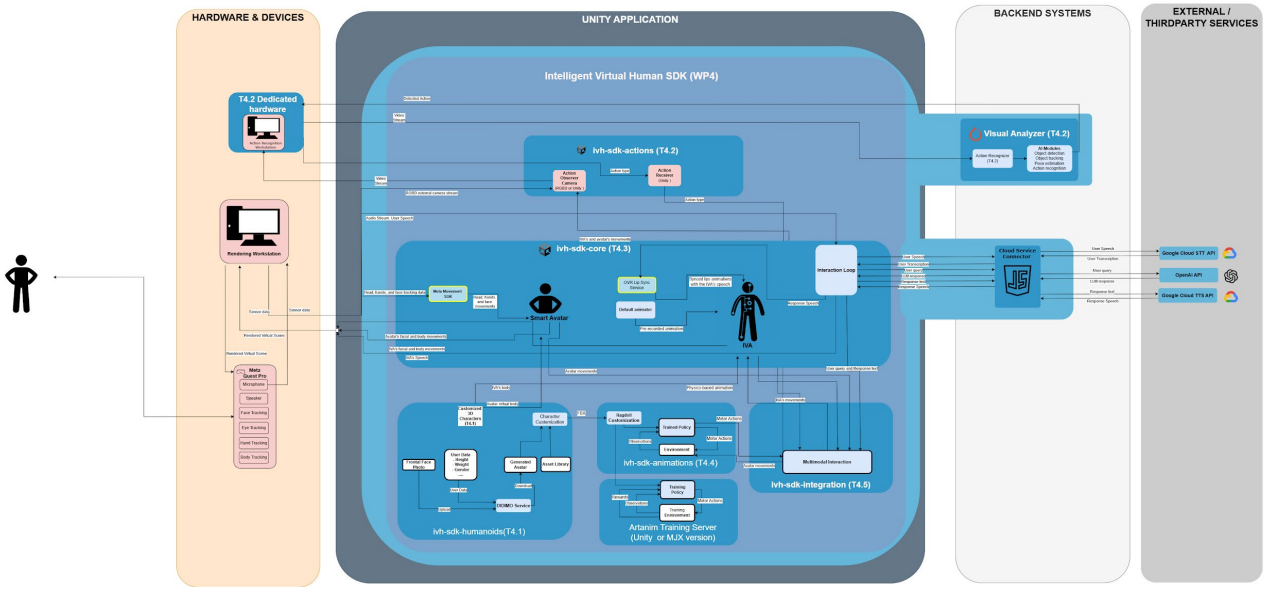


Figure 1: WP4 system architecture

3. Related Work

Intelligent Virtual Agents (IVAs) are autonomous characters designed to interact naturally with humans in virtual environments. Recent advances in artificial intelligence have significantly enhanced their capabilities across several dimensions. IVAs typically integrate natural language processing for conversation, computer vision for environmental awareness, and generative models for behavior generation. Notable developments include improved emotional intelligence through facial expression recognition and generation (Kuo et al., 2018), context-aware dialogue systems (Liu et al., 2019), and sophisticated decision-making architectures (Markel et al., 2023). Their

effectiveness depends heavily on the seamless integration of verbal and non-verbal communication channels, including gaze, gestures, and full-body movements.

Previous studies have shown that humans maintain social rules in the presence of more human-like virtual agents (Mostajeran et al., 2022). Bailenson et al. (Bailenson et al., 2001), for example, showed that people in virtual environments keep a larger distance to virtual agents than to virtual objects. They also associate more human-like characteristics such as being alive, calm, intelligent, and friendly to virtual agents in XR (Mostajeran et al., 2020). Moreover, research has shown that IVAs can be used to elicit human emotions such as psycho-social anxiety (Mostajeran et al., 2020) or facilitate their cognitive (Kruse et al., 2023) or physical task performances (Mostajeran et al., 2022).

On the other hand, avatars are referred to as virtual characters that are controlled by real users and are used for self-representation in virtual worlds (Freeman & Maloney, 2021). Previous studies have shown that the sense of presence and embodiment can be improved when users are provided with avatars (Steed et al., 2016). Avatars can facilitate generating the body-ownership illusion which arises when users have a sense of ownership over the virtual body that they have received in the virtual world, despite the certain knowledge that the virtual body is not their real body (Maselli & Slater, 2013). When using a humanoid avatar, users typically receive an upper body representation which can be controlled with limited input including a head-mounted display (HMD) and hand controllers. However, recent research has proposed using Smart Avatars (SA) (Freiwald et al., 2022) which can perform complex movements and express natural behavior despite having limited system input. For instance, users with SAs can have continuous full-body human representations for noncontinuous locomotion in XR. In addition, if the users teleport, their SAs would imitate their assigned user's real-world movements and autonomously navigate to their user when the distance between them exceeds a certain threshold. Thus, the observers could observe a natural human walk for that user instead of instant jumps caused by the teleportation.

A number of methods have been employed to create realistic humanoid 3D models that can be used as both agents and avatars. This may include complex technical setups, such as multi-view camera domes (Chu et al., 2020), or AI-based approaches, such as generative neural network architectures and diffusion models (e.g., autoencoders (Lombardi et al., 2018)), and Neural Rendering. The combination of these methods can potentially overcome the limitations of current solutions (Unity ZIVA[1], Unreal MetaHuman[2], or SoulMachines[3]), such as disturbing gaps in the perception process leading to uncanny valley effects (Mori et al., 2012), particularly noticeable for facial animations given that the human neural system is extremely sensitive for processing faces (Kanwisher et al., 2002).

Regarding interactive character animation, recent years have shown considerable progress in the use of machine learning techniques for both kinematic (see, for example, (Zhang et al., 2018), (Starke et al., 2022) and physics-based controllers (see, for example, (Peng et al., 2018), (Bergamin et al., 2019), (Hassan et al., 2023), (Lee et al., 2021), (Zhang et al., 2023). These have made these techniques more amenable to the creation of IVAs for VR experiences. It also allows exploring the perceived quality of character animation in VR experiences (Debarba et al., 2020) and introduces a different way to investigate open questions in motor neuroscience (Llobera & Charbonnier, 2023).

[1] https://unity.com/blog/news/update-about-ziva
[2] https://www.unrealengine.com/en-US/metahuman
[3] https://www.soulmachines.com/

There are however still numerous issues to address in the field. Current techniques can generate physically plausible movements (Peng et al., 2018), (Park et al., 2019), but achieving the nuanced expressivity and stylistic variations typical of human motion remains difficult (Park et al., 2019), (Wang et al., 2010). The mastering of body language is still a challenge, maintaining a consistent character style across different behaviors (Aristidou et al., 2017), (Smith et al., 2019). To enable interactions with embodied users, real-time performance is mandatory to react with the lowest latency (Van Welbergen et al., 2010). Both users and environmental interactions present exciting opportunities for future research in combining physics-based approaches with data-driven methods to create more sophisticated and believable virtual characters (Bergamin et al., 2019), (Hassan et al., 2023), (Starke et al., 2019).

4. WP4 Status

WP4 aims to deliver a toolkit (a.k.a. SDK or API) called Intelligent Virtual Human (IVH) whose goals are i) to provide real-time photorealistic humanoid 3D models based on cost-efficient technology, which can represent users (avatars) or agents (IVA), ii) to support natural and multimodal communication and interaction via speech, facial expressions, gaze, and full-body animations, and iii) to move from simple human-human communication to hybrid forms of interaction including multiple real and artificial users represented by smart avatars and IVAs. This toolkit is developed for the Unity3D game engine and comprises five modules (i.e., T4.1-5) that are explained in this section. Figure 2 shows an overview of the IVH toolkit with examples of each task.

Figure 2: An overview of the WP4's IVH SDK with examples of each task

4.1. T4.1 Virtual Humanoid 3D Models (DIDIMO)

4.1.1. Developments & key achievements

This task is focused on providing real-time photorealistic humanoid 3D models based on cost-efficient technology, which can represent users (avatars) or agents (IVA) to use in virtual reality applications, more specifically in the Use cases defined in the PRESENCE project: Professional Meeting, Manufacturing, Cultural Heritage and Health.

The development scope inside the WP4 framework included:

- The compilation of the technical requirements necessary for the creation of the 3D Humanoids in terms of model, rig, and textures, including garments and aesthetics for the use cases.
- The technical definition of the deliverables for pipeline integration.
- The creation of prototypes for pipeline testing purposes.
- The analysis of requirements for the different use cases - mainly aesthetic requirements that are being compiled with the use cases owners
- The final delivery of Avatars and IVA Unity Package upload to the WP4 framework repository with all the use cases characters defined in the previous steps (except one for, because it still lacks definition).

Architecture. The architecture is based on an on-line service provided by Didimo to generate humanoids from a selfie. There are two different modules:

The **Didimo Service**: it generates Humanoid 3D virtual characters that can be used in two different ways:

Smart Avatars (SAs): they are representations of Users. The Users need to upload a frontal facial photo (Dias et al. (2022), Dias et al. (2024)) and some general description of user characteristics : Weight, Height, Gender, etc.

Intelligent Virtual Agents (IVAs): they are characters controlled by computer agents. Since they do not need to represent a real user, they can be generated by just giving some general description - including Weight, Height, Gender and Ethnicity - or completely randomized. Note: we can also use synthetic photos to generate these IVAs.

The **Customization Service**: it provides the ability to customize the generated avatar in terms of weight/height and customizing it with hairs and garments assets, chosen by the user.

There are two Architectural outputs of T4.1. One output is directed for ivh-sdk-core module (T4.3) framework (uploaded in the WP4 framework repository as prefab Unity format), and it is the character that will be used in the final use case application. The second output is directed to ivh-sdk-animations module (T4.4) and is used for animation training purposes (defined in the T4.4 Full Body Animation and Interaction).

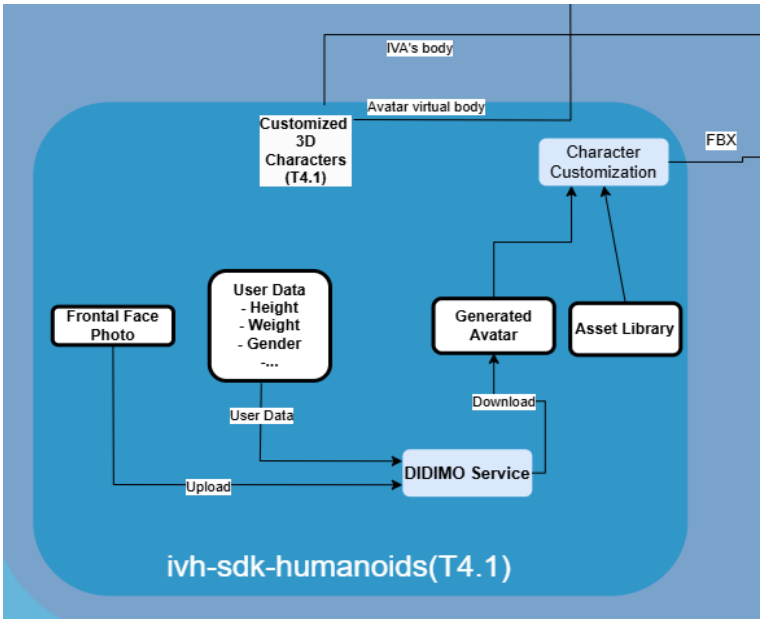


Figure 3: T4.1 architecture

Current Release Description

Figure 4: Project Use Case characters

The current release of task 4.1 is **Version 2.0.3** of the Humanoid 3D Model Package.

It Includes all Project Use Case characters (except Time Witness for the UC2.2 Cultural Heritage), and it is integrated in the WP4 framework repository (Unity package with prefabs of the characters).

The latest developments for task 4.1 were related to custom shading and lightning for improving visuals of the characters inside the VR environment.

4.1.2. KPIs status

Regarding *KPI 4.1: Delivery of efficient 3D humanoid model generation pipeline*: We have provided all the required Humanoid Characters for the initial integration and implementation of the four Use

Cases as Unity prefabs. Next, we will provide a service, so that new Humanoids can be generated to represent users or to populate new Demonstrators - for example, creating avatars with the project members.

Regarding *KPI 4.3: Integration of multimodal interaction capabilities*: We have given support related to facial animation methods. Specifically, the use of the F.A.C.S. (Facial Acting Coding System) to describe facial emotions.

Regarding *KPI 4.4: Deliver a set of APIs to integrate virtual humans with the two other pillars in the two demonstrators*: We have delivered our humanoid characters inside the provided template Unity packages, to be available to all WP4 and WP5 partners for integration.

4.1.3. Deviations & Mitigation Plan (if applicable)

Not applicable, as we haven't made major deviations to the planned work.

4.2. T4.2 Motion Tracking and Action Classification (JRS)

4.2.1. Developments & key achievements

Introduction

This module is responsible for detecting and classifying the actions of virtual humans (either avatars or virtual agents) in the scene in realtime. Unfortunately, the machine learning capabilities available within Unity are severely limited. For example, the *Barracuda* Unity package for neural networks does not support modern Transformer-based architectures. On the other hand, Python is the de-facto standard for implementing AI-power applications as one can employ very powerful Python packages for deep learning (like Pytorch) and computer vision (like MMDetection or MMPose).

We therefore developed a novel framework that combines two components: (1) A *Visual Analyzer* Python application that receives a live video stream and does the actual real-time action classification and sends the classification results to Unity, and (2) a *JrsVision* Unity package that renders a video live stream from a certain virtual camera viewpoint (which will by processed by the Visual Analyzer), receives the result of the action classification via REST API and associates them with the virtual humans in the scene.

Figure 5 illustrates the two components of the JRS framework for real-time action classification, consisting of the Visual Analyzer (the two windows showing the bounding box & detected action for the virtual human as well as its pose/skeleton) and the JrsVision Unity package (which receives the detected actions, matches them with the virtual humans and optionally shows the detected action as a text overlay in the Unity scene).

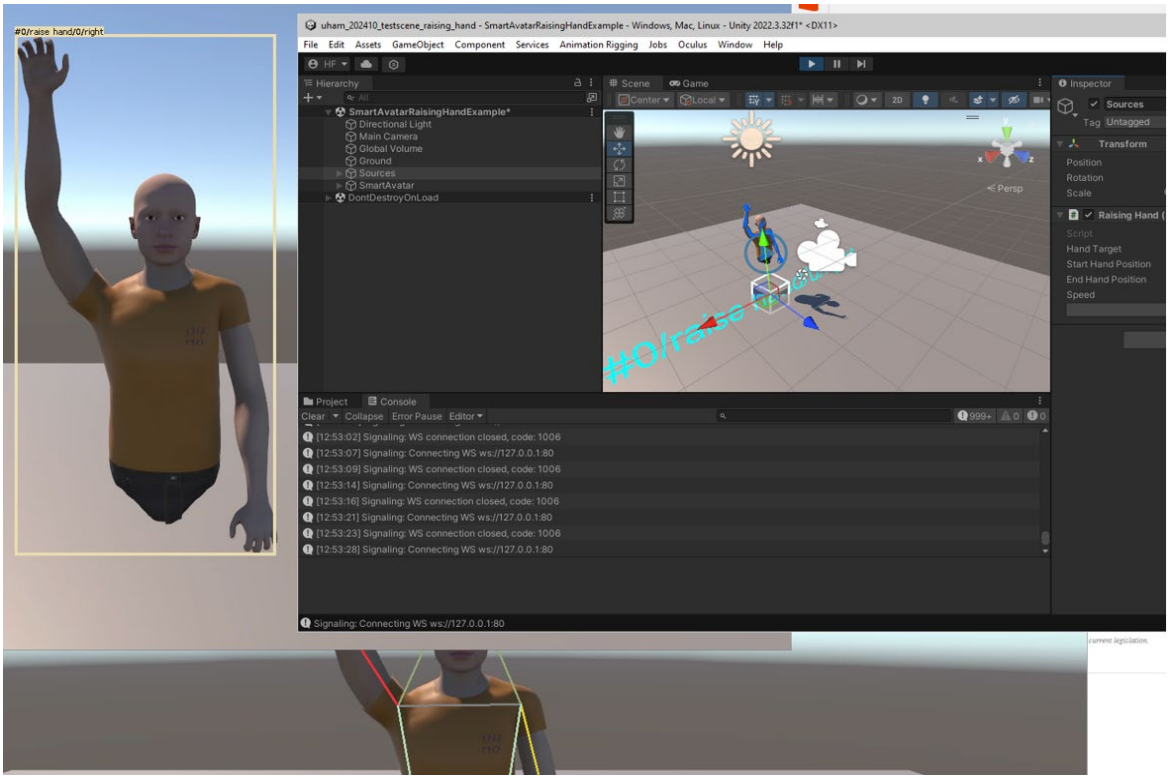


Figure 5: Illustration of the two components of JRS' framework for real-time action classification

Visual Analyzer

The *Visual Analyzer* Python application does the processing of the input video stream received from Unity in multiple steps. First, all virtual humans are detected with a deep learning based object detector and tracked with an optical flow based method. In parallel, for all detected humans their 2D pose (skeleton) is calculated with a deep-learning based pose estimation algorithm. The actions of all virtual humans are now calculated with both handcrafted and deep-learning based action recognition algorithms from the trajectory of its 2D poses within an analysis period of roughly two seconds. All detected actions are then sent back via REST API to the PC where Unity runs. In the following, we provide more details on the individual processing steps.

In the first step of the Visual Analyzer pipeline, all virtual humans visible in the current frame of the video live stream are detected (bounding boxes) and tracked. This is done by combining the *Scaled-YoloV4* object detector with an optical-flow based tracking method (using *TV-L1* optical flow). For best performance, both the neural network for object detection and the variational TV-L1 optical flow run on the GPU using CUDA and furthermore both steps are done in parallel via multi-threading. This makes the **object detection and tracking** real-time capable, performance measurements show that it takes on average 25 milliseconds per frame on a system with a RTX 3090 GPU. More details can be found in our *Omnitrack* paper (Fassold et al., 2019).

After that, for all detected virtual humans their 2D pose (skeleton) is calculated, as we need this as input for the subsequent skeleton-based action recognition. For the **human pose estimation,** we employ the recently proposed *RTMPose* algorithm (Jiang et al, 2023). It relies on the detected human bounding boxes and returns for each bounding box the estimated human pose (skeleton) in the *COCO* format (17 keypoints for the skeleton). RTMPose adopts *CSPNeXt* as the network backbone,

which shows a good balance of speed and accuracy and is furthermore deployment-friendly. We use the *RTMPose* implementation from the *MMPose* library [4]. From the available pretrained model variants there, we choose the *RTMPose-m* variant because it strikes a good balance between quality and runtime. For best performance, the pose estimation step is done in parallel to the main computation, in a separate *worker process* (using the Python Multiprocessing[5] package). The pose estimation takes roughly 8-10 milliseconds for each virtual human, so we achieve real-time performance for up to five virtual humans visible in the scene. More details can be found in our *LiveSkeleton* paper (Fassold et al., 2024).

For the **skeleton-based action recognition**, we rely on a combination of both handcrafted and deep learning based action recognition algorithms. We denote each of those as an *action expert*, where each action expert is responsible for detecting a certain set of predefined actions.For performance reasons, each action expert is running in parallel to the main application in a separate worker process. Each action expert retrieves the trajectory of the 2D poses for one virtual human over an analysis period of roughly two seconds and returns the detected action for it as the result. We employ two action experts which are detecting complementary actions. The first action expert uses a *hand-crafted approach* in order to detect the action “raise hand”. It does so via a simple heuristic which checks whether the elbow is located vertically higher than the center of the torso and whether the orientation of the forearm is approximately vertical. The second action expert uses a state of the art *deep learning based approach* to detect the 50 single-person actions from the NTU-60 dataset [6]. This covers actions like pickup, clapping, cross hands, hand waving, kicking something, jumping up and shaking head. We employ the *PoseConv3D* algorithm (Duan et al., 2022) which relies on a 3D CNN calculated from the 2D pose’s heatmaps. Compared to common graph convolutional network based methods, PoseConv3D is more effective in learning spatiotemporal features, more robust against pose estimation noises, and generalises better in cross-dataset settings. We use the PoseConv3D implementation which is available from the *MMAction2* [7] library. The runtime of the PoseConv3D algorithm is roughly 40-50 milliseconds for each virtual human, due to the complex 3D CNN model which is employed here. Note this does not hinder the real-time performance of the application, as it is done in parallel in a separate worker process and furthermore it is sufficient to do the action recognition a few times per second. Initial experiments on Unity scenes recordings provided by Artanim and I2CAT show that both action experts are able to detect the actions of the virtual humans robustly, as can be seen in Figure 6 below. We have designed the action recognition component very flexible, so that it can be extended by additional action experts (e.g. based on large vision-language models) in the future if needed.

[4] https://github.com/open-mmlab/mmpose
[5] https://docs.python.org/3/library/multiprocessing.html
[6] https://rose1.ntu.edu.sg/dataset/actionRecognition/
[7] https://github.com/open-mmlab/mmaction2

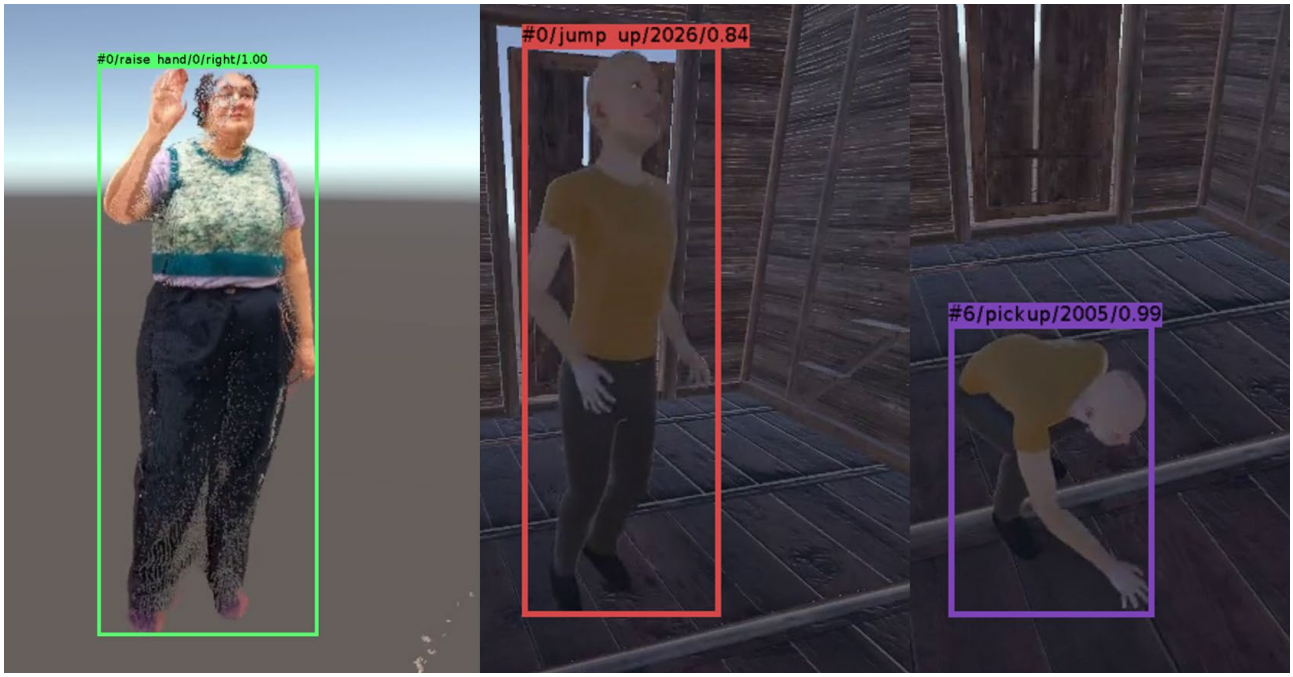


Figure 6: Illustration of successfully detected actions for virtual humans

JrsVision Unity package

The *JrsVision* Unity package is responsible for creating a video livestream from a dedicated Unity camera (we will call it observer camera in the following as it observes the actions of the virtual humans in the Unity scenes) and receiving the detected actions from the Visual Analyzer application via REST API and matching them with the virtual humans in the Unity scene.

In order to provide all this functionality, we developed several Unity components:

⇒ The *Stream Camera Capture* component generates a low-latency video stream (MPEG-TS, H.264 codec) from the view of the Unity camera to which it is attached to. This live video stream is analyzed in real-time by the Visual Analyzer python application. The low-latency video stream is generated by utilizing the *ffmepg* tool [8] embedded in a subprocess. The MPEG-TS transport protocol has several unique properties [9] (like defined latency) which make it well suited for creating a low-latency video livestream.

⇒ The *Collider Location History* component tracks the bounding box position (over the last few seconds) of the colliders which have been attached to Unity game objects of a certain tag (we use tag 'VirtualHuman' by default). This is necessary for the matching step (in *Action Receiver* component), as even with a low-latency MPEG-TS video stream there will still be input video stream latency of ~ 100-200 milliseconds.

[8] https://www.ffmpeg.org/
[9] https://www.obe.tv/why-does-mpeg-ts-still-exist/

⇒ The *Listener Rest Api* component implements a REST Service which handles the REST calls (REST messages) generated by the Visual Analyzer Python application.

⇒ The *Action Receiver* component parses the received REST messages (which contains the detected actions for the virtual humans) and matches them with the Virtual Human game objects in the Unity scene.

⇒ The *Action Visualizer* component is optional. When attached to the camera, then a text overlay will be shown over the virtual humans which shows their current action, which can be useful for debugging.

In order to provide action classification in the Unity scene, one has to create a static Unity camera (the *observer camera*) in the scene and place it in a way so that the virtual humans for which the actions shall be classified are visible from the field of view of the observer camera. After that, the JrsVision components listed above are added to the observer camera. Additionally, the Visual Analyzer application has to be started on the processing PC/Notebook.

After that, the result of the action classification for all virtual humans in the scene can be fetched by continuously polling the *Action Receiver* component. The JrsVision package alternatively provides also a callback-based interface via the *Action Info Listener* component. It has to be attached to the Virtual Human game object for which the action classification shall be done and provides the action for this virtual human via a callback function.

4.2.2. KPIs status

Regarding KPI 4.2: *Delivery of natural multi-user communication and collaboration capabilities for virtual humans with physics-based full-body animation*: A first prototype of the JRS action classification framework was provided, which is able to detect one action (‘raise hand’) with a heuristic method and additionally 60 actions (from NTU-60 dataset) with a deep learning based algorithm. The algorithm is real-time capable even for multiple virtual humans in the scene. After integration into the WP4 SDK, the action classification framework will help with creating a more natural and interactive way of communication between multiple users.

Regarding KPI 4.4: *Deliver a set of APIs to integrate virtual humans with the two other pillars in the two demonstrators:* We have provided a release of both components (Visual Analyzer / JrsVision) of our action classification framework, with documentation on how to install them and integrate them into scenes. Furthermore, a demo Unity project showing how the JrsVision Unity package can be used has been provided.

4.2.3. Deviations & Mitigation Plan (if applicable)

Not applicable.

4.3. T4.3 Speech and Facial Interaction (UHAM)

4.3.1. Developments & key achievements

This task focuses on enhancing the communication capabilities of the IVHs through speech and facial interaction utilizing the 3D humanoid characters developed in the T4.1. Our implementation is modular and can support a diverse set of 3D humanoids (e.g., Character Creator humanoid models)

as long as they follow a humanoid skeleton and provide facial blendshapes, necessary for facial expressions and lipsyncing technologies. The configuration also offers an easy to use setup for developers, where they just need to drag and drop a compatible 3D humanoid into the respective field in our IVH SDK and the application will configure the model and assign the necessary scripts automatically, so that it will work smoothly with other parts of the SDK. Figure 7 shows a close-up look into T4.3 architecture.

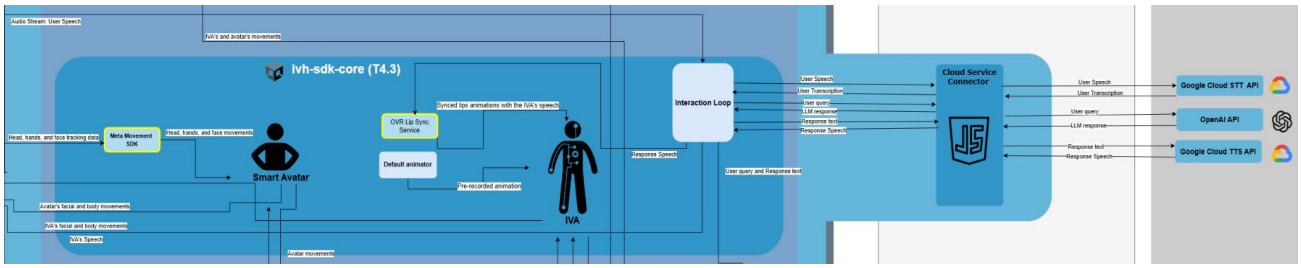


Figure 7: T4.3 architecture

This task aims to achieve two main objectives using the 3D humanoid models:

1) Creation of SAs, which embody users in XR environments, and
2) Development of IVAs, which are AI-powered computer-generated characters.

These avatars and agents enhance interactivity and engagement in XR applications, utilizing advanced AI technologies for natural communication. As it can be seen in Figure 7: T4.3 architecture, we have developed methods for natural communication between humans, SAs and IVAs based on processing spoken language and facial expressions. For this purpose, we used Meta Quest Pro HMD and its controllers to track users' head, face, eyes, and hands.

Smart Avatars (SAs)

In the first step, we created SAs based on the 3D humanoid models. To do so, we used only the upper-body part of these models which can be mapped using the Meta Movement SDK[10] to the tracking information received from the Meta Quest Pro HMD and its controllers. Since our current technology does not allow leg and foot tracking, a full-body representation for SAs is still not available. The ARKit blend shape standard is also compatible with Meta's Movement SDK which enables facial expression for SAs.

Intelligent Virtual Agents (IVAs)

To create IVAs we used the full-body of the humanoid models from T4.1. We did not change anything in the mesh, texture or bone structure of these models to create IVAs. As these are controlled via computer, the full-body can be used and animated. To enable natural and multi-modal interaction between users and IVAs, we have integrated AI-based services into our application. In particular, we have employed speech-to-text and text-to-speech synthesis technologies as well as large language models (LLMs) to provide conversational capabilities for the IVAs. This allows for contextually relevant and coherent dialogues between users and IVAs. We have integrated the following services

[10] https://developers.meta.com/horizon/documentation/unity/move-overview/

and our Service Connector application (based on JavaScript) facilitates connecting to each of these services:

- Speech-to-text (STT): Google API STT
- Text-to-speech (TTS): Google API TTS
- Large language models (LLMs): OpenAI's chat creation endpoint using models such as GPT-4o, GPT-4o-mini, or GPT3.5-turbo

We use a multimodal prompting strategy, sending the user's message to the LLM foundation models as a combination of text and system prompts. The text input, representing the user's message, is obtained via STT services like the Google Cloud API. The system prompt specifies the goal of the conversation and includes a list of potential actions and facial expressions that the LLM can trigger in its response. The action animations are sourced from Adobe's Mixamo gesture pack[11].

In our implementation, the IVAs are capable of expressing six basic emotions (happiness, sadness, anger, disgust, fear, and surprise) in different intensities and durations according to their conversation with the user. This is done by changing their corresponding facial blend shapes based on the action units of the facial action coding system (FACS) (Ekman and Friesen, 1978).

Furthermore, the LLM are prompted to generate structured outputs, responding to the user's query with both text and appropriate non-verbal behaviors, such as facial expressions. The text response is converted to speech using a TTS service (e.g., Google Cloud API). This audio is then synchronized with the agent's speech using the Oculus OVR Lip Sync[12], which aligns the speech with a standard 1:1 viseme set and modifies 15 viseme blendshapes. A single animator manages both the IVA's actions and facial expressions, while two distinct animation layers and animation masks separate the head animation from the body animation, ensuring coherent and synchronized movements.

We developed a configuration script in Unity to streamline the setup by connecting the IVA to predefined animation controllers, AI services, and lip-sync tools in a single step. This workflow accelerates IVA creation, making them quickly deployable for various XR applications.

4.3.2. KPIs status

KPI 4.1, Delivery of an efficient 3D humanoid model generation pipeline: for this KPI, we provided a modular implementation which can support a diverse set of 3D humanoids as long as they follow a humanoid skeleton and provide facial blendshapes, necessary for facial expressions and lipsyncing technologies. The configuration also offers an easy to use setup for developers, where they just need to drag and drop a compatible 3D humanoid into the respective field in our IVH SDK and the application will configure the model and assign the necessary scripts automatically, so that it will work smoothly with other parts of the SDK.

KPI 4.2: Delivery of natural multi-user communication and collaboration capabilities for virtual humans with physics-based full-body animation and KPI 4.3: Integration of multimodal interaction capabilities: for these KPIs, we provided virtual humans that are capable of natural communication

[11] https://www.mixamo.com/

[12] https://developers.meta.com/horizon/downloads/package/oculus-lipsync-unity/

and interaction with the users via speech and gaze and show facial expressions.

KPI 4.4: Deliver a set of APIs to integrate virtual humans with the two other pillars in the two demonstrators: we have provided the template Unity packages for T4.1-5 as well as example Unity projects and documentations. All the implementations described in this section are already available to all WP4 and WP5 partners.

1.1.1. Deviations & Mitigation Plan (if applicable)

Not applicable.

4.4. T4.4 Full Body Animation and Interaction (ARTANIM)

4.4.1. Developments & key achievements

This task is concerned with bringing physics-based animation methods to virtual reality environments and working to make these methods usable within the context and practices of virtual reality content production.

These systems are based on approaching continuous physical control as a model-free deep reinforcement learning problem, and therefore train a physical controller offline, which can then be used in real time within the physics update loop of the virtual reality simulation (see diagram on the left).

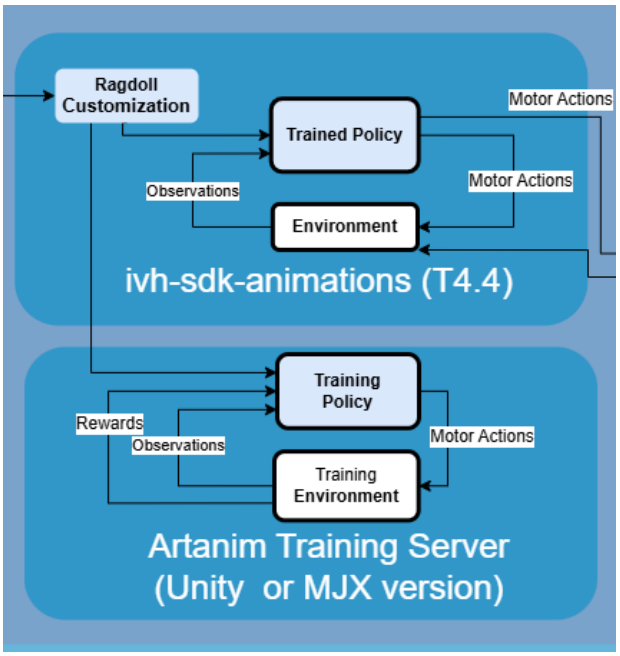


Figure 8: T4.4 architecture

Building from an initial package setup published by UHAM (18th september 2024), we have published 2 releases:

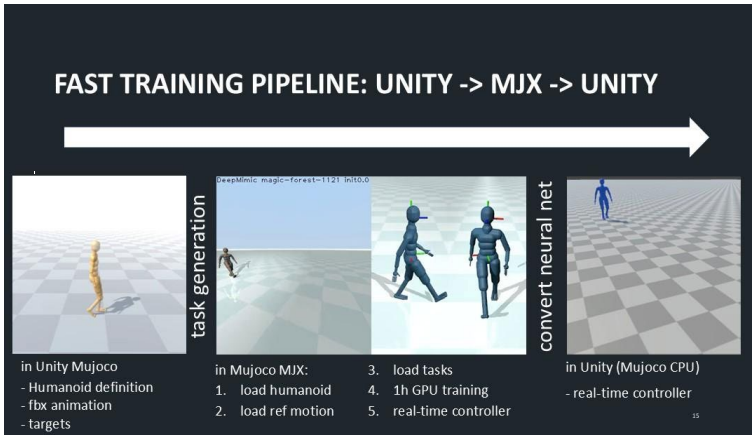
- Release v0.0.3 contained two benchmarks for physics-based controllers based on a state-based controller (DReCon (Bergamin et al., 2019)) and a stateless controller (DeepMimic (Peng et al., 2018)). The aim of these benchmarks was to allow a direct comparison of the

benefits and drawbacks of each approach for our use case. We also demonstrated the capabilities of the scaling scripts by showcasing the DReCon implementation adapted to a Didimo character.

- Release v0.0.4 integrated a commercial motion matching implementation, available as a plugin in Unity, with the DReCon implementation applied to the Didimo character, and three demonstration scenes. The decision to integrate a commercial implementation of motion matching was taken after preliminary tests with a simpler motion matching implementation, when we noticed that the commercial version was considerably better at blending the animations of the reference dataset.

Release v0.0.4 reflects our efforts generating a controller that could be used to control a virtual character in proximal distances. We have also focused on DReCon, which is a state-based controller, to make sure we have the maximum animation fidelity, and that the quality of the movement is therefore potentially better.

At the moment the main challenges remaining are improving the training data and improving the speed and size of the training process. To address this last challenge, fundamental to scale our training process and make our approach robust, it is important to highlight that, in parallel with these releases, we have worked on improving our training pipeline (see diagram, on the left). Indeed, exporting the ragdoll and motion data to Mujoco (Todorov et al., 2012) MJX allows compiling a training environment with Jax (Frostig et al., 2018), which in turns allows increasing the training



speed one order of magnitude, and doing so in a much more stable environment for training. We have demonstrated the process for one simple animation, a loop of a character walking. To keep the data export and import simple we choose to implement a stateless controller (DeepMimic). The next step is testing the result with skinned characters and analysing the result in terms of control quality, flexibility, and also animation fidelity.

Figure 9: T4.4 Training Pipeline

In parallel, we have begun implementing an Adversarial Motion Prior (AMP) (Peng et al., 2021) framework in MJX to improve motion naturalness and diversity. The framework uses a discriminator network to distinguish real from generated motions, guiding the policy network to produce more human-like movements. Our implementation leverages MJX and JAX for computational efficiency. Initial AMP experiments using the CMU mocap database show promising results, but we still have to debug and test the suitability of the approach. Next steps include scaling up training and leveraging

the extended motion dataset, advancing our goal of more natural physics-based character animations.

4.4.2. KPIs status

Regarding KPI 4.1, *Delivery of an efficient 3D humanoid model generation pipeline,* we have delivered a set of scripts that allow scaling the ragdoll to match the proportions of the skinned humanoid character. This implies training the ragdoll controller with the right proportions and allows mapping back to the character the movements created by the physics controller, and doing so with a one to one correspondence that clearly reflects the quality of the movements created by the controller.

Regarding KPI 4.2: *Delivery of natural multi-user communication and collaboration capabilities for virtual humans with physics-based full-body animation,* we have implemented a basic proof of concept integrating a physics-based controller with the outcome of task 4.3 (see task 4.5 below for a more detailed report). We have also delivered a framework allowing us to train, generate training data, and load the results of training in real-time controllers that are integrated into Unity.

Regarding KPI 4.4: *Deliver a set of APIs to integrate virtual humans with the two other pillars in the two demonstrators,* we have provided a release of our package, with documentation on the main scripts to call, and how to integrate it in other scenes, and tested with UHAM the ease with which this could be adopted.

4.4.3. Deviations & Mitigation Plan (if applicable)

There haven’t been major changes to the planned work

4.5. T4.5 Multimodal Interaction (UHAM)

4.5.1. Developments & key achievements

The objective of this task is to integrate the results from T4.1-4 to provide IVHs, which can represent real users and/or IVAs for realistic multimodal communication and interaction between humans and agents. This task seeks to combine the individual contributions presented in the previous tasks of WP4 into a coherent representation, keeping in mind the requirements presented by the use-case scenarios.

As previously described in T4.3, the current state of development has successfully integrated the 3D humanoid models from T4.1 into T4.3 and has created configurable models to be used as SAs or IVAs. We use a multimodal prompting strategy, sending the user's message to the LLM foundation models as a combination of text and system prompts. This results in a multimodal interaction between the user and IVAs using speech and gaze.

Moreover, and as mentioned briefly in T4.4, we could successfully integrate T4.1 and T4.3 with the physics based animation developed in T4.4. We confirmed that following the steps described in the README file was sufficient to get a humanoid controlled by a physics-based controller running with the conversational modules. Therefore, in the next step, T4.2 needs to be integrated into the IVH SDK.

After the integration of all tasks, we are planning to create a technical demo to showcase this integration. Figure 2 (V. Multimodal Interaction) shows an exemplary scenario of this integration. Here, the user has a smart avatar. The 3D model is for the embodiment of the user has been taken from the 3D models created in T4.1 for the cultural heritage use case. This model shows a female character with a casual outfit. In this demo scene, the user sees a standing IVA in the distance and greets the agent via speech while raising their hand (combining T4.1 and T4.3). By hearing this and detecting the raised hand (T4.2), the agent approaches the user, respects the social distance (T4.4), and greets the user back. The conversation can continue (T4.3). If the user changes the social distance and comes closer to the agent, the agent will move back and resume the socially acceptable distance (T4.4).

4.5.2. KPIs status

KPI 4.1, Delivery of an efficient 3D humanoid model generation pipeline: for this KPI, we provided a modular implementation which can support a diverse set of 3D humanoids as long as they follow a humanoid skeleton and provide facial blendshapes, necessary for facial expressions and lipsyncing technologies. The configuration also offers an easy to use setup for developers, where they just need to drag and drop a compatible 3D humanoid into the respective field in our IVH SDK and the application will configure the model and assign the necessary scripts automatically, so that it will work smoothly with other parts of the SDK.

KPI 4.2: Delivery of natural multi-user communication and collaboration capabilities for virtual humans with physics-based full-body animation and KPI 4.3: Integration of multimodal interaction capabilities: we have provided IVAs that are capable of natural communication and interaction with the users via speech and gaze and show facial expressions. They also work with the physics based animations developed in T4.4.

KPI 4.4: Deliver a set of APIs to integrate virtual humans with the two other pillars in the two demonstrators: we have provided the template Unity packages for T4.1-5 as well as example Unity projects and documentations. All the implementations of the APIs are already available to all WP4 and WP5 partners for integration.

4.5.3. Deviations & Mitigation Plan (if applicable)

Although this task started officially at the same time as T4.1-T.4, it could not be practically progressed before other tasks. The definition of this task explicitly describes a dependency to all other tasks in WP4. To begin with all tasks, including T4.5, UHAM created the necessary repositories, Unity package and example templates and dependency between the packages for a smooth integration of all packages into one. However, as T4.1-4 are done by different partners, T4.5 could not be integrated with all at once and with the same amount of effort. As explained above, T4.5 has already integrated T4.1, T4.3, and T4.4. Next is T4.2 which needs more efforts to be integrated and will be done in the next step.

5. Ethical Considerations

Regarding **T4.1** Virtual Human 3D Models, Didimo uses the end user's facial image to generate a 3D high fidelity avatar or Didimo, that is a realistic representation of them and that we deliver to them for their onward use.

Our legal basis for processing their personal information, of which there are various legal bases on which we may rely include:

Informed Consent: where you have given us (our customer if you are an authorised user) clear consent for us to process your personal information (including your biometric data) for a specific purpose. Informed consent requires that individuals aged 16 and over, who are of sound mind, fully understand and voluntarily agree to the terms and implications of a decision before proceeding.

Contract: where our use of your personal information is necessary for a contract we have with you, or because you have asked us to take specific steps before entering into a contract

Legitimate interests: where our use of your personal information is necessary for our legitimate interests or the legitimate interests of a third party (unless there is a good reason to protect your personal information which overrides our legitimate interests)

Scientific Research: where our use of your personal information (including biometric or Likeness Data) is necessary to allow us to carry out research and development into our machine learning processes, our development and improvement of our AI and algorithms, provided that we implement all relevant and legally required safeguards.

Further details could be found in Didimo's Privacy Policy.

With respect to the **action classification** algorithm developed by JRS in **T4.2**, we took care to consider potential ethical considerations during the development of the algorithm, especially with the AI modules employed within the algorithm. The algorithm employs three different neural networks in total, specifically (1) a neural network (Scaled-YoloV4) for object/person detection, (2) a neural network (RTMPose) for 2D pose / skeleton estimation and (3) a neural network (PoseConv3D) for pose-based action recognition. Note all used AI models have been taken from widely available and thoroughly reviewed open-source repositories, and the AI models have been trained on popular open-source training datasets (like MS-COCO for object detection). The action recognition model detects a set of roughly 50 actions from the NTU-60 dataset, with actions like waving hand, pickup something, jump up, thumbs up, throw and more. None of the detected actions might reveal sensitive personal information about the (virtual) human, and a potential misdetection of the action won't lead to a critical behaviour / reaction of other virtual humans in the scene.

In addition, in **T4.3**, we cannot fully control the output given by the LLM. ChatGPT has policies in place that filters out certain content, such as hate speech, violence, harassment, illegal activities, self-harm or harm to others or misinformation. Nevertheless, there are ways to bypass this filtering and GPT is not customized to individual tolerances and preferences. Therefore, the IVAs might say inappropriate sentences or hurtful content, which might impact the users. All users will be informed about this risk beforehand.

To display the IVHs, we use Meta products (e.g., Meta Quest 3 or Meta Quest Pro). In this case, Meta is a third-party provider that complies with GDPR and provides VR systems. To further maintain data protection, we could make a data deletion request to Meta upon users' requests. Meta HMDs are state-of-the-art and are currently the most affordable and usable; especially as stand-alone HMDs without cables.

Another third-party service that we use is OpenAI's GPT which enables natural conversations between the human user and the virtual agent. Data might be kept for up to 30 days and will be deleted afterwards. The data will not be used for training future models of OpenAI. We are using OpenAI's GPT because it currently shows the highest performance – cost ratio for us, since we do not plan to train and host our own LLM. The usage of OpenAI's GPT was allowed by the CTO of University of Hamburg for research purposes in our project.

We also use the Google TTS API to transcribe what the user is saying. This information is then sent on to the OpenAI LLM. Through Google TTS, original audio files are not saved, they are just used for transcription. In our settings, the audio files are not used to improve the speech recognition of the STT API. Temporary meta data will be saved for around five days, but this does not include the data of the customers.

Regarding **task T4.4**, concerned with bringing physics-based character animation techniques to VR, we do not detect ethical considerations regarding the use of data or the training methods used.

The data we use is recorded at our premises in a motion capture facility, using actors (or researchers adopting the role of actors), who explicitly accept that their behavior (walking, grasping an object, etc.) may be used in a virtual reality production. This is common practice in the video game and virtual reality production industry and poses no ethical considerations beyond common labor law.

Regarding the training procedure, we use a combination of code from two sources: code that is open source and publicly available with licenses that are quite liberal for research and commercial purposes, together with code developed in-house for this purpose. The entirety of the training is done in house on servers bought at the beginning of the project. We therefore don't see major considerations regarding the data, procedures and result of our ongoing efforts.

6. Preliminary Evaluation

We presented and provided a version of our IVH toolkit which contained only one virtual humanoid model (T4.1) and speech interaction from T4.3 to the Computer Science students of the Department of Computer Science at the University of Hamburg. They used this version to develop their Unity projects about IVHs for their Master's project. Four groups of students were working on different topics, all including IVAs. Their task was to develop a research study on I) natural interruption techniques, II) referencing scene objects in a conversation with an IVA, III) non-verbal communication, and IV) the ability of an IVA to demonstrate physiotherapeutic exercises to users. We used an 11-point Likert scale to capture their prior experience with Unity development and virtual reality before using our tool. We also captured their experience with our IVH toolkit via standard questionnaires and open-ended questions. Four students (three men and one woman) with an average age of 27 (SD=6.93) participated in our survey. One student entered 91 for their age and thus, was excluded from the mean calculation of the age. We incorporated the responses from all four participants in the evaluation of the remaining questions and questionnaires. They rated their prior experience with Unity development on average 6.75 (SD=2.5) and their prior experience with VR on average 4.5 (SD=1.29). The rest of this section reports on the results of this preliminary evaluation using standard usability and user experience questionnaires and open-ended questions.

○ System Usability Scale (SUS)

The first questionnaire that we used was SUS (Brooke, 1996) which has 10 items and measures the usability of a system on a scale of 0-100. Our participants' responses to this questionnaire gave us an average SUS score of 56.88 (SD=6.88). This indicates an OK to Good usability and a marginal acceptability score.

○ AttrakDiff

The second questionnaire was AttrakDiff (Hassenzahl et al., 2003) which consists of 28 items grouped into four sub-scales: Pragmatic Quality (PQ), Hedonic Quality relating to Identity (HQ-I), Hedonic Quality concerning Stimulation (HQ-S), and overall Attractiveness (ATT). This questionnaire helps evaluate overall user satisfaction and enjoyment. It has been also used in previous XR research for evaluating UX with XR systems (Horst & Dörner, 2019). The results show neutral to positive evaluations for all four sub-scales: Pragmatic Quality (M= .89, SD=.69), Hedonic Quality-Identity (M= .71, SD=.57), Hedonic Quality-Stimulation (M= .39, SD=.47), and Attractiveness (M= .46, SD= .36). Figure 10 shows the mean values for all four sub-scales.

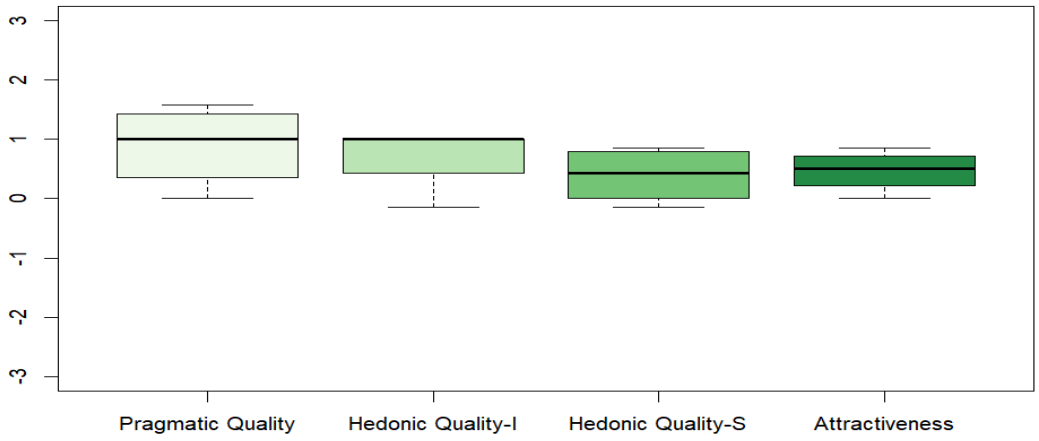


Figure 10: Mean values relating the AttrakDiff sub-escales

○ Extended Technology Acceptance Model (TAM2)

The next questionnaire was TAM2 (Venkatesh & Davis, 2000) which was employed to evaluate user acceptance and resistance to technologies. We used five sub-scales of this questionnaire to measure Intention to Use, Perceived Usefulness, Perceived Ease of Use, Output Quality (e.g., “The quality of the output I get from the system is high.”), Result Demonstrability (e.g., “I have no difficulty telling others about the results of using the system.”). As depicted in Figure 11, the evaluation of all scales was neutral to positive: Intention to Use (M= 4.5, SD= 1.41), Perceived Usefulness (M=4.81, SD= 1.28), Perceived Ease of Use (M=4.88, SD=1.15), Output Quality (M= 4.5, SD=.76), Result Demonstrability (M= 4.44, SD=1.26).

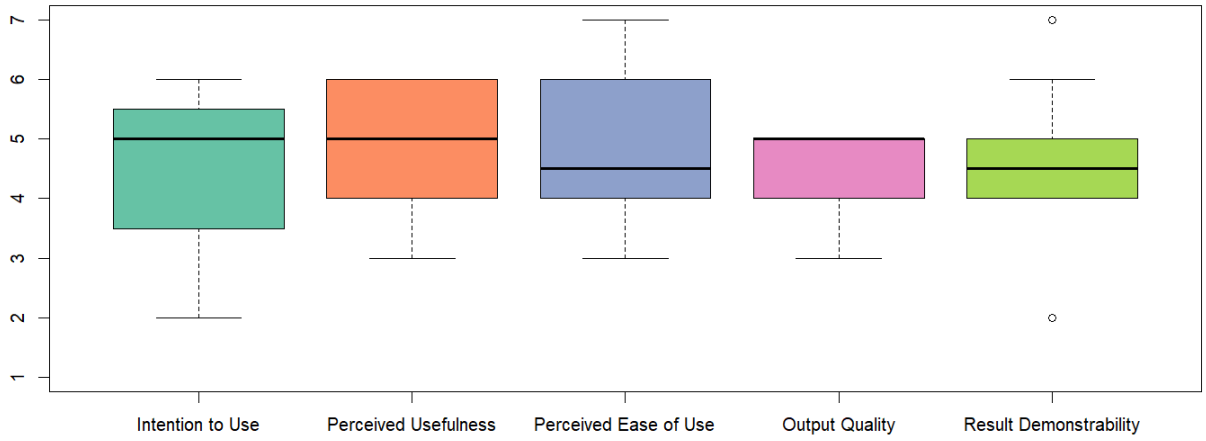


Figure 11: Mean values relating to Extended Technology Acceptance Model

○ Open-ended questions

At the end of the survey, we asked several open-ended questions to capture further views of the participants about their use of our toolkit. The first question asked about their general experience with the toolkit. Two participants wrote that they used it for their Master's project, one of them also for their bachelor thesis and they were happy about it. Another user wrote, “The intelligent virtual agent is currently missing some controls, such as the left arm up and down, which is a little frustrating, but otherwise, it's easy to use”.

The next question asked which part of the toolkit they liked or disliked. One user wrote, “I liked the fundamental functionality of the agent as these worked fairly well, the only problem is that the sound wasn't always recognized”. Another user wrote that they liked that ``many parts of the toolkit are controllable", while another user liked “the 3D Animation world and disliked the usability of the toolkit”.

We also asked what they would like to see added and their answers included “more models”, “more styles, hair, clothes for the avatar”, and “higher usability when animating and building objects”. We did not receive any responses regarding our question about what they wish to be removed or changed from the toolkit.

To our question of whether they would work with the toolkit again, we received three answers, and all were a definite yes. Two users also wrote that they would recommend our toolkit to others, one of them mentioned the recommendation would be for specific use cases. Another user wrote a “yes and no” (maybe) response to this question.

7. Outlook

In this deliverable, we presented the first iteration of the implementation of our Intelligent Virtual Human Toolkit (a.k.a. SDK) within WP4 which consists of five modules each of which delivers the

efforts in each WP4 task and contributes to producing realistic human-to-human and human-to-agent interactions in XR.

T4.1 provides a set of diverse virtual humanoid 3D models in terms of represented gender, age, and ethnicity, to be used out of the box for multiple use cases including professional collaboration, manufacturing training, health care, and cultural heritage. The current set of models has several limitations which will be addressed by diversifying the body shape and size, increasing visual fidelity, improving rendering conditions - scene, lights, and shaders, and improving body deformation for more complex muscular movements like arm twisting or secondary shoulder/clavicle motion. Furthermore, T4.1 will deliver a pipeline for creating humanoid 3D models based on users' photographs in the next iteration.

T4.2 facilitates detecting and classifying the actions of virtual humans in the virtual world in real-time. Using computer vision AI-based techniques, this module can already detect the action when a virtual human raises a hand in real-time as well as the 60 actions from the NTU-60 dataset. This will be integrated in the future to provide a multimodal interaction between users and agents. In addition, this module will be further improved to include more actions and visual representations of virtual humans such as the ones being holoported in real-time using 3D reconstruction techniques (WP2).

T4.3 provides speech and facial interactions for SAs and IVAs using AI-based services (such as speech-to-text, LLM, and text-to-speech). This allows for contextually relevant and emotionally intelligent dialogues between users and IVAs.

The techniques for interactive character animation that we are exploring are typically used in robotics and in physics simulation engines. They therefore use ragdoll-like characters, made of rigid or soft bodies assembled with joints that are actuated. However, these techniques are rarely used with skinned characters. In turn, both video games and virtual reality users use systematically skinned characters. It is an open question if these techniques can render the quality of movement that is expected when we compare these with more traditional interactive character animation techniques (typically, kinematic techniques) that are *de facto* considered industry standards. We plan to study in detail whether this is the case as a complement to our development efforts. An additional implicit assumption of our efforts in Body Animation and Interaction is that adopting physics-based interactive character animation techniques will help bring more life-like movement and dynamics to IVAs. This is an assumption that we plan to evaluate in terms of the comfort and plausibility of the VR environment, as perceived by VR users (Slater et al., 2022).

Finally, the results of our preliminary evaluation with four participants indicated an OK to Good usability and a marginal acceptability score. The poor usability was also mentioned in the additional comments of the participants which needs to be improved in the future. Our toolkit at the time of evaluation contained one humanoid 3D model from Module I (which represented both an SA and an IVA) with one idle animation and basic speech-based interaction between the user and the IVA from Module III (i.e., facial and emotional expressions were not included). As a result, participants wished to see more models and ready-to-use animations included in the toolkit. This has been partially addressed in terms of ready-to-use 3D models for specific use cases included in Module I and will be further improved in the future by including a pipeline for creating 3D models from simple camera photos and videos by the developer users themselves. We also observed neutral to positive evaluations for all sub-scales of AttrakDiff and TAM2 questionnaires. This means that improvements need to be made to both the task-oriented and hedonistic qualities of our toolkit. Further comments

from the participants revealed that despite all limitations, they would work with the toolkit again and would also recommend it to others.

We will continue improving our toolkit to make it more usable for creating XR solutions featuring IVHs. For future research, we will conduct several experiments to study the effects of interaction with single or multiple IVHs on users. For instance, we will study the effects of multi-modal interaction with IVHs in various XR scenarios including cultural heritage where IVHs represent tour guides and embody tourists, a manufacturing training scenario where both trainers and trainees are embodied as SA and receive assistance from an IVA, a health care scenario where IVHs help in reduction of medical procedure anxiety, and a professional collaboration scenario where users used IVHs to remotely participate in meetings in metaverse.

7.1. Planned Experiments

Table 2 summarizes some of the future planned experiments of WP4. Each partner leads one or two of these experiments and is involved in collaborative experiments.

Table 2: Test for list of tables

Title of study	Lead	Other involved partner(s)	Study purpose and relation to the project objectives
Evaluation of Intelligent Virtual Human SDK	UHAM	DIDIMO, JRS, Artanim	Usability and UX evaluation of the IVH SDK by developers
Multimodal Interaction with IVAs	UHAM	All WP4 partners	Evaluation of multimodal interaction with IVAs by end-users
Evaluation of the action recognition algorithm	JRS	-	Test the quality and robustness of the algorithm for human action recognition
Smart Avatar Likeness	Didimo		Test and evaluate the likeness of 3D Humanoid Characters to the user input photo. It is Intended for Smart Avatars (SAs)
Human Humanoid Cooperation	Artanim		Study under which conditions VR users feel at ease collaborating with humanoid, physics-based characters

7.2. Planned Publications

We are planning to submit the results of the above-mentioned experiments as well as five additional and in-progress behavioural studies at UHAM to various venues like ACM SUI, INTERACT, IEEE ISMAR, ICAT-EGVE, IEEE VR, ACM CHI. Furthermore, we have an ACM UIST 2025 submission in hand which will report on UHAM's work on T4.3 and T4.5.

JRS plans to publish the real-time action recognition algorithm at AI / computer vision conferences like IEEE ICIP, IEEE ISM and ACM Multimedia.

8. Abbreviations and definitions

8.1. Abbreviations

AI	Artificial Intelligence
HMD	Head-mounted display
IVA	Intelligent Virtual Agent
IVH	Intelligent Virtual Human
LLM	Large language model
SA	Smart Avatar
STT	Speech-to-text
TTS	Text-to-speech

8.2. Definitions

Avatar	A virtual character that is controlled by a real user and is used for self-representation in virtual worlds.
Virtual Agent	An autonomous character designed to interact naturally with humans in virtual environments.
Virtual Human	An umbrella term covering both virtual agents and avatars.

9. References

Aristidou, A., Zeng, Q., Stavrakis, E., Yin, K., Cohen-Or, D., Chrysanthou, Y., & Chen, B. (2017). Emotion control of unstructured dance movements. *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*, 1--10. 10.1145/3099564.309956

Bailenson, J. N., Blascovich, J., Beall, A. C., & Loomis, J. M. (2001). Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence: Teleoperators & Virtual Environments*, *10*(6), 583--598.

Bergamin, K., Clavet, S., Holden, D., & Forbes, J. R. (2019). DReCon: data-driven responsive control of physics-based characters. *ACM Transactions On Graphics (TOG)*, *38*(6), 1--11. https://doi.org/10.1145/3355089.335653

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), 4--7.

Chu, H., Ma, S., De la Torre, F., Fidler, S., & Sheikh, Y. (2020). Expressive telepresence via modular codec avatars. *Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23--28, 2020, Proceedings, Part XII 16*, 330--345.

Debarba, H. G., Chagué, S., & Charbonnier, C. (2020). On the plausibility of virtual body animation features in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, *28*(4), 1880--1893. 10.1109/TVCG.2020.3025175

Dias M., Coelho P., Figueiredo R., Carvalho R., Orvalho V., & Roche A (2024). Creating infinite characters from a single template: How automation may give super powers to 3d artists. In

ACM SIGGRAPH2024 Talks, pp. 1–2. 2024. 2. https://doi.org/10.1145/3641233.3664339

Dias M., Roche A., Fernandes M., & Orvalho V. (2022). High-fidelity facial reconstruction from a single photo using photo-realistic rendering. In ACM SIGGRAPH 2022 Talks, pp. 1–2. 2022. 1, 2 https://doi.org/10.1145/3532836.3536273

Duan, H. et al (2022). Revisiting Skeleton-based Action Recognition. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Fassold, H. (2019). OmniTrack: Real-time detection and tracking of objects, text and logos in video. *IEEE International Symposium on Multimedia (IEEE ISM).*

Fassold, H. (2024). LiveSkeleton: High-Quality Real-Time Human Tracking and Pose Estimation. *IEEE International Symposium on Multimedia (IEEE ISM).*

Freeman, G., & Maloney, D. (2021). Body, avatar, and me: The presentation and perception of self in social virtual reality. *Proceedings of the ACM on human-computer interaction*, *4*(CSCW3), 1--27.

Freiwald, J. P., Schmidt, S., Riecke, B. E., & Steinicke, F. (2022). The continuity of locomotion: Rethinking conventions for locomotion and its visualization in shared virtual reality spaces. *ACM Transactions on Graphics (TOG)*, *41*(6), 1--14.

Frostig, R., Johnson, M. J., & Leary, C. (2018). Compiling machine learning programs via high-level tracing. *Systems for Machine Learning*, *4*(9).

Hassan, M., Guo, Y., Wang, T., Black, M., Fidler, S., & Peng, X. B. (2023). Synthesizing physical character-scene interactions. *ACM SIGGRAPH 2023 Conference Proceedings*, 1--9. https://doi.org/10.1145/3588432.359152

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. *Mensch & Computer 2003: Interaktion in Bewegung*, 187--196.

Horst, R., & Dörner, R. (2019). Virtual reality forge: Pattern-oriented authoring of virtual reality nuggets. *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology*, 1--12.

Jiang, T. et al. (2023). RTMPose: Real-time multi-person pose estimation based on MMPose. ArXiv, abs/2303.07399.

Kanwisher, N., McDermott, J., & Chun, M. M. (2002). The fusiform face area: a module in human extrastriate cortex specialized for face perception.

Kruse, L., Mostajeran, F., & Steinicke, F. (2023). The Influence of Virtual Agent Visibility in Virtual Reality Cognitive Training. *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*, 1--9.

Kuo, C.-M., Lai, S.-H., & Sarkis, M. (2018). A Compact Deep Learning Model for Robust Facial Expression Recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2202-22028. 10.1109/CVPRW.2018.00286

Lee, S., Lee, S., Lee, Y., & Lee, J. (2021). Learning a family of motor skills from a single motion clip. *ACM Transactions on Graphics (TOG)*, *40*(4), 1--13. https://doi.org/10.1145/3450626.34597

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*. 10.48550/arXiv.1907.11692

Llobera, J., & Charbonnier, C. (2023). Physics-based character animation and human motor control. *Physics of Life Reviews*. https://doi.org/10.1016/j.plrev.2023.06.012

Lombardi, S., Saragih, J., Simon, T., & Sheikh, Y. (2018). Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)*, *37*(4), 1--13.

Markel, J. M., Opferman, S. G., Landay, J. A., & Piech, C. (2023). GPTeach: Interactive TA Training with GPT-based Students. *Proceedings of the Tenth ACM Conference on Learning @ Scale*. 10.1145/3573051.3593393

Maselli, A., & Slater, M. (2013). The building blocks of the full body ownership illusion. *Frontiers in human neuroscience*, *7*, 83.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, *19*(2), 98--100.

Mostajeran, F., Balci, M. B., Steinicke, F., Kühn, S., & Gallinat, J. (2020). The effects of virtual audience size on social anxiety during public speaking. *2020 IEEE conference on virtual reality and 3D user interfaces (VR)*, 303--312.

Mostajeran, F., Burke, N., Ertugrul, N., Hildebrandt, K., Matov, J., Tapie, N., Zittel, W. G., Reisewitz, P., & Steinicke, F. (2022). Anthropomorphism of virtual agents and human cognitive performance in augmented reality. *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 329--332.

Mostajeran, F., Reisewitz, P., & Steinicke, F. (2022). Social facilitation and inhibition in augmented reality: performing motor and cognitive tasks in the presence of a virtual agent. *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 323--328.

Mostajeran, F., Steinicke, F., Ariza Nunez, O. J., Gatsios, D., & Fotiadis, D. (2020). Augmented reality for older adults: exploring acceptability of virtual coaches for home-based balance training in an aging population. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1--12.

Park, S., Ryu, H., Lee, S., Lee, S., & Lee, J. (2019). Learning predict-and-simulate policies from unorganized human motion data. *ACM Transactions on Graphics (TOG)*, *38*(6), 1--11. 10.1145/3355089.335650

Peng, X. B., Abbeel, P., Levine, S., & Van de Panne, M. (2018). Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, *37*(4), 1--14. https://doi.org/10.1145/3197517.320131

Peng, X. B., Ma, Z., Abbeel, P., Levine, S., & Kanazawa, A. (2021). AMP: adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, *40*(4), 1-20. 10.1145/3450626.3459670

Slater, M., Banakou, D., Beacco, A., Gallego, J., Macia-Varela, F., & Oliva, R. (2022). A separate reality: An update on place illusion and plausibility in virtual reality. *Frontiers in virtual reality*, *3*, 914392. https://doi.org/10.3389/frvir.2022.914392

Smith, H. J., Cao, C., Neff, M., & Wang, Y. (2019). Efficient neural networks for real-time motion style transfer. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, *2*(2), 1--17. 10.1145/334025

Starke, S., Mason, I., & Komura, T. (2022). Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)*, *41*(4), 1--13. https://doi.org/10.1145/3528223.3530178

Starke, S., Zhang, H., Komura, T., & Saito, J. (2019). Neural state machine for character-scene interactions. *ACM Transactions on Graphics*, *38*(6), 178. 10.1145/3355089.3356505

Steed, A., Frlston, S., Lopez, M. M., Drummond, J., Pan, Y., & Swapp, D. (2016). An 'in the wild'experiment on presence and embodiment using consumer virtual reality equipment. *IEEE transactions on visualization and computer graphics*, *22*(4), 1406--1414.

Todorov, E., Erez, T., & Tassa, Y. (2012). MuJoCo: A physics engine for model-based control. *IEEE/RSJ International Conference on Intelligent Robots and Systems (2012)*, 5026-5033. 10.1109/IROS.2012.6386109

Van Welbergen, H., Van Basten, B. J., Egges, A., Ruttkay, Z. M., & Overmars, M. H. (2010). Real time animation of virtual humans: a trade-off between naturalness and control. *Computer Graphics Forum*, *29*(8), 2530--2554. 10.1111/j.1467-8659.2010.01822.x

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, *46*(2), 186--204.

Wang, J. M., Fleet, D. J., & Hertzmann, A. (2010). Optimizing walking controllers for uncertain inputs and environments. *ACM Transactions on Graphics (TOG)*, *29*(4), 1--8. 10.1145/1778765.1778810

Zhang, H., Starke, S., Komura, T., & Saito, J. (2018). Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics (TOG)*, *37*(4), 1--11. https://doi.org/10.1145/3197517.320136

Zhang, Y., Gopinath, D., Ye, Y., Hodgins, J., Turk, G., & Won, J. (2023). Simulation and retargeting of complex multi-character interactions. *ACM SIGGRAPH 2023 Conference Proceedings*, 1--11. https://doi.org/10.1145/3588432.359149

10. Annex I: project External Ethics Advisor report

ETHICS ADVISOR REPORT

PRESENCE - A toolset for hyper-realistic and XR-based human-human and human-machine interactions (Grant Agreement n. 101135025)

Period: 1st July 2024 – 30th June 2025

APPOINTMENT OF AN EXTERNAL ETHICS ADVISOR

The PRESENCE consortium proposed the appointment of an External Ethics Advisor (EEA) at the time of the proposal. In May 2024, Joana Porcel was appointed as the project's EEA to assess the ethical aspects of the work carried out in the project and to provide independent recommendations. In particular, the following deliverables will be review and reports will be prepared (June 2024, June 2025, December 2026):

Year 1:
D1.1- Human Centred Development Phase I - Foundations, Requirements and Initial Planning
D7.2- Ethics Framework and Data Management Plan I

Year 2:
D1.2- Human centred Development Phase II - Intermediate User Testing, Presence Evaluation, Ethics, Trust & Privacy
D4.1- Virtual humans technologies report I
D7.4- Ethics Framework and Data Management Plan II

Year 3:
D1.3- Human centred Development Phase III – Final User Testing, Presence Evaluation, Ethics, Trust & Privacy
D4.2- Virtual humans technologies report II
D7.6- Ethics Framework and Data Management Plan III

Additional input in the set-up of the experiments, e.g., the content of the informed consent, the nature of the requested participation, data management practices or provided incentives is also foreseen.

MEETINGS and OTHER CONSULTATIONS

During the Period, the beneficiaries and the EEA have met on:

DATE	ATTENDEES	OBSERVATIONS
July 10th, 2024	- PRESENCE Consortium - Joana Porcel (ISGlobal). EEA	The EEA attended the consortium meeting on July 2024. The presentation is attached to this report.
February 5th, 2025	- Louise Hallström, Researcher VUB - Joana Porcel (ISGlobal). EEA	Consultation if obtaining ethical approval from an ethical board is required for the study where the general public (students from the Vrije Universiteit Brussel) will try the First Playable app, a VR application developed within the project. This test session will last approximately 30 minutes and is designed to gather feedback on user experience and usability, using questionnaires prepared by the partner UHAM.

		The researcher is advised that in my role as EEA, I cannot give ethics approval. Approval, if required, must come from an accredited committee. I recommend to contact the university committee that reviews social science or technology projects to check whether this study requires approval under local regulations. Also check if the involvement of students requires specific approval from the university. The informed consent form was also reviewed. It is very well prepared. The only advice I gave was about the possibility of anonymising the data in case participants do not need to be contacted again, including the removal of the voice recordings.

DOCUMENTS REVISED

D4.1- Virtual humans technologies report I (Version 0.5; 18-02-2025).

This deliverable describes the objectives, key performance indicators (KPIs), and essential tasks concerning the technological pillars of smart avatars (SAs) and intelligent virtual agents (IVAs).

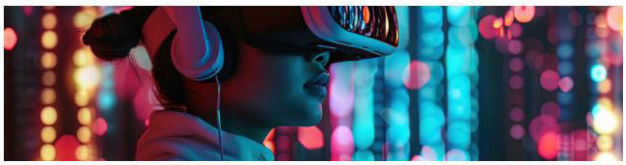
The document includes a section on the Ethical Considerations, including privacy issues and the legal basis for personal data processing (see Privacy policy of Didimo; https://privacy.didimo.co/privacy-policy/). An analysis of the different tools, platforms used in these activities in terms of privacy is also included.
The researchers inform that certain content of hate speech, violence, harassment, etc. may occur when using ChatGPT. As well as informing participants that this may happen, the team is advised to monitor closely and take steps to minimise its occurrence.
Elements of age, gender and ethnicity diversity in the design of the 3D human models are considered.
The following manuscript is recommended: Gerlich, Michael. 2024. Societal Perceptions and Acceptance of Virtual Humans: Trust and Ethics across Different Contexts. Social Sciences 13: 516. https://www.mdpi.com/2076-0760/13/10/516

ETHICS ISSUES ON EXTENDED REALITY

The consortium is recommended to explore the training modules of the iRECs project and to recommend young researchers and other relevant project staff to undergo this targeted training.

iRECs - Training modules on Extended Reality

Welcome To Extended Reality: Technology Basics.
https://classroom.eneri.eu/node/259



Welcome To Extended Reality: Ethics Issues.
https://classroom.eneri.eu/node/266

Signed:

PORCEL CARBONELL JUANA ALICIA -

Firmado digitalmente por PORCEL CARBONELL JUANA ALICIA xxxxxxxxxx
Fecha: 2025.02.24 11:56:06 +01'00'