Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI-Assisted Tutorial Generation for Immersive Training in Virtual and Mixed Reality

Bálint György Nagy

nagy.balint.gyorgy@vik.bme.hu

Budapest University of Technology and Economics

Péter Szögi

HSN Lab Ltd.

János Dóka

Budapest University of Technology and Economics

Bence Bihari

Budapest University of Technology and Economics

László Kopácsi

German Research Centre for Artificial Intelligence

Balázs Sonkoly

Budapest University of Technology and Economics

Research Article

Keywords: Human-centered computing, Human computer interaction (HCI), Natural Language Processing (NLP), AI for education, AI-assisted tutorial generation, Mixed reality training systems

Posted Date: July 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10374382/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AI-Assisted Tutorial Generation for Immersive Training in Virtual and Mixed Reality

Bálint György Nagy [1,2*], Péter Szögi [3], Bence Bihari [1], János Dóka [1,2], László Kopácsi [4], Balázs Sonkoly [1,2]

[1]Department of Telecommunications and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Hungary.
[2]HUN-REN-BME Cloud Applications Research Group, Hungary.
[3]HSN Lab Ltd., Hungary.
[4]German Research Center for Artificial Intelligence (DFKI), Germany.

*Corresponding author(s). E-mail(s): nagy.balint.gyorgy@vik.bme.hu;

Abstract

Training in industrial environments is often expensive and time-consuming, and the creation of new training materials remains a major bottleneck, especially in domains where procedures change frequently. Virtual reality (VR) and mixed reality (MR) technologies provide promising environments for immersive procedural training, but their broader adoption is limited by the expertise, time, and effort required to author high-quality content.

In this paper, we present an AI-assisted tutorial generation framework for immersive training in VR and MR environments. The proposed approach is based on expert demonstrations: an expert performs and explains a task once, while the system captures multimodal information, including speech, hand movements, head pose, and gaze information. Speech-to-text and large language model components are then used to generate a structured step-by-step tutorial. The resulting training material includes textual and voice-based instructions, as well as a time-aligned 3D avatar replay of the expert's recorded movements, enabling trainees to observe how each subtask should be performed.

The framework was evaluated in two complementary scenarios. In VR, we implemented an industrial engine assembly tutorial on the VIROO platform using Meta Quest 2 headsets and investigated the effect of expert replay on tutorial execution through questionnaires and automatically logged task completion times. In MR, we evaluated the AI-assisted training material generation workflow using

a HoloLens 2 headset, focusing on usability, user experience, learning progress, and content generation accuracy.
The results indicate that expert replay helps clarify correct execution techniques, spatial positioning, object orientation, and required movements. Moreover, the MR evaluation demonstrates the feasibility of generating reusable immersive tutorials from expert demonstrations. Overall, the proposed framework supports scalable and instructorless immersive training for industrial and educational applications.

Keywords: Human-centered computing, Human computer interaction (HCI), Natural Language Processing (NLP), AI for education, AI-assisted tutorial generation, Mixed reality training systems

1 Introduction

Industrial training increasingly requires workers to acquire practical procedural skills while production processes, tools, and safety regulations evolve continuously. As a result, organizations must update training materials frequently, creating substantial costs in both content development and instructor effort. Although virtual reality (VR) and mixed reality (MR) technologies have demonstrated considerable potential for improving procedural learning, the creation of immersive training content remains a major obstacle to their widespread adoption. Developing high-quality tutorials typically requires expertise in 3D modeling, interaction design, and software development, making the authoring process both time-consuming and expensive.

Existing immersive training systems primarily focus on improving the learner's experience rather than simplifying the creation of training content. In industrial environments, where workflows and equipment often change, manually updating immersive tutorials can become a significant maintenance burden. Consequently, reducing the effort required to transform expert knowledge into reusable immersive training material has become an increasingly important research challenge.

Recent advances in artificial intelligence provide new opportunities to address this problem. Automatic speech transcription, large language models (LLMs), and multimodal data processing enable expert demonstrations to be converted into structured instructional material with substantially less manual intervention. At the same time, modern head-mounted displays provide accurate tracking of head, hand, and gaze movements, allowing expert demonstrations to capture not only verbal explanations but also the spatial and temporal characteristics of task execution. However, these capabilities are often studied or applied separately, and their combination into a practical end-to-end workflow for generating reusable immersive tutorials from expert demonstrations remains an open challenge.

In this paper, we present an AI-assisted framework for generating immersive procedural tutorials from expert demonstrations. The proposed system captures synchronized speech, head movement, hand tracking, and gaze information while an expert performs a task. Speech recordings are automatically transcribed and processed by an LLM to generate structured step-by-step instructions, while the recorded multimodal

data are synchronized with the generated tutorial to provide replayable expert demonstrations. The resulting tutorials can be deployed both in VR, where learners practice within a fully virtual environment, and in MR, where digital guidance is superimposed onto the real workspace.

This paper extends our previous work on AI-assisted mixed reality training systems. An initial demonstration of the concept was presented in [1], where we introduced the basic idea of using artificial intelligence to support mixed reality training material creation. A later conference version [2] described the mixed reality implementation of the framework in more detail, focusing on AI-supported tutorial generation, speech-based authoring, and the use of expert demonstrations for scalable training material creation.

Compared with these earlier publications, the present work substantially extends the system and its evaluation in several directions. First, we implemented the tutorial playback and expert replay module in a fully virtual reality environment. For this purpose, we used the VIROO platform [3], an enterprise-oriented XR platform that supports the creation and deployment of immersive training and simulation applications in virtual environments. This allowed us to evaluate the proposed tutorial concept in a controlled VR setting using Meta Quest 2 headsets. Second, we conducted a dedicated user study to analyze how participants perceived the VR tutorial system and whether the avatar-based expert replay provided added value compared with a standard VR tutorial using conventional guidance elements. Third, we extended the evaluation with automatically logged execution-time measurements, SUS and UEQ questionnaires, and a custom expert replay questionnaire. In this way, the present paper goes beyond the earlier MR-focused studies by analyzing both the usability of the VR implementation and the practical contribution of expert replay to immersive procedural training.

The main contributions of this paper are as follows:

- We present an AI-assisted tutorial generation framework for immersive procedural training based on expert demonstrations and multimodal data capture (gaze information, head and hand tracking).
- We implement and evaluate the learner-facing tutorial experience in VR using an industrial engine assembly scenario on the VIROO platform with Meta Quest 2 headsets.
- We investigate the added value of expert replay through subjective questionnaire data and automatically logged tutorial execution times.
- We evaluate the mixed reality component of the framework using a HoloLens 2 headset, focusing on usability, user experience, learning progress, and accuracy of the automatically generated instructional content.
- We discuss the complementary roles of VR and MR for scalable immersive training, where VR provides a controlled environment for procedural practice, while MR enables context-aware training and rapid tutorial authoring based on expert demonstrations.

2 Related Work

Training and onboarding have long been recognized as resource-intensive activities in both industrial and educational environments. Conventional training methods typically rely on instructors, printed documentation, videos, or manually prepared digital learning materials, making them difficult to scale and expensive to maintain. According to Tran et al. [4], the development of a single hour of training material may require between 50 and 300 hours of preparation. This considerable authoring effort becomes an even greater challenge in industrial settings, where procedures, equipment, or safety regulations frequently change and training materials must therefore be updated continuously.

2.1 Artificial Intelligence for Learning Content Generation

Recent advances in artificial intelligence have significantly expanded the capabilities of educational technologies. AI-based systems can personalize learning by monitoring learner progress, adapting instructional content, and providing individualized feedback in real time [5]. Such capabilities are particularly valuable in self-directed learning environments where continuous instructor supervision is impractical.

Large language models have further increased the potential for automating instructional design. GPT-4 and related models are capable of generating learning objectives, lesson plans, assessment questions, and educational feedback. Sridhar et al. [6] demonstrated that GPT-4 can produce pedagogically meaningful learning objectives for higher education, while Yadav [7] proposed an LLM-supported framework for scalable evidence-based instructional design. Although these studies demonstrate the effectiveness of LLMs for generating textual educational material, their application to immersive procedural training remains relatively limited.

2.2 Virtual and Mixed Reality Training

VR and MR have become important technologies for procedural training because they enable interactive and immersive learning experiences that would be difficult or unsafe to reproduce using conventional methods. Previous studies have shown that MR-based instruction can improve learner engagement and knowledge retention and may outperform traditional presentation-based teaching in certain educational scenarios [8–11].

Immersive technologies are particularly valuable in domains where real-world practice is hazardous, expensive, or difficult to organize, including healthcare, manufacturing, emergency response, and military training [12–14]. Unlike VR, which requires the complete virtual recreation of the training environment, MR overlays digital information directly onto the physical workspace. This characteristic makes MR especially attractive for industrial training because existing equipment and workspaces can be reused without extensive virtual content creation.

Commercial learning platforms such as Immersive4Learning [15] and ClassVR [16] demonstrate the growing adoption of immersive educational technologies. However, these systems primarily focus on delivering VR learning experiences and provide limited support for rapid authoring of industrial MR training scenarios.

2.3 Demonstration-Based Learning and Expert Replay

Many procedural tasks require learners to understand not only what action should be performed but also how it should be executed. Static instructions, arrows, or visual highlights often fail to communicate movement trajectories, hand orientations, or the timing of complex operations. Consequently, expert demonstrations represent an important source of procedural knowledge.

Kuo [17] proposed a motion capture system that records body movements, facial expressions, voice, and object states for educational VR tutorials through humanoid avatars. Similarly, VR-Replay [18] captures verbal and non-verbal communication for asynchronous collaboration by replaying avatar movements, gestures, and speech within a virtual environment. These studies demonstrate that replaying expert behavior can communicate procedural knowledge that is difficult to convey using textual or static visual instructions alone.

2.4 Immersive Content Authoring

The creation of immersive tutorials remains one of the principal barriers to adopting VR and MR training systems. Developing such applications typically requires expertise in interaction design, 3D content creation, software development, and instructional design.

Several recent approaches aim to reduce this authoring effort. EditAR [19] introduces a digital twin-based workflow for generating AR, VR, and video instructions from a single recorded demonstration. Video2MR [20] automatically converts conventional instructional videos into MR experiences using AI-based motion reconstruction. The Interaction Design Toolkit for Physical Task Guidance [21] provides structured methods for translating real-world workflows into MR applications by incorporating multimodal interaction techniques such as gestures, gaze, and speech.

Our previous work also addressed this authoring bottleneck in the context of AI-assisted MR training. In an initial demonstration paper, we introduced the basic concept of an AI-powered mixed reality training platform in which an expert performs and explains a complex procedure once, and the system uses the recorded demonstration to generate a step-by-step tutorial [1]. In a subsequent conference paper, we extended this concept into a more complete MR implementation, including speech-based authoring, LLM-based step segmentation, hand, head, and gaze recording, expert motion playback, and an evaluation of usability, user experience, learning progress, and content generation accuracy [2]. These works demonstrated the feasibility of AI-supported tutorial generation in mixed reality, but they focused primarily on the MR authoring workflow and did not investigate the learner-facing value of expert replay in a fully virtual environment.

Commercial systems such as Altoura [22] and VirtualSpeech [23] further demonstrate the integration of AI and immersive technologies for onboarding, communication training, and procedural guidance. Nevertheless, these platforms generally rely on manually authored content and predefined instructional flows, limiting their adaptability when training materials must be created or updated rapidly.

Overall, existing research demonstrates that AI, VR, and MR can substantially improve the delivery of immersive learning experiences. However, comparatively less attention has been devoted to reducing the effort required to create immersive procedural tutorials. In particular, there remains a lack of integrated frameworks that capture expert demonstrations, automatically structure procedural knowledge using AI, and generate learner-ready VR and MR tutorials while preserving replayable expert movements. Building on our previous MR-focused work, the framework presented in this paper extends AI-assisted tutorial generation toward a broader XR setting by supporting both MR-based authoring and VR-based tutorial execution. In addition, the present work specifically evaluates the added value of avatar-based expert replay in VR, which was not investigated in our earlier studies.

3 Main Ideas

The main idea of the proposed work is to reduce the effort required to create immersive procedural training tutorials while preserving the practical advantages of VR- and MR-based learning. Existing training practices and available immersive training solutions indicate that there is still a lack of tools that are simultaneously cost-effective, rapid to use, and accessible to non-technical instructors. This limitation is particularly relevant for small-scale or highly specialized workflows, where the development of a dedicated training application by an external company would not be economically justified. Nevertheless, such workflows can still benefit substantially from immersive guidance, especially when learners must understand spatial relations, tool usage, object manipulation, or procedural sequences.

The proposed framework is designed around a single-demonstration authoring concept. Instead of requiring instructors to manually design every training step, create all tutorial elements from scratch, or script the complete interaction flow, the system enables a teacher or domain expert to record a workflow by performing it once. During this demonstration, the platform captures multiple complementary data streams, including hand movements, head pose, gaze direction, and spoken instructions. These modalities describe not only what the expert says, but also how the task is physically performed and which objects, locations, or movements are relevant at each stage of the procedure.

The recorded demonstration is subsequently processed using a set of AI-assisted modules. First, the expert's spoken instructions are transcribed into text using speech-to-text technology. The resulting transcript is then analyzed by a large language model (LLM), which identifies the logical structure of the demonstrated procedure and decomposes the overall task into smaller, manageable sub-tasks. This step is essential because raw recordings are rarely suitable for direct instructional use: learners require a clear sequence of actions, concise explanations, and explicit transitions between procedural steps. By automatically structuring the demonstration, the system transforms an expert performance into an organized tutorial.

The generated training material combines textual, auditory, visual, and spatial information. For each sub-task, the platform provides a short textual description, a spoken instruction generated through text-to-speech, and visual guidance derived

from the recorded expert behavior. A key element of this guidance is the time-aligned 3D replay of the expert's movements through an avatar representation. This allows trainees to observe the correct movement, orientation, positioning, and timing of each action directly in the immersive environment. Such multimodal presentation is especially useful in tasks where the correct execution technique is difficult to convey through text, static cues, or highlights alone.

The framework is designed to support both VR and MR deployments, although the two modalities differ in their practical requirements. In MR, training can take place either at the final workplace or in a dedicated test room, where the relevant tools, parts, and physical objects are physically present and can be manipulated directly. This reduces the need to recreate the entire environment virtually and makes MR especially suitable for context-aware guidance and rapid tutorial authoring in real-world settings. In contrast, VR provides a fully virtual training environment, where the workspace, tools, and task-relevant objects must be modeled in advance. These virtual objects also need to be prepared with appropriate interaction logic, such as grasping, placement, snapping, or validation behavior. Consequently, VR deployment requires a higher initial preparation effort. However, once a virtual representation of a factory, workstation, or training scenario is available, the proposed system can reuse this environment and enable experts to generate new tutorials more easily by demonstrating additional procedures.

From the learner's perspective, the resulting tutorial functions as an interactive step-by-step guide. After selecting the desired training material, the trainee can progress through the generated sub-tasks at an appropriate pace. The system provides contextual instructions directly in the training environment, allowing the learner to connect each explanation with the corresponding object, location, or action. In VR, this enables a controlled and repeatable training scenario where expert replay and tutorial guidance can be evaluated systematically. In MR, the same concept can be applied directly in the physical workspace, reducing the need to recreate the entire environment virtually and supporting faster content creation for real-world procedures.

Overall, the proposed approach shifts the focus of immersive training development from manual content creation to AI-assisted generation based on expert demonstrations. By integrating multimodal recording with speech processing, large language models, and synchronized 3D replay of expert actions, the framework aims to make immersive training content easier to produce, update, and reuse. This can increase the feasibility of VR- and MR-based education not only in large industrial settings, but also in smaller educational or professional contexts where conventional custom development would be too costly or time consuming.

4 Our XR training platform

The proposed Extended Reality (XR) training platform was developed using the Unity 3D Engine. As depicted in Figure 1, the system architecture is composed of four principal components: the *XR Training Material Generator*, the *Data Provider*, and the *Virtual/Mixed Reality Trainer*. The *XR Training Material Generator* is implemented for HoloLens 2 headsets using MRTK3 [24] in Unity, and enables domain experts to

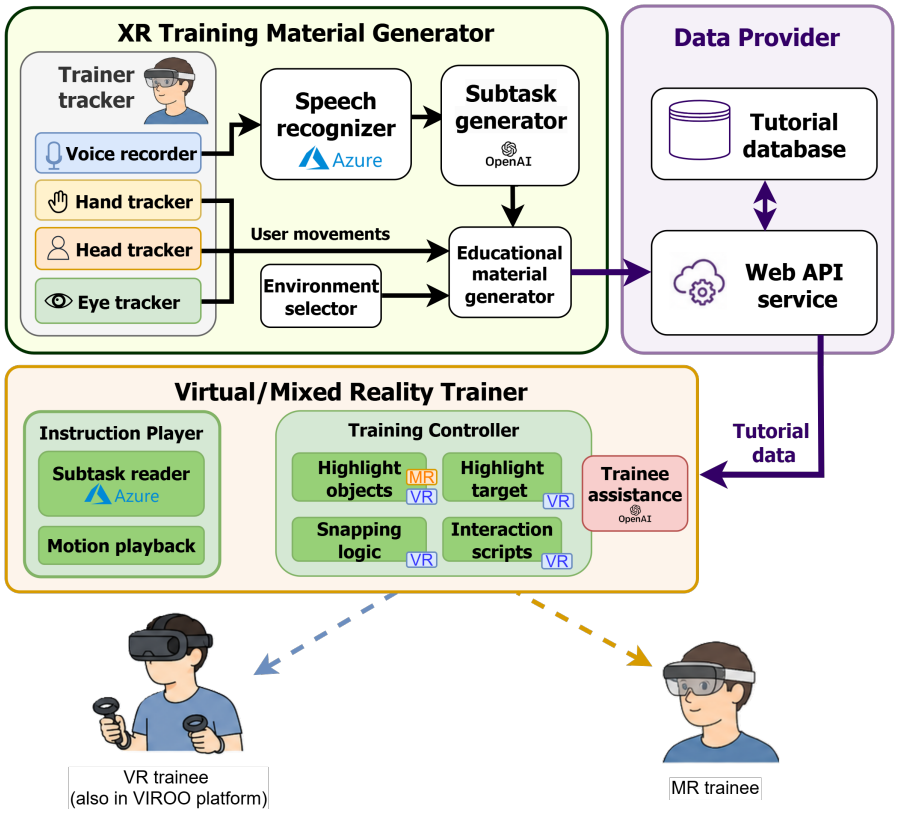


Fig. 1: Proposed training platform with artificial intelligence support

record demonstrations and generate instructional content. The *Data Provider* serves as the central repository for storing and managing the generated training materials. The two trainer modules provide the learner-facing tutorial experience in different immersive environments. The mixed reality trainer is implemented in Unity using MRTK3, while the virtual reality trainer is implemented in Unity within the VIROO [3] environment.

When creating new instructional material, the system guides the expert through a structured workflow comprised of several distinct stages. The complete workflow for content creation is illustrated in Figure 2. The initial step involves selecting the operational context or working scene relevant to the task at hand. This is handled by the environment loader, which allows users to import predefined virtual scenes, including scenes originally created for VR tutorials. In this case, the tutorial can be authored and executed in a fully virtual environment. Alternatively, if no predefined scene is loaded, the system operates in MR mode and uses the physical workspace as the basis of the tutorial. The real environment can be dynamically captured through spatial scanning, which is accomplished via Unity's AR Mesh Manager and enables spatial mapping of the user's surroundings. In cases where the procedure is independent of any specific environment, this step can also be bypassed. Once the environment is

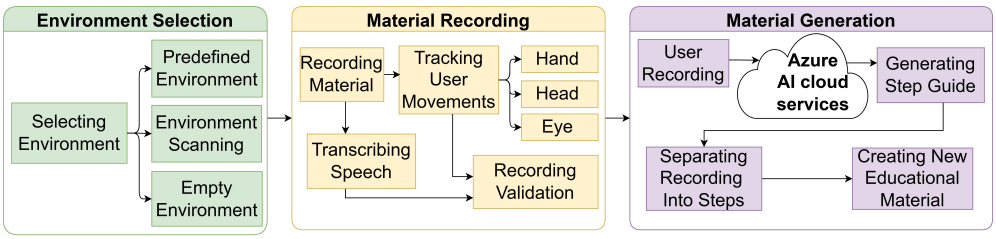


Fig. 2: System flow diagram

established, either by loading a virtual scene or by scanning the real workspace, it can be repositioned and aligned with the reference frame of the tutorial to improve spatial coherence.

Following the workflow presented in the previous section, the expert's demonstration recording module implements multimodal data acquisition by synchronizing speech, hand tracking, eye tracking, and head pose information throughout the recording session. Speech is transcribed in real time using Azure AI Cognitive Services [25]. Hand motion is captured through the MRTK Hands Aggregator Subsystem, which provides the positions of 26 joints for each hand at regular intervals. Eye gaze is recorded using a custom subsystem built on the MRTK Gaze Interactor, which determines the intersection of the user's gaze ray with objects in the scene. Head pose information is obtained directly from the HoloLens 2 tracking system, including both position and orientation. All recorded modalities are timestamped using a common time reference, enabling accurate temporal alignment during post-processing. This synchronization allows the AI pipeline to associate specific movement segments with the corresponding procedural steps extracted from the transcribed verbal instructions.

Upon completing the recording, the expert can review and validate the captured actions, and also has the opportunity to refine the transcribed narration. To support the generation of pedagogically effective instructions, we integrate Azure OpenAI's chat completion service. By feeding the time-aligned transcription into a prompt-engineered language model, we obtain a structured and comprehensible step-by-step guide, suitable even for novice trainees. The recorded user demonstration is then segmented according to these instructional steps by aligning time-stamps, enabling a progressive and synchronized learning experience. In the prompt, we explicitly informed the language model that the input text may contain minor errors, as it was generated through speech dictation. The model was instructed to correct these issues during processing. The full version of the prompt is provided in Appendix.

Once finalized, the instructional material is uploaded and stored in the *Data Provider* unit via a secure web-based API. This API also serves as an interface for authorized users, who can subsequently access a web-based editing platform to modify individual sub-tasks and their associated instructional text, ensuring the material remains up to date and adaptable.

The virtual reality and mixed reality trainer modules follow the same general tutorial logic, but differ in their implementation and in the type of guidance they provide. In the MR trainer, the tutorial is executed in the physical workspace. Therefore, the



Fig. 3: Playback of the expert's recorded movements in VR

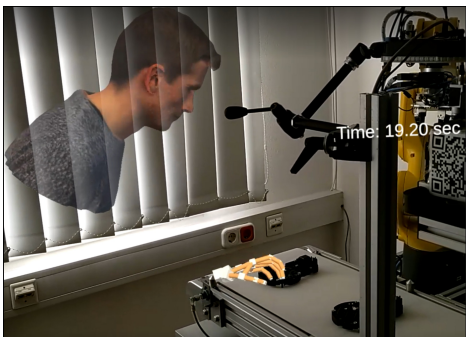


Fig. 4: Playback of the expert's recorded movements in MR

system assumes that all required tools, parts, and objects are placed in their initial positions before the tutorial begins. Since the trainee interacts with real objects, additional virtual representations of these objects are not required, and the tutorial can guide the learner through the procedure using instructions and expert replay. In contrast, the VR trainer operates in a fully virtual environment, where the relevant objects must be modeled and prepared in advance. This enables a more explicit form of spatial guidance, including object highlighting to help trainees identify the objects they need to use, and target highlighting to indicate where an object should be moved or placed during manipulation tasks. In addition, the VR implementation requires auxiliary interaction scripts, such as grabbable object behavior and snapping logic, to support object manipulation and placement validation within the virtual scene.

Trainees interact with the platform through the immersive trainer modules. From a catalog of available tutorials, users can select the relevant training material and initiate the associated training session. The *Instruction Player* sub-component presents the generated tutorial as a step-by-step immersive guide. For each subtask, it can display the corresponding textual or spoken instruction together with the time-aligned replay of the expert's recorded motion. This replay is visualized through a 3D avatar, providing a reference model that helps the trainee understand not only what action should be performed, but also how it should be executed. The same replay concept can be used in both fully virtual and mixed reality environments, as illustrated by the VR and MR playback examples in Figures 3 and 4, respectively.

Users may then select their preferred level of guidance from three modes: *guided*, *assisted*, and *unassisted*. The guided mode offers comprehensive support, including task descriptions, virtual cues, and expert replay. The assisted mode provides selective help upon request, while the unassisted mode minimizes intervention and allows users to perform the procedure with limited guidance.

The training session is orchestrated by a dedicated component known as the *Training Controller*. This controller manages the progression between tutorial steps, synchronizes the displayed instruction with the corresponding segment of the expert replay, and provides the interface through which different guidance levels can be selected. The architecture also allows future extensions in which trainee motion could

be continuously compared with the recorded expert trajectory to provide automatic feedback or adaptive assistance during task execution.

A possible future extension of the MR trainer is the integration of object-level and target-based highlighting directly in the physical workspace. However, this would require dynamic scene understanding, object recognition, and the generation of correctly positioned 3D highlights in real time. These operations are computationally demanding for standalone MR headsets, and could therefore benefit from edge/cloud resources in future versions of the system.

5 Measurement Methodology

The evaluation was designed as a two-stage study. First, we investigated the training-support capabilities of the proposed concept in a controlled VR environment, with a particular focus on the usefulness of expert replay during procedural assembly tasks. Second, we evaluated the mixed reality implementation of the system, where the emphasis shifted from tutorial execution support to rapid training material generation and the usability of the complete AI-assisted authoring workflow.

Across both evaluation phases, we used standardized questionnaires to assess the perceived quality of the system. General usability was measured using the System Usability Scale (SUS) questionnaire [26]. Broader user experience was evaluated using the User Experience Questionnaire (UEQ) [27], which assesses six dimensions: attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. These instruments allowed us to evaluate not only whether the system was considered usable, but also whether the overall training experience was perceived as understandable, efficient, predictable, engaging, and innovative.

5.1 Virtual Reality Evaluation Protocol

The first evaluation phase assessed the virtual reality implementation of the proposed framework. The study was conducted on the VIROO platform using an industrial six-cylinder engine assembly scenario. The overall assembly process was divided into several tutorial modules representing coherent procedural units, including piston installation, camshaft carrier assembly, turbo preparation, and turbo installation. These subtasks were selected because they require procedural knowledge, accurate object manipulation, spatial positioning, and the correct sequencing of assembly operations. Each tutorial consisted of approximately 20 instructional steps.

Two versions of the tutorial were evaluated. The baseline condition provided conventional instructional guidance through visual highlights, textual and spoken instructions, contextual cues, and execution feedback. In the experimental condition, these guidance elements were complemented by replaying an expert's recorded movements through a virtual avatar. The purpose of the expert replay was to illustrate not only which object should be manipulated, but also how each action should be performed, including hand movements, body posture, object orientation, and spatial placement.

The VR evaluation involved 19 participants, who completed the tutorials using Meta Quest 2 headsets within the VIROO-based virtual environment. Since age was

collected using predefined intervals, demographic information is reported by age group rather than mean age. Five participants belonged to the 18–24 age group, five to the 25–29 group, three to the 30–34 group, two to the 35–44 group, and four to the 45–54 group.

To obtain objective performance measures, logging scripts were integrated into the VR application. These automatically recorded the completion time of each tutorial module as well as the total execution time, enabling a direct comparison between the baseline and expert replay conditions.

Subjective user experience was assessed using the System Usability Scale (SUS) and the User Experience Questionnaire (UEQ), described in the previous section. In addition, a custom questionnaire was designed to evaluate the perceived usefulness of the expert replay functionality. The questionnaire investigated whether the avatar-based demonstration improved participants' understanding of the required actions, clarified the execution technique, increased confidence during task execution, facilitated comprehension of the procedural sequence, and would be preferred in future training scenarios. Participants were also invited to provide qualitative feedback regarding situations in which expert replay was particularly beneficial or unnecessary.

The combination of objective performance measures and subjective user feedback enabled a comprehensive evaluation of the contribution of expert replay to immersive procedural training.

5.2 Mixed Reality Evaluation Protocol

The second evaluation phase investigated the mixed reality implementation of the proposed framework, with particular emphasis on the complete AI-assisted tutorial generation workflow. Whereas the VR evaluation focused on the learner-facing training experience and the contribution of expert replay, the MR evaluation examined both the effectiveness of the generated tutorials and the usability of the authoring process itself.

To support this evaluation, we developed a simulated aircraft maintenance scenario. Participants first completed a maintenance tutorial consisting of 27 procedural steps using the *Mixed Reality Trainer*. The tutorial included inspection tasks such as verifying the movement of the flap and aileron, and examining the landing gear, brakes, and cable connections. After completing the learning session, participants used the *XR Training Material Generator* to create a simplified maintenance tutorial containing ten key procedural steps. This two-phase workflow allowed the complete authoring pipeline and the learner-facing training application to be evaluated within a single experiment. The MR evaluation involved 11 participants with a mean age of 33.45 years ($SD = 7.31$ years).

Several complementary evaluation methods were employed. The usability of both the *XR Training Material Generator* and the *Mixed Reality Trainer* was assessed using the System Usability Scale (SUS), while overall user experience was measured with the User Experience Questionnaire (UEQ).

To evaluate learning effectiveness, participants completed a knowledge assessment immediately before the training session and again 24 hours after training. The difference between the pre- and post-training results was used to quantify knowledge acquisition and retention.

The reliability of the AI-assisted authoring workflow was evaluated by monitoring errors during tutorial generation. In particular, we analyzed the accuracy of task segmentation and the error rate of the speech-to-text module, providing quantitative measures of the robustness of the automated content generation pipeline.

The SUS and UEQ results obtained during the VR and MR evaluations are reported independently and are not intended for direct comparison. Although both studies evaluated components of the proposed framework, they addressed different research questions. The VR evaluation concentrated on the learner-facing tutorial experience and the impact of expert replay in an engine assembly task, whereas the MR evaluation additionally assessed the AI-assisted tutorial authoring workflow. Consequently, the subjective questionnaire results reflect different aspects of the system in the two experiments.

6 Evaluation

The evaluation was organized into two complementary parts. First, we evaluated the VR implementation of the tutorial concept, focusing on the learner-facing training experience and the added value of expert replay during procedural assembly tasks. In this phase, participants completed VR tutorials with and without expert replay, while the system automatically logged subtask and total execution times. After the training session, participants evaluated the VR tutorial through usability and user experience questionnaires, as well as a custom questionnaire specifically targeting the expert replay functionality.

The second part of the evaluation focused on the MR implementation and the complete AI-assisted tutorial generation workflow. In this phase, participants first completed a knowledge assessment, then performed a 27-step tutorial created by the authors using the system, and finally created their own step-by-step guide in order to explore the main functionalities of the training material generator. Afterwards, they evaluated the system's usability and user experience, and completed the knowledge assessment again on the following day. This setup allowed us to assess not only the learner-facing immersive training experience, but also the usability and reliability of the AI-assisted authoring workflow.

6.1 Effect of expert replay on tutorial execution in VR

The questionnaire results indicate that the expert replay feature provided a meaningful additional layer of guidance during the VR tutorial. In the expert replay condition, conventional virtual cues, such as highlights, holograms, voice instructions, and textual descriptions, were complemented by the replay of an expert's recorded movements through an avatar. This allowed participants not only to identify the target objects and locations, but also to observe how the task should be performed in practice.

Table 1 summarizes the participants' preferences in the final comparative questionnaire. The results show that the expert replay version was preferred in several aspects. Most notably, the correct execution technique was perceived as clearer with expert replay by the majority of participants. In addition, more participants reported higher confidence and a stronger preference for learning with the expert replay version in the future.

Table 1: Summary of participant preferences in the final comparison between the standard tutorial and the expert replay tutorial.

Question	Expert replay	Standard	No preference
Overall preference	10	4	5
Higher confidence	11	6	2
Clearer correct technique	16	2	1
Easier to understand subtask order	6	5	8
Preferred for future learning	12	5	2

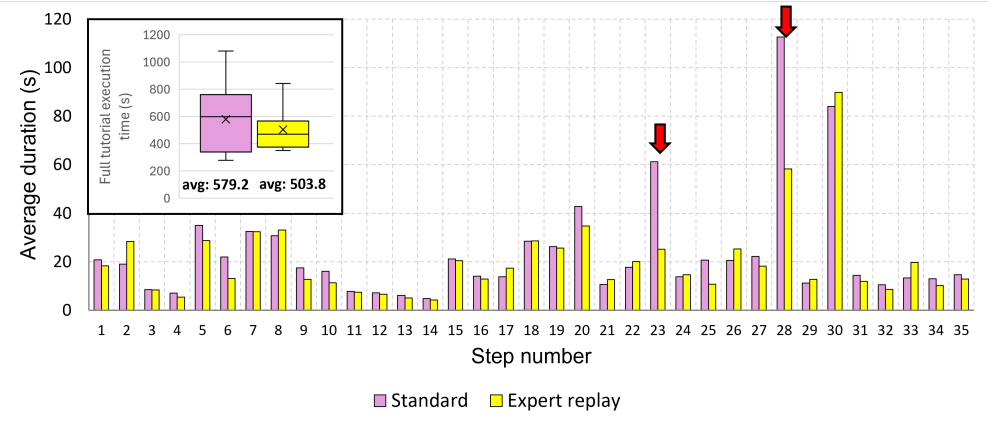


Fig. 5: Comparison of tutorial step execution time

The qualitative feedback also shows that expert replay was not equally useful in every situation. Some participants reported that the avatar could become unnecessary or distracting when the task was already clear, when similar steps were repeated, or when the avatar occluded the working area. This was especially relevant during repeated subtasks, such as inserting similar components or tightening screws, where users had already understood the required action.

Based on these findings, expert replay appears to be most valuable as an adaptive or optional guidance mechanism rather than as a continuously enforced element. Several participants suggested that the replay should either start automatically only during the first demonstration, be manually available on demand, or be triggered when the

system detects user hesitation or repeated failed attempts. Therefore, a promising design direction is to combine automatic expert replay for unfamiliar or complex steps with manual replay controls for later use.

In summary, the questionnaire results suggest that expert replay enhances the tutorial mainly by improving the clarity of correct technique and spatial positioning. It helps users understand not only what should be done, but also how the action should be executed. However, its usefulness depends on task complexity and user experience.

In addition to the subjective feedback, the execution time measurements also indicate a practical benefit of expert replay. As shown in Figure 5, the average tutorial execution time decreased from 579.2 s in the standard tutorial condition to 503.8 s when expert replay was available. This corresponds to a reduction of 75.4 s, or approximately 13.0%. The median execution time also decreased from 598.5 s to 470.0 s, suggesting that the avatar-based expert demonstration helped participants complete the procedure more efficiently. This improvement is consistent with the questionnaire responses, where participants reported that expert replay boosted confidence, decreased mental load and helped in understanding what to do.

In addition to the total tutorial completion time, Figure 5 also presents the average execution time of the individual subtasks in the standard and expert replay conditions. The results show that expert replay generally supported faster task execution across the tutorial steps. However, the most notable differences can be observed in the highlighted cases marked with red arrows. These subtasks were less intuitive for participants, even with the standard VR guidance elements such as object and target highlights. In these cases, users required considerably more time to understand and complete the required action in the standard tutorial condition. When expert replay was available, the avatar demonstration provided substantial additional support by showing the correct movement, object orientation, and placement procedure. As a result, the average execution time of these subtasks was reduced by approximately half, indicating that expert replay is especially beneficial for complex or ambiguous manipulation steps.

6.2 VR user experience and system usability

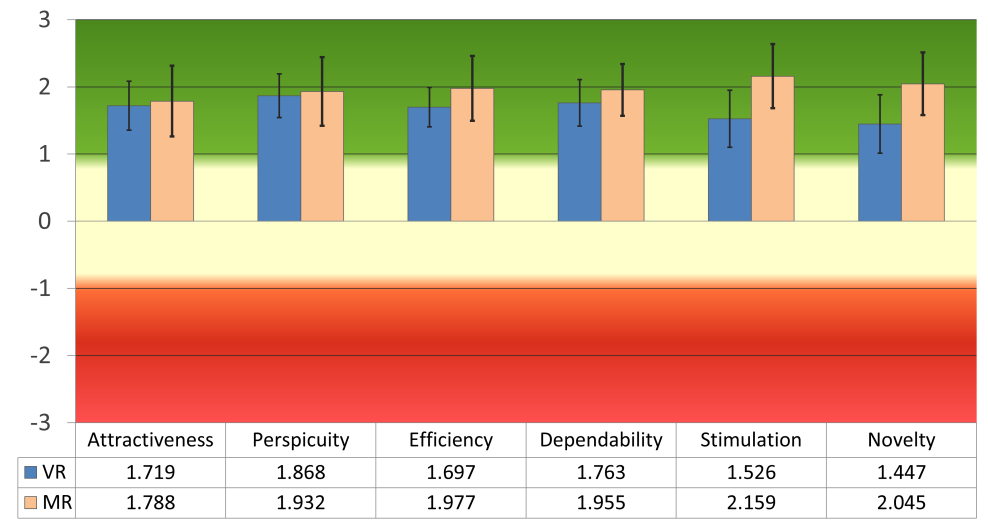
The usability of the VR tutorial system was assessed using the System Usability Scale (SUS). As shown in Figure 7, the VR system achieved a mean SUS score of 81.84 (SD = 8.89), with individual scores ranging from 65 to 95 and a median value of 85. According to common SUS interpretation guidelines, scores above 68 are considered above average, while scores above 80 indicate excellent perceived usability. Therefore, the results suggest that participants found the VR tutorial system highly usable.

The distribution of the scores further supports this interpretation. Most participants rated the system between 75 and 95 points, indicating consistently positive usability feedback across the sample. Although two participants gave lower scores of 65, the overall tendency remained clearly positive. These results show that the VR tutorial implementation, including the guided step-by-step structure and expert replay functionality, was generally perceived as easy to use and suitable for immersive procedural training.

Figure 6 presents the User Experience Questionnaire results of the VR tutorial evaluation. The results show that all six UEQ dimensions received positive ratings, with mean values clearly above the positive threshold of 0.8. This indicates that participants perceived the VR tutorial system as understandable, efficient, reliable, and generally pleasant to use.

Among the evaluated dimensions, *Perspicuity*, *Dependability*, *Attractiveness*, and *Efficiency* achieved the highest scores, suggesting that users found the tutorial easy to understand, predictable, and effective for completing the assembly tasks. The *Stimulation* and *Novelty* dimensions received slightly lower values compared to the other scales, but they also remained within the positive range. This suggests that while the VR tutorial was perceived as engaging and sufficiently innovative, its strongest advantage was the clarity and usability of the step-by-step guidance rather than novelty alone.

The confidence intervals also remain within the positive region for all UEQ dimensions, which indicates a consistent overall user experience across participants. These results support the applicability of the VR tutorial implementation as a usable and understandable immersive training environment for procedural assembly tasks.



	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
VR	1.719	1.868	1.697	1.763	1.526	1.447
MR	1.788	1.932	1.977	1.955	2.159	2.045

Fig. 6: Mean scores and confidence intervals of the UEQ Measurements

6.3 Mixed reality user experience and system usability

Figure 6 presents the UEQ results for both the VR and MR evaluations. As described in the measurement methodology, the same standardized questionnaire was used in both phases. The MR implementation achieved positive results across all UEQ scales and showed slightly higher mean values than the VR implementation. This difference should not be interpreted as a direct comparison between two equivalent versions of

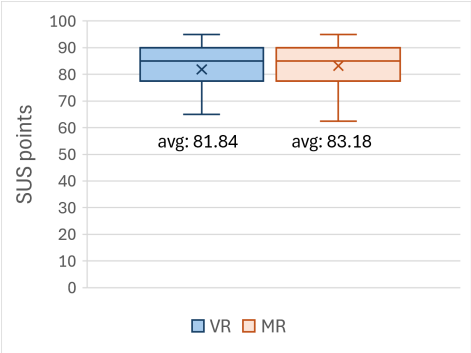


Fig. 7: System usability scale points

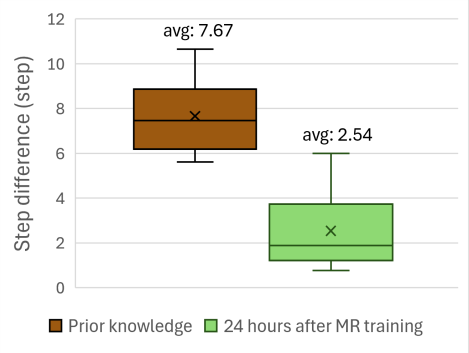


Fig. 8: Knowledge test: Position deviation scoring

the same task, since the two evaluations differed in both scenario and evaluated functionality. The VR test focused on the learner-facing tutorial experience and the effect of expert replay in an engine assembly scenario, while the MR test also included the tutorial generation process and evaluated the complete AI-assisted authoring workflow.

With this limitation in mind, the MR results suggest that participants perceived the mixed reality workflow particularly positively. The difference was especially visible for *Stimulation* and *Novelty*, where the MR scores increased from 1.526 to 2.159 and from 1.447 to 2.045, respectively. This may be explained by the combination of real physical objects, contextual digital guidance, expert replay, and the possibility of creating new tutorial content directly in the mixed reality environment.

The usability of the two implementations was assessed using the System Usability Scale (SUS), as shown in Figure 7. SUS scores are calculated on a 0–100 scale, where higher values indicate better perceived usability. Both implementations achieved high SUS scores, with a mean score of 81.84 for VR and 83.18 for MR. The median value was 85.00 in both cases, indicating that participants generally rated both systems as highly usable.

According to standard SUS interpretation guidelines, scores above 68 are considered above average, while scores above 80 indicate excellent usability. Therefore, both the VR and MR implementations can be interpreted as having excellent perceived usability. The MR implementation achieved a slightly higher mean score, but the distributions are largely comparable. This suggests that the proposed tutorial concept remained usable across both immersive environments, while the MR version provided a modest additional improvement in perceived usability and a more pronounced improvement in user experience according to the UEQ results.

6.4 Learning progress

To evaluate the learning effect of the MR-based training session, participants completed a knowledge assessment before the tutorial and again 24 hours after the training. The delayed post-test was used instead of an immediate post-test in order to obtain

an indication of short-term retention rather than only measuring the immediate recall of the presented instructions.

The assessment focused on the correct order of the maintenance procedure. Participants were asked to arrange the individual steps of the aircraft maintenance workflow into the correct sequence. Instead of evaluating only the number of exactly correct answers, we applied a position deviation scoring method. This metric measures how far each step was placed from its correct position in the reference sequence. Therefore, a lower score indicates a better understanding of the procedural order, while a higher score indicates larger deviations from the correct sequence.

This scoring method was selected because procedural knowledge is not only reflected by whether a participant remembers a specific step, but also by whether they understand where that step belongs in the overall workflow. For example, placing a step only one position away from its correct location represents a smaller error than moving it to a completely different part of the procedure. The position deviation score therefore provides a more fine-grained measure of learning progress than a simple correct/incorrect evaluation.

As shown in Figure 8, the average position deviation decreased after the MR training session. This indicates that participants were able to reconstruct the correct procedural order more accurately after using the immersive tutorial. The result suggests that the MR-based training experience supported the acquisition and retention of the aircraft maintenance procedure, at least in terms of understanding the sequence of the required subtasks.

6.5 Step generation errors

Participants were asked to generate their own step-by-step tutorial using spoken input. This task enabled us to examine the number of transcription errors produced by the speech-to-text (STT) system, how many of these errors could be corrected by a large language model (LLM), and how many erroneous steps the LLM itself generated. The prompt explicitly informed the LLM that minor errors might be present in the input text.

During the evaluation, it was observed that the performance of the STT system varied across participants. While it operated flawlessly for the majority —introducing no errors— some participants experienced a noticeable decline in accuracy, with 4 to 5 errors detected in their transcripts. On average, the STT system introduced 1.36 errors during the short instructional input (typically comprising 15–20 sentences). After LLM-based correction, the average error count decreased to 0.18. The LLM effectively corrected minor grammatical or pluralization errors, but it was unable to resolve cases where the STT had misrecognized a word entirely.

In the tutorial generation task, which typically consisted of ten steps, the LLM added or removed a step in 36% of cases. These alterations were not disruptive, they typically involved the merging or splitting of logically connected steps rather than introducing or omitting core content.

7 Conclusion and Outlook

This paper presented an AI-assisted tutorial generation framework for immersive procedural training in virtual and mixed reality environments. The proposed approach is based on a single-demonstration authoring concept, where an expert performs and explains a task once, while the system captures speech, hand movements, head pose, and gaze information. Speech-to-text and large language model components are then used to transform the recorded demonstration into a structured step-by-step tutorial. In addition to textual and voice-based instructions, the generated material preserves the expert's recorded motion as a time-aligned 3D avatar replay, allowing trainees to observe how each subtask should be executed.

The framework was evaluated in two complementary scenarios. In the VR evaluation, we investigated the learner-facing tutorial experience and the added value of expert replay during an industrial assembly task. The results show that expert replay was particularly useful for clarifying correct execution techniques, spatial positioning, object orientation, and required movements. This was also reflected in the objective timing measurements, where the average tutorial execution time decreased from 579.2 s to 503.8 s when expert replay was available. The questionnaire responses further indicated that participants preferred expert replay especially for complex or unfamiliar steps, while also suggesting that replay should be optional or adaptive in simpler and repetitive tasks.

The MR evaluation focused on the complete AI-assisted tutorial generation workflow, including content creation, immersive playback, usability, user experience, learning progress, and generation accuracy. The results indicate that both VR and MR implementations achieved positive user experience and high perceived usability. The MR evaluation also showed that participants were able to use the system not only for following a generated tutorial, but also for creating new instructional material based on an expert demonstration. The learning assessment showed improvement after MR-based training, while the analysis of speech-to-text and LLM-based step generation errors demonstrated that the proposed workflow can produce usable tutorial content with limited manual correction.

Overall, the results support the feasibility of using expert demonstrations and AI-assisted processing to create reusable immersive tutorials. The proposed framework addresses two important challenges in immersive training. First, it supports trainees during procedural task execution by combining step-by-step instructions with spatial guidance and time-aligned expert replay. Second, it reduces the effort required to create new training materials by shifting the authoring process from manual tutorial development toward demonstration-based content generation.

The results also highlight the complementary roles of VR and MR in immersive procedural training. VR provides a controlled, safe, and repeatable environment for practicing procedures and evaluating tutorial guidance. However, VR deployment requires a higher initial preparation effort, because the training environment, tools, and task-relevant objects must be modeled in advance. In addition, virtual objects typically need auxiliary interaction scripts, such as grabbable behavior, snapping logic, placement validation, and other task-specific interaction mechanisms. Therefore, the same level of authoring-time reduction is more difficult to achieve in VR when the virtual

environment has to be created from scratch. Nevertheless, once a virtual representation of a factory, workstation, or training scenario is available, the proposed framework can reuse this environment and enable experts to generate additional tutorials with substantially less manual effort.

In contrast, MR allows tutorial authoring and training to take place directly in the physical workspace or in a dedicated test room where the real tools and components are already present. This reduces the need to recreate the complete environment virtually and makes MR especially suitable for procedures where physical object manipulation and workplace-specific context are important. At the same time, the current MR implementation assumes that the required tools and components are placed in their initial positions before the tutorial begins. Unlike the VR implementation, it does not yet provide object-level or target-based highlighting in the physical environment, because this would require dynamic scene understanding, object recognition, and the real-time generation of correctly positioned 3D highlights.

Future work will focus on extending the system with adaptive guidance and intelligent monitoring capabilities. In particular, we plan to investigate how trainee movements can be compared with the recorded expert trajectory in real time to detect hesitation, execution errors, or potentially unsafe actions. This would enable the system to trigger expert replay only when needed, provide context-sensitive feedback, and further improve the safety, efficiency, and scalability of immersive procedural training. Another important future direction is the integration of object-level and target-based guidance in MR. Since real-time scene understanding and spatial annotation can be computationally demanding for standalone MR headsets, future versions of the system could benefit from edge or cloud resources to support more advanced environment interpretation and adaptive guidance generation.

Funding

This project has received funding from *i*) the CHIPS Joint Undertaking as part of the European Union's Horizon Europe research and innovation programme, SMARTY Project (grant agreement No. 101140087), and *ii*) project No. 2024-1.2.4-KDT-2024-00002 supported by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, under the 2024-1.2.4-KDT funding scheme. This work was also supported by *iii*) the AIXRIG Project, which has received funding from the European Union's Horizon Europe research and innovation action programme via the Open Call 2 – MASTER issued and executed under the MASTER-XR Project, grant agreement No. 101093079. B. Sonkoly was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Compliance with Ethical Standards

- **Conflict of Interest**: The authors declare that they have no conflict of interest.
- **Informed Consent:** Informed consent was obtained from all participants prior to their participation in the study. Throughout the entire research process, the principle of protecting participants' privacy was strictly observed, and all personal information was anonymized.

References

[1] Bihari, B., Nagy, B.G., Dóka, J., Sonkoly, B.: Ai-powered mixed reality: Revolutionizing training methodologies. In: 2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), pp. 634–635 (2024). https://doi.org/10.1109/ISMAR-Adjunct64951.2024.00186

[2] Nagy, B.G., Bihari, B., Sonkoly, B.: Ai-driven mixed reality for scalable and adaptive training solutions. In: Nakamatsu, K., Kountcheva, R., Patnaik, S. (eds.) Virtual Reality and Visualization Based on AI Technologies, pp. 279–291. Springer, Cham (2026)

[3] Virtualware: VIROO: The Enterprise XR Platform. https://viroo.virtualwareco.com/. Accessed: 2026-06-10 (2026)

[4] Tran, K.-N., Lau, J.H., Contractor, D., Gupta, U., Sengupta, B., Butler, C.J., Mohania, M.: Document chunking and learning objective generation for instruction design. arXiv preprint arXiv:1806.01351 (2018)

[5] Yaseen, H., Mohammad, A.S., Ashal, N., Abusaimeh, H., Ali, A., Sharabati, A.-A.A.: The impact of adaptive learning technologies, personalized feedback, and interactive ai tools on student engagement: The moderating role of digital literacy. Sustainability **17**(3), 1133 (2025)

[6] Sridhar, P., Doyle, A., Agarwal, A., Bogart, C., Savelka, J., Sakr, M.: Harnessing LLMs in Curricular Design: Using GPT-4 to Support Authoring of Learning Objectives (2023). https://arxiv.org/abs/2306.17459

[7] Yadav, G.: Scaling Evidence-based Instructional Design Expertise through Large Language Models (2023). https://arxiv.org/abs/2306.01006

[8] Petruse, R.E., Grecu, V., Gakić, M., Gutierrez, J.M., Mara, D.: Exploring the efficacy of mixed reality versus traditional methods in higher education: A comparative study. Applied Sciences **14**(3), 1050 (2024)

[9] Zhao, G., Fan, M., Yuan, Y., Zhao, F., Huang, H.: The comparison of teaching efficiency between virtual reality and traditional education in medical education: a systematic review and meta-analysis. Annals of translational medicine **9**(3), 252 (2021)

[10] Guha, P., Lawson, J., Minty, I., Kinross, J., Martin, G.: Can mixed reality technologies teach surgical skills better than traditional methods? a prospective randomised feasibility study. BMC Medical Education **23**(1), 144 (2023)

[11] Daling, L.M., Schlittmeier, S.J.: Effects of augmented reality-, virtual reality-, and mixed reality–based training on objective performance measures and subjective evaluations in manual assembly tasks: a scoping review. Human factors **66**(2),

589–626 (2024)

[12] Haj-Bolouri, A., Katende, J., Rossi, M.: Gamified immersive safety training in virtual reality: a mixed methods approach. Journal of Workplace Learning (2024)

[13] Pedram, S., Ogie, R., Palmisano, S., Farrelly, M., Perez, P.: Cost–benefit analysis of virtual reality-based training for emergency rescue workers: a socio-technical systems approach. Virtual Reality **25**(4), 1071–1086 (2021)

[14] Xie, B., Liu, H., Alghofaili, R., Zhang, Y., Jiang, Y., Lobo, F.D., Li, C., Li, W., Huang, H., Akdere, M., *et al.*: A review on virtual reality skill training applications. Frontiers in Virtual Reality **2**, 645153 (2021)

[15] ARVRTech: Immersive4Learning. https://arvrtech.eu/immersive-4-learning/ (2024)

[16] Education, A.: ClassVR. https://www.classvr.com/ (2024)

[17] Kuo, T.-L.: Humanoid avatar motion capture recording system in virtual lab for educational tutorials. In: 2017 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC), pp. 193–199 (2017). https://doi.org/10.23919/PNC.2017.8203543

[18] Wang, C.Y., Drumm, L., Troup, C., Ding, Y., Stevenson Won, A.: Vr-replay: Capturing and replaying avatars in vr for asynchronous 3d collaborative design. In: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 1215–1216 (2019). https://doi.org/10.1109/VR.2019.8797789

[19] Chidambaram, S., Reddy, S.S., Rumple, M., Ipsita, A., Villanueva, A., Redick, T., Stuerzlinger, W., Ramani, K.: Editar: A digital twin authoring environment for creation of ar/vr and video instructions from a single demonstration. In: 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 326–335 (2022). https://doi.org/10.1109/ISMAR55827.2022.00048

[20] Ihara, K., Monteiro, K., Faridan, M., Kazi, R.H., Suzuki, R.: Video2mr: Automatically generating mixed reality 3d instructions by augmenting extracted motion from 2d videos. In: Proceedings of the 30th International Conference on Intelligent User Interfaces, pp. 1548–1563 (2025)

[21] Caetano, A., Aponte, A., Sra, M.: An interaction design toolkit for physical task guidance with artificial intelligence and mixed reality. arXiv preprint arXiv:2412.16892 (2024)

[22] Inc., A.: Altoura: AI-Powered Learning Platform. Accessed: 2025-04-24 (2025). https://www.altoura.com/

[23] Ltd., V.: VirtualSpeech: Soft Skills Training in VR with AI Feedback. Accessed: 2025-04-24 (2025). https://virtualspeech.com/

[24] Microsoft: Mixed Reality Toolkit 3. https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk3-overview/ (2024)

[25] Microsoft: Azure AI Services. https://azure.microsoft.com/en-us/products/ai-services (2024)

[26] Brooke, J., *et al.*: Sus-a quick and dirty usability scale. Usability evaluation in industry **189**(194), 4–7 (1996)

[27] Schrepp, M.: User experience questionnaire handbook. All you need to know to apply the UEQ successfully in your project **10** (2015)

Appendix: The prompt used for material generation

"You are tasked with converting speech descriptions provided by a work expert into a detailed technical work step guide. The input will consist of step descriptions along with start time labels in the format [ss.SS] (seconds and hundredths of a second). Your objectives are as follows: Step-by-Step Guide Creation: Transform the speech descriptions into a clear and concise step-by-step guide, the speech may also contain speech detection errors, you need to correct so it will be semantically correct. Each step should be numbered sequentially, exactly in this format: Step n: Step description [timestamp]. Time Labels: Include start time label at the end of each step formatted exactly in [ss.SS] to facilitate the slicing of a demonstration video. Ensure the time labels accurately reflect the timing provided in the input. Technical Style: Write in a formal, technical style suitable for a professional audience, ensuring clarity and precision in the language used. If there is no user input, ask for another input, but do not include extra text beside asking for the step guide."