# AI-Driven Mixed Reality for Scalable and Adaptive Training Solutions

Bálint György Nagy[1,2][0000−0001−9917−952X], Bence Bihari[1,3][0009−0002−2286−9471] and Balázs Sonkoly[1,2][0000−0002−4640−388X]

[1] HSN Lab, 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] evopro systems engineering Ltd., Hungary

**Abstract.** Training in industrial environments is often expensive and time-consuming. While self-learning approaches, customized curricula, and automation offer promising solutions, the creation of new training materials remains a major bottleneck—especially in rapidly evolving industries where processes change frequently. Mixed reality (MR) and virtual reality (VR) technologies introduce a new generation of immersive learning platforms, but their widespread adoption is limited by the expertise and effort required to generate content.
In this paper, we present a novel AI-powered system designed to simplify and accelerate the development of MR-based training applications. Using a HoloLens 2 headset, an expert needs to demonstrate and explain a complex task only once. Our system captures this demonstration, interprets the instructions using various AI techniques, and automatically produces a step-by-step tutorial. The resulting guide includes time-aligned video snippets and descriptive text, allowing trainees to visually follow each subtask. This approach significantly reduces the cost and complexity of training content creation, opening the door to scalable and efficient instructorless learning. The system was evaluated in terms of usability, user experience, and the accuracy of the generated content, the results of which confirm the correct functioning and applicability of our approach.

**Keywords:** Human-centered computing · Human computer interaction (HCI) · Natural Language Processing (NLP) · Generative artificial intelligence · AI-assisted tutorial generation · Mixed reality training systems

## 1 Introduction

Training and education are crucial and challenging tasks for companies and schools. It has been shown that, in many cases, using virtual reality (VR) or mixed reality (MR) technologies can be more effective than traditional teaching methods. These technologies allow learners to engage in interactive and immersive learning experiences, which helps them better understand and retain information through active participation in the learning process than teaching with

traditional PowerPoint presentations [14]. Furthermore, VR and MR technologies are particularly useful for simulating dangerous situations [7] in a training environment. This makes their use in industries like healthcare or the military much safer and more cost-effective [13]. Besides, these immersive technologies pave the way towards the fully instructorless, self-learning solutions making use of customized curriculum and automation in the teaching process. However, the cost of current MR headsets and the creation of new learning content pose additional challenges and require special expertise which could hinder the technology adaptation. The pricing issue of the devices will hopefully be resolved over time. But creating custom training software for specific processes will still be expensive, making it primarily feasible in industrial environments.

There are a few VR learning platforms on the market, such as Immersive4Learning [1], or ClassVR [5] but they do not support MR glasses, and creating teaching materials on these is a time-consuming task. Additionally, mixed reality is often more suitable for training. For instance, if there is a training room equipped with all the necessary tools and components for the task, the learning process will be more effective than if we were only assembling something virtually. Moreover, the guidance via MR is also crucial, especially in a trainerless environment.

We can conclude that creating a mixed reality training application is currently a very costly and time-consuming task, which is unaffordable in many areas, although the usability of the technology is unquestionable. We aim to develop a solution to this problem by leveraging current AI technologies, thereby making the benefits of mixed reality more accessible in education.

## 2 Related Work

Training and onboarding remain persistent challenges for both industry and educational institutions, primarily due to the inefficiencies of traditional instructional methods. These approaches often require extensive time and resources, with the development of a single hour of training material demanding between 50 and 300 hours of effort [17]. Recent studies have also explored the integration of artificial intelligence to enable personalized, adaptive learning experiences [20]. AI-enhanced platforms can monitor learner behavior in real time and adjust feedback or guidance dynamically, thereby reducing the need for human supervision. This combination has shown promise for scaling training delivery while maintaining effectiveness and learner satisfaction.

Recent advancements in generative AI (particularly large language models such as GPT-4) have demonstrated strong potential in automating aspects of instructional design. Sridhar et al.[16] investigated the effectiveness of GPT-4 in generating learning objectives for an introductory AI course. Their findings show that the model was capable of producing relevant and pedagogically aligned objectives, suggesting that such models can support structured curriculum design. Similarly, Yadav[19] proposed a framework for using LLMs to scale instructional design processes by generating lesson plans, questions, and feedback prompts.

While the use of generative models has primarily focused on web-based education, their application in immersive and spatial learning environments remains underexplored.

MR-based instruction offers immersive, interactive experiences that promote learner engagement and support better knowledge retention, particularly in procedural and hands-on training scenarios common in vocational, technical, and safety-critical domains. By enabling realistic and repeatable practice through contextualized virtual environments, MR systems can outperform conventional methods [21, 6, 4]. Although MR is already used in many application areas such as First responder training, medical training, military training and workforce training [18], its broader adoption remains constrained by the high time and cost requirements associated with developing tailored training programs. Reducing these barriers could unlock the potential for MR to be deployed more widely across both industrial and educational settings.

A key challenge in deploying MR-based training remains the creation of educational content, which traditionally requires significant time, domain knowledge, and technical expertise. Several efforts have emerged to address this bottleneck. For instance, Video2MR [8] proposes a method to automatically convert 2D instructional videos into 3D MR training experiences using AI-driven motion reconstruction. Similarly, toolkits such as the Interaction Design Toolkit for Physical Task Guidance [3] provide structured support for mapping real-world workflows into MR applications by leveraging gesture, gaze, and speech inputs.

Several VR-based learning platforms are available on the market, such as Immersive4Learning [1] and ClassVR [5]. However, these solutions typically lack support for mixed reality headsets, and the process of developing educational content for them can be both time-consuming and labor-intensive. Moreover, MR often proves to be a more appropriate medium for training scenarios that involve interaction with the physical environment. Commercial systems like Altoura [9] and VirtualSpeech [10] illustrate how AI and XR technologies can be used to build scalable training platforms for onboarding, soft skills, and procedural guidance. These platforms typically use speech recognition, behavior analysis, and predefined instructional flows to support users in immersive environments. However, most of these systems still rely heavily on manually authored content and scripted interactions, limiting their flexibility and ease of deployment.

Despite these advancements, few existing solutions offer a seamless authoring pipeline that automatically transforms expert demonstrations into structured MR training content. Our approach addresses this gap by introducing an AI-supported, semi-automated platform that records expert hand, head, and gaze movements along with verbal instructions, then processes this data into step-by-step immersive training modules. This contribution is focusing not only on learner-facing delivery mechanisms but also on simplifying and accelerating the authoring process itself.

## 3 Main ideas

After conducting a thorough analysis of the industry landscape and existing needs, we identified a significant gap in the availability of cost-effective, user-friendly, and rapid solutions tailored for training or educational purposes in mixed reality (MR) environments. Our research highlighted that, in many scenarios—particularly in smaller-scale or highly specialized workflows—it is not economically viable to commission custom training applications from development companies. However, these workflows could still greatly benefit from the immersive and interactive capabilities that mixed reality technology offers in educational contexts.

To address this need, our primary objective is to develop an AI-supported mixed reality training platform that enables instructors or domain experts to create training content in an intuitive and efficient manner. Specifically, the system is designed so that an instructor, wearing an MR headset, only needs to perform a complex workflow once. During this demonstration, our platform captures multimodal data, including hand and head movements as well as gaze direction, while simultaneously processing spoken instructions through integrated AI components.

This recorded data is then used to automatically generate detailed step-by-step training materials, which may include segmented video clips, textual descriptions, and spatial cues. These materials allow learners to visually and interactively follow the relevant steps of the workflow at their own pace.

In practice, the instructor's spoken instructions are first transcribed using speech-to-text technology. The resulting text is then analyzed by a large language model, which decomposes the overall task into manageable sub-tasks and organizes them into a structured learning module. When trainees engage with this module, they can explore each sub-task with the assistance of text-to-speech playback for instructions, visual highlights showing where the instructor was looking, and real-time 3D replays of the instructor's hand and head movements during the demonstration. This comprehensive, multimodal feedback greatly enhances the learning experience and effectiveness of MR-based training.

## 4 Our MR training platform

We have developed and implemented our Mixed Reality (MR) training platform using the Unity 3D Engine, building upon the capabilities provided by the Mixed Reality Toolkit 3 (MRTK3)[12]. As depicted in Figure 1, the system architecture is composed of three principal components: the *Training material generator* module, where domain experts in manufacturing processes are able to author instructional content, the *data provider* unit, which serves as the central repository for storing and managing these materials, and the *immersive trainer* module, which enables end users to engage in guided training experiences using the created content.

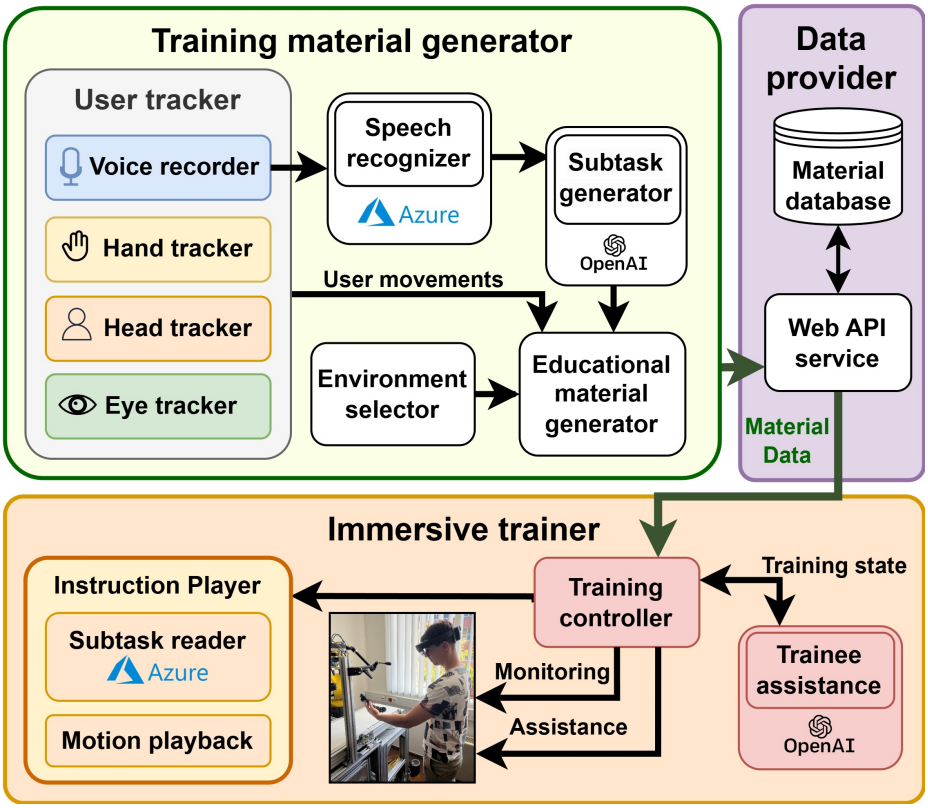


Fig. 1: Our training platform with artificial intelligence support

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. In cases where the procedure is independent of any specific environment, the system allows this step to be bypassed. To define the environment, users can either load a predefined virtual model or dynamically capture their physical surroundings by scanning the real-world space. The latter is accomplished via Unity's AR Mesh Manager, which enables spatial mapping of the user's physical environment. Once the environment is established, it can be repositioned to best align with the real-world reference frame, enhancing spatial coherence.

Following this, the application facilitates the recording of the expert's demonstration, a crucial phase in generating the tutorial. During this process, the system continuously tracks and records the user's hand, eye, and head movements. In parallel, it performs real-time speech-to-text conversion using Azure AI Cognitive Services [11]. Hand motion data is obtained through the MRTK Hands Aggregator Subsystem, capturing the positions of 26 joints for each hand at reg-

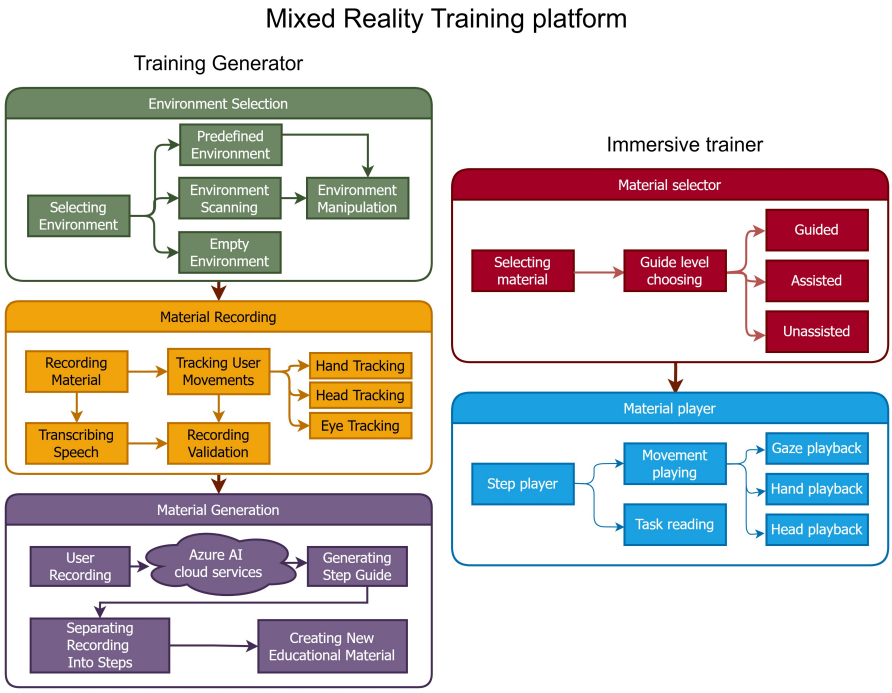


Fig. 2: System flow diagram

ular intervals. For eye tracking, we developed a dedicated subsystem utilizing the MRTK Gaze Interactor, which computes the intersection point of the gaze ray with scene objects. The head posture is derived from the MR headset's positional and rotational data. It is important to highlight that each collected data point is associated with a timestamp, so that during later post-processing, we can determine which segment of the movement corresponds to a given subtask.

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

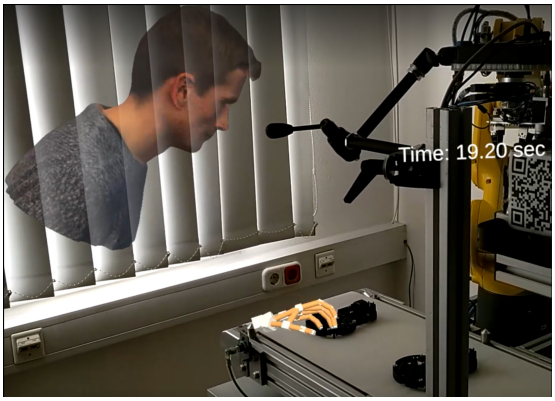


Fig. 3: Playback of the expert's recorded movements

to modify individual sub-tasks and their associated instructional text, ensuring the material remains up to date and adaptable.

Trainees interact with the platform through the immersive trainer module. From a catalog of available tutorials, users can select the relevant training material and initiate the associated training session. For first-time access, an *instruction player* sub-component replays the expert's demonstration to provide a reference model for the trainee as shown in Figure 3. 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 and virtual cues. The assisted mode provides selective help upon request, while the unassisted mode minimizes intervention, offering feedback only when the user deviates significantly from the expert trajectory.

The training session is orchestrated by a dedicated component known as the *training controller*. This controller continuously evaluates real-time data from the *trainee monitoring* subsystem, delivering immediate feedback and managing transitions between instructional steps. Once the user successfully completes a task, the next segment of the guide is automatically activated, ensuring a seamless and adaptive training experience.

## 5 Measurement Methodology

To test the system, we assembled a test group consisting of 11 participants with an average age of 33.45 years (SD = 7.31 years). During the evaluation, we separately assessed the usability of the *Training material generator* and the *Immersive trainer* components. Usability was measured using the System Usability Scale (SUS) questionnaire [2], a standardized and widely adopted tool for obtaining quick feedback on the general usability of a system. Additionally, the User Experience Questionnaire (UEQ)[15] was utilized to provide a more comprehen-

sive assessment of various user experience dimensions, such as attractiveness, efficiency, and dependability.

The educational effectiveness of the *Immersive trainer* was evaluated through a learning outcome measurement. Participants completed a knowledge test both before and again 24 hours after the training session, allowing us to quantitatively assess their learning progress.

During the use of the *Training material generator*, we monitored errors occurring during the content generation process. Special attention was given to the success rate of task segmentation, as well as the error rate of the speech-to-text module. Through these measurements, we aimed to gain a comprehensive understanding of the system's reliability and the accuracy of the content generation workflow.

In order to conduct the evaluation, we designed a simulated aircraft maintenance scenario. Participants were initially tasked with learning a comprehensive maintenance procedure consisting of 27 individual sub-tasks using the Immersive Trainer component, including actions such as checking the motion of the flap and aileron, and inspecting the landing gear, brakes, and cables. Following the completion of the associated pre- and post-training knowledge assessments, participants were instructed to use the Training Material Generator module to create a simplified procedure, reduced to 10 key steps. This dual-phase evaluation approach enabled us to systematically assess the learning effectiveness provided by the immersive training environment, as well as the usability, efficiency, and reliability of the automated training material generation process.

## 6 Evaluation

The participants first completed a knowledge assessment, then performed a 27-step tutorial created by us using the system, and finally created their own step-by-step guide to explore all the functionalities of the system. Afterwards, they evaluated the system's usability and user experience, and completed the knowledge assessment again the following day.

### 6.1 User experience and system usability

Figure 4 presents the results of the User Experience Questionnaire (UEQ), showing the mean scores and their corresponding 95% confidence intervals across six dimensions: *Attractiveness*, *Perspicuity*, *Efficiency*, *Dependability*, *Stimulation*, and *Novelty*. All measured scales received positive evaluations, with each mean score exceeding the neutral threshold of 0.8.

Among the dimensions, *Stimulation* and *Novelty* achieved the highest scores (both slightly above 2.0), suggesting that users perceived the system as engaging and innovative. The scales of *Efficiency*, *Dependability*, and *Perspicuity* also scored well, with mean values approaching or exceeding 2.0. These results indicate that participants found the system to be easy to understand, reliable, and effective in supporting task performance. *Attractiveness*, while scoring slightly

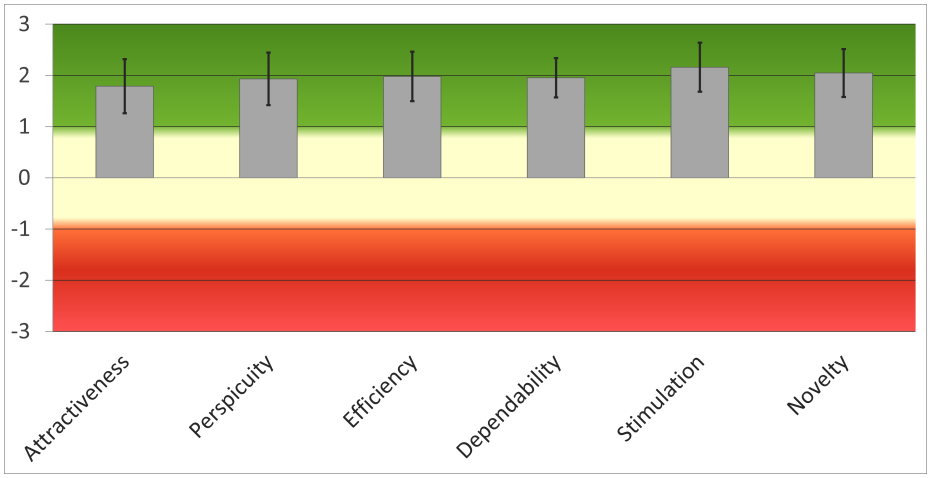


Fig. 4: Mean scores and confidence intervals of the UEQ Measurements

lower than the other dimensions, still achieved a clearly positive rating (1.79), implying that the overall impression of the system was favorable.

The confidence intervals for all dimensions remain well above the neutral evaluation range, and none overlap the negative zone. This consistency suggests a high degree of agreement among users regarding the system's positive qualities. The colored background reflects the UEQ benchmark interpretation: values above 0.8 are considered *positive*, those between $-0.8$ and 0.8 are *neutral*, and values below $-0.8$ are *negative*. Based on these results, the evaluated system demonstrates a strong and uniformly positive user experience across all assessed dimensions.

The usability of the system was assessed using the System Usability Scale (SUS), which consists of ten statements rated on a 5-point Likert scale. SUS scores are calculated by converting each participant's responses to a 0–100 scale, where higher scores indicate better perceived usability.

The calculated SUS scores are illustrated in Figure 5. They ranged from 62.5 to 95, with a mean score of 84.25 (SD = 9.77). According to standard SUS interpretation guidelines, scores above 68 are considered above average, while scores exceeding 80 reflect excellent usability. Thus, the system under evaluation was perceived as highly usable by the participants. The relatively low standard deviation suggests that user responses were consistent, indicating a uniformly positive usability experience across the sample.

### 6.2 Learning progress

To evaluate the effectiveness of the training, we administered a test to the participants immediately before they performed the training task and again 24 hours after completing it. The second test was intentionally scheduled 24 hours later

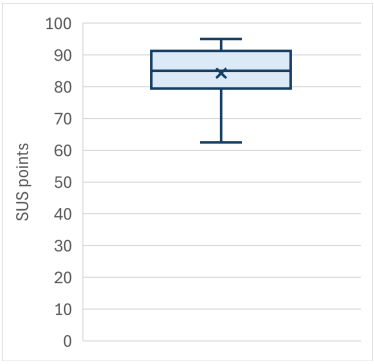


Fig. 5: System usability scale points


Fig. 6: Knowledge test: Position deviation scoring

rather than immediately after the training to assess information retention in the participants' long-term memory. The task involved arranging the steps of an aircraft maintenance procedure in the correct order. During evaluation, we did not primarily focus on the number of correct responses (although significant improvements were observed in this regard), but rather utilized a position deviation scoring method. This scoring approach measured the average deviation of the participants' answers from the correct sequence. The test results, illustrated in Figure 6 , clearly indicate substantial improvements resulting from the mixed reality training.

### 6.3 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 dis-

ruptive, they typically involved the merging or splitting of logically connected steps rather than introducing or omitting core content.

## 7 Conclusion

In conclusion, this paper introduced an artificial intelligence-based solution designed to significantly simplify and accelerate the creation of mixed reality training and educational materials. With our system, preparing instructional content can be completed within minutes, whereas traditional methods typically require months. The instructional material is generated by capturing a task demonstration from an expert, after which the system automatically creates a step-by-step mixed reality tutorial using speech-to-text and large language model technologies. The generated training content does not include privacy-sensitive information; instead, an avatar replaces the expert during playback, and instructions for subtasks are provided via text-to-speech.

We have evaluated the performance of our system by assessing user experience through the User Experience Questionnaire (UEQ), evaluating usability with the System Usability Scale (SUS), and examining the error rates associated with artificial intelligence processes. Should errors occur during generation, they can be easily corrected through a web interface. The results were highly positive, confirming the validity and effectiveness of our system's concept.

As part of future work, we plan to incorporate an AI-enhanced monitoring component capable of continuously tracking the learner's physical actions and movements. This module would provide real-time feedback and issue alerts if incorrect or potentially dangerous actions are detected. Such functionality could further improve the safety, effectiveness, and adaptability of the training experience in high-risk or precision-critical environments.

## Acknowledgments

This project has received funding from the CHIPS Joint Undertaken as part of the European Union's Horizon Europe research and innovation programme, SMARTY Project, grant agreement No. 101140087. B. Sonkoly was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

## References

1. ARVRTech: Immersive4Learning. https://arvrtech.eu/immersive-4-learning/ (2024)
2. Brooke, J., et al.: Sus-a quick and dirty usability scale. Usability evaluation in industry **189**(194), 4–7 (1996)
3. 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)

4. 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)
5. Education, A.: Classvr. https://www.classvr.com/ (2024)
6. 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)
7. Haj-Bolouri, A., Katende, J., Rossi, M.: Gamified immersive safety training in virtual reality: a mixed methods approach. Journal of Workplace Learning (2024)
8. 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)
9. Inc., A.: Altoura: Ai-powered learning platform (2025), https://www.altoura.com/, accessed: 2025-04-24
10. Ltd., V.: Virtualspeech: Soft skills training in vr with ai feedback (2025), https://virtualspeech.com/, accessed: 2025-04-24
11. Microsoft: Azure ai services. https://azure.microsoft.com/en-us/products/ai-services (2024)
12. Microsoft: Mixed reality toolkit 3. https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk3-overview/ (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. 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)
15. Schrepp, M.: User experience questionnaire handbook. All you need to know to apply the UEQ successfully in your project **10** (2015)
16. 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
17. 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)
18. 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)
19. Yadav, G.: Scaling evidence-based instructional design expertise through large language models (2023), https://arxiv.org/abs/2306.01006
20. 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)
21. 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)

## 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."