Augmented Reality Action Summaries for Situated Instructions of Repetitive Tasks

Lucchas Ribeiro Skreinig*
Graz University of Technology
Ana Stanescu
Graz University of Technology
Peter Mohr
Graz University of Technology
Christoph Ebner
Graz University of Technology
Markus Tatzgern
Salzburg University of Applied Sciences
Dieter Schmalstieg
University of Stuttgart
Denis Kalkofen†
Graz University of Technology

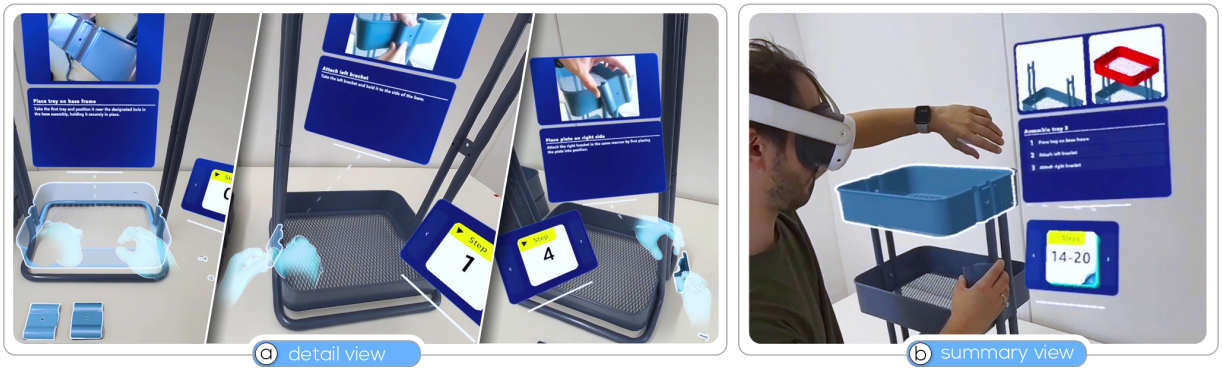

Figure 1: Adaptive AR instructions provide (a) a detailed view, showing each procedural step using AR hand animations and, optionally, a video demonstration on a virtual billboard, and (b) a summary view, presenting a summary visualization of recurring actions or action groups. The system presents the detailed view the first time an action occurs and the summary view thereafter. From the summary view, users can interactively request the detailed view of a single action in case more guidance is needed.

Abstract

Augmented reality instructions help people understand how to operate, assemble, or maintain objects by overlaying visual guidance directly onto the physical environment. However, repeatedly presenting detailed instructions can negatively impact performance. We introduce an interactive AR instruction system that provides detailed guidance on the first occurrence of actions and a summary visualization for repeated actions. We use a multimodal large language model to automatically derive information about actions, groups of actions, and recurring patterns, thereby increasing the scalability of the AR guidance system. A pilot user study shows that summary visualizations are preferred over repetitive step-by-step instructions.

Index Terms: Augmented reality, visualization, adaptive instructions.

1 Introduction

User manuals help people understand how to operate, assemble, or maintain objects. However, traditional paper-based or video-based instructions require users to infer 3D actions from 2D depictions, which can increase cognitive load and lead to errors [24]. Several studies have shown that augmented reality (AR) guides can reduce the cognitive load required to follow user manuals by embedding visual instructions in the physical environment [9, 13, 27].

AR instruction systems commonly use step-by-step visualizations, which present individual actions sequentially and independently of recurrence. However, cognitive load theory suggests that unnecessarily repeated instructions can increase workload and reduce task performance. In particular, research on the *expertise reversal effect* [11] shows that visual instructions can become ineffective or even have a negative effect for users who have already seen them. These users rely on existing knowledge and therefore require less guidance. When an AR system presents instructions they already know, they must either process redundant information or actively ignore the AR guidance. Given that AR instructions are typically designed to be prominent and to draw attention [23], both responses can increase cognitive load for experienced users.

We introduce a novel AR visualization system based on skill acquisition theory [17]. We designed an interactive guidance system that assumes users' knowledge increases with practice. To avoid repetitive AR instructions, we present detailed guidance only the first time an action appears. For subsequent occurrences of the same action, we show only its final state instead of demonstrating *how* it should be performed. This allows users to see *where* the action should be performed while avoiding redundant guidance on *how* to perform it. If users later require additional support, they can interactively request detailed guidance for that specific action.

In many user manuals, repetitive actions occur as groups, i.e., a set of actions that must be repeated. We present groups of repetitive actions using structural diagrams [8] showing the final states of all actions within a group. However, as situated structural diagrams usually consist of several parts, they may clutter the user's workspace. Therefore, we present AR structural diagrams as semi-transparent in-situ overlays. To compensate for the reduced visibility of details due to the transparent AR overlay, we complement the diagrams with high-resolution before-and-after images on a billboard positioned near the AR structural diagram. See Figure 1b for an example of an AR structural diagram and Figure 1a for the

*e-mail:lucchas.ribeiroskreinig@tugraz.at
†e-mail: kalkofen@tugraz.at

corresponding detailed view at a different location.

Manually creating step-by-step instructions and annotating groups is a demanding manual task. Authors must review action sequences, define boundaries, label steps, and construct consistent groups, limiting scalability and practical adoption. Therefore, we demonstrate how a multimodal large language model (LLM) can automatically extract procedural steps from narrated expert demonstrations and organize them into an interactive AR guide with summary visualizations. Our system builds on prior work on authoring by demonstration [18, 20, 15], extending it by incorporating verbal narrations to guide the LLM-based organization process.

In summary, our work makes the following contributions:

- We introduce action summaries as part of a novel interactive AR instruction system for repetitive actions. Our approach avoids redundant guidance on *how* to perform actions, while continuing to indicate *where* the action should be performed using AR.
- We demonstrate how a multimodal large language model can be used to automatically identify actions, groups of actions, and repetitive action patterns.
- We report on a pilot user study that compares AR user manuals that provide step-by-step instructions with AR user manuals that additionally include interactive action summaries.

2 Related Work

Our system supports both detailed step-by-step instructions and summary visualizations for repetitive actions in AR. Accordingly, we review prior work on visualization techniques for AR instruction systems and approaches for automatically authoring detailed instructions and summarizing related task steps.

2.1 Visualizing instructions

Multiple consecutive, similar steps can be summarized using instructional diagrams [8]. Action diagrams combine multiple steps into a single representation, often using glyphs to indicate individual actions. A common example is exploded views [1, 10, 12, 25]. Structural diagrams show the object configuration before and after performing the actions, indicating which parts are manipulated and their resulting locations. Action diagrams are recommended for assembly instructions in printed user manuals [8] because they more explicitly communicate which parts are manipulated and they provide an abstraction of how to manipulate them. We instead use structural diagrams for AR summary visualizations because they introduce less visual clutter by omitting information about *how* the action is performed. We emphasize manipulated parts with a colored highlight (red in the examples in this paper) in the 2D structural diagram displayed on the billboard next to the object.

Several approaches have also explored adaptive instructions. For example, Tatzgern et al. [26] identify redundant data to adaptively optimize view-dependent exploded views, and Anderson et al. [2] progressively reduce guidance. While these automatically adapt instructions, we allow users to interactively reveal details, letting them choose the amount of information that fits their needs.

2.2 Authoring instructions

Pongnumkul et al. [16] highlight the importance of temporal segmentation to effectively guide users through tutorials. Recent approaches have explored using an LLM to segment and summarize narrations. For example, Gao et al. [6] address the challenges of summarizing text by implementing a framework that extracts individual steps from text, then summarizes them using an LLM. Gunturu et al. [7] propose RealitySummary, a reading assistant tool that scans texts via OCR and produces spatially situated summaries using an LLM. Shi et al. [19] author AR instruction steps by prompting an LLM to generate instructions using voice and spatial input, which are labeled based on predefined action labels.

Our system builds on these ideas using a multimodal LLM for step segmentation and summary. Audio narration captured during an expert demonstration is automatically transcribed into text. Prior work has explored the combination of 3D demonstration and narration. For example, Kong et al. [14] pair demonstrations with transcribed instructions for AR guidance. Our system uses the transcribed text for task segmentation and summarization. Following related work on grounded instruction execution, which emphasizes aligning language with real-world tasks for reliable action generation [5], we complement incomplete narration by enhancing the LLM input with a synthetic video generated from the motion-captured VR performance of an expert user.

3 System Overview

Repetitive actions must first be identified before they can be presented as individual steps or summarized instructions in AR. We designed our system as an end-to-end solution for authoring and visualizing instructions.

3.1 Authoring

To increase the scalability of our AR guidance system, we automatically derive actions and groups of recurring actions from expert demonstrations. We capture these demonstrations in a virtual reality (VR) environment, where we record 3D trajectories of the expert's head and hands, virtual object motion data, and synchronized audio narration. We chose to record such data from interactions with a digital twin instead of real objects because VR provides reliable object pose data and avoids the challenges of real-time object recognition and tracking in physical environments. During demonstrations, experts manipulate virtual objects using a pinch gesture.

However, VR interactions alone may not capture sufficient detail for complex manipulations, especially with small or non-rigid objects. Therefore, our system also supports recording video snippets through the headset's outward-facing camera. We implemented our prototype as a standalone application for the *Meta Quest 3*, which provides six-degree-of-freedom tracking for the hands and head, video-see-through AR, and recording audio and video.

The captured data alone does not provide the structure required for our instruction system. We therefore use a multimodal LLM to segment and organize demonstrations into labeled actions and groups suitable for replay and navigation. Based on preliminary tests, we selected *Gemini Flash* because it provided the most reliable action segmentation and labeling from audio input.

We transcribe the expert narration and render an egocentric video from the recorded motion data. The transcript and video frames are provided to the multimodal LLM, which identifies distinct steps and organizes them hierarchically. Each step receives a concise label, detailed description, and reference to the corresponding narration segment. Per-word timestamps and transcript links preserve temporal alignment between recorded actions and generated instructions.

State-of-the-art instruction generation methods using LLMs rely on implicit knowledge encoded in model parameters during training. While this approach has proven effective in multimodal interfaces [21, 19], it may hallucinate incorrect instructions when input is incomplete. Therefore, we instruct the LLM to describe only observed actions rather than generate instructions from scratch.

3.2 Visualization

Our system provides different visualizations for step-by-step instructions. Individual actions are shown through spatially registered 3D animations of objects and hands (Figure 1a), directly captured during expert demonstrations. For more complex actions or interactions with non-rigid objects, which may be unsuitable for VR demonstration, we visualize only hand actions and provide additional detail through a filmed video displayed on a virtual billboard near the object of interest (Figure 2b). The expert decides

Figure 2: Interaction. (a) Visual occlusion is reduced by fading virtual overlays as the user's hand approaches the target area. Right: The target outline is preserved to indicate its position. (b) Groups of recurring actions are summarized in structural diagrams that indicate the planned end state. Additionally, our system shows a summarized list of the steps from which the users can jump to the detail view for a specific action, which shows how the actions need to be performed. This enables users to interactively switch between summarized and detailed instructions.

during authoring which actions require a filmed demonstration. We represent groups of actions as before-and-after images and a situated, semi-transparent 3D digital twin model of the after-state (Figure1b). While the images provide high visual fidelity, the 3D-registered model conveys the group's placement within the physical environment. Before-and-after images are rendered from recorded 3D data using a canonical off-axis, downward-facing camera that focuses on the parts of a group [4].

Per-step information, such as the step number, label, and description, is additionally displayed on a billboard near the object of interest. For groups, the billboard shows substep labels instead of a full description. Users can interactively adjust the instruction detail by expanding groups into individual steps by selecting the icon next to the step number, or by directly selecting a step in the summarized list. Then they can optionally return to the summarized view. To reduce occlusions between real and virtual elements, AR visualizations become transparent as the user's hand approaches the object. However, to indicate its position, we preserve the outline of the virtual object (Figure 2a).

4 Evaluation

We conducted a within-subjects usability pilot study comparing two conditions: *ungrouped*, where each individual action was presented, and *grouped*, which used the AR action summary visualization after users completed the first sub-assembly with detailed guidance. In both conditions, participants followed instructions to assemble sets of repeating sub-assemblies and to place them on the main object at different locations (Figure 2a). The object was designed and 3D printed for the study to avoid relying on prior user knowledge, and the tasks required spatial reasoning and hand–object interaction in AR. Conditions were balanced such that users constructed the first set of four sub-assemblies under one condition and the second set under the other to mitigate order effects.

The study was conducted using the *Meta Quest 3* HMD, which allowed participants to view AR instructions while interacting with physical objects. The study was approved by the ethics review board of *Graz University of Technology*.

We recruited 12 participants (3 female, 8 male, one undisclosed) through social media and personal networks. Participants were between 25 and 45 years old (M=31.7, SD=5.5). They first completed a brief training task to familiarize themselves with the AR setup, followed by both experimental conditions. After completing the conditions, participants took part in a semi-structured interview and completed a demographic questionnaire.

Results Qualitative feedback from post-study interviews revealed a clear preference among all participants for the *grouped* condition. Seven participants in the ungrouped condition tried to skip steps or asked for grouping *without being prompted*, indicating a natural tendency to mentally aggregate repeated actions.

Eight found the ungrouped, repeated steps "tedious" or "needlessly cumbersome". In contrast, the *grouped* condition received consistently positive feedback. Seven participants described the workflow as "natural", with details being preferred during the first initial exposure, then grouping once the task was understood.

Six participants reported that grouped instructions reduce cognitive load and uncertainty, while in the ungrouped condition they "always had the fear something changes", while grouping took "mental load from you". We further noted that the grouped condition did not limit user engagement. Instead, participants often assembled the sub-assembly using their own strategies rather than strictly following instructions. Only three participants expanded the groups to their component steps when encountering a grouped visualization for the first time, suggesting that the summarized view was generally sufficient after users had seen the detailed steps once.

Across both conditions, situated visualizations were identified as the most helpful modality by ten participants, as they "show exactly where to put parts" and "eliminate the need for manual measurements" to find the right location; one participant stated to be able to "do the task with only in-place visualization". Videos were used as a *complementary* modality by nine participants, particularly for conveying motion and fine-grained details such as rotations, while eight participants reported that they used text instructions less frequently and primarily for confirmation. Overall, participants found the system intuitive, easy to use, and beneficial.

Discussion User feedback confirms that instruction grouping reflects the users' natural tendency to organize steps into structures. The strategy of first showing detailed step-by-step instructions before switching to a grouped representation aligns with users' comments about how they would organize sub-assembly actions: carefully following instructions first, then recalling what has been previously done. Repeatedly presenting already introduced steps reportedly introduced perceived insecurities about the procedure, leading participants to verify identical assembly steps. Participants appeared to develop their own solution paths when detailed steps were not provided, indicating that grouped instructions can provide scaffolding for learning procedural tasks, as users not only recall steps, but also rearrange them into new patterns.

5 Future Work

The results of our study suggest that our summary visualizations align with the natural workflow of the users when following procedural instructions with repeated actions. We believe that adapting instructions to the user's needs is important for effective AR guidance, and that our approach takes a step forward in that direction.

Future work should investigate more complex procedures, including assembly tasks that contain symmetrical structures and varying levels of repetition. We aim to understand how effective grouping remains when repeated steps are only partially similar, or when similar actions produce different outcomes. We will explore when action summarization is beneficial and whether different grouping strategies are needed for different types of procedures. We also plan to evaluate LLM-based approaches for automatically identifying and grouping related steps, including comparisons between different models and grouping methods.

Additionally, we plan to investigate automatic selection of instruction presentation based on factors such as workload estimation [3] or error detection [22]. We will explore interaction techniques that enable smooth transitions between different levels of instruction detail while minimizing the need for manual adjustments. Beyond immediate task performance, future studies should examine whether gradually grouping and summarizing instructions in sets of repetitive actions influence learning and retention compared to traditional step-by-step instructions over longer periods.

Acknowledgments

This work was supported by the Alexander von Humboldt Foundation funded by the German Federal Ministry of Education and Research.

References

[1] M. Agrawala, D. Phan, J. Heiser, J. Haymaker, J. Klingner, P. Hanrahan, and B. Tversky. Designing effective step-by-step assembly instructions. *ACM Transactions on Graphics (TOG)*, 22(3):828–837, July 2003. doi: 10.1145/882262.882352

[2] F. Anderson, T. Grossman, J. Matejka, and G. Fitzmaurice. YouMove: enhancing movement training with an augmented reality mirror. In *Proc. ACM Symposium on User Interface Software and Technology (UIST)*, pp. 311–320, 2013. doi: 10.1145/2501988.2502045

[3] S. S. Bhat, C. Dobbins, A. Dey, and O. Sharma. Multi-modal classification of cognitive load in a vr-based training system. In *Proc. International Symposium on Mixed and Augmented Reality (ISMAR)*. doi: 10.1109/ISMAR59233.2023.00065

[4] V. Blanz, M. J. Tarr, and H. H. Bülthoff. What object attributes determine canonical views? *Perception*, 1999. doi: 10.1068/p2897

[5] A. Brohan et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning*, pp. 287–318. Pmlr, 2023.

[6] S. Gao et al. Summarizing procedural text: Data and approach. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022. doi: 10.18653/v1/2022.findings-emnlp.162

[7] A. Gunturu et al. Realitysummary: Exploring on-demand mixed reality text summarization and question answering using large language models. 2025. doi: 10.1145/3694907.3765933

[8] J. Heiser et al. Identification and validation of cognitive design principles for automated generation of assembly instructions. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, p. 311–319, 2004. doi: 10.1145/989863.989917

[9] S. Henderson and S. Feiner. Exploring the benefits of augmented reality documentation for maintenance and repair. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 17(10):1355–1368, 2010. doi: 10.1109/TVCG.2010.245

[10] D. Kalkofen, M. Tatzgern, and D. Schmalstieg. Explosion diagrams in augmented reality. In *2009 IEEE Virtual Reality Conference*, pp. 71–78, 2009. doi: 10.1109/VR.2009.4811001

[11] S. Kalyuga, P. Ayres, P. Chandler, and J. Sweller. The expertise reversal effect. *Educational Psychologist*, 38(1):23–31, 2003. doi: 10.1207/S15326985EP3801_4

[12] B. Kerbl, D. Kalkofen, M. Steinberger, and D. Schmalstieg. Interactive disassembly planning for complex objects. *Computer Graphics Forum*, 34(2):287–297, 2015. doi: 10.1111/cgf.12560

[13] B. M. Khuong, K. Kiyokawa, A. Miller, J. J. La Viola, T. Mashita, and H. Takemura. The effectiveness of an ar-based context-aware assembly support system in object assembly. In *2014 IEEE Virtual Reality (VR)*, pp. 57–62, 2014. doi: 10.1109/VR.2014.6802051

[14] J. Kong et al. Tutoriallens: Authoring interactive augmented reality tutorials through narration and demonstration. 2021. doi: 10.1145/3485279.3485289

[15] M. Patel et al. AssemblifyVR: An authoring system tool for instructional mechanical assembly animations. *Computers & Education: X Reality*, 8:100132, 2026. doi: 10.1016/j.cexr.2025.100132

[16] S. Pongnumkul et al., M. Dontcheva, W. Li, J. Wang, L. Bourdev, S. Avidan, and M. F. Cohen. Pause-and-play: automatically linking screencast video tutorials with applications. In *Proc. ACM Symposium on User Interface Software and Technology (UIST)*, p. 135–144. Association for Computing Machinery, New York, NY, USA, 2011. doi: 10.1145/2047196.2047213

[17] A. Renkl and R. K. Atkinson. Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational Psychologist*, 38(1):15–22, 2003. doi: 10.1207/S15326985EP3801_3

[18] C. Rolim, D. Schmalstieg, D. Kalkofen, and V. Teichrieb. [poster] design guidelines for generating augmented reality instructions. In *2015 IEEE International Symposium on Mixed and Augmented Reality*, pp. 120–123, 2015. doi: 10.1109/ISMAR.2015.36

[19] J. Shi, R. Jain, S. Chi, H. Doh, H.-g. Chi, A. J. Quinn, and K. Ramani. Caring-ai: Towards authoring context-aware augmented reality instruction through generative artificial intelligence. In *Proceedings of CHI Conference on Human Factors in Computing Systems*, 2025. doi: 10.1145/3706598.3713348

[20] L. R. Skreinig, P. Mohr, B. Berger, M. Tatzgern, D. Schmalstieg, and D. Kalkofen. Immersive authoring by demonstration of industrial procedures. In *Proceedings - 2024 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2024*, pp. 1293–1302. Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ISMAR62088.2024.00146

[21] S. Srinidhi, E. Lu, and A. Rowe. XaiR: An XR Platform that Integrates Large Language Models with the Physical World. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 759–767, 2024. doi: 10.1109/ISMAR62088.2024.00091

[22] A. Stanescu, P. Mohr, F. Thaler, M. Kozinski, L. R. Skreinig, D. Schmalstieg, and D. Kalkofen. Error management for augmented reality assembly instructions. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 690–699, 2024. doi: 10.1109/ISMAR62088.2024.00084

[23] B. V. Syiem et al. Impact of task on attentional tunneling in handheld augmented reality. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411764.3445580

[24] A. Tang et al. Comparative effectiveness of augmented reality in object assembly. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*. Association for Computing Machinery, New York, NY, USA, 2003. doi: 10.1145/642611.642626

[25] M. Tatzgern, D. Kalkofen, and D. Schmalstieg. Compact explosion diagrams. In *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*, NPAR '10, p. 17–26. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1809939.1809942

[26] M. Tatzgern, D. Kalkofen, and D. Schmalstieg. Dynamic compact visualizations for augmented reality. In *2013 IEEE Virtual Reality (VR)*, pp. 3–6, 2013. doi: 10.1109/VR.2013.6549347

[27] Z. Yang et al. Influences of augmented reality assistance on performance and cognitive loads in different stages of assembly task. *Frontiers in Psychology*, Volume 10 - 2019, 2019. doi: 10.3389/fpsyg.2019.01703