CARING-AI: Towards Authoring Context-aware Augmented Reality INstruction through Generative Artificial Intelligence

Jingyu Shi[*]
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, USA
shi537@purdue.edu

Rahul Jain[*]
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, USA
jain348@purdue.edu

Seunggeun Chi[*]
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, USA
chi65@purdue.edu

Hyungjun Doh
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, USA
hdoh@purdue.edu

Hyung-gun Chi
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, USA
chi45@purdue.edu

Alexander J. Quinn
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana, USA
alexanderjquinn@gmail.com

Karthik Ramani
School of Mechanical Engineering
Purdue University
West Lafayette, Indiana, USA
ramani@purdue.edu

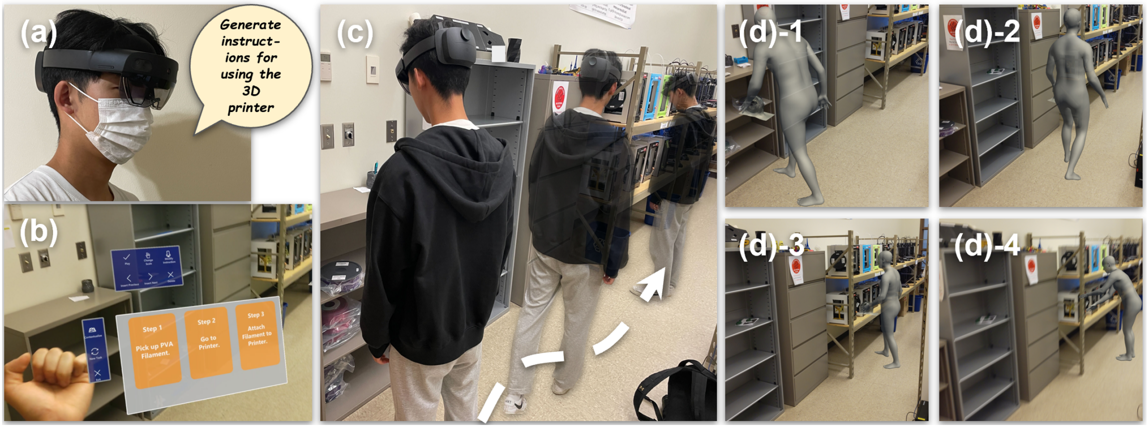


Figure 1: An overview of CARING-AI system authoring workflow. CARING-AI enables authors to create contextualized AR instructions through generative AI. (a) Using CARING-AI, authors first speak their intended instruction content, (b) then the corresponding step-by-step instructions are generated in text. Authors interact with the interface to modify the textual instructions and group them. (c) Then the authors provide contextual information to the instructions by walking in the environment and taking screenshots with the AR HMD. (d) Finally, CARING-AI generates step-by-step humanoid avatar demonstrations of the AR instruction situated in the context.

Abstract

Context-aware AR instruction enables adaptive and in-situ learning experiences. However, hardware limitations and expertise requirements constrain the creation of such instructions. With recent developments in Generative Artificial Intelligence (Gen-AI), current research tries to tackle these constraints by deploying AI-generated content (AIGC) in AR applications. However, our preliminary study with six AR practitioners revealed that the current AIGC lacks

[*] Three authors contributed equally to this research.

CHI '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713348

contextual information to adapt to varying application scenarios and is therefore limited in authoring. To utilize the strong generative power of GenAI to ease the authoring of AR instruction while capturing the context, we developed CARING-AI, an AR system to author context-aware humanoid-avatar-based instructions with GenAI. By navigating in the environment, users naturally provide contextual information to generate humanoid-avatar animation as AR instructions that blend in the context spatially and temporally. We showcased three application scenarios of CARING-AI: Asynchronous Instructions, Remote Instructions, and Ad Hoc Instructions based on a design space of AIGC in AR Instructions. With two user studies (N=12), we assessed the system usability of CARING-AI and demonstrated the easiness and effectiveness of authoring with Gen-AI.

CCS Concepts

• **Human-centered computing → Mixed / augmented reality**.

Keywords

Augmented Reality, Generative Artificial Intelligence

ACM Reference Format:
Jingyu Shi, Rahul Jain, Seunggeun Chi, Hyungjun Doh, Hyung-gun Chi, Alexander J. Quinn, and Karthik Ramani. 2025. CARING-AI: Towards Authoring Context-aware Augmented Reality INstruction through Generative Artificial Intelligence. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3706598.3713348

1 Introduction

Augmented Reality (AR) instructions provide an interactive and immersive learning experience by rendering digital content onto physical environments and enabling visualization of complex concepts or procedures. With such instructions, end-users explore various scenarios and practice skills in a more realistic and context-rich setting. Due to their vast capabilities and their potential to enhance user engagement [125], facilitate learning [135], and improve performance [24, 43] in various contexts, AR instructions have gained considerable attention in a range of fields.

In the manual task instruction domains, humanoid avatars are the preferred options of visualization [13, 22], because they can convey spatial and temporal instructions on complex sequences of tasks, such as machine tasks, assembly tasks, manual skill learning, and medical training. Prior works thrived to optimize the authoring of animated humanoid avatars in AR. Beyond regular animation workflows supported by software such as Unity [117], Unreal Engine [35], or Blender [32], research has proposed diverse methodologies to overcome the requirement of expertise in both the subject matter of the instructions and the programming for animation [22]. A promising method is Authoring/Programming by embodied Demonstration (PbD, i.e. creating or editing humanoid animation in AR environments by physically interacting or demonstrating actions in the real world). PbD have the advantages such as realistic animation [57, 126], code-less efficiency [23], engagement [2, 7, 81], interactivity [50, 122], and learning gain [135] in AR instruction applications. Despite the benefits and simplicity for the authors, PbD is still subject to real-world human motion (i.e. the authors have to physically present and demonstrate) and requires complex hardware setups and re-setups for Motion Capture (MoCap) such as cameras or motion sensors. Therefore, authoring with PbD systems is limited in varying contexts ad hoc.

The development of Generative Artificial Intelligence (Gen-AI) has brought AI-generated content (AIGC) into the discussion of authoring AR instructions [45], considering its potential to eliminate expertise barriers and hardware requirements. With this rapid growth of Gen-AI power, content creation in various modalities can be democratized to higher levels [12, 77]. Users are enabled to generate desired content by simply prompting via intuitive modalities (e.g. textual conversation [9, 56, 95, 96] and reference image [94, 98, 99]). Many ongoing research and discussions have identified opportunities for deploying AIGC in AR for its power of abstracting human knowledge and a wide range of I/O modalities [12, 113].

In pursuit of the design space of AIGC in AR instructions, research is faced with the challenge that Gen-AI lacks the contextual and background information to be deployed into real-world applications [77]. In the scope of AR instruction, contextual information is a critical metaphor, where spatial-temporal information of the instruction is to be blended in the context of the users. A taxonomy of context-awareness in AR instruction, that many prior works [38, 92, 120] converge towards, encompasses three key aspects: the human, environment, and system.

Building on this existing knowledge, we aim to fill the gap between state-of-the-art Gen-AI and context-aware AR humanoid avatar instructions. Specifically, our research is motivated to explore (1) What context information does AI-generated humanoid avatar animation lack for AR instructions? (section 3) (2) How can this missing contextual information be delivered to Gen-AI? (section 4) and (3) What insights can we gain from our designs to further foster developments towards the use of AIGC in AR? (section 9)

From a preliminary expert interview, we summarize the design goals for naturally providing contextual information to AIGC incorporating user interactions in the authoring process. We then present CARING-AI, an AR system enabling authoring contextualized humanoid avatar animation for AR instructions. Given a textual description of the task to instruct, CARING-AI generates step-by-step textual instructions that can be modified by the users and further generates motion that animates humanoid avatars as the visual cues in the instructions. After giving the textual instructions to animate, authors navigate and scan the environment with an AR Head-Mounted Device (HMD). Then, CARING-AI temporally and spatially adapts the AI-generated instructions to the human, environment, and system context of the task.

Our contributions are four-fold:

- A code-less and Mocap-free workflow for authoring animated humanoid avatar instructions in AR with Gen-AI, contextually aware of the human, environment, and system.
- A diffusion-model-based algorithm to temporally smooth sequences of individually generated humanoid motions.
- An AR interface for authoring AR instructions from textual input describing the tasks, avatars' trajectory, and FOV.
- A series of studies evaluating the performance of our system and assessing the efficiency of creating AR animation with Gen-AI compared with a baseline PbD method.

2 Related Work

2.1 AR Instruction

AR instruction refers to the use of AR technology for instructional purposes, such as visualizing complex concepts, exploring various scenarios, practicing skills, and providing real-time feedback.

Our use of the AR instruction metaphor is grounded in real-world applications in diverse domains including assembly [22, 33, 64], education [34, 49, 71, 80, 89, 119], manufacturing [91], logistics [78], IoT [111, 129] and domestic applications [10, 40, 54, 123, 123].

Our scope focuses on the visualization techniques of animated humanoid avatars in authoring AR content for tasks that convey spatio-temporal instructions to the end-users. Through our wide literature review of AR instructions, we conclude that the information conveyed by AR instructions can be categorized into three types:

Spatial information refers to the geographical or spatially-related data such as the location of certain objects or the occurrence of interactions. Spatial information is usually visualized by 3D models [34, 41, 48, 130], overlaying data [40, 54], and visual cues such as arrows and lines [10, 64, 112].

Temporal information refers to the time-related data such as the order, synchronization, or timing of the movement or occurrence in the AR. Temporal information can be visualized through textual descriptions of order or procedural [36], animation [33, 89], video [11], or sequential overlays [112].

Spatio-temporal information refers to the information that encompasses both spatial and temporal descriptions of an event, an interaction, or movement in AR, explicitly addressing the change of spatial data in a temporal interval. Spatio-temporal information can be visualized in AR by combining spatial and temporal methodologies. When spatio-temporal information depicts a human motion or their interaction with the environment, it is better visualized in the animated humanoid avatars [14, 22, 46, 121], where the end-users of the content can learn through following the avatars.

2.2 Authoring AR Content

Authoring AR content refers to the process where designers explicitly assign spatial behaviors of the virtual components to the physical world [93]. Programming-based authoring tools enable authors to create AR content through programming languages and mathematical modeling [32, 35, 117]. Authoring by programming creates a precise AR experience, however, at the cost of requiring authors' expertise in both the subject matter and programming. Moreover, it isolates the authors from the target environment where the AR applications emerge, depriving the spatio-temporal connection to the target environment of the authors.

To tackle the challenges above, prior arts propose the concept of immersive authoring, where the author can create AR content by interacting with both the virtual components and the physical world [66]. To immersively author humanoid avatar animation, prior work has applied methods based on embodied demonstration to authoring. Through embodied methods, designers can create human movement and interactions with objects by simply demonstrating [14, 22, 23, 73, 100, 120, 121]. However, authoring through demonstration is subject to the hardware needed for Mocap [22, 121]. In addition, it requires the author to be physically interacting with the environment, which is often not possible or even needed. For example, the environment may be remote for the author, the environment itself is virtual, the concept that is being demonstrated is not physically plausible or imaginary, or costly for various reasons.

To overcome the barriers of expertise requirement, hardware limitation, and physical interactions, researchers have investigated the uses of AI-generated content (AIGC) in AR applications. Early works are limited by the modalities and generating power of Gen-AI and, therefore, focus on only a bounded area. For example, Generative Adversarial Networks (GAN) are capable of generating images based on a given text or image input. It has been deployed in visual tasks such as fashion design [109, 133], rendering a realistic shadow [69], reconstructing an occluded human body [21] or virtual objects [132] or generating new virtual objects [59, 114].

With the recent development in Gen-AI technology, methodologies have enabled content generation in a wider range of modalities (e.g. text-to-text by Models such as Generative Pre-trained Transformer (GPT) and its successors [9, 56, 95, 96], T5 [97], and BERT [27], text-to-image by large vision models [1, 94, 98, 99, 107] and by Diffusion Models [42, 87, 104, 108, 115], text-to-3D [72], image-to-text [94], etc.) with faster and better-generated quality [28].

The uniqueness of Gen-AI arises from the fact that it can **generate novel content**, rather than inferencing and acting on existing data or knowledge bases and **choosing existing content via an if-else rule database** [37].

The recent developments that have demonstrated the out-of-ordinary capabilities of Gen-AI have inspired and enabled our work to embed AIGC into AR applications. We present related ongoing research (i.e. non-peer-reviewed reports) as well as some recently published papers to differentiate the key aspects of our approach. To the best of our knowledge, the capabilities we have demonstrated in AIGC for AR are new and are to be still explored from both the design space and applications viewpoints. Hu et al. [45] explored the design space of AIGC + AR applications through an interview, and concluded with several discussions regarding the user, environment, and function of the AR application. Lv et al. [77] concluded that context is a key consideration in giving prompts to Large Language Models (LLM). Soliman et al. [116] envisioned using Gen-AI in ARGC for its wide range of modalities. Chen et al. [18] implemented an LLM-based AR system that incorporates spatial and contextualized information to generate textual instruction in the AR application. However, these prior works deal with textual instructions, while ours focuses on humanoid animation to provide spatio-temporal instructions with the avatar. A recent survey by Chamola et al. [16] investigated the capabilities of existing Gen-AI methodologies and summarized the characteristics of possible AIGC + Metaverse applications via clustering the methodologies. Their research pointed out a key insight towards the prospect of Gen-AI in Metaverse: generating 3D content for Metaverse applications (AR in our scope) via Gen-AI needs the incorporation of contextual information. This insight is also aligned with the recent works such as those of Huang et al. [47] and Shi et al. [113]. They recognized the missing "contextual memory" and designed a knowledge interactive agent to identify the missing knowledge and pass it to the Gen-AI model to ground the model in contextual applications.

Motivated by the prior works, we position our work to fill the gap between the AI-generated humanoid avatar animation and AR instructional applications, by contextualizing the generated content via author interactions.

2.3 Context-aware AR Applications

The metaphor of *context-awareness* has been a significant area of interest among both researchers and practitioners. Lee et al. [65] define context awareness as the ability of a system to *apply the patterns given the constraints imposed by the real world*. An established taxonomy [38, 92] categorizes context-awareness into three types:

- ***Human Context***, where the AR systems recognize the humans (users and non-users), take into consideration their profiles [46, 61], status [8, 10, 110, 112], or their interactions [25, 29, 71, 76, 131] and adjust the AR components accordingly.
- ***Environmental Context***, where the AR systems perceive the surroundings of the users and understand the presence and absence of physical objects [19, 40, 54, 58, 123], temporal primitives [105, 106], or digital representations of the scenario [20, 25, 36, 63, 68, 83, 93, 110, 121], and adjust their components correspondingly.
- ***System Context***, where the AR systems are aware of their input/output [61] or their own states [29, 112, 131] in the realities, and adapt to these contexts.

In the scope of AR instructions, all three categories of context awareness are essential. With human-context awareness, systems are capable of adapting the instructional content according to the users' performance to maximize the learning gain [46]. Besides, understanding human motion enables the system actively to decide which steps in the instructions are best to be visualized to the users [29]. The location of the visualization in AR also relies on the human context [25, 71, 112, 131]. On the other hand, environmental context also plays a key role in AR instructions. For instance, instructing hand-object interactions in AR requires the overlays of 3D models of the objects to be aligned with the physical world for visual cues [22, 40]. The environment also possesses rich semantic information that determines the content of the instruction [121]. Moreover, the system context helps to decide the procedures in AR instructions by recognizing the states of the instructions and timely transiting to the subsequent ones [22].

Grounded on the prior works and the three categories above, we discuss how we can contextualize the AIGC in AR instructions.

3 Preleminary Study and Design Rationale

3.1 Preliminary Study

To better understand AI-generative content (AIGC) for AR instructions, we conducted a study with six participants (P1-P6, four males and two females) who have prior experience in creating AR applications for procedural instruction. All Participants were academic researchers from different disciplines: Electrical and Computer Engineering (3), Computer Science (2) and Mechanical Engineering (1). The mean age of participants was 29.5 and all of them had at least 4 years of experience in creating AR/VR/MR applications.

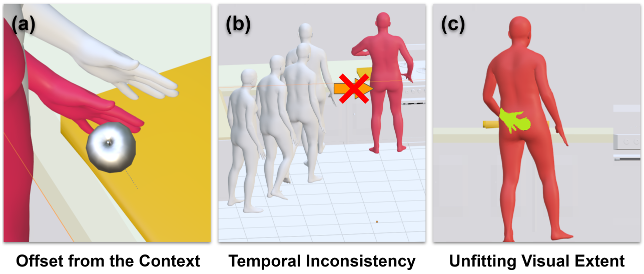


Figure 2: Problems of AI-generated humanoid avatar animation identified in the preliminary study (a) the offset between the generated content and the context, i.e. the interaction is not spatially aligned with the object, (b) the temporal inconsistency, i.e. the generated motion is not temporally connected, and (c) the unfitting visualization extend, i.e. the generated avatars are not of the best scale to convey the instructions (full-body v.s. half-body v.s. hand-only)

Procedure: We showed a seven-step humanoid avatar animation instruction task to the participant, generated by the state-of-the-art Generative AI algorithm GMD [55]. The animation is generated from the textual input of *"cutting an apple"* and contains the following steps: 1) Go to the cutting board, 2) Take the apple with the left hand, 3) Put the apple on the cutting board, 4) Go towards the knife area, 5) Take the knife with the right hand, 6) Go to the cutting board, 7) Cut the apple with a knife.

After participants watched the content, three authors interviewed them for 30 to 60 minutes with inductive and open-ended questions. In addition to their opinion on the quality of the shown animation, we asked general questions about the challenges of creating an AR avatar tutorial, the quality of the content, and the potential gap between the characteristics of demonstrations in AR instructions and AIGC. The interviews were recorded, transcribed, and coded by the same three authors. Each author reviewed the transcripts and summarized an initial set of design goals. Three authors merged to discuss each other's design goals and concluded a refined version by eliminating redundant points and including as many exclusive points as possible. The analysis provides the following insights and the Design Goals (**DG**) listed below:

DG 1) Spatially Aware Content The need for AIGC to be grounded in the real world for AR applications is evident. The AIGC should be aware of the user's real-world environment which includes objects, their locations, and surfaces. All participants pointed out that spatial information is important to transfer virtual content into the physical world for AR applications (P1-P6). Additionally, the AIGC should provide avatar demonstrations subject to the users' vicinity where specific interactions and objects are located (P1, P2, P5). "*The tutorial should include an avatar demonstration of manipulating a virtual object, when real and virtual are overlaid for a better understanding of the content.*" - P2

DG 2) Transition Continuity The AIGC should be smooth when transitioning from one event or interaction to another. All the users mentioned that the content shown was not continuous and there were sudden breaks between the interactions. "*All the actions present were looking separate and there was no connection between*

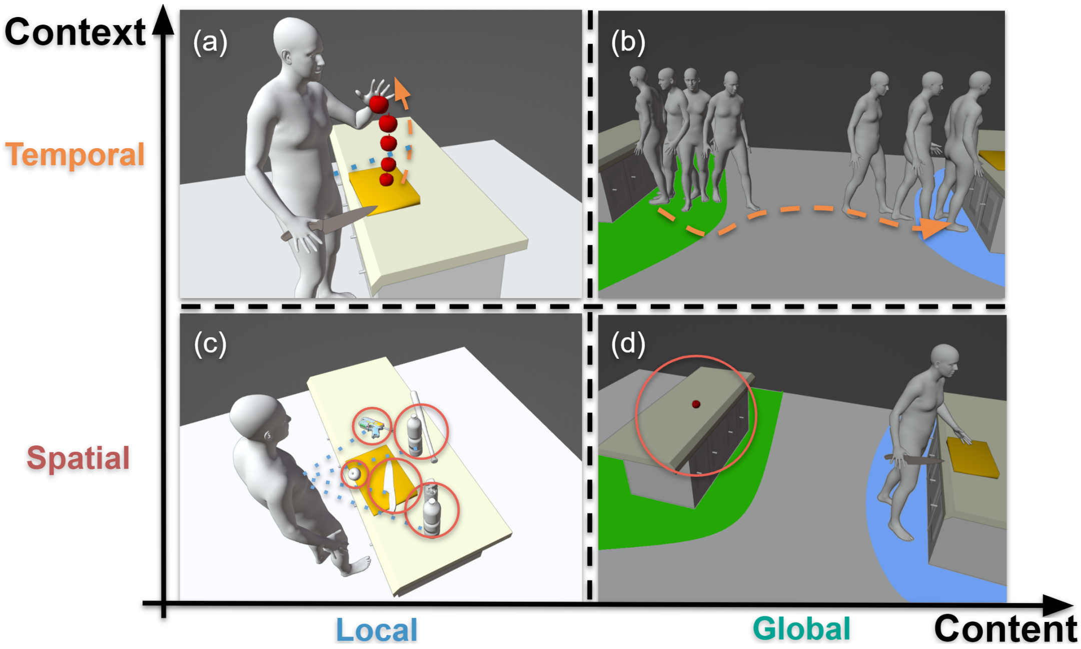


Figure 3: Our consideration of the design space of AIGC in AR instructions is composed of two dimensions: context and content. An AR instruction can be either temporal or spatial based on the contextual information it conveys, either local or global, based on the scale of the content it contains.

the actions" - P1. Participants also mentioned that it was difficult to create continuous and smooth AR avatar instructions with the currently available technology (P2, P4). "*In my case, I created step-by-step small steps for creating tutorials by avatar demonstrations of assembly*" - P2.

DG 3) Scale of Content The AIGC should include different scales of demonstration adaptive to the different scales of the content in terms of the movement, focusing on different parts of the instructions. This can be achieved by giving users the freedom to decide whether they prefer to see the whole body (third-person view) or just the hands (First-person view) of an avatar(P1, P5). Moreover, this will also decide the scale of the avatar and virtual objects present in the scene (P2, P3). "*Author should have the freedom to watch the content in the visualization method they preferred.*" - P3.

DG 4) Flexibility in Modifications of the Content The AR tutorials should contain flexibility in editing, recreating, or removing the content (P1, P4, P5), which is not enabled by the Gen-AI models themselves without designated interactions with the user. Participants from their prior experience also mentioned that modification in AR tutorials is time-consuming and requires a lot more effort (P1, P6). *"I created an AR avatar tutorial for a mechanical assembly task and it took me a lot of time to make the content"* - P6. Participants acknowledge the use of the AI model in creating the tutorials because of less coding effort (P1). "*It amazes me that these tutorials are just created from the text. This will make AR content creation easy and fast*" - P1

3.2 Design Space

From prior works [6, 17, 44, 51] and our study findings, we conclude that the current methods of creating AR instructions from AI-Generated Content (AIGC) are sophisticated and cumbersome. The four aforementioned design goals are key to grounding AIGC in AR instructions. Most participants agreed that Gen-AI is a powerful tool that can be used to create AR avatar motions, per intuitive and efficient interaction techniques designated to utilize the generative power (**DG 4**). We also found that context and content are the most important aspects for AIGC to be used in creating avatar instructions for AR applications. From the context side, the Gen-AI model should understand the physical space and their elements which includes recognizing specific locations, objects, landmarks, and their relations (**DG 1**). The content can be either an event or interaction and should be presented temporally consistent to the user (**DG 1**). Moreover, the scale of the content also matters when it comes to the efficiency of the instructions (**DG 3**). To this end, we identify context and content as two essential dimensions of the design space of AIGC in AR instructions, as shown in Figure 3. The first dimension is the context, which can be either spatial or temporal:

- **Spatial context:** It refers to the information related to the physical environment which involves location, objects, and their interactions.
- **Temporal context:** It refers to the synchronization and timing of information conveyed by the AIGC.

The second dimension is the content in AR, which can be either global or local:

- **Local content:** It refers to the specific content of the instruction constrained in the users' immediate vicinity, which is to be depicted in low-level details in the AIGC instruction.
- **Global content:** It refers to the broader perspective of the content relating to the overall scope of the task, describing the high-level goals of steps.

We further explore the AIGC in AR instructions located in each of the quadrants divided by the two dimensions above.

Local-spatial instructions explain users' closest vicinity information about the objects, locations, their semantic information, and relation with each other **(DG 1)**. Such instructions locate and align the 3D object models and humanoid avatars with the corresponding physical objects or areas.

Local-temporal instructions reveal the timely order of interactions between the avatars and the vicinity. Such instructions illustrate step-by-step how-to for each interaction or action with temporal consistent transitioning from one to another **(DG 2)**.

Global-spatial: instructions depict the approximate whereabouts of the objects, areas, or interactions that are positioned outside the local vicinity. In contrast to local-spatial instructions, global-spatial instructions posit the content approximately in a space rather than detailing the exact location in the space **(DG1)**.

Global-temporal: instructions guide the end-users from one space into another and change the vicinity of the end-users with temporally consistent transitions **(DG2)**.

We built the CARING-AI system based on the design space decomposition above, addressing the design goals that we have derived.

4 CARING-AI System

We developed the CARING-AI system that allows authors to generate and contextualize avatar animation instructions in AR. Based on the discussion above, we derived the following features in our system: 1) Allowing authors to create textual instruction with editable features (**DG 4**), 2) Scanning the environment to get spatial context information (**DG 1**), and 3) An authoring interface for visualization and editing of the generated content (**DG 2, 3, 4**). In this section, we discuss the implementations of the algorithms and modules of CARING-AI and the present our interface.

4.1 System Overview

CARING-AI consists of the following steps as shown in Figure 4:

1) Refining textual instructions. The user provides a task description to ChatGPT [88], which returns the step-by-step textual instructions to perform the task. The user can then further modify or correct the generated textual instructions.

2) Scanning the environment. The user moves in the physical environment to scan the objects, locations, and areas, as well as record their trajectory, which will be used to provide spatial context information to the system.

3) Generating avatar instructions. The system takes refined textual input from Step 1 to generate avatar instructions based on the **design space** discussed in subsection 3.2. The generated instructions are also grounded by the context information provided in Step 2.

4) Visualization and Editing. The user can view and edit the AI-generated instructions.

4.2 Textual Instructions

This module allows users to refine textual instructions for a task using a large language model (LLM), namely ChatGPT API [88]. Given a user-intended task to instruct, CARING-AI prompts [128] the ChatGPT API to refine the user description of the task into a sequence of step-by-step predefined action labels, which are presented in the HumanML3D dataset [39] (a large computer vision benchmark dataset), by specifically asking *"detailed step-by-step instructions of the [task name]"*. The purpose of this step is to align the terminology of the textual instructions with the available action labels from the dataset to ensure precise generation by the model.

After the refined instructions are generated, users can make necessary adjustments, add more details, or remove information to ensure the instructions align with their specific needs (for example if the object in the textual instruction is not present in the environment). The finalized instructions will then be used to generate the avatar motion for the task.

4.3 Scanning the Physical Space

CARING-AI utilizes HoloLens2 AR-HMD [79] as the front-end platform. In order to capture the spatial context of the environment, such as objects, their locations, and semantic meaning as shown in Figure 5, the user navigates and scans the environment with the HMD and starts the **Scan mode** in the interface. The user walks around from and to contexts where actions happen, and scans the entire required environment. Upon entering the **Scan mode**, CARING-AI records the surroundings by taking RGB images of the HMD FOV and starts recording the global trajectory of the user (built-in SLAM). The RGB images are passed to an object detection algorithm [101] (30ms per image) to get the semantic classification and relative location of the objects. The RGB images and the object information are further passed to the state-of-the-art 6 DoF algorithm, MegaPose 6D [62] to obtain the 6 DoF information of the objects. Then CARING-AI overlays virtual objects onto the real object based on 6DoF information. This information of objects is then used to generate avatar motion with detected and overlayed objects.

4.4 Generating the motion

After getting the textual instruction and spatial information from the user, we generate the avatar motion utilizing a Gen-AI model. Specifically, we modified the state-of-the-art text-to-motion AI model (MDM [118]), to generate the global-spatial-context-aware motion (subsubsection 4.4.1), local-spatial-context-aware motion (subsubsection 4.4.2), and temporal-context-aware motion (subsubsection 4.4.3), covering our design space of AR avatar instructions shown in Figure 3.

4.4.1 Global-Spatial-Context-Aware Generation. As discussed in **DG 1**, it is key to contextualizing the generated animation for AR instructions. To tackle this challenge, we exploit the idea of Guided Motion Diffusion (GMD) [55]. On top of other motion generation diffusion models, GMD can generate humanoid motion data, using text descriptions and location cues as the conditions to guide the

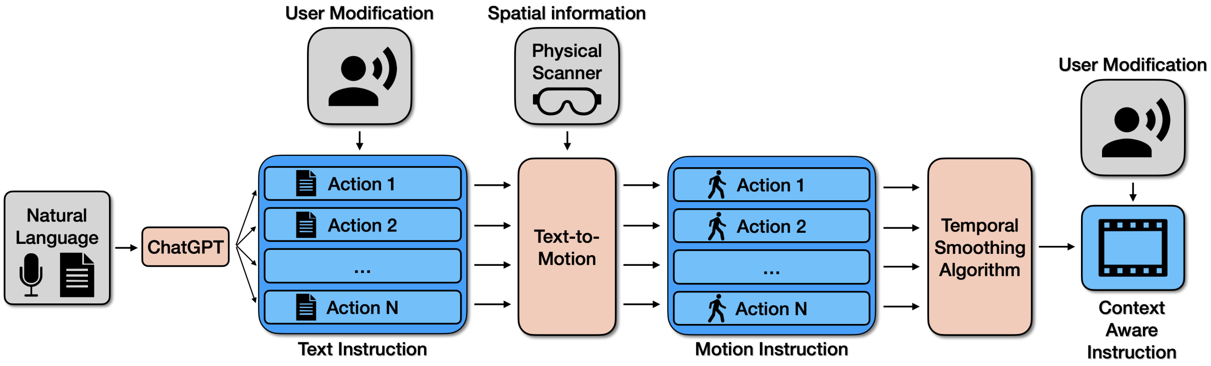


Figure 4: The overall pipeline of the CARING-AI system. Users start by generating textual instructions by speech or text. These instructions will be further grounded in the context of the users by scanning the environment. With context, instructions are used to generate humanoid avatar motion to demonstrate the instructions, blended in AR.

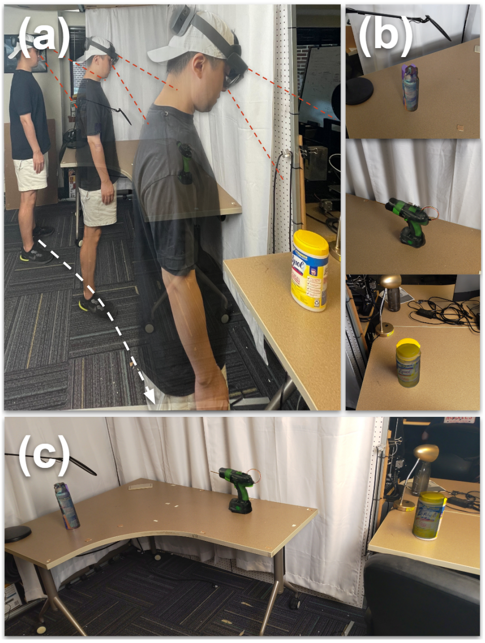


Figure 5: Our methodology for obtaining the contextual information. For global information, users walk from one location to another to provide trajectories (a). For spatial information, users look at the local objects and take screenshots (b, c). This contextual information will be used to generate humanoid avatar motions that are aware of the spatial context for global and local content.

generation. However, GMD does not support the generation of sequences of multiple actions. To address this challenge, we modified the architecture of the Motion Diffusion Model (MDM) [118] (which is also used by GMD as their base model to include trajectories) as shown in Figure 7 and applied the GMD method to generate the humanoid motion with trajectory guidance. We use the trajectories

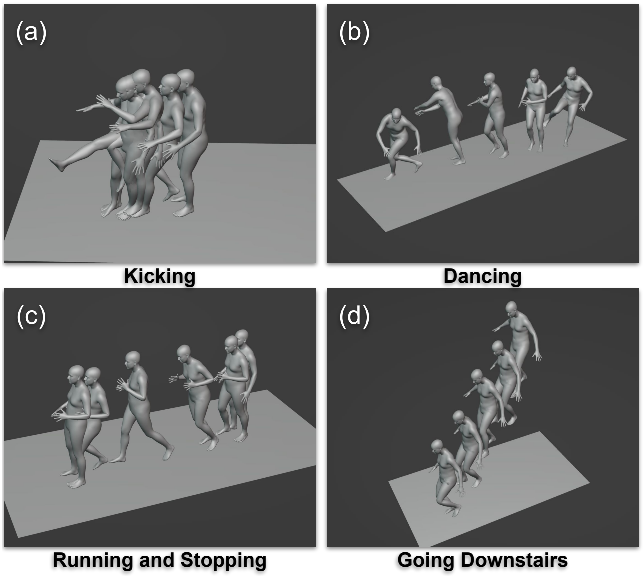


Figure 6: Some examples of our motion generation models. The motion can be local (a) or global (b, c, d, i.e. from one place to another)

recorded in the **Scan mode** as the conditions to the diffusion model to provide global spatial information to the generated motion.

4.4.2 Local-Spatial-Context-Aware Generation. CARING-AI is also capable of conveying local spatial information in the instructions, by generating the motion of the hand with 3D virtual objects overlaid on the physical objects using another motion diffusion model for hand and object interaction [15]. As described in subsection 4.3, in the **Scan mode**, CARING-AI obtains the location information of the objects in the physical environment and overlays 3D models of them in AR. To make sure that the generated avatar interacts with the objects correctly, we ask the users to exit the **Scan mode** at the end of their trajectory while looking (with the HMD) at the objects that they intend to interact with at the step. In this way, we

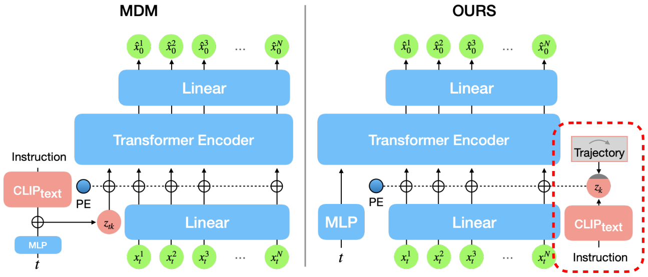


Figure 7: The comparison of the diffusion training overview is between the Motion Diffusion Model (MDM) (*left*) and ours (*right*). The MDM conditions motion frames by placing z_{tk} at the first location, while our conditions motion frames by adding z_k to each motion embedding. For simplicity, we have omitted the random masking of the text embedding used for classifier-free diffusion guidance.

guarantee to record the object 6 DoF information relative to the last global location of the trajectory.

4.4.3 Temporal-Context-Aware Generation. As discussed in **DG 2**, temporal smoothness is key to the sense of continuity in AR instructions. The original MDM model is designed to generate only a single action by conditioning the instruction into the whole sequence at once. Generated motions exhibit discontinuity in transition segments because they are produced independently, without incorporating information about the start and end of each instruction as shown in Figure 2 (b).

To address this challenge, we modified MDM to condition instructions to each frame, allowing them to generate multiple action sequences jointly. We visualized the architecture and modification in Figure 7. For the sampling process, we generate multiple actions by adding distinct text conditions, represented by z_k, to the frames. For example, for three actions each 60 frames long, we applied different z_k values across the ranges: 1–60, 61–120, and 121–180 frames.

However, due to the limitation of the frame length of the training dataset, the quality of the motion drops empirically when the frame number exceeds 196. Further, we designed a temporal smoothing algorithm to generate an unlimited length of smooth avatar motion and applied it after the generation of motions. As illustrated in Figure 8, the temporal smoothing function, (denoted as f) aims to mitigate the discontinuity among the transitional segments of motion (K^1 and K^2, where K represent two transition segments). Each of the transition segments comprises a length of L frames. We also set the weight function α_t to define the ratio for combining the two transition segments. For this purpose, we employed the shifted sigmoid function for α_t, given by $\alpha(t) = \frac{1}{1+e^{-(t-(L/2)}}$, to serve as our smoothing mechanism. Consequently, the resultant mixed frames, represented as $\tilde{K}_t$, can be expressed as

$$\tilde{K}_t = f(K_t^1, K_t^2, \alpha_t) = \alpha_t K_t^1 + (1 - \alpha_t) K_t^2. \tag{1}$$

Then, to keep the length of the generation action length, we extended its length twice with linear interpolation sampling.

$$\hat{K}_t = \tilde{K}_{x_0} + \frac{\tilde{K}_{x_1} - \tilde{K}_{x_0}}{x_1 - x_0}(x - x_0), \tag{2}$$

where x is $\frac{L-1}{2L-1}t$, x_0 is $\lfloor \frac{L-1}{2L-1}t \rfloor$, x_1 is $\lceil \frac{L-1}{2L-1}t \rceil$, $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ indicate the ceiling and the floor operator, respectively.

Algorithm 1 Temporal smoothing

```
INPUT: $K_t^1, K_t^2$                                   ▷ Transition segments
       $\alpha_t$                                       ▷ Temporal smoothing function
OUTPUT: $\hat{K}$                                       ▷ New transition segments
for $t = 0, 1, ..., L-1$ do                             ▷ Temporal smoothing
    $\tilde{K}_t = \alpha_t K_t + (1 - \alpha_t) K_t$
end for
for $t = 0, 1, ..., 2L-1$ do                            ▷ Linear interpolation
    $x \leftarrow \frac{L-1}{2L-1}t, x_0 \leftarrow \lfloor \frac{L-1}{2L-1}t \rfloor, x_1 \leftarrow \lceil \frac{L-1}{2L-1}t \rceil$

    $\hat{K}_t = \tilde{K}_{x_0} + \frac{\tilde{K}_{x_1} - \tilde{K}_{x_0}}{x_1 - x_0}(x - x_0)$

end for
```

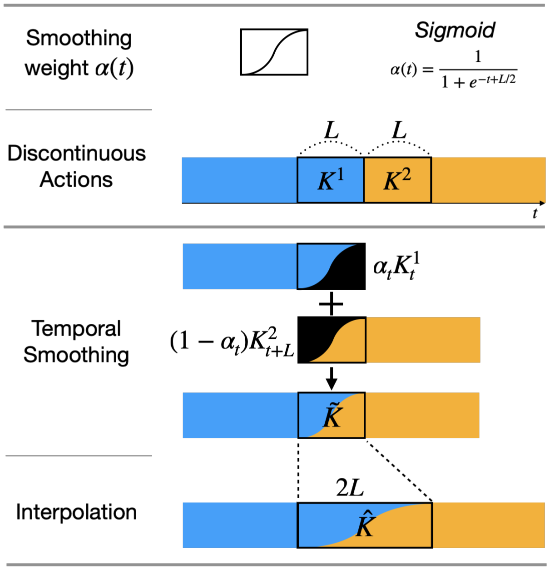
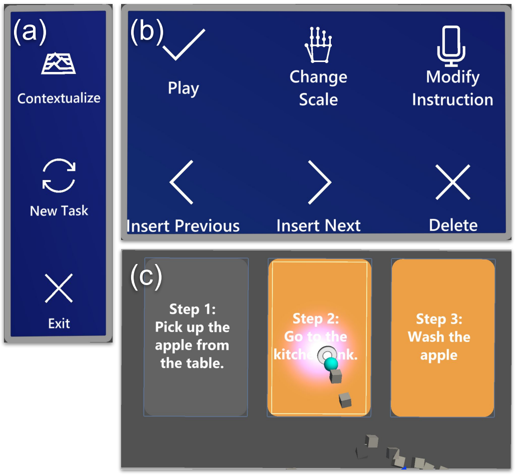


Figure 8: The illustration of the temporal smoothing algorithm of CARING-AI

4.5 AR Interface

To achieve **DG 3** and **DG 4**, We introduce an AR interface that includes all the functions discussed above and additional functions such as visualization, editing, and modifying the content. The authoring system for CARING-AI consists of four modes: 1) **Task mode** to get users the step-by-step instructions, 2) **Scan mode** to ground the instructions in the context, 3) **Author mode** to design and edit textual instruction and avatar motion content, and 4) **View mode** to examine the authored AR avatar instructions. The AR menu is always present in the user's view on the left hand so that users can easily access the all functions of the current mode and also switch between them.



Step 1: Pick up the apple from the table.

Step 2: Go to the kitch nk.

Step 3: Wash the apple

Figure 9: AR User Interface of CARING-AI. In (a), users can start authoring a new task or start contextualizing the instructions. In (c), they can see the generated textual instructions and select a single or multiple of the instructions. In (b), they can choose to view the humanoid avatar animation (*Play* button), change the scale of the humanoid avatar (*Change Scale* button), modify the textual instruction by speech (*Modify Instruction* button), insert new instructions (*Insert Previous* and *Insert Next* buttons), and delete the selected instructions (*Delete* button).

As shown in Figure 9 (a), the user first starts by providing the task description using a voice command by clicking the *New Task* button to enter the **Task mode**. The user speaks to the system to specify their task, then the system generates textual instructions shown in the instruction panel Figure 9 (c). When the users select one step from the panel, they can insert new instruction steps, delete the selected ones, or modify them.

Then, the user selects and groups several steps that happen in a global context (i.e., steps that happen at the same location in the space, for example, in Figure 9 (c), *Step2: go to the kitchen sink* and *Step3: wash the apple* belong to the same global context), with the selected instructions highlighted in yellow. After selection, the user clicks the *Contextualize* button and enters **Scan mode** to scan the physical environment. In the **Scan mode**, the user simply walks in their physical space to mark the global location for the current group (e.g., in Figure 9, the user walks to the kitchen sink) and ends contextualizing the current group by taking a screenshot while looking at the contextual environment (e.g. looking at the sink with the apple and knife visible in the scenario). Upon object detection, CARING-AI then overlays 3D virtual objects on the corresponding physical objects which users can see and adjust the 6 DoF with built-in freehand interactions. Iteratively, the user groups and contextualizes the rest of the instructions. The contextualized instruction panels are highlighted in green while the users are still allowed to revisit and edit.

Upon the completion of contextualizing all steps, the user enters the **Author mode**. The user can click the *Modify Instruction* button to modify the instruction and regenerate animation for a specific step, or click the *Change Scale* button to change the visualization scale of the selected step. The available scales of the visualization are full-body avatars and hand-object avatars (i.e. only the hands, the forearms, and the objects are rendered).

Meanwhile, the user can enter **View mode** by clicking the *Play* button. This mode visualizes the currently selected instruction by rendering the generated context-aware avatar animation in the HMD.

4.6 Software and Hardware Implementation

We implement CARING-AI using Hololens 2 [79] with built-in SLAM tracking for AR experiences. CARING-AI interface was developed in Unity 3D on a local PC (Intel core i7-9700K CPU, 26 GHz, 128 GB RAM). During the scanning mode, we use a resolution of 1280 x 720 for the RGB image. The images are then processed in a local PC for object detection and the 6 DoF estimation algorithm for overlaying the virtual 3D on the real object. We used the Mixed Reality toolkit (MRTK) for the interactions of hands with the virtual objects and the interface. For 6 DoF of the object, we used the pre-trained MegaPose6D [62] model, which can estimate 6 DoF of objects in the wild. For object detection, we used the detection model [101] pre-trained on ImageNet [26]. We fine-tuned the object detection algorithm which is used in finding the spatial context for the content. The training of object detection was performed on objects dataset collected for used cases and user study purposes. For each object class, we collected 600 images. The 3D scans of the objects were also collected and stored in the database for the 6DoF algorithm and virtual object overlays in physical. As mentioned in section 4.4, we used the pre-trained Guided Diffusion Model [55] as the motion generation model on the HumanML3D [39] dataset. The action classes from the dataset are further used in the user study and for the demonstration. One batch of motion generation takes time of 36 seconds, with one NVIDIA RTX A6000 GPU.

5 Quantitative Evaluation

In this section, we assess the efficacy of our context-aware generative AI approach in real-life scenarios by comparing it with a baseline (GMD [55]). As a preliminary step, we evaluated our modified diffusion model algorithm Figure 7 compared to GMD quantitatively. We chose GMD as our baseline for comparison because GMD is a state-of-the-art model based on MDM. This study assesses the modified model's performance in generating humanoid animation, which is the backend algorithm of our system.

5.1 Evaluation

5.1.1 Baseline. We used a pre-trained model of GMD to compare our algorithm. GMD [55] is pre-trained with the HumanML3D dataset, which is annotated human motion data. The dataset has 22 joints $|\mathcal{J}| = 22$ following the skeleton representation of the HumanML3D dataset [39]. The HumanML3D dataset encompasses 14,616 motions, paired with 44,970 descriptions that are comprised of 5,371 unique words. The combined duration of all motions is

Table 1: Task and instructions

Task	Instructions
Charging a Phone	Get the charger; Insert the cable into the phone; Plug the charger into an outlet
Turning on the TV	Pick up the remote; Point it at the TV; Press the power button
Closing a Window	Approach the window; Grasp the handle or sash; Push to close
Starting a Computer	Sit in front of the computer; Press the power button; Wait for it to boot up.
Exercising	Crawl; Run; Band Push; Crawl to Stand
Reading a Book	Walk to the bookshelf; Choose a book; Go to the living room; Sit on the couch or chair;
Closing a Window	Approach the window; Grasp the handle or sash; Push or slide to close
Eating an apple	Approach to the table; Pick up the remote; Eat the apple; Move back; Turn around; Leave the kitchen
Use a 3D printer	Pick up PVA; Go to printer; Attach Filament to printer; Start printer
Making Tea	Boil the water; Place a cup on the table; Pick the pot; Pour boiling water into the cup.

28.59 hours. On average, each motion spans 7.1 seconds, and each description contains 12 words.

5.1.2 Metrics. To validate the performance of our model, we constructed 10 practical scenarios Table 1 using both the baseline method [55] and our context-aware approach. Our evaluation has been done in two dimensions: spatial and temporal context awareness. For assessing temporal context awareness, we quantified the motion discontinuity between consecutive instructions. A heightened awareness of the temporal context by the AI should result in reduced discontinuities in the generated instructions. The motion distance across frames was computed following the [3]. We calculate the transition distance, which calculates the joint distance of two transition frames.

$$d_{\text{temporal}} = \frac{1}{|\mathcal{K}||\mathcal{J}|} \sum_{K \in \mathcal{K}} \sum_{J \in \mathcal{J}} ||J_{K^{last}} - J_{K^{first}}||_2, \tag{3}$$

where $\mathcal{K}$ is the set of the two consecutive indices of transition frames $(K^{last}, K^{first}) \in \mathbb{N}^2$, which is composed of the last frame of the previous action K^{last} and the first of the next action K^{first}. The number of transitions is equal to substituting one from the number of instructions $|\mathcal{K}| = |\mathcal{A} - 1|$. $\mathcal{J}$ is the set of joints, containing the 3D location of joints at the transition, $J_{K^*} \in \mathbb{R}^3$ as elements. The human skeleton data we used has 22 joints $|\mathcal{J}| = 22$ following the skeleton representation of the HumanML3D dataset.

In terms of spatial context awareness, we gauged the proximity between the avatar and the object specified in the instruction. The absence of spatial context often results in instructions that position the avatar at a considerable distance from the target object, potentially leading to user confusion. We employed the mean Euclidean distance to measure the spatial alignment within the frames of interest.

$$d_{\text{spatial}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} ||J_t^{xy} - O_t||_2, \tag{4}$$

where $J_t^{xy}, O_t \in \mathbb{R}^2$ indicates the 2D X, Y coordinates of the root joint and target keypoint at the t-th frame, respectively. $\mathcal{T}$ is the set of frames that is spatially conditioned by target keypoint O_t.

Method	Transition distance↓ (m)	Absolute distance↓ (m)
GMD	0.15	0.08
Ours	0.03	0.09

Table 2: Transition Distance is the comparison of the discontinuity with and without temporal smoothing methods. The lower the better. Absolute distance is the average distance between the avatar and the key points. Distance under 0.1 m is considered as plausible motion [55]

5.2 Procedure

To evaluate our developed algorithm performance, we choose 10 practical scenarios Table 1 commonly found in real-world tasks. These tasks have more than two instructions and are performed at varied locations covering our design space which makes them suitable for evaluating our algorithm and comparing it with the baseline. To get the instructions for the task, three authors individually provided the task description to ChatGPT and noted the instructions. Then the authors discuss to finalize the steps of the instructions. Additionally, one of them wears hololens to get the spatial context for the algorithms. After generating the text instructions, we input them into a Text-to-Motion generator, resulting in motion instructions. To evaluate our approach, we compared our motion instructions with those from the GMD[55], one of the state-of-the-art algorithms in Text-to-Motion generation. For a consistent comparison, we kept the length of each instruction the same in 90 frames.

5.3 Results and Analysis

In this section, we detail the results of our preliminary evaluations. We highlight the transitional gap between two consecutive frames measured in meters (m). As illustrated in Table 2, our approach ensures smooth frame transitions. GMD [55] exhibits a transition distance of 0.15m when frames are simply concatenated. In contrast, our method substantially decreases this transition distance to 0.03m ($p < 0.05$), eliminating any motion discontinuity. Additionally, Table 2 showcases the spatial alignment. The distance is determined between the avatar's center and the guided keypoint, assuming an avatar height of 175cm. Our method produces results closely aligned with GMD, generating plausible motion with an error margin under $0.1m$ ($p < 0.05$) [55], while also capable of producing smooth and varied actions in one seamless operation. The *Exercising* task shows the highest spatial error because it contains the instruction to *run*, which represents the most sudden motion among all instructions. Meanwhile, the *Starting a Computer* task has the lowest error due to its fewer movements. We observed that the quality of hand motion generated by both GMD and our method is subpar. Instructions involving hand-object interactions especially exhibit awkward hand gestures. For instance, the *pickup* motion doesn't adequately display grabbing gestures.

6 User study 1: Usability

We conducted a user study to qualitatively evaluate the usability of our system. We qualitatively evaluated all the steps used in our system as well as the quality of the human motion generation. We invited 12 users (10 males and 2 females) from the technical university. All of them have prior experience in AR/VR applications using tablets, AR screens, and head-mounted devices. CARING-AI is designed to help both experts and non-experts create human avatar motion. The users are from graduate and undergraduate programs and their age ranges from 19 to 30. None of the users have used our system and have had no knowledge about it before. The entire study took one hour - two hours and each user was compensated with a 15 USD e-gift card. The study was conducted in an indoor environment. After the user arrived, we provided a brief overview of the study. Then the users were asked to sign the consent form only when they were comfortable in performing the user study. After that, we explained the entire CARING-AI system workflow and each function in the UI. Out of 12 users, 5 users had prior experience with developing or using Hololens. The users were given enough time to get comfortable with the CARING-AI system before the study was officially started. Also, some of the users with no experience were provided with a built-in Hololens tutorial to learn the basics. As our study focus was on the usability of the system and user experience on the generated content, we asked the users to complete a System Usability Scale (SUS) and a 5-point scale Likert-type questionnaire followed by 20-minute post-session conversation-type interviews to provide subjective feedback about CARING-AI.

6.1 Procedure

We evaluated the performance of our system and let users generate avatar motions for the tasks **Cutting an Apple**. The study took place in a kitchen environment. The task was chosen because it involves multiple steps and different locations. The task is suitable for evaluating context-generated content and other system components like interface. The users were tasked to generate step-by-step instructions for the task from ChatGPT. The most common steps found in the task as shown in Figure 10 are

(1) Walking to the cut board (Global, Temporal and Spatial),
(2) Pick up an apple on the table (Local, Temporal and Spatial),
(3) Pick up the knife (Local, Temporal and Spatial),
(4) Cut the apple (Local, Temporal and Spatial),
(5) Put down the knife (Local, Temporal and Spatial),
(6) Eat the apple (Local, Temporal),
(7) Move back (Global, Temporal),
(8) Turn around (Global, Temporal),
(9) and Leave the kitchen (Global, Temporal and Spatial).

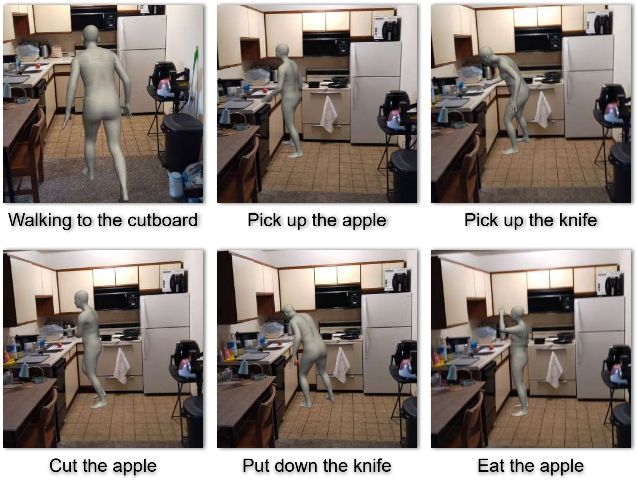


Figure 10: Examples of humanoid animation generated in User Study 1.

Then the user scans the environment and takes the screenshots at different locations. After that user aligns the virtual objects on the real object if they are not properly aligned by the system. And finally, the user uses the CARING-AI interface to generate the motions.

6.2 Results and Analysis

We analyzed responses to a 5-point scale Likert-type questionnaire, SUS, and transcribed the interview from the user.

6.2.1 Textual Instructions. We qualitatively evaluate textual step-by-step instruction generated from ChatGPT. In general, users preferred the step-by-step instructions generated from the ChatGPT to be relevant to the task. *"P1: I think I don't need to modify the instructions. They were correct and right for the task."*. However, some users modified a few steps little for their instructions. Many users acknowledge the visualization of a graph representation of step-by-step instructions and agree that the interactions with the graph are easy to use and simple to follow (Q1: AVG = 4.08, SD = 1.00). *"P5: The process of creating the instructions was easy and quick."*

6.2.2 Context Aware Instruction by Avatar. Through post-study interviews and designated Likert-scale questionnaires with the users, we qualitatively evaluate the context in the content generated by CARING-AI during the user study. Many of the users stated that the avatar was performing the actions with the object at the correct location (Q7: AVG = 4.42, SD = 0.51). As a piece of evidence, P3 commented in the interview *"P3: I was actually surprised by the way Avatar went to the exact position and performed the activity."* Another user mentioned *" P2: I liked that I could see the avatar move towards the apple and the fluid and connected motion"*. The majority of the users were satisfied by the actions performed by the avatar using the virtual objects (Q4: AVG = 4.25, SD = 0.97). As P9 commented

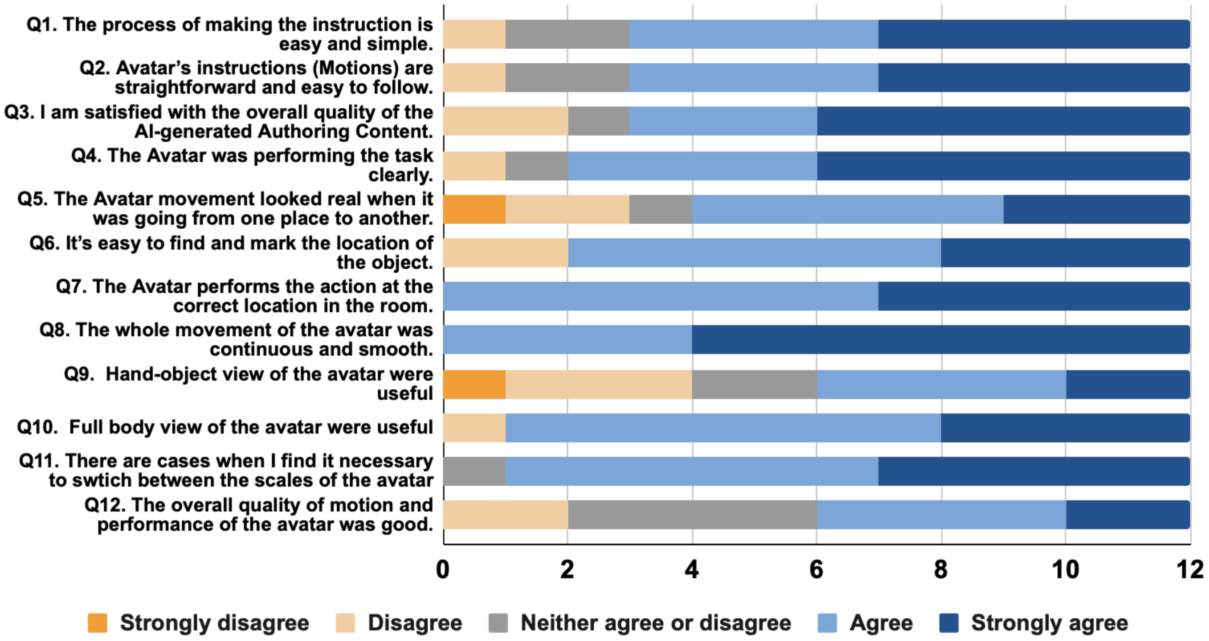


Figure 11: Likert-type questionnaire results from User Study 1.

in the interview on *"P9: The action demonstrated by the avatar was with the right object." However, a few users raised concerns about accurate avatar hand and virtual object interactions, such as P12 "P12: It is not clear to me why the hand was not grabbing the object and it automatically sticks to the hand."* We discuss this limitation in more detail in section 9. Users acknowledge that the motion of the avatar from one place to another looks real (Q5: AVG = 3,58, SD = 1.31), such as P11 *"P11: I can't believe that the avatar movement exactly looks as if a real human is walking. I should say this is too cool."* Users found a smooth transition of the avatar motion between the instructions (Q8: AVG = 4.67, SD = 0.49). P8 pointed out that the transition by our smoothing algorithm made the animations seamless and the breaks between animations hard to identify, *"P8: It was hard for me to draw the boundary between the instructions when I was looking at the avatar motion."*

6.2.3 Overall System Usability and Utility. The overall system Likert scale results are shown in Figure 11. Context from the user is the foundation of our generated animation and most of the users were satisfied and comfortable with taking screenshots during the scanning of the environment process (Q6: AVG= 4.00, SD = 1.04), as P8 commented *"P8: I didn't find any difficulty in moving around and taking screen pictures."* Further, users also found the alignment of real and virtual objects was accurate. P4 commented that the accurate alignment contributed to their overall experience *"P4: I think the virtual model was approximately over the top of the real for many objects and the visualization being a 3D rendering definitely helps my experience".* The CARING-AI system interface was appreciated by the users. The positive feedback from the users on the usability of the interface is mainly attributed to the easiness of using it, as P2 commented: *"P2: In my opinion, I find the UI very straightforward and easy to use."* Moreover, users find it easy to switch between full-body pose and only hand (Q2: AVG= 4.08, SD = 1.00). The final avatar motion instruction generated from CARING-AI received a positive response from the user after watching the final generation of instructions (Q12: AVG= 4.42, SD = 0.51).

Regarding utility, many users reported positive regarding utilizing CARING-AI in creating AR instruction tutorials of human demonstration. P7 with previous experience of authoring AR instructions in Unity positively commented on the efficiency when utilizing CARING-AI *"P7: I have developed an AR instruction by coding in Unity and it took me several days to make it. I wish this thing was developed earlier so that I could have used it."* Some users needed more features to display such as text, and icons for object movement directions along with just demonstrations, such as P9 *"P9: For the base level, it is okay but I think it would have been better if your system provided visual cues showing the movement of the object"* We discuss the limitation in more detail in section 9. For the system usability, the users agree that the system is usable (SUS: M = 83.21 out of 100 and SD = 7.34). A score above 70 is practically considered "Good" usability and an 85-and-beyond score is considered "excellent" as mentioned in [4, 5].

7 User Study 2: Interaction

To evaluate the interaction design of our system compared wt the baseline programming by Demonstration, we conducted an additional within-subject comparative user study (N=12) between CARING-AI and a baseline PbD method. The purpose of this study is to assess the novel interaction proposed in CARING-AI and compare the user feedback on the interactions with that from the existing methods (PbD). To make a reasonable PbD baseline, we followed a similar approach [102] and built our setup, where the humanoid animation is captured by a third-person-view RGB camera.

The participants (8 males and 4 females) are six novices and six experts in developing AR applications. They were recruited and compensated as in User Study 1. Users were asked to author AR instructions with both CARING-AI and PbD, counterbalanced by 6 participants authoring with CARING-AI first followed by PbD, and the other 6 participants with PbD and then CARING-AI. The entire study took 1 to 2 hours. We followed the same protocol for explaining our system as in User Study 1. To quantitatively evaluate our system, We asked the users to fill out NASA TLX [85] and a five-point Likert-type questionnaire (Figure 17). This questionnaire is designed to collect qualitative evaluations of the users on the efficiency and accuracy of both authoring methods (Q2-5), as well as the quality of the final output (Q1). Additionally, a 15-minute semi-structured interview was conducted for each participant. In post-processing of the study data, we calculated error rates during interactions and time spent in creating animation.

7.1 Procedure

The user study was performed in a living room and users were asked to perform three tasks: organizing the living room, watering a plant, and hammering a nail to a door. We specifically chose tasks that require human motion that can be guided by humanoid avatar animation in AR. Also, all chosen tasks involved hand-object interactions with different objects and took place at various global locations, which are:

(1) Walk to the sofa. Place the water bottle on the sofa. Pick up the book.
Walk to the chair. Put down the book on the chair.
(2) Pick up the mug. Walk to the plant. Pour the water into the plant.
(3) Walk to the door. Pick up the hammer. Hammer the nail on the door.

With CARING-AI, the users first scanned the environment by moving around and taking screenshots of the locations where local actions were to happen. After that, users aligned the virtual object with their real counterparts. Then, users generated the final animated instructions following the workflow of CARING-AI as in User Study 1. Until satisfied, the users could regenerate or adjust the animation with CARING-AI. The examples of generated AR animation with CARING-AI in this study are shown in Figure 12.

With PbD, the users first manually aligned the virtual objects with their real counterparts. Then, users wrote down the instructions for each task and performed the task in the environment. During this process, the user's actions were recorded by four camera setups, each capturing one global location in the tasks (sofa, chair, plant, and door). We followed prior work [48] to calibrate the camera setups and align them with the AR HMD to obtain accurate camera coordinates. The recorded videos are then passed into a video-to-3D algorithm [84] to convert the demonstrated motion into presentable 3D humanoid animation assets. To execute the video-to-3D algorithm, users are first required to segment both the human and the object using the segmentation module from [60]. The users then situated the animation assets in AR with an HMD, by moving the assets to align with the physical environment. Until satisfied, the users could redo the tasks and adjust the animation assets. The examples of generated AR animation with PbD in this study are shown in Figure 13.

For fair comparison of the avatar quality, both PbD and CARING-AI used the full SMPL-X [90] model as the humanoid avatars as shown in Figure 12 and Figure 13 (c)-1.

7.2 Results and Analysis

We obtained the data from the user study, including (1) the Error Rates (We manually counted the number of times each user modified the instruction, re-performed a task, or re-adjusted the animation assets), (2) the time performance in minutes taken by each user to complete the authoring tasks, and (3) NASA TLX scores. We then confirmed if the normality assumption is followed in each collected data group with a Shapiro-Wilk test, followed by a paired t-test if normally distributed, or a Wilcoxon Signed-Rank test otherwise. We then analyzed and discussed the results as follows.

7.2.1 Task Load: CARING-AI v.s. PbD. Since only data from Effort scores are normally distributed in both PbD and CARING-AI setups ($p_{PbD} = 0.051$, $p_{Ours} = 0.159$, henceforth, we conducted paired t-tests for Effort and Wilcoxon Signed-Rank tests for the rest, as shown in Figure 14. The results showed that users experienced significantly less Mental Demand with CARING-AI ($M_{Ours} = 2.666$, $SD_{Ours} = 0.651$, $M_{PbD} = 3.250$, $SD_{PbD} = 0.965$, $p = 0.025$,$z = -2.242$) compared to PbD. Also, users reported significantly higher Physical Demand ($M_{Ours} = 2.416$, $SD_{Ours} = 0.514$, $M_{PbD} = 3.333$, $SD_{PbD} = 1.073$, $p = 0.046$, $z = -2.001$) in PbD than in CARING-AI. The less Mental Demand with CARING-AI can be attributed to a shorter workflow with no consideration of the camera position (as we will also discuss in the next subsubsection), while the less Physical Demand with CARING-AI can be attributed to the physical easiness of creating animation with only text instructions compared to that of demonstrating the actions to the cameras. Additionally, users felt more confident in their performance in completing tasks with CARING-AI ($M_{Ours} = 3.833$,$SD_{Ours} = 0.834$,$M_{PbD} = 2.916$, $SD_{PbD} = 0.996$,$p = 0.026$,$z = -2.222$). The better performance scores can relate to the less Error Rates and shorter task time in the next subsubsection. No significant differences were observed in Temporal Demand ($z = -0.560$,$W = 14 > W_{critical} = 3$) or Frustration ($z = -0.280$,$W = 16 > W_{critical} = 3$) between the two systems.

7.2.2 Error Rate and Time Performance. The collected data was normally distributed in both Error Rate ($p_{PbD} = 0.515$, $p_{Ours} = 0.242$) and Time Performance ($p_{PbD} = 0.487$, $p_{Ours} = 0.987$). The reported Error Rates were high in PbD as compared to CARING-AI as shown in Figure 16 ($p = 0.034$). During the study, we mainly observed that some participants re-did the tasks with PbD multiple times because the cameras had been occluded from a proper view to generate accurate animation. In practical scenarios, this problem can worsen since the camera setup has to be relocated and re-calibrated to tackle the occlusion problem of PbD authoring. Redoing the demonstration also added much more mental and physical demand as we showed in the NASA TLX results. For the total time taken, users finished all tasks quicker with CARING-AI compared to PbD ($p = 0.001$). This was because performing the actions took longer as

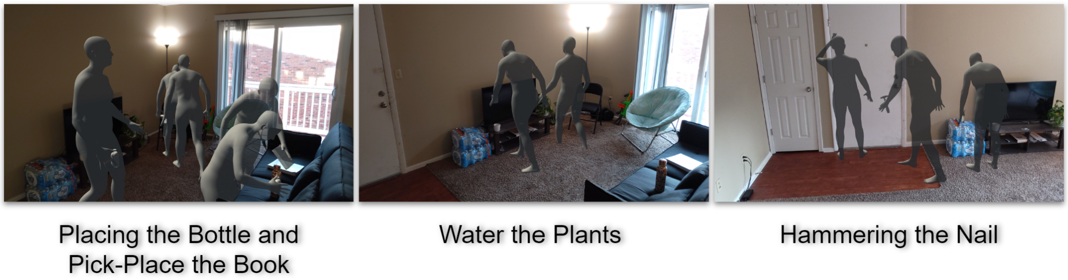


Figure 12: AR animation generated by users using CARING-AI in User Study 2.

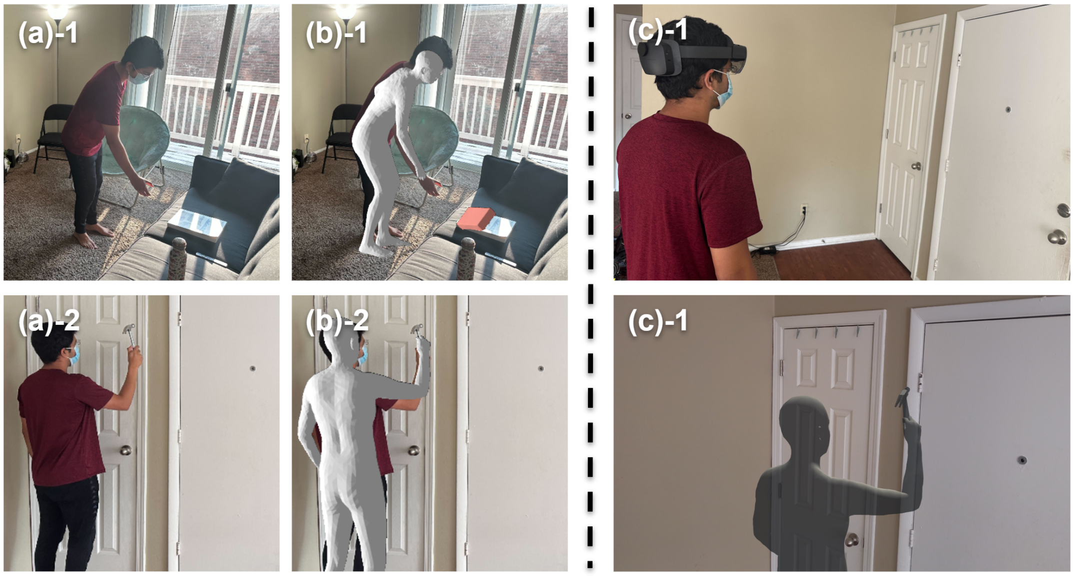


Figure 13: AR animation generated by users using our PbD baselines in User Study 2. (a) users demonstrating the task, (b) the generated 3D animation assets from the camera captures, and (c) the users viewing and adjusting the 3D animation assets with the AR HMD. The differences between the assets in (b) and (c) are due to rendering methods.

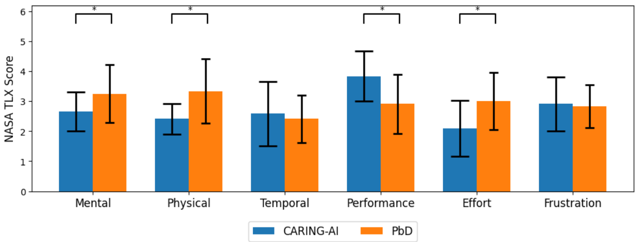


Figure 14: NASA TLX Scores, where * denotes $p < 0.05$

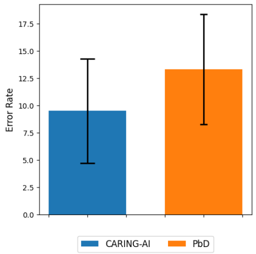


Figure 15: Average Error Rates Calculated in User Study 2, with CARING-AI and PbD

compared to adjusting the text or the animation itself. Also, more Error Rates meant more numbers of times re-demonstrating.

7.2.3 Subjective Ratings. We analyzed the questionnaire results from users' feedback and conducted interviews with the participants (Figure 17). After confirming the normality of the rating data, we further performed paired t-tests to check the significance of the comparison. Users preferred the quality of the final animation instruction generated from CARING-AI(Q1: $p < 0.05$), as P7 commented in the interview *"I like the overall animation quality from the first system (Ours)" (P7).* Users found editing animation through text much easier than demonstrating (Q4 $p < 0.05$). *"When I create*

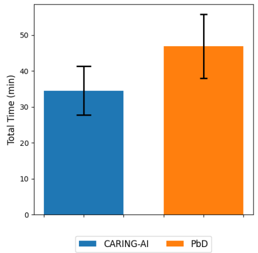


Figure 16: Average Time Spent in Authoring Task, with CARING-AI and PbD

the instruction and I don't like it, I would prefer some easier way to edit like the second system (Ours) than performing the task again" (P2). This aligns with the results from NASA TLX and quantitative evaluations. The easiness of authoring with CARING-AI can also be attributed to the smoother learning curve of text-to-animation models as compared to PbD or demonstration-to-animation methods as pointed out in [113]. Users reported better feedback regarding hand-object rendering in CARING-AI (Q5 $p < 0.05$) as compared to PbD, attributed to the error in detecting hand-object interaction in PbD, which results in the degenerated user experience as described by P3 *"I don't know but the hand was not actually grabbing the object in the first system (PbD) but in the second system It was much better" (P3).* The quality of hand-object rendering was partially influenced by the occlusion during the study. Also, the rendered interactions in PbD were not situated in the environment correctly, and users reported that they had adjusted the animations more in the final stage. We found no significant difference in controlling the avatar (Q2) and editing the textual instruction (Q3).

8 Discussion

8.1 Contextual Awareness in CARING-AI

Context plays a vital role in AR applications [121]. Despite the potential of Gen-AI in creating 3D content [118], a key limitation lies in its lack of contextual awareness in AR as identified in the preliminary study. One of the core design goals of CARING-AI is to bring context awareness to AI-generated humanoid avatar instructions in AR. With the evaluation of the CARING-AI (section 6 and section 5), we look back at the conclusion drawn in the preliminary study and seek the reason why context-awareness is necessary in AR instruction, i.e., *What does context-awareness bring to AR instructions?* We highlighted how users in the study emphasized the importance of precise positioning and action performance in humanoid avatar animations (subsubsection 6.2.2) and demonstrated how CARING-AI effectively addressed the need for accuracy in both positioning and animation. Our findings indicate that users, as the authors of the instructions, they are aware of the context. The actions that they intend or anticipate the receivers of the instructions to take are based on the context, i.e., the authors give instructions based on the context. *"P3: When I wanna instruct somebody to do something, I want them to know exactly the objects and actions. This is very important when in a complex task where students can pick the wrong stuff and act anyway and make mistakes."* The author's context-awareness is the essence of instructions, as many prior works pointed out as "the prior to function as an instruction" [127] or "the flexibility and accommodation to external constraints for designing an instruction" [82]. In short, there are specific *"where"* and *"what"* the authors intend to convey in the instructions. By preserving and representing the author's context awareness, CARING-AI enables the core functions of instructions, which were previously missing in AI-generated humanoid avatar animation.

8.2 CARING-AI Excels in Instructing: *How*?

Given the capability of CARING-AI preserving and representing context-awareness in AI-generated humanoid avatar animation. We have witnessed the preference and positive ratings of the users for both our full-body avatar and hand-object avatar animation. Yet, we noticed some comments from the users addressing the necessity of using humanoid avatar instructions in some scenarios. Some users mentioned inconvenience brought by the use of humanoid avatars. *"P8: The avatar moving backward was not visible to me. When the movement [of the avatar] is out of my vision, I think there are better ways to tell me to look at the avatar or tell me what to do."* In addition to *"where"* and *"what"*, *"how"* the instructions can be conveyed to the users is also important. As discovered in prior works [13, 46], in AR tutoring, learners prefer half-body avatars for spatial interactions (interactions that require large spatial navigation before proceeding), full-body avatars for body-coordinated interactions (interactions that require coordination among learners' body, hands, and eyes). We further bring hand-object avatars for local interactions in CARING-AI. CARING AI changes form of avatar based on scale of the task. However, some of the users mentioned that avatar forms should be based on designation and details of the actions. Furthermore, as AR instructions are not limited to the form of humanoid avatars, we conclude that non-avatar AR can also be included in the CARING-AI pipeline as a means of visualization. As P8 commented, non-avatar AR instructions excel avatar instructions in the cases where no particular body gesture is required or the visualization of humanoid avatars is not visible to the users. This conclusion highly aligns with the findings of prior study [13].

8.3 Other Modalities of AI-generated Instructions

Humanoid avatar motion along with additional cues helps in learning content [46]. As previously discussed, to further develop CARING-AI into a comprehensive AR instruction system, we envisioned future versions with other AI-generated modalities such as (1) visual cues [67, 70], e.g. arrows, bounding boxes, lines, etc., (2) contextualized textual instructions [18], (3) images [72, 94], (4) audio [74], and (5) videos [31]. We argue that our design space of AR instructions and the pipeline of CARING-AI apply to the other modalities of AR cues and instructions as well, since the spatial locations/placements, as well as the temporality of the cues, are key design considerations in the prior works referenced above, and can be situated through CARING-AI's pipeline, where the authors walk through the context and assign the cues by taking contextual snapshots.

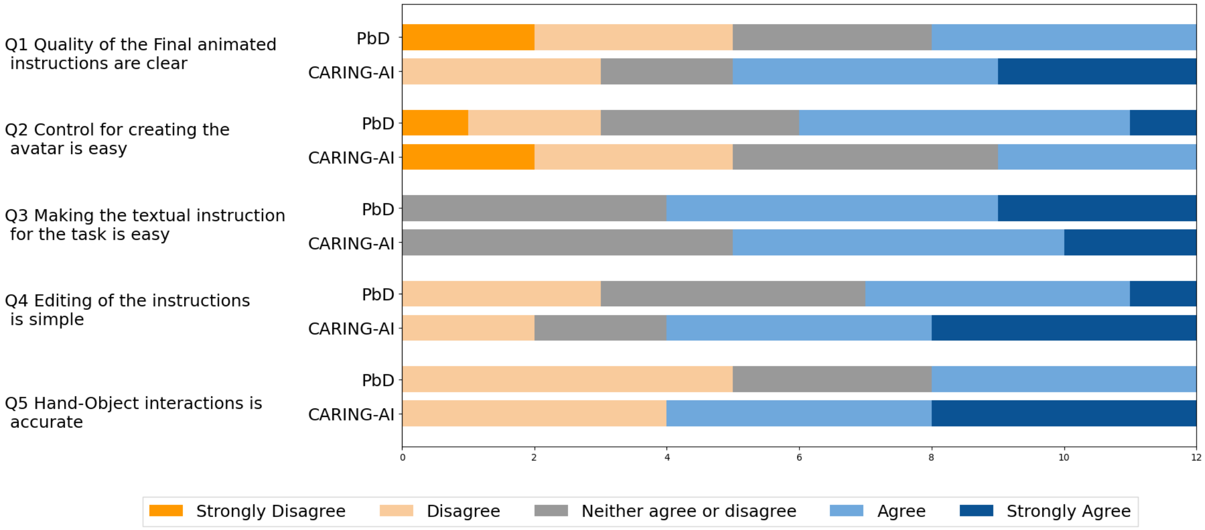


Figure 17: Subjective Likert-scale Ratings of the easiness, animation quality, and interactions of CARING-AI and PbD

8.4 Authoring with CARING-AI v.s. PbD

Authoring by real-time demonstration or embodied often requires bulky hardware setup, which limits the mobility of the end-users due to the size of the devices [121]. In our User Study 2, a camera setup has to be built for the baseline PbD method, while CARING-AI does not require complex hardware setup and allows users to create instructions without programming knowledge and physical presence. CARING-AI can help users generate instructions even at remote locations without performing the actions and movements (subsection A.3). As P7 commented on the efficiency in creating AR instructions *"this system, I think, can become very effective in creating remote instructing like without even physically present."*. Instructions from PbD are also tied to the environment or context in which users performed demonstrations to create the content. On the other hand, instructions generated from CARING-AI can be effectively adapted to various environments settings because of context-aware modeling, concluded from our observations of the user feedback on the quality of the generated content as shown in Figure 17. As the results suggest, the CARING-AI pipeline performs better than PbD in generating instructions with fewer mistakes and faster and with less cognitive load.

9 Limitations, and Future Work

In this section, we discuss the limitations of CARING-AI identified from development, user study, and the analysis of the study process. Deriving from such, we propose future directions that can contribute to the topic of GenAI in AR instructions.

9.1 Object Representation and Interactions

One of the limitations of CARING-AI lies in its ability to handle complex hand-object interactions. We apply an additional module to render hand-object interactions, which focuses solely on visualizing the hand and object rather than the entire body. However, this module tackles only rigid objects and does not render high-fidelity hand-object interaction, particularly limiting the use cases requiring complexity and dexterity in manual tasks. Complex interactions involve hands engaging with objects that are articulated, segmented, foldable, or deformable, whereas the objects in our system are strictly rigid. Thus, the system cannot represent actions that involve changing the shape or form of an object such as tying a shoelace or folding a cloth. Nevertheless, such constraints are attributed to the limitation of the Gen-AI algorithms applied in the pipeline, while the overall workflow of CARING-AI remains effective in capturing and presenting the context information to the generated content. While algorithmic development remains relatively unexplored in the AI field, we foresee this limitation can be addressed in future work by incorporating more generalized state-of-the-art algorithms and datasets, such as those introduced in [30, 75, 134], to enable high fidelity rendering of hand-object interactions in AR instruction.

Lastly, CARING-AI currently does not support object-object interactions. This limitation stems from the aforementioned challenges in hand pose plausibility and rigid object representations. Without the ability to depict detailed hand-object interactions and object articulations, representing interactions between multiple objects becomes unfeasible. However, we believe that exploring object-object interactions offers a promising direction for future research, providing a richer and more comprehensive understanding of interactions in virtual environments.

9.2 Generalizability

Like all other deep-learning-based methodologies, the performance of our motion generation model is subject to the training process [52, 53]. Nevertheless, the model we used has been pre-trained on a large-scale motion dataset [39] containing 14,616 motions and 44,970 descriptions composed of 5,371 distinct words, which fulfills the requirement for our use cases and study. In our study, we emphasize the HCI design and the workflow bridging AR applications and Gen-AI, rather than contributing to the existing algorithms of Gen-AI by trying to outperform them. To this end, we further argue that the generalizability (more types) and scalability (more detailed motion) of this method are promising. Firstly, prior works have demonstrated the capabilities of large generative models on large-scale datasets [103]. We envision the scale of this method will be further improved upon datasets with wider ranges of action labels being fed into the Gen-AI model (e.g. task-specific motion datasets in each domain). Secondly, the size and complexity of the model in our implementation are constrained by our hardware condition, particularly the GPU sizes. With a better (empirically more complex) model, we expect the quality and the details of the generated content to improve.

Put simply, our methodology focuses on the HCI design for AIGC in AR, maintains its applicability with the current ideology of Gen-AI, and is generalizable as long as the plugged-in Gen-AI is generalizable.

In addition to the generalizability of the algorithm, we also acknowledge that the findings of the formative studies are derived from academic researchers, which could be further refined and expanded with diverse perspectives from industrial practitioners. Moreover, the avatars used in our paper are sex-neutral, however, unclothed human avatar representations from [90], which can be replaced with inclusive and realistically rendered avatars for more user-friendly and family-friendly use.

9.3 Software and Hardware Constraints

One of the major constraints imposed by our hardware condition is the time performance. It has been reported in subsection 4.6 that generating a batch of motion takes 36 seconds (i.e. anything between 1 and the batch size takes 36 seconds). Even though 36 seconds of latency seems beyond the cost of real-time performance, batch processing guarantees that users can render their desired avatar instructions once altogether, given a batch size of 128 in our setup, which is, in all cases of our study, more than the users' expectation of the number of interactions in the demonstration. Moreover, to address this problem of computational cost in the future, we anticipate methodologies such as utilizing cloud services for data transferring and computation, parallel programming for the generation, and usage of better GPUs (high computational).

Another one of our hardware constraints comes from our implementation platform, Hololens 2. With a field of view (FOV) of (43°×29°), users cannot experience a fully immersive AR environment as content might not be visible outside this boxed area. For AR authoring and consuming, this poses a challenge. Users have to be acutely aware of this constraint to ensure that critical interactive elements or information are positioned within this limited space. Under the circumstances when the humanoid avatar is close to the users, motion outside the FOV is not visible. This problem may not influence the quality of the generated content itself but induce biases in the evaluation of the user study, such as more negative feedback due to the jeopardized user experience.

10 Conclusion

In this work, we present CARING-AI, an AR authoring system that enables users to author AR instructions with contextualized humanoid avatar movement generated by Gen-AI. We first discussed with experts in AR authoring in a preliminary interview, aiming to identify the gap between current AI-generated humanoid avatars and AR instruction applications. Based on the insights gained from the discussion, we further characterized the design space for context-aware AR instructions from AI-generated content with two dimensions, namely context (spatial or temporal) and content (local or global). We then proposed a workflow for contextualizing AI-generated AR instruction with three major steps: (1) generating and modifying textual instructions, (2) contextualizing by traversing and scanning the environment, and (3) generating and smoothing humanoid avatar animation. We further showcased three application scenarios for authoring AR instructions with CARING-AI: asynchronous, remote, and ad hoc instruction. We evaluated the performance of CARING-AI with a preliminary quantitative evaluation focusing on the model performance and the quality of the AIGC, followed by a user study evaluating the qualitative performance and overall usability of CARING-AI as an AR authoring system through complimentary qualitative user feedback. Eventually, we discuss the limitations of the current version of CARING-AI and further envision the opportunities and promising future research directions our work has revealed. We believe our work is capable of opening up and contributing to the discussion of the broad topic of AIGC in AR applications.

Acknowledgments

We wish to thank all the reviewers for their invaluable feedback. This work is partially supported by the NSF under the Future of Work at the Human-Technology Frontier (FW-HTF) 1839971. We also acknowledge the Feddersen Distinguished Professorship Funds and a gift from Thomas J. Malott. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency.

References

[1] 2023. Midjourney. https://www.midjourney.com/ Accessed: 2023/08/02.

[2] Fahad M Alblehai. 2022. Individual Experience and Engagement in Avatar-Mediated Environments: The Mediating Effect of Interpersonal Attraction. *Journal of Educational Computing Research* 60, 4 (2022), 986–1007.

[3] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. 2022. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*. IEEE, 414–423.

[4] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

[5] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[6] Majid Behravan and Denis Gracanin. 2024. Generative Multi-Modal Artificial Intelligence for Dynamic Real-Time Context-Aware Content Creation in Augmented Reality. In *Proceedings of the 30th ACM Symposium on Virtual Reality Software and Technology*. 1–2.

[7] Daniel Black. 2017. Why can I see my avatar? Embodied visual engagement in the third-person video game. *Games and Culture* 12, 2 (2017), 179–199.

[8] Doug A Bowman, Joseph Gabbard, Daniel Auerbach, Nazila Roofigari-Esfahan, Kathryn Britt, Cory I Ilo, and Keerthana Adapa. 2022. BuildAR: a proof-of-concept prototype of intelligent augmented reality in construction. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 508–512.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[10] Jacky Cao, Xiaoli Liu, Xiang Su, Sasu Tarkoma, and Pan Hui. 2021. Context-aware augmented reality with 5G edge. In *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 1–6.

[11] Yuanzhi Cao, Anna Fuste, and Valentin Heun. 2022. MobileTutAR: a Lightweight Augmented Reality Tutorial System using Spatially Situated Human Segmentation Videos. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–8.

[12] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226* (2023).

[13] Yuanzhi Cao, Xun Qian, Tianyi Wang, Rachel Lee, Ke Huo, and Karthik Ramani. 2020. An exploratory study of augmented reality presence for tutoring machine tasks. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[14] Yuanzhi Cao, Tianyi Wang, Xun Qian, Pawan S. Rao, Manav Wadhawan, Ke Huo, and Karthik Ramani. 2019. GhostAR: A Time-Space Editor for Embodied Authoring of Human-Robot Collaborative Task with Augmented Reality. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 521–534. https://doi.org/10.1145/3332165.3347902

[15] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. 2024. Text2HOI: Text-guided 3D Motion Generation for Hand-Object Interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1577–1585.

[16] Vinay Chamola, Gaurang Bansal, Tridib Kumar Das, Vikas Hassija, Naga Siva Sai Reddy, Jiacheng Wang, Sherali Zeadally, Amir Hussain, F Richard Yu, Mohsen Guizani, et al. 2023. Beyond Reality: The Pivotal Role of Generative AI in the Metaverse. *arXiv preprint arXiv:2308.06272* (2023).

[17] Vinay Chamola, Gaurang Bansal, Tridib Kumar Das, Vikas Hassija, Siva Sai, Jiacheng Wang, Sherali Zeadally, Amir Hussain, Fei Richard Yu, Mohsen Guizani, et al. 2024. Beyond reality: The pivotal role of generative ai in the metaverse. *IEEE Internet of Things Magazine* 7, 4 (2024), 126–135.

[18] Chen Chen, Cuong Nguyen, Jane Hoffswell, Jennifer Healey, Trung Bui, and Nadir Weibel. 2023. PaperToPlace: Transforming Instruction Documents into Spatialized and Context-Aware Mixed Reality Experiences. *arXiv preprint arXiv:2308.13924* (2023).

[19] Long Chen, Wen Tang, Nigel John, Tao Ruan Wan, and Jian Jun Zhang. 2018. Context-aware mixed reality: A framework for ubiquitous interaction. *arXiv preprint arXiv:1803.05541* (2018).

[20] Long Chen, Wen Tang, Nigel W John, Tao Ruan Wan, and Jian J Zhang. 2020. Context-Aware Mixed Reality: A Learning-Based Framework for Semantic-Level Interaction. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 484–496.

[21] Hyung-gun Chi, Seunggeun Chi, Stanley Chan, and Karthik Ramani. 2023. Pose Relation Transformer Refine Occlusions for Human Pose Estimation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 6138–6145.

[22] Subramanian Chidambaram, Hank Huang, Fengming He, Xun Qian, Ana M Villanueva, Thomas S Redick, Wolfgang Stuerzlinger, and Karthik Ramani. 2021. Processar: An augmented reality-based tool to create in-situ procedural 2d/3d ar instructions. In *Designing Interactive Systems Conference 2021*. 234–249.

[23] Subramanian Chidambaram, Sai Swarup Reddy, Matthew Rumple, Ananya Ipsita, Ana Villanueva, Thomas Redick, Wolfgang Stuerzlinger, and Karthik Ramani. 2022. 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)*. 326–335. https://doi.org/10.1109/ISMAR55827.2022.00048

[24] Athanasios Christopoulos, Nikolaos Pellas, Justyna Kurczaba, and Robert Macredie. 2022. The effects of augmented reality-supported instruction in tertiary-level medical education. *British Journal of Educational Technology* 53, 2 (2022), 307–325.

[25] Shakiba Davari, Feiyu Lu, and Doug A Bowman. 2022. Validating the benefits of glanceable and context-aware augmented reality for everyday information access tasks. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 436–444.

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[28] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 8780–8794. https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

[29] Mitchell Doughty, Karan Singh, and Nilesh R Ghugre. 2021. Surgeonassist-net: Towards context-aware head-mounted display-based augmented reality for surgical guidance. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*. Springer, 667–677.

[30] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. 2023. ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12943–12954.

[31] Jianwei Fei, Zhihua Xia, Peipeng Yu, and Fengjun Xiao. 2021. Exposing AI-generated videos with motion magnification. *Multimedia Tools and Applications* 80 (2021), 30789–30802.

[32] Blender Foundation. 2023. Blender - Open Source 3D Creation Software. Retrieved 12.09.2023 from https://www.blender.org/

[33] Leonardo Frizziero, Christian Leon-Cardenas, Marco Freddi, Alessandro Grassoni, and Alfredo Liverani. 2023. Augmented reality applied to design for disassembly assessment for a volumetric pump with rotating cylinder. *Production & Manufacturing Research* 11, 1 (2023), 2199815.

[34] Alec Guerrero Gallardo, María Lucía Barrón Estrada, Ramón Zatarain Cabada, Mese Z Giannina Dalle, and Aldo Uriarte Portillo. 2022. EstelAR: an Augmented Reality Astronomy learning tool for STEM students. In *2022 IEEE Mexican International Conference on Computer Science (ENC)*. IEEE, 1–8.

[35] Epic Games. 2023. Unreal Engine. Retrieved 12.09.2023 from https://www.unrealengine.com/

[36] Michele Gattullo, Alessandro Evangelista, Vito M Manghisi, Antonio E Uva, Michele Fiorentino, Antonio Boccaccio, Michele Ruta, and Joseph L Gabbard. 2020. Towards next generation technical documentation in augmented reality using a context-aware information manager. *Applied sciences* 10, 3 (2020), 780.

[37] Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchan. 2023. ChatGPT is not all you need. A State of the Art Review of large Generative AI models. *arXiv preprint arXiv:2301.04655* (2023).

[38] Jens Grubert, Tobias Langlotz, Stefanie Zollmann, and Holger Regenbrecht. 2016. Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE transactions on visualization and computer graphics* 23, 6 (2016), 1706–1724.

[39] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5152–5161.

[40] Francisco Gutiérrez, Katrien Verbert, and Nyi Nyi Htun. 2018. PHARA: an augmented reality grocery store assistant. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. 339–345.

[41] Jaylin Herskovitz, Yi Fei Cheng, Anhong Guo, Alanson P Sample, and Michael Nebeling. 2022. XSpace: An Augmented Reality Toolkit for Enabling Spatially-Aware Distributed Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 6, ISS (2022), 277–302.

[42] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239 [cs.LG]

[43] Melynda Hoover, Jack Miller, Stephen Gilbert, and Eliot Winer. 2020. Measuring the performance impact of using the microsoft hololens 1 to provide guided assembly work instructions. *Journal of Computing and Information Science in Engineering* 20, 6 (2020), 061001.

[44] Yongquan Hu, Wen Hu, and Aaron Quigley. 2023. Towards Using Generative AI for Facilitating Image Creation in Spatial Augmented Reality. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 441–443.

[45] Yongquan Hu, Mingyue Yuan, Kaiqi Xian, Don Samitha Elvitigala, and Aaron Quigley. 2023. Exploring the Design Space of Employing AI-Generated Content for Augmented Reality Display. *arXiv preprint arXiv:2303.16593* (2023).

[46] Gaoping Huang, Xun Qian, Tianyi Wang, Fagun Patel, Maitreya Sreeram, Yuanzhi Cao, Karthik Ramani, and Alexander J Quinn. 2021. Adaptutar: An adaptive tutoring system for machine tasks in augmented reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[47] Qiuyuan Huang, Jae Sung Park, Abhinav Gupta, Paul Bennett, Ran Gong, Subhojit Som, Baolin Peng, Owais Khan Mohammed, Chris Pal, Yejin Choi, et al. 2023. ArK: Augmented Reality with Knowledge Interactive Emergent Ability. *arXiv preprint arXiv:2305.00970* (2023).

[48] Rahul Jain, Jingyu Shi, Runlin Duan, Zhengzhe Zhu, Xun Qian, and Karthik Ramani. 2023. Ubi-TOUCH: Ubiquitous Tangible Object Utilization through Consistent Hand-object interaction in Augmented Reality. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–18.

[49] Heeyoon Jeong and Gerard Jounghyun Kim. 2023. Table2Table: Merging "Similar" Workspaces and Supporting Adaptive Telepresence Demonstration Guidance. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 402–406.

[50] Dongsik Jo, Ki-Hong Kim, and Gerard Jounghyun Kim. 2015. SpaceTime: adaptive control of the teleported avatar for improved AR tele-conference experience. *Computer Animation and Virtual Worlds* 26, 3-4 (2015), 259–269.

[51] BRENNAN Jones, Y Xu, MARY ANNE Hood, MOHAMMAD SHAHIDUL Kader, and HAMID Eghbalzadeh. 2023. Using generative ai to produce situated action recommendations in augmented reality for high-level goals.

[52] Daniel Justus, John Brennan, Stephen Bonner, and Andrew Stephen McGough. 2018. Predicting the computational cost of deep learning models. In *2018 IEEE international conference on big data (Big Data)*. IEEE, 3873–3882.

[53] Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, et al. 2019. A study of BFLOAT16 for deep learning training. *arXiv preprint arXiv:1905.12322* (2019).

[54] Cholmin Kang, Inhwa Yeom, Amirsaman Ashtari, Woontack Woo, and Junyong Noh. 2023. ARbility: re-inviting older wheelchair users to in-store shopping via wearable augmented reality. *Virtual Reality* (2023), 1–18.

[55] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. 2023. GMD: Controllable Human Motion Synthesis via Guided Diffusion Models. *arXiv preprint arXiv:2305.12577* (2023).

[56] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* (2019).

[57] Do Yuon Kim, Ha Kyung Lee, and Kyunghwa Chung. 2023. Avatar-mediated experience in the metaverse: The impact of avatar realism on user-avatar relationship. *Journal of Retailing and Consumer Services* 73 (2023), 103382.

[58] Minji Kim, Kyungjin Lee, Rajesh Balan, and Youngki Lee. 2023. Bubbleu: Exploring Augmented Reality Game Design with Uncertain AI-based Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[59] Sangpil Kim, Hyung-gun Chi, and Karthik Ramani. 2021. Object synthesis by learning part geometry with surface and volumetric representations. *Computer-Aided Design* 130 (2021), 102932.

[60] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv:2304.02643* (2023).

[61] Sarah Krings, Enes Yigitbas, Ivan Jovanovikj, Stefan Sauer, and Gregor Engels. 2020. Development framework for context-aware augmented reality applications. In *Companion Proceedings of the 12th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. 1–6.

[62] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. 2022. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870* (2022).

[63] Jérémy Lacoche and Eric Villain. 2022. Prototyping context-aware augmented reality applications for smart environments inside virtual reality. In *GRAPP 2022*.

[64] Traian Lavric. 2022. *Methodologies and tools for expert knowledge sharing in manual assembly industries by using augmented reality.* Ph. D. Dissertation. Institut Polytechnique de Paris.

[65] Benjamin Lee, Michael Sedlmair, and Dieter Schmalstieg. 2023. Design Patterns for Situated Visualization in Augmented Reality. *arXiv preprint arXiv:2307.09157* (2023).

[66] Gun A. Lee, Gerard J. Kim, and Mark Billinghurst. 2005. Immersive Authoring: What You EXperience Is What You Get (WYXIWYG). *Commun. ACM* 48, 7 (jul 2005), 76–81. https://doi.org/10.1145/1070838.1070840

[67] Jaewook Lee and Andrew Lan. 2023. SmartPhone: Exploring Keyword Mnemonic with Auto-generated Verbal and Visual Cues. In *International Conference on Artificial Intelligence in Education*. Springer, 16–27.

[68] Wanwan Li, Changyang Li, Minyoung Kim, Haikun Huang, and Lap-Fai Yu. 2023. Location-Aware Adaptation of Augmented Reality Narratives. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[69] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. 2020. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8139–8148.

[70] Jen-Shuo Liu, Barbara Tversky, and Steven Feiner. 2022. Precueing Object Placement and Orientation for Manual Tasks in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (2022), 3799–3809. https://doi.org/10.1109/TVCG.2022.3203111

[71] Shi Liu, Peyman Toreini, and Alexander Maedche. 2022. Designing Gaze-Aware Attention Feedback for Learning in Mixed Reality. In *Proceedings of Mensch und Computer 2022*. 503–508.

[72] Vivian Liu, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 2023. 3DALL-E: Integrating text-to-image AI in 3D design workflows. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 1955–1977.

[73] Ziyi Liu, Zhengzhe Zhu, Enze Jiang, Feichi Huang, Ana M Villanueva, Xun Qian, Tianyi Wang, and Karthik Ramani. 2023. InstruMentAR: Auto-Generation of Augmented Reality Tutorials for Operating Digital Instruments Through Recording Embodied Demonstration. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[74] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. 2020. Novice-AI music co-creation via AI-steering tools for deep generative models. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[75] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. 2023. Humantomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978* (2023).

[76] Yan Luo, Fang Liu, Yingying She, and Baorong Yang. 2023. A context-aware mobile augmented reality pet interaction model to enhance user experience. *Computer Animation and Virtual Worlds* 34, 1 (2023), e2123.

[77] Zhihan Lv. 2023. Generative artificial intelligence in the metaverse era. *Cognitive Robotics* 3 (2023), 208–217. https://doi.org/10.1016/j.cogr.2023.06.001

[78] Rafael Maio, André Santos, Bernardo Marques, Carlos Ferreira, Duarte Almeida, Pedro Ramalho, Joel Batista, Paulo Dias, and Beatriz Sousa Santos. 2023. Pervasive Augmented Reality to support logistics operators in industrial scenarios: a shop floor user study on kit assembly. *The International Journal of Advanced Manufacturing Technology* (2023), 1–19.

[79] Microsoft. 2021. HoloLens 2. https://www.microsoft.com/en-us/hololens/hardware. Accessed on September 12, 2023.

[80] Kyzyl Monteiro, Ritik Vatsal, Neil Chulpongsatorn, Aman Parnami, and Ryo Suzuki. 2023. Teachable Reality: Prototyping Tangible Augmented Reality with Everyday Objects by Leveraging Interactive Machine Teaching. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[81] Alexis Morris, Jie Guan, Nadine Lessio, and Yiyi Shao. 2020. Toward mixed reality hybrid objects with iot avatar agents. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 766–773.

[82] Gary R Morrison, Steven J Ross, Jennifer R Morrison, and Howard K Kalman. 2019. *Designing effective instruction.* John Wiley & Sons.

[83] Fabian Muff and Hans-Georg Fill. 2022. A Framework for Context-Dependent Augmented Reality Applications Using Machine Learning and Ontological Reasoning.. In *AAAI Spring Symposium: MAKE*.

[84] Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. 2024. Joint Reconstruction of 3D Human and Object via Contact-Based Refinement Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[85] NASA NASA. 1986. Task load index (tlx) v. 1.0 manual. *NASA, NASA-Ames Research Center Moffett Field* (1986).

[86] Michael Nebeling, Katy Lewis, Yu-Cheng Chang, Lihan Zhu, Michelle Chung, Piaoyang Wang, and Janet Nebeling. 2020. XRDirector: A Role-Based Collaborative Immersive Authoring System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376637

[87] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. arXiv:2112.10741 [cs.CV]

[88] OpenAI. 2021. ChatGPT: A large-scale generative language model. https://www.openai.com/research/chatgpt. Accessed on September 12, 2023.

[89] Xingyu Pan, Mengya Zheng, Xuanhui Xu, and Abraham G Campbell. 2021. Knowing your student: Targeted teaching decision support through asymmetric mixed reality collaborative learning. *IEEE Access* 9 (2021), 164742–164751.

[90] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10975–10985.

[91] Marius Preda and Traian Lavric. 2023. Augmented Reality Training in Manufacturing Sectors. In *The Digital Twin*. Springer, 447–496.

[92] Xun Qian. 2023. *Explore the Design and Authoring of Ai-driven Context-aware Augmented Reality Experiences.* Ph. D. Dissertation. Purdue University.

[93] Xun Qian, Fengming He, Xiyun Hu, Tianyi Wang, Ananya Ipsita, and Karthik Ramani. 2022. Scalar: Authoring semantically adaptive augmented reality experiences in virtual reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[94] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

[95] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[96] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[97] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG]

[98] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 [cs.CV]

[99] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. arXiv:2102.12092 [cs.CV]

[100] Vedapalle Sri Sai Swarup Reddy. 2022. *An Exploration of the Virtual Digital Twin Capture for Spatial Tasks and its Applications.* Ph. D. Dissertation. Purdue University.

[101] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[102] Patrick Reipschläger, Frederik Brudy, Raimund Dachselt, Justin Matejka, George Fitzmaurice, and Fraser Anderson. 2022. Avatar: An immersive analysis environment for human motion data combining interactive 3d avatars and trajectories. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–15.

[103] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[104] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 10684–10695.

[105] ZHAO Ruizhi. 2021. Context-Aware AR Scheduling Assistant Using Personalized Avatars. (2021).

[106] Manjul Singh Sachan and Roshan L Peiris. 2022. Designing Augmented Reality Based Interventions to Encourage Physical Activity During Virtual Classes. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts.* 1–6.

[107] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487 [cs.CV]

[108] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 36479–36494. https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf

[109] Asangika Sandamini, Chamodi Jayathilaka, Thisara Pannala, Kasun Karunanayaka, Prabhash Kumarasinghe, and Dushani Perera. 2022. An Augmented Reality-based Fashion Design Interface with Artistic Contents Generated Using Deep Generative Models. In *2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer).* IEEE, 104–109.

[110] Tim Scargill. 2021. Context-Aware Markerless Augmented Reality for Shared Educational Spaces. In *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct).* IEEE, 469–472.

[111] Timothy Scargill, Ying Chen, Sangjun Eom, Jessilyn Dunn, and Maria Gorlatova. 2022. Environmental, user, and social context-aware augmented reality for supporting personal development and change. In *2022 IEEE conference on virtual reality and 3d user interfaces abstracts and workshops (VRW).* IEEE, 155–162.

[112] Arne Seeliger, Raphael P Weibel, and Stefan Feuerriegel. 2022. Context-Adaptive Visual Cues for Safe Navigation in Augmented Reality Using Machine Learning. *International Journal of Human–Computer Interaction* (2022), 1–21.

[113] Jingyu Shi, Rahul Jain, Hyungjun Doh, Ryo Suzuki, and Karthik Ramani. 2023. An HCI-Centric Survey and Taxonomy of Human-Generative-AI Interactions. *arXiv preprint arXiv:2310.07127* (2023).

[114] Domen Šoberl. 2023. *Mixed Reality and Deep Learning: Augmenting Visual Information Using Generative Adversarial Networks.* Springer Nature Switzerland, Cham, 3–29. https://doi.org/10.1007/978-3-031-27166-3_1

[115] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. arXiv:1503.03585 [cs.LG]

[116] M Soliman and MK Al Balushi. 2023. Unveiling destination evangelism through generative AI tools. *ROBONOMICS: The Journal of the Automated Economy* 4, 54 (2023), 1.

[117] Unity Technologies. 2023. Unity3D. Retrieved 12.09.2023 from https://unity.com/

[118] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).

[119] Aldo Uriarte-Portillo, María Blanca Ibáñez, Ramón Zatarain-Cabada, and María Lucía Barrón-Estrada. 2023. Comparison of using an augmented reality learning tool at home and in a classroom regarding motivation and learning outcomes. *Multimodal Technologies and Interaction* 7, 3 (2023), 23.

[120] Tianyi Wang. 2022. *Supporting the Design and Authoring of Pervasive Smart Environments.* Ph. D. Dissertation. Purdue University.

[121] Tianyi Wang, Xun Qian, Fengming He, Xiyun Hu, Ke Huo, Yuanzhi Cao, and Karthik Ramani. 2020. CAPturAR: An augmented reality tool for authoring human-involved context-aware applications. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology.* 328–341.

[122] Xuanyu Wang, Yang Wang, Yan Shi, Weizhan Zhang, and Qinghua Zheng. 2020. Avatarmeeting: An augmented reality remote interaction system with personalized avatars. In *Proceedings of the 28th ACM International Conference on Multimedia.* 4533–4535.

[123] X Wang, AWW Yew, Soh-Khim Ong, and Andrew YC Nee. 2020. Enhancing smart shop floor management with ubiquitous augmented reality. *International Journal of Production Research* 58, 8 (2020), 2352–2367.

[124] Zeyu Wang, Cuong Nguyen, Paul Asente, and Julie Dorsey. 2021. Distanciar: Authoring site-specific augmented reality experiences for remote environments. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–12.

[125] Maheshya Weerasinghe, Aaron Quigley, Klen Čopič Pucihar, Alice Toniolo, Angela Miguel, and Matjaž Kljun. 2022. Arigatō: Effects of Adaptive Guidance on Engagement and Performance in Augmented Reality Learning Environments. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (2022), 3737–3747.

[126] Florian Weidner, Gerd Boettcher, Stephanie Arevalo Arboleda, Chenyao Diao, Luljeta Sinani, Christian Kunert, Christoph Gerhardt, Wolfgang Broll, and Alexander Raake. 2023. A Systematic Review on the Visualization of Avatars and Agents in AR & VR displayed using Head-Mounted Displays. *IEEE Transactions on Visualization and Computer Graphics* (2023).

[127] AndrÉ M Weitzenhoffer. 1974. When is an "instruction" an "instruction"? *International Journal of Clinical and Experimental Hypnosis* 22, 3 (1974), 258–269.

[128] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382* (2023).

[129] Hui Ye and Hongbo Fu. 2022. ProGesAR: Mobile AR Prototyping for Proxemic and Gestural Interactions with Real-world IoT Enhanced Spaces. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–14.

[130] Hui Ye, Jiaye Leng, Chufeng Xiao, Lili Wang, and Hongbo Fu. 2023. ProObjAR: Prototyping Spatially-aware Interactions of Smart Objects with AR-HMD. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–15.

[131] Xuyue Yin, Xiumin Fan, Wenmin Zhu, and Rui Liu. 2018. Synchronous AR assembly assistance and monitoring system based on ego-centric vision. *Assembly Automation* 39, 1 (2018), 1–16.

[132] Kyongsik Yun, Thomas Lu, and Edward Chow. 2018. Occluded object reconstruction for first responders with augmented reality glasses using conditional generative adversarial networks. In *Pattern Recognition and Tracking XXIX*, Vol. 10649. SPIE, 225–231.

[133] Zhenjie Zhao and Xiaojuan Ma. 2018. A Compensation Method of Two-Stage Image Generation for Human-AI Collaborated In-Situ Fashion Design in Augmented Reality Environment. In *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR).* 76–83. https://doi.org/10.1109/AIVR.2018.00018

[134] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. 2023. CAMS: CAnonicalized Manipulation Spaces for Category-Level Functional Hand-Object Manipulation Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 585–594.

[135] Zhengzhe Zhu, Ziyi Liu, Youyou Zhang, Lijun Zhu, Joey Huang, Ana M Villanueva, Xun Qian, Kylie Peppler, and Karthik Ramani. 2023. LearnIoTVR: An End-to-End Virtual Reality Environment Providing Authentic Learning Experiences for Internet of Things. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–17.

A Application Scenario

With CARING-AI, users are enabled to author context-aware humanoid avatar animation for AR instructions that can be adaptively deployed into various application scenarios (**AS**). Our primary goal of presenting the **AS** is to demonstrate that CARING-AI can capture and convey context information identified in design space. Specifically, we showcase three major scenarios where CARING-AI demonstrates its ability to grant code-less and Mocap-free authoring (**AS-1**, **AS-2**, and **AS-3**), create content that is to be deployed in different time primitives or via different platforms (**AS-1**, **AS-3**), and adapt to varying contexts (**AS-2**).

A.1 AS-1: Asynchronous Instructions

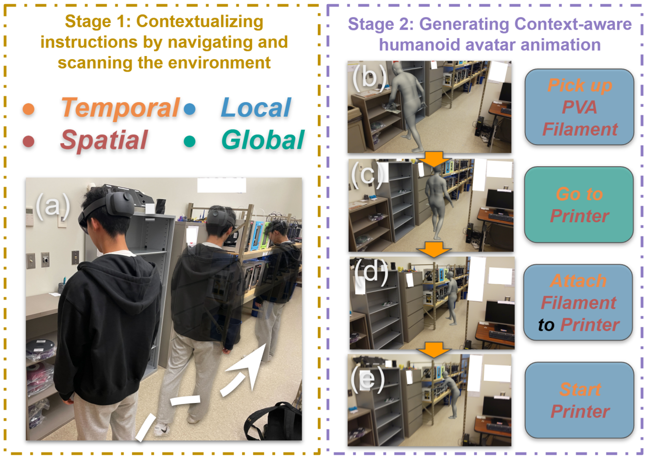

Stage 1: Contextualizing instructions by navigating and scanning the environment
Temporal
Local
Spatial
Global
(a)
Stage 2: Generating Context-aware humanoid avatar animation
(b)
Pick up PVA Filament
(c)
Go to Printer
(d)
Attach Filament to Printer
(e)
Start Printer

Figure 18: CARING-AI for authoring asynchronous Instructions. A senior lab researcher (a) leaves an AR memo for his colleague on how to use a 3D printer. He simply walks around the printing lab using CARING-AI to contextualize the textual instructions, capturing the location of the PVA filament and the printer. (b) The corresponding humanoid animation is generated according to the step-by-step instructions. CARING-AI is capable of handling AR instructions of diverse content and context, namely spatial or temporal context, and local or global content.

Asynchronous instructions are the most common case in the applications of AR instructions, where the authors create the content prior to the consumption of the AR experiences [18, 22]. CARING-AI naturally supports asynchronous instructions and situates AI-generated humanoid avatar animation into the physical world contextually. Here, we showcase a scenario in a research lab, where a senior researcher (Tom, the author) would like to leave an AR memo for his junior colleague (Jerry, the consumer) to instruct him on how to operate a 3D printer. Tom creates and modifies the text instructions with the help of CARING-AI, then provides context to the system by walking up to the locations and taking snapshots of the environment as shown in Figure 18. He informs Jerry to get the printing materials and then go to a specific 3D printer to print a product. Later, when Jerry arrives in the laboratory, he follows the step-by-step AR memo from Tom to start working on the product.

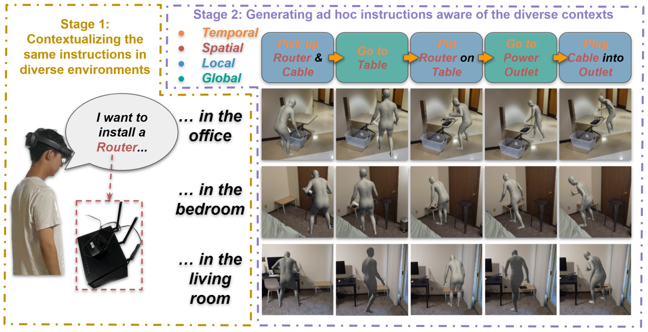


Figure 19: CARING-AI for Ad Hoc AR Instructions. For the same task (e.g. installing a router), the instructions vary across diverse contexts. By using CARING-AI, users simply need to scan the environment to provide contextual information to the system. CARING-AI will generate humanoid avatar animation that blends into different physical realities.

A.2 AS-2: Ad Hoc Instruction Creation

CARING-AI enables authoring AR instructions through Gen-AI by contextualizing the instructions. In this scenario, we showcase how CARING-AI enables authoring ad hoc instructions in changing contexts with simplified user interactions. As a technician from the lab, Tom would like to teach his colleague how to install a router as shown in Figure 19. The instructions are fairly simple and easy to understand. However, the detail of the steps varies across environments, e.g. in the office, the bedroom, or the living room, because the locations of the router and the outlet vary. With the same protocol to be visualized, Tom simply has to contextualize the protocol in different places, assigning the locations of the objects by traversing the rooms. As a result, Tom authors different avatar animations for diverse contexts with the same instruction protocol.

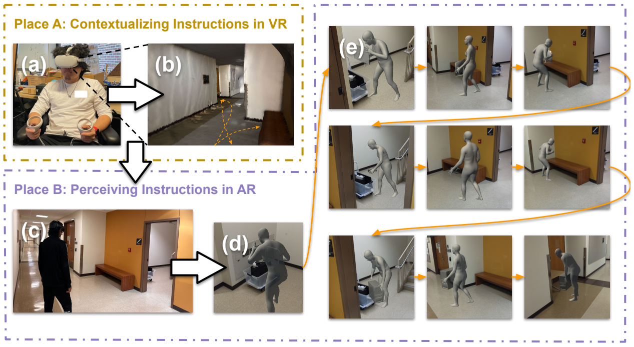
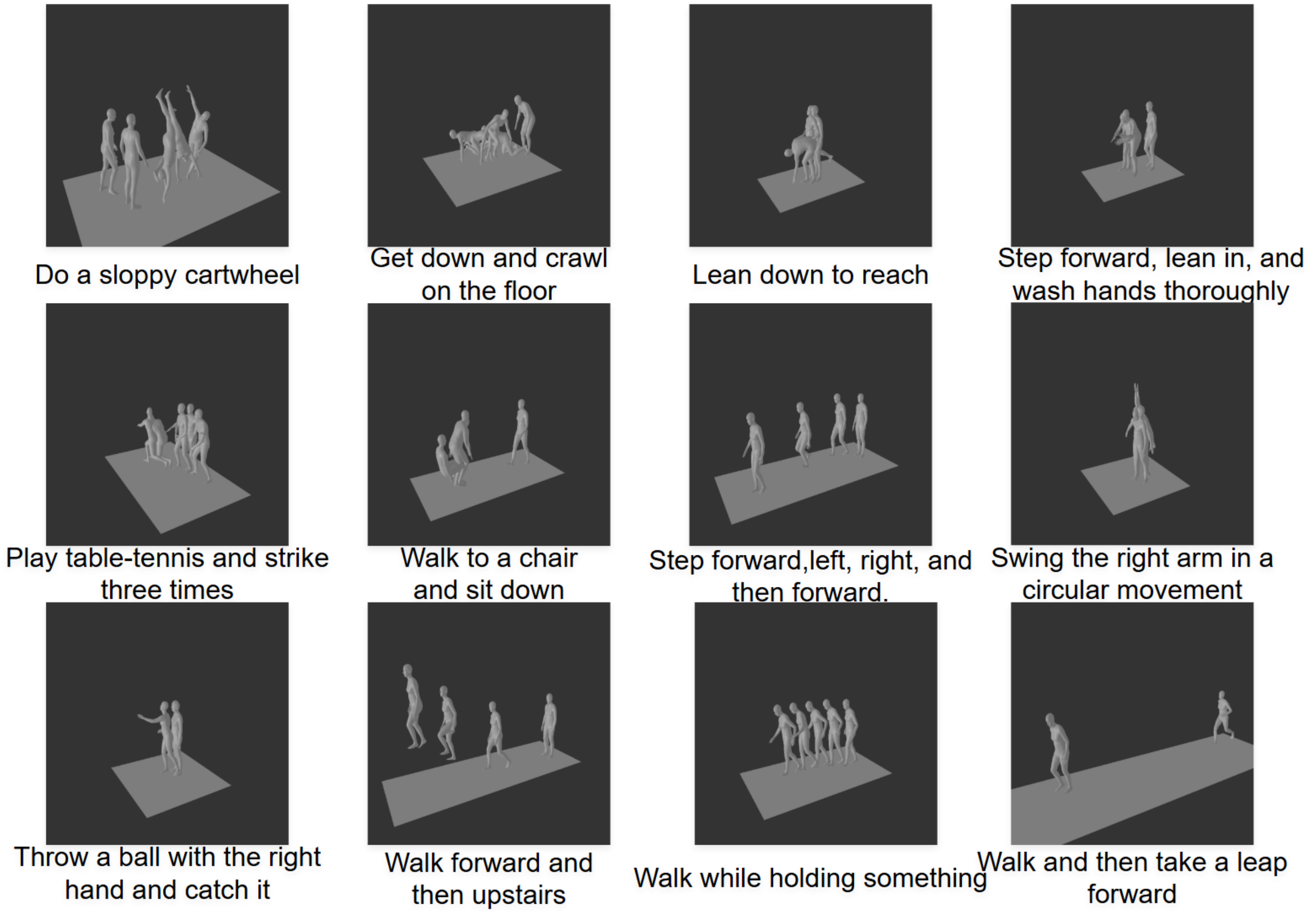

Place A: Contextualizing Instructions in VR
(a)
(b)
Place B: Perceiving Instructions in AR
(c)
(d)
(e)

Figure 20: CARING-AI's capability of authoring Remote Instructions. In this scenario, a delivery man is asking for the destinations of the packages. (a) A lab member is giving contextual information through a pre-scanned scenario in VR. (b) We built a mock-up VR scenario to record the locations and correspond them back into the physical reality. Once the instructions are contextualized, the delivery man can view the humanoid instructions on delivering the packages (c, d).



Do a sloppy cartwheel

Get down and crawl on the floor

Lean down to reach

Step forward, lean in, and wash hands thoroughly

Play table-tennis and strike three times

Walk to a chair and sit down

Step forward,left, right, and then forward.

Swing the right arm in a circular movement

Throw a ball with the right hand and catch it

Walk forward and then upstairs

Walk while holding something

Walk and then take a leap forward

Figure 21: We showcase more examples of humanoid animation generated from our backend algorithm. Specifically, the animations generated are guided by a textual description of the motion, rendered with humanoid avatars (in our implementation, SMPL [90]). Each animation clip presents a short sequence of human motion and represents a step in a given AR instruction.

A.3 AS-3: Remote Instructions

In this scenario, CARING-AI is deployed in a remote instruction task. We showcased how CARING-AI can adapt to context information of diverse modalities and liberate the authors from demonstrating in the actual physical environment. Toodles, the deliverywoman of the building, arrives in the lab with new devices to be allocated Figure 20 (c). Noticing no one is in the lab, Toodles contacts Jerry, asking about the allocation of the devices. Jerry, who is not present at the lab, confirms the devices and their checkout points (i.e. where they are to be placed). Jerry then enters a pre-scanned point-cloud map of the lab in Virtual Reality (VR), where he authors the instructions in VR using CARING-AI by navigating the map and taking screenshots Figure 20 (a, b) (We built a mock-up VR program to record Jerry's locations in VR and correspond them to the physical reality). CARING-AI generates humanoid avatar instructions according to the contextual information provided. The authored AR instructions are then sent to Toodles, who follows the avatar demonstrations to allocate the devices to different locations Figure 20 (d, e). In this case, we see that CARING-AI is capable of authoring synchronous remote instructions. It also showcases the possibilities of authoring AR experiences in VR with CARING-AI with aligned contextual information between physical reality and VR. The alignment of context is subsumed here as described and inspired by many prior works [86, 93, 124]

B More Generated Examples

In this section, we showcase more examples generated from our backend diffusion model as shown in Figure 21 and situated humanoid animation by CARING-AI through our pipeline as shown in Figure 22. Given a textual prompt, our motion diffusion model can generate high-fidelity humanoid avatar motion. With the user-provided context, specifically object location and motion trajectory, the CARING-AI system can situate the generated animations in the space and temporally smooth them for a seamless user experience.

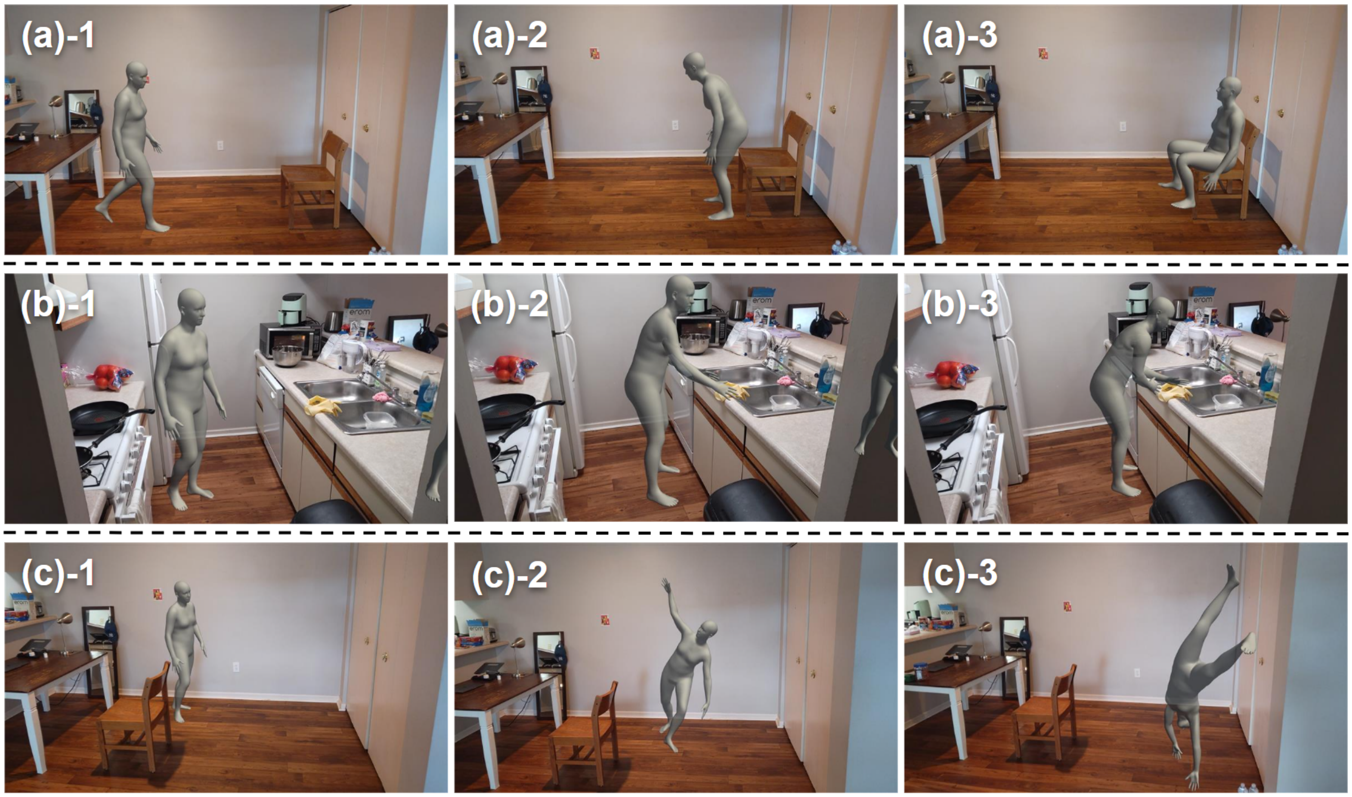


Figure 22: We showcase more examples of humanoid animation contextualized by CARING-AI, (a) Walk to the chair and sit down, (b) Walk to the sink, lean in, and wash hands thoroughly, and (c) do a a sloppy cartwheel around the chair. All animations are generated by prompting the CARING-AI with text, scanning the environment to mark the object, and passing user trajectories to the generative model.