HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World
Supplementary Material

Xin Wang[1*] Taein Kwon[1,2*] Mahdi Rad[1] Bowen Pan[1†] Ishani Chakraborty[1] Sean Andrist[1]
Dan Bohus[1] Ashley Feniello[1] Bugra Tekin[1†] Felipe Vieira Frujeri[1] Neel Joshi[1] Marc Pollefeys[1,2]
[1]Microsoft [2] ETH Zurich

This supplementary material shows additional qualitative results and sample visualization of our dataset. Also, we show details about our annotation method, data capture procedure, and data analysis. Furthermore, we will report implementation details, and qualitative results. Additional qualitative results and visualization of our dataset can be found at https://holoassist.github.io/.

S.1. Interactive Assistive Task Completion

HoloAssist features a unique interactive assistive task completion setting where instructors intervene during the task completion process if task performers make mistakes or get confused. In Figure S1, we show additional illustrations of how different users may experience different intervention moments while completing the same task. The diversity and complexity of the intervention moments and types suggest the challenges of building an interactive AI assistant.

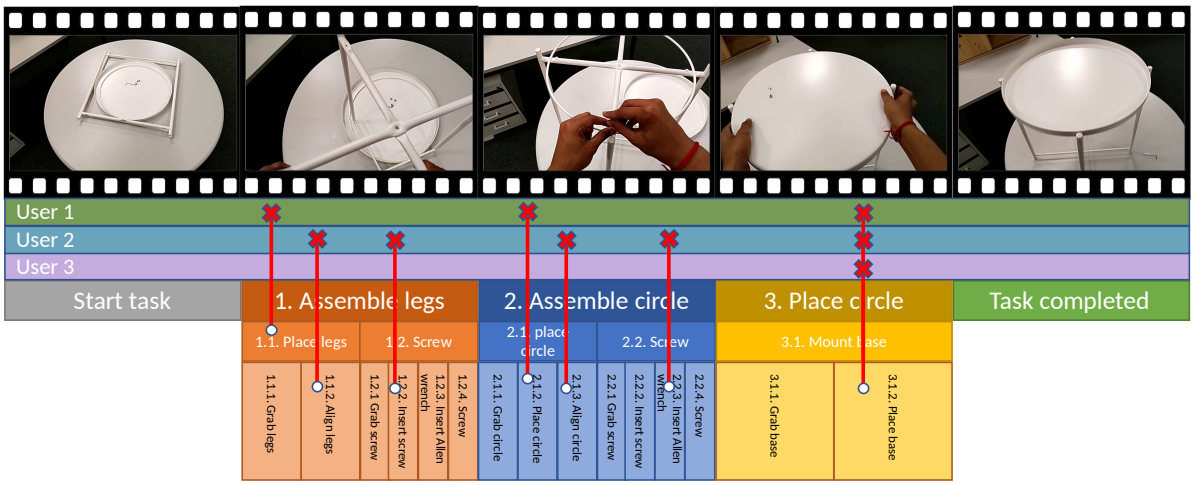

Figure S1: Intervention-actions structure. This figure shows the tree structure of different levels of structures (high, mid, or low-level instructions) for the `Assemble tray table` task, together with the intervention moments indicated by `red crosses` for three different *task performers*. As the figure illustrates, different task performers need different levels of supervision at different moments, which shows the importance of our dataset for building an interactive AI assistant.

S.2. Data

S.2.1 Instructor Training

Before data collection starts, we go through a training process for the instructors to get familiar with the objects and tasks. We provide the instructors with detailed guidelines for the example tasks as shown in Table S1. The steps in the task instructions are mainly used as a

*Co-first authors; †Work done at Microsoft

reference and the instructors do not need to strictly follow the suggestions. They can change the order, add or skip tasks, and change the way they instruct the process as they see fit during the data capture.

Objects	Example Tasks	Comments
GoPro	- Change Battery - Change micro sd card - Turn on - Turn off - Put on a strap - Change the strap to handheld grip (or tripod) - Remove the handheld grip (or tripod)	Turning on/off is controlled by the button on the side of the GoPro not the top of device and you need to press $\sim$3 seconds to make it happen.
Nintendo	- Change controller from the handheld mode to pad/joy-con comfort grip - Change controller from pad/joy-con comfort grip to two joy-cons - Change controller from two joy-cons back to the handheld mode - Stand Nintendo using kick stand - Change game card - Change micro sd card - Turn on - Turn off (hold for 3 seconds and power off) - just one second is only sleeping mode	1. The task always starts from the handheld mode. Also, you need to turn off the device entirely before starting the data capture. 2. For two joy-cons, task performers need to lock them using the white button. 3. The strip always faces towards the wrist side.
DSLR	- Detach lens - Attach lens - Detach a lens cover - Attach a lens cover - Turn on - Turn off - Change battery - Change SD card	1. When you turn on the camera, please wait until the screen is on so we can capture the visual cues. 2. For some DSLR models, make sure to put a microSD card in the SD card adapter - or the camera may not be turned on properly.
Espresso Machine	- Add coffee beans - Make a cup of coffee - Add milk/cream and stir - Empty drip tray	
Nespresso / Capsule Machine	- Please turn on the machine before starting capture - Add water - Make a cup of coffee (put a capsule) - make sure the capsule went to the capsule container - Empty the capsule container into the trashcan - Empty drip tray	The coffee is hot so please make sure that the participants dispose the waste into the sink can directly.
Printer (Big)	- Wake up from sleep mode (Turn on) - Go into sleep mode (Turn off) - Add/load paper to tray 1 (Please say tray 1 explicitly) - Change black printer cartridge	

Printer (small)	- Load paper and pull out the tray. - Copy a black-and-white paper (make sure the subject removes the paper at this stage) - Change (install) color and black printer cartridge (Don't turn off the printer at this stage) - Close everything	1. Reset the machine first. 2. Details for steps 1 and 3 are on Page 4 of the manual. You can instruct the process according to the manual. 3. If you can hear noise after changing the cartridge, please ignore it and just close everything. 4. When you start, the power button should not be blinking. Press the cancel button if things go wrong. 5. When copying the paper, please place the paper on the right-bottom of the machine.
Small furniture from IKEA	For IKEA furniture, please follow the steps in the manual. - Assemble - Disassemble	Assemble and disassemble should be in different videos
Objects in the lab	Instructors decide the setup procedures or steps for manipulating the objects.	

Table S1: Task instructions for instructors to get familiar with the objects. Instructors are free to add or skip tasks, change orders, and use their own way of completing the tasks during the data collection.

S.2.2 IRB Approval for Data Collection

The data collection process of HoloAssist was reviewed and approved by IRB before the work started. The data capture mostly takes place in public areas (*i.e.*, offices and labs) and avoids capturing human faces. The participants have reviewed and signed the consent forms and information sheet so they are fully aware of the data capture process and future usage. The consent form that is approved by IRB and the information sheet can be provided upon request.

S.3. Data Capture Platform

In Table S2, we provide additional information about the various streams available in the HoloAssist dataset.

Spatial coordinate systems. The data capture platform utilizes a spatial anchor approach to establish a world coordinate system. All sensor modalities that are spatial in nature (head pose, gaze direction, hand poses, camera poses) are expressed with respect to the world coordinate system. The basis vector interpretation for coordinate system axes follows the convention of X=Forward, Y=Left, and Z=Up.

Camera intrinsics are represented with several parameters, including a 3×3 intrinsics matrix that converts camera coordinates (in the camera's local space) into normalized device coordinates (NDC) ranging from -1 to +1, radial and tangential distortion parameters, focal length, the principal point, and the image width and height.

Sensor	Stream	Description	Representation	FPS
Color Camera	Color Image	Image captured by front-facing HoloLens 2 camera	RGB image 896×504 pixels	29.5 Hz
	Color Camera Intrinsics	Intrinsic parameters for front-facing HoloLens 2 camera	Camera intrinsics parameters (listed above)	29.5 Hz
	Color Camera Pose	Spatial pose for color camera location	4×4 coordinate system matrix	29.5 Hz
Depth Camera	Depth Image	Depth image captured by HoloLens 2 depth camera in AHAT mode (provides pseudo-depth with phase wrap beyond 1 meter)	16bpp Grayscale image 504×504 pixels	32.4 Hz
	Depth Camera Intrinsics	Intrinsic parameters for HoloLens 2 depth camera	Camera intrinsics parameters (listed above)	32.4 Hz

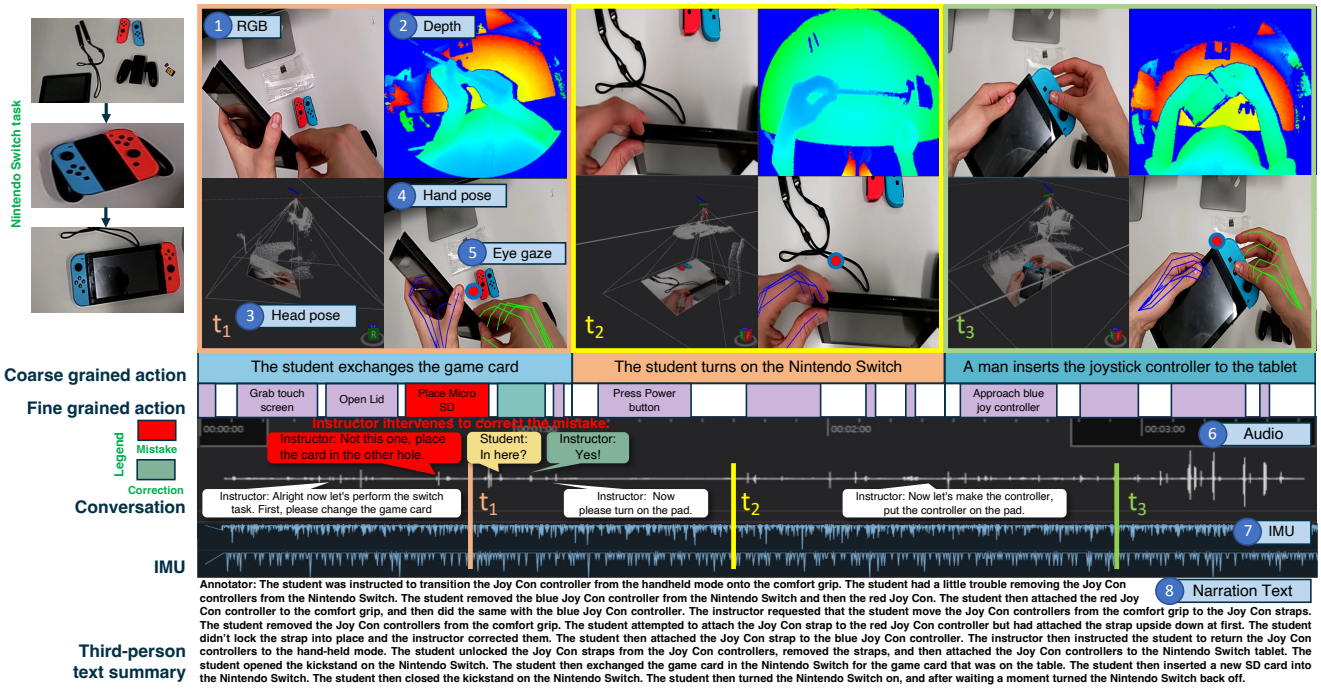

Figure S2: HoloAssist includes action and conversational annotations, in addition to text summaries of the videos, to indicate the mistakes and interventions in task completion. *mistake* or *correct* attributes are associated with each fine-grained action. A purpose label is associated with every conversation to indicate the type of verbal intervention.

	Depth Camera Pose	Spatial pose for depth camera location	4×4 coordinate system matrix	32.4 Hz
Audio	Audio	Audio stream captured by the microphone array	Single-channel, 48Khz, IEEE Float Wave Format	48 KHz
IMU	Accelerometer	X-, Y-, and Z-axis inertial force in m/s^2	3×1 vector	1.1 KHz
	Gyroscope	X-, Y-, and Z-axis angular momentum in rad/s	3×1 vector	6.4 KHz
	Magnetometer	X-, Y-, and Z-axis magnetic flux density in microteslas	3×1 vector	49.5 Hz
Head	Head Pose	Spatial pose for the user's head	4×4 coordinate system matrix	17.2 Hz
Eye Gaze	Gaze Direction	Spatial direction of the user's gaze	3×1 vector for position and 3×1 vector for direction	30.7 Hz
Hands	Hands Pose	Spatial poses for the user's left and right-hand joints	4×4 coordinate system for each of the 26 left-hand joints and 26 right-hand joints	43.2 Hz

Table S2: Details about the 7 sensor modalities captured live during data collection. The 8th modality, text, is added after the data collection.

S.4. Annotation

S.4.1 Annotation Definition

Figure S2 provides an overview of the HoloAssist task and annotation structure. There are four event types in total for annotations: text summaries, coarse-grained Actions, fine-grained actions, and conversation.

Text summary. A summary of all activities performed by the task performer throughout the video. This annotation will describe all the actions across the video. The video annotator should use the Coarse-Grained Action annotations and the Conversation annotations to help build the Narration annotation.

Coarse-grained actions. A coarse-grained action is a step in the process that the task performer performs to help complete the task. Note that often the instructor will state the task before the task performer does the action–this is *not* considered part of the Coarse-Grained Action. Examples of coarse-grained actions include loading paper into a printer, screwing furniture legs into place, turning on a device, etc.

Fine-grained actions. Fine-grained actions are the indivisible movements of the task performer during a coarse-grained action. These actions typically last for around 1-2 seconds or shorter, depending on the action. These will also include a designation of whether the action was correct or whether it was corrected later on by the task performer themselves or the instructor. Examples of fine-grained actions include the hand approaching button, hand pressing button, a task performer walking across the room, etc.

Conversation. A Conversation annotation denotes the time frame when either the instructor or the task performer is speaking. The annotation specifies the content of the conversation.

S.4.2 Annotation Structure

The annotations follow the following format.

Narration annotation structure.

- **Long-form description**: Use multiple sentences and make this as long as is necessary to be exhaustive. There are a finite number of scenarios across all videos, so make sure to call out the distinctive changes between videos, in particular, mistakes that the task performer makes in the learning process that are either self-corrected or corrected by the instructor.
- **Example**: A man operates a big office printer. The instructor provides directions on how to turn on and load paper into the big office printer. The man turns on the printer and then turns it off. He then loads paper into the first drawer of the printer and replaces the first black cartridge from left to right in the printer. The instructor corrects the man on where to place the first black cartridge. The man moves the black cartridge from its current position to the correct location. Note: The time stamps for this annotation will always start at 0 and end at the end of the video.

Coarse-grained action annotation structure.

- **Coarse-Grained Action Sentences**: A factual statement describing the interaction that the collector/camera wearer is performing with a digital device and the software on the device.
  - Example 1: A man changes the battery of the bright green GoPro.
  - Example 2: A woman attaches the leg to a chair.
- **Coarse-Grained Action Verb**: This verb was part of the Coarse-Grained Action sentence.
  - Example 1: Change
  - Example 2: Attach
- **Coarse-Grained Action Adjective**: This is the adjective(s) that helps distinguish the noun from other similar items. This field is optional if the noun is unique enough on its own.
  - Example 1: bright green
  - Example 2: [blank]
- **Coarse-Grained Action Noun**: This is the generic noun that is part of the Coarse-Grained Action sentence.
  - Example 1: GoPro
  - Example 2: Leg

Fine-grained action annotation structure.

- **Fine-Grained Action Verb**: This is the verb that occurred during the Fine-Grained Action in the video.
  - Example 1: Change
  - Example 2: Attach
- **Fine-Grained Action Adjective**: This is the adjective(s) that helps distinguish the noun from other similar items. This field is optional if the noun is unique enough on its own.
  - Example 1: Bright Green

  - Example 2: [blank]
- **Fine-Grained Action Noun**: This is the generic noun that is part of the Fine-Grained Action in the video.
  - Example 1: GoPro
  - Example 2: Leg
- **Fine-Grained Action Attribute**: This attribute denotes whether the Fine-Grained action was the correct action and if it was the incorrect action, whether it was corrected, and by whom. You only need to select a single option from the list of options.
  - Correct action
  - Wrong action, corrected by instructor verbally
  - Wrong action, corrected by performer
  - Wrong action, not corrected
  - Others

Conversation annotation structure.

- **Conversation Transcriptions**: Transcribe the conversation into texts.
- **Conversation Attribute**: Select an option that best describes the purpose of the speech. This is limited to the individual speaking and does not include any pause time waiting for a response.
- **Intervention Types**:
  - Instructor-start-conversation: Describing high-level instruction
  - Instructor-start-conversation: Opening remarks
  - Instructor-start-conversation: Closing remarks
  - Instructor-start-conversation: Adjusting to capture better quality video
  - Instructor-start-conversation: Confirming the previous or future action
  - Instructor-start-conversation: Correct the wrong action
  - Instructor-start-conversation: Follow-up instruction
  - Instructor-start-conversation: Other
  - Instructor-reply-to-task performer: Confirming the previous or future action
  - Instructor-reply-to-task performer: Correct the wrong action
  - Instructor-reply-to-task performer: Follow-up instruction
  - Instructor-reply-to-task performer: other
  - task performer-start-conversation: ask questions
  - task performer-start-conversation: others

S.4.3 Label Statistics

Fine-grained action annotations. In Figure 6 in the main paper, we present the distributions of the fine-grained actions. We can see that the fine-grained actions follow a long-tail distribution. This is partly due to the open-world nature of the interaction and scenes. Also, there are cases where the same action can be referred to with different expressions. For example, 'screw screw' and 'screw hex-cap-screw' are similar phrases. We asked the annotators to revisit the distributions of the nouns and verbs when the vocabulary is constructed and merge the ones referring to the same actions but with different expressions. Still, this is a natural outcome of human annotations, and future research on addressing the label confusion issue is needed.

Coarse-grained action annotations. The coarse-grained actions are usually defined as high-level steps in the task, usually lasting around 30 seconds. In Figure 7 in the main paper, we show the distributions of the coarse-grained actions and verbs and nouns of the actions. We can see that the coarse-grained actions follow a long-tail distribution similar to the fine-grained actions.

Examples of conversation transcriptions and third-person text summary. Here we provide examples of the text summaries and the conversation transcriptions. We can see that the third-person text summaries capture the key events in the video, while the conversation transcripts are more interactive. HoloAssist provides both materials and can be useful in different contexts and applications.

Third-person text summary	Conversation transcription
A task performer operated a capsule coffee machine. The task performer grabbed a glass of water, and poured it into the water container. The task performer grabbed a coffee capsule and inserted it into the capsule slot. The task performer pulled the lever to use the capsule. With this, the task performer made a coffee cup. The instructor asked the task performer to empty the capsule container. The task performer withdrew the drip tray, then, removed the capsule. And then inserted the drip tray back.	task performer: That one task performer: Start already Instructor: Add water Instructor: Grab a capsule and make a cup of coffee Instructor: Right Instructor: now, empty the capsule tray Instructor: Throw out the coffee Instructor: Put it on the table Instructor: Put it back Instructor: Your task is done press stop.
The task performer grabbed the water container and took it to the sink to fill it with water. Then the task performer grabbed a coffee capsule and a cup. The task performer placed the cup on the drip tray and inserted the capsule into the capsule slot pushing it down. The task performer reinserted the capsule container by lowering the drip tray as well. Then the task performer reinserted the coffee capsule into the capsule slot. The task performer lowered the lever and placed the cup on the grid. Then pressed the right button to start preparing the cup of coffee. Finally, the task performer removed the drip tray and the capsule container from the espresso machine and separated them. Then the task performer placed the capsule container on the table and carried the drip tray to the sink to empty it. Later the task performer inserted the capsule container back into the drip tray and finally inserted them into the espresso machine.	Instructor: Hello, and thank you for agreeing to take part in our study. In front of you, you will see an espresso coffee machine we will be using for this task. First please load water into the machine Instructor: Please as you put the water facing down lift the lid. Instructor: Please lift the lid Instructor: Please lift the top and slide it down Instructor: Please lift the top as you slide it down. Instructor: Next, please make a cup of coffee Instructor: Please, please return the capsule, please Instructor: Please grab the capsule from the capsule area. Instructor: Please remove the cup Instructor: Please lift up words the door and to the, please afterward please pull please lift it Instructor: Please lift the expansion, and pull it from the machine Instructor: Please reinsert the capsule. Instructor: Please push down on the metal Instructor: push it towards me now Instructor: Please make sure the capsule is part of the Instructor: Please make sure to retrain the capsule, to drop the capsule to the capsule area Instructor: Next, please empty the capsule tray Instructor: Lastly, please empty the drip tray in the sink Instructor: Please take it out of the machine as before Instructor: Please remove the plastic, the capsule tray Instructor: That concludes everything for this task, you can now press the stop button

The task performer began the task by moving the parts of the Nintendo on the table and then grabbing it to change the game card for another. He then opened the Nintendo kick stand and grabbed a micro-SD card, which he inserted into the SD slot. Then he closed the kickstand. The task performer, following the instructions of the instructor, opened the kickstand behind the Nintendo and stood it on the table. The man turned on the Nintendo by pressing the power button. Then, following the instructions of the instructor, he turned off the console by pressing the power button for three seconds and pressing the touch screen of the console. The man grabbed the touch screen and tried to place the blue controller on it, until the instructor corrected him. Then he placed both controls on the sides of the touch screen of the console. The task performer removed the controllers from the touch screen and inserted them on the comfort grip, one on each side, as instructed by the instructor. Then he removed the joy controllers from the comfort grip and then, correcting the errors he had when placing the blue joy with controller, he correctly placed each of them on the joy with straps as the final action of the task.	task performer: Capturing Instructor: Alright now let's perform the switch task. First, please change the game card task performer: In here? Instructor: Yes task performer: Game card there you go. task performer: Done Instructor: Okay Instructor: Now please change the SD card. task performer: There is no SD card Instructor: Please insert one Instructor: Okay Instructor: Now please make the path to stand up. Instructor: Open the support where the SD card is. Instructor: And now please turn on the pad Instructor: And now please turn it off. Instructor: You need to press more than three seconds to turn it off. Instructor: Alright Instructor: Now let's make the controller, put the controller on the pad. Instructor: No, the blue one go left task performer: Done Instructor: Okay Instructor: Now let's change the controller to the joystick. task performer: Change the... so I should remove it from the pad? Instructor: Yes, remove it first. task performer: And then? Instructor: Put it down the joystick task performer: On this? Instructor: Joystick task performer: Yeah, this is the joystick, right? task performer: It's done Instructor: Now change it to another stick task performer: So I should? okay. Instructor: Take it out first. task performer: Okay Instructor: Okay. Instructor: Now you can press stop.

S.5. Experiment Details

S.5.1 Implementation Details

ViT and TimeSformer. In our action recognition, anticipation, mistake detection, and intervention prediction benchmarks, we use the vanilla Vision Transformers [2] and the TimeSformer [1] with divided space-time attention as the base model architecture. Specifically, we use the `vit_base_patch16_224` defined in the TimeSformer codebase. The models are trained for 15 epochs with a base learning rate of 0.01. The learning rate is divided by 10 at epochs 11 and 14. We sample 8 frames as inputs depending on the need of the benchmarks. For example, in the action recognition benchmark, we sample 8 frames from the input action clip. For action anticipation and intervention prediction, we sample the frames from N seconds before the current action clip. For mistake detection, we sample 8 frames from the beginning of the sequence until the end of the current action clip. The number of frames may affect the model performance, and we use 8 to fit our compute resources better.

Seq2Seq model. We use a Seq2Seq model for 3D hand forecasting. We followed the basic Seq2Seq model in the literature [5] with modifications. We train the Seq2Seq network using ADAM optimizer [4] with a learning rate of 0.1. We implement the model with 512 hidden dimensions and 3 LSTM layers. We set the output dimension of the decoder as 156, the same as the number of the hand joint dimension. For the objective function, we use an L2 distance loss. We uniformly sample 8 frames as inputs in 3 seconds before the fine-grained action starts and validate hand forecasting with 1.5 seconds of hand poses from the start of the fine-grained action. If the invalid

flag comes out of the device or the hand location is further than 1.5 from the head position, we consider this as the wrong hand pose, and we do not evaluate them.

- **Encoder.** All inputs passed MLP with the same dimension number as the inputs except for the RGB images before the LSTM layers. For RGB images, we use the pre-trained ResNet [3] with 1000 output dimensions to extract RGB features. Then we concatenate features from MLP and ResNet and feed them into the LSTM layers.
- **Decoder.** For inputs of the LSTM layers, we use the same MLP and the ResNet architecture as the one in the Encoder. Finally, We use the outputs of the LSTM layers as the input to an additional MLP to predict hand joints.

S.5.2 3D Hand Pose Forecasting Visualization

In Figure S3, we provide visualizations of the Seq2Seq model used in the main paper on the 3D hand pose forecasting task. As we can see from the figure, the accuracy of the hand pose forecasting still has a lot of room for improvement.

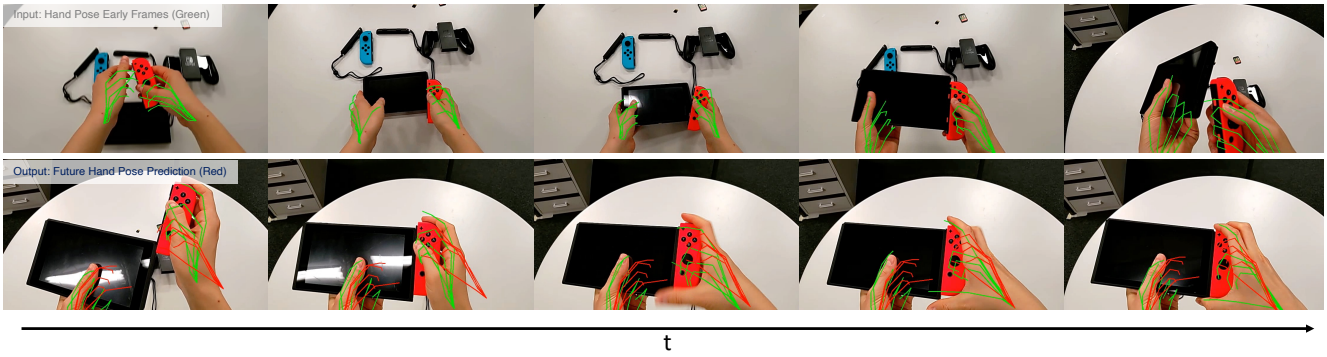

Figure S3: **3D hand pose forecasting qualitative results**. The input to the model is the historical ground-truth hand joint positions (*i.e.*, hand poses, visualized in Green) with RGB images in the multimodal setting (RGB images are optional). Then the output is the prediction of future hand poses visualized in Red.

S.6. Data Sample Visualization

In Figure S4, we provide a visualization of the session sequences in the dataset. Here we visualize the head pose, point cloud, RGB images, and depth images in the figure. The hand poses and eye gazes are omitted here, and you can refer to the figures in the main paper for reference.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 8

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 9

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 8

[5] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 8

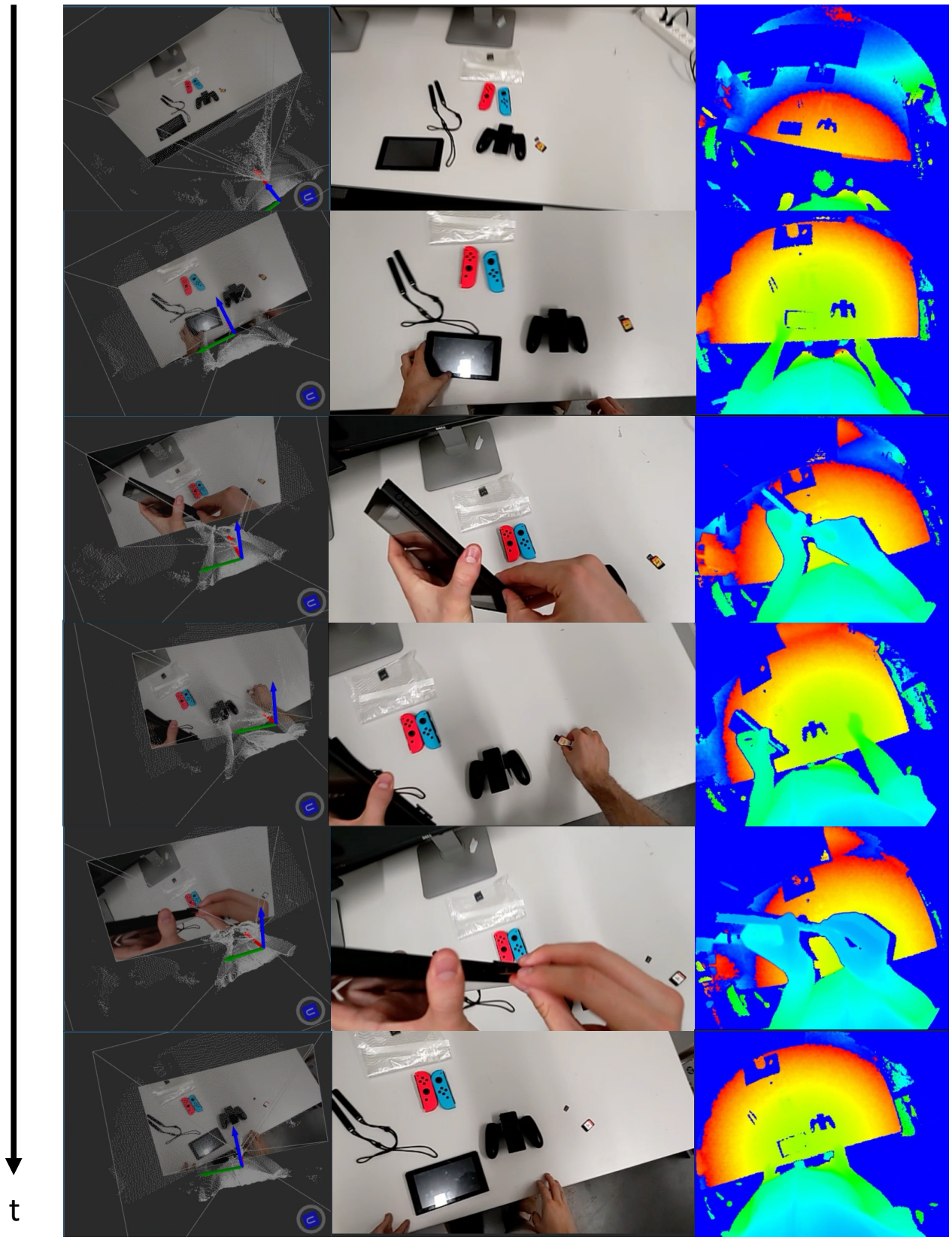

Figure S4: **Sample Data in HoloAssist**. The **left** column shows the point cloud and head poses. The **middle** column shows the RGB images. The **right** column shows the depth images.