AssistQ: Affordance-centric Question-driven Task Completion for Egocentric Assistant (Supplementary Material)

Benita Wong⋆, Joya Chen*, You Wu*, Stan Weixian Lei,
Dongxing Mao, Difei Gao, and Mike Zheng Shou†

Show Lab, National University of Singapore
benitawong@u.nus.edu {joyachen97,mike.zheng.shou}@gmail.com

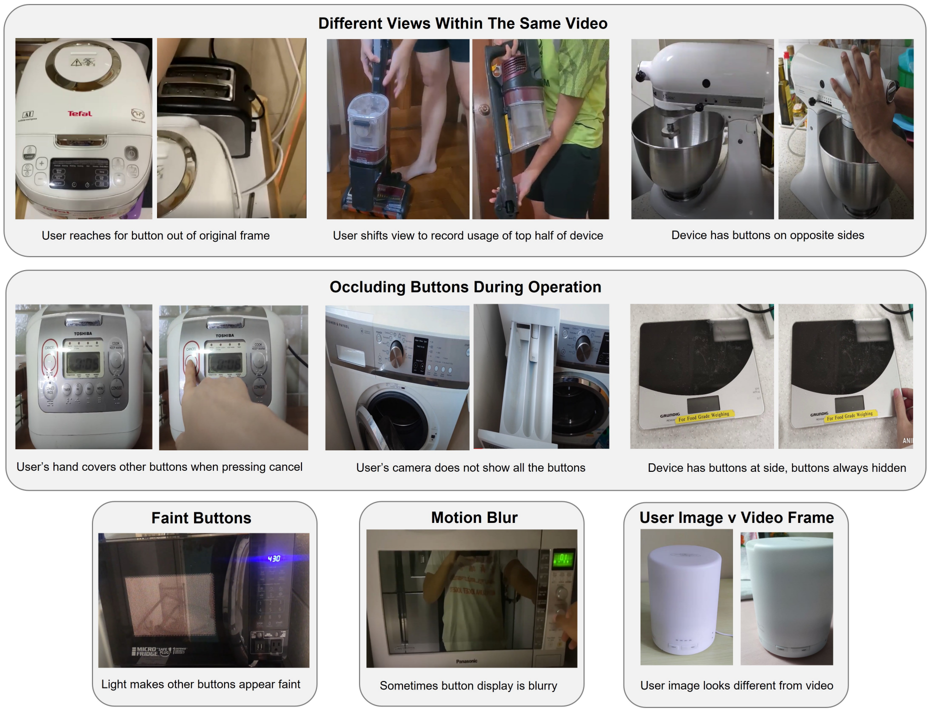
1 Dataset Quality Control

To ensure the quality of videos and scripts, we worked with participants closely (*e.g.* dissemination of comprehensive submission guidelines, checking of device before filming, hands-on recording guidance, review of transcripts). Researchers also conduct a quality check on the recorded videos before acceptance. In fact, 2-3% of videos have been rejected due to undesired quality (e.g. unclear video instructions). To ensure the quality of QA pairs and annotation, AssistQ also adopted all 4 quality control measures referenced in TGIF-QA [3]: gold standard annotation to help annotators to understand requirements, rejection and reviews (each video has been audited by at least 1 more person). The high quality of our data can be further confirmed by the high human performances at 95.8% recall@1.

2 Dataset Collection Source

We noticed that many virtual world simulators could help to produce a clean dataset, but we hold on collecting data in the real world. The main reason is that ultimately we want to deploy such AI assistant models in real environments such as smart glasses. To this end, at least we need to have testing data in real environment. Regarding training data, we agree it is a promising research direction to leverage a virtual environment. However, due to the significant domain gap between the existing virtual environments (*e.g.,* AI2Thor [4], AI Habitat [7]) and the real env, real-world training data at the scale provided by us is needed to adapt model trained in virtual to real. In this paper's scope, we focus on models trained only on the real environment; we consider it as future work to pre-train in the virtual environment first and then fine-tuning it on real-world data.

⋆ Equal Contribution. † Corresponding Author.



Different Views Within The Same Video

User reaches for button out of original frame

User shifts view to record usage of top half of device

Device has buttons on opposite sides

Occluding Buttons During Operation

User's hand covers other buttons when pressing cancel

User's camera does not show all the buttons

Device has buttons at side, buttons always hidden

Faint Buttons

Light makes other buttons appear faint

Motion Blur

Sometimes button display is blurry

User Image v Video Frame

User image looks different from video

Fig. 1. Some challenging examples in our AssistQ datasets.

3 Some Dataset Challenges

There are some interesting challenges in our dataset. These challenges are summarized in Figure 1. Some challenges include (a) changing views within the same instructional video; (b) occlusion of buttons due to hand placement, narrow views or hidden buttons; (c) faint buttons from lighting; (d) motion blurs from ego-motion or out-of-focus cameras; as well as; (e) user images taken under different angles and lighting from the video, reiterating the richness and complexity of AssistQ.

4 Dataset Comparison

Table 1 compares AssistQ to related datasets. As we can see, AssistQ is unique compared to others: (1) the video content is sourced from real-world, everyday situations in the egocentric perspective, (2) the question are affordance-centric *i.e.* how to use and execute functions of appliances, and multi-modal to more closely reflect the inputs that an AI assistant (*e.g.* AR glass) receives from the user. (3) Answers are a sequence of actions for the user to execute. AssistQ also has longer videos on average compared to most datasets.

Dataset	Video Source	Video Type	Questions	Question Modality		Answers	#Clips	#QA	Ave Dur (s)
				Text	Visual				
MovieQA [8]	Movie	Third-person	Factoid	✓	-	Single-step	408	14,944	202.7
TVQA [5]	TV show	Third-person	Factoid	✓	-	Single-step	21,793	152,545	76.2
AVSD [1]	Crowdsourced	Third-person	Factoid	✓	-	Multi-round	11,816	118,160	30
Embodied QA [2]	Simulated env.	Egocentric	Factoid	✓	-	Multi-step	-	5,281	-
AssistQ (ours)	Crowdsourced	Egocentric	Affordance	✓	✓	Multi-step	100	531	115

Table 1. Comparison of AssistQ with related datasets.

MovieQA [8]. The goal of MovieQA is to understand story plots in movies. 408 subtitled movies were collected together with their *Wikipedia* synopsis and *imsdb*/Described Video Service (DVS) scripts where available. In the first round of annotation, the annotators were shown the plot synopsis only and asked to create any number of QA pairs that can be localized to a set of sentences in the synopsis. Naturally, this led to questions that were plot-focused and less reliant on visual information. In the second round of annotation, annotators were asked to create 5 multiple-choice answers (1 right, 4 wrong) based on the synopsis and questions. Finally, each sentence of the synopsis was aligned to time-stamps on the video clips ($\sim$ 200s in length); the video clip and aligned QA pairs then formed the benchmark (Figure 2).

Q: How does E.T. show his happiness that he is finally returning home?
A: His heart lights up

Q: Why do Joy and Jack get married that first night they meet in Las Vegas?
A: They are both vulnerable and totally drunk

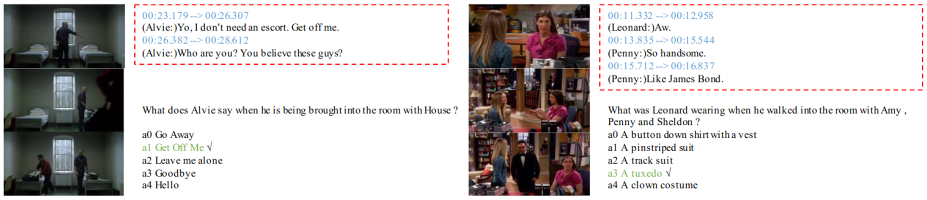
Q: Why does Forrest undertake a three-year marathon?
A: Because he is upset that Jenny left him

Q: How does Patrick start winning Kat over?
A: By getting personal information about her likes and dislikes

Fig. 2. Examples from the MovieQA dataset. 18% of questions were about *who*, followed by 12.4% about *why* and 9.7% about *what.* As seen in the examples, questions typically focused on plot developments with little reference to visual signals.

TVQA [5]. Similar to MovieQA, TVQA was created to understand human-centric plots in videos. 6 TV shows were segmented into 60/90-second clips, accompanied by subtitles and aligned transcripts. Annotators were shown the video clip and aligned subtitles, and encouraged to create questions in a 2-part format: [What/How/Where/Why/. . .] ___________ [when/before/after] ___________. The second part served to localize the question to the relevant moment in the clip and ground the question in visual signals, such as *What was House saying before he leaned over the bed?*. Annotators provided 5 multiple-choice answers (1 right, 4 wrong) and annotated time-stamps of the exact video portion required to answer the question (Figure 3).

AVSD [1]. In Audio Visual Scene-Aware Dialog (AVSD), an agent is given an input video, a dialog history and a follow-up question, and its goal is to generate a correct response (Figure 4). 11,816 videos of everyday human activ-



00:23.179 --> 00:26.307
(Alvie:)Yo, I don't need an escort. Get off me.
00:26.382 --> 00:28.612
(Alvie:)Who are you? You believe these guys?

What does Alvie say when he is being brought into the room with House ?

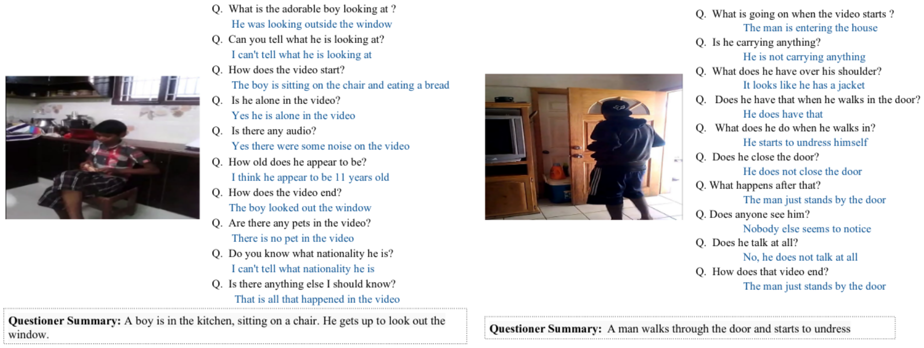
a0 Go Away
a1 Get Off Me √
a2 Leave me alone
a3 Goodbye
a4 Hello

00:11.332 --> 00:12.958
(Leonard:)Aw.
00:13.835 --> 00:15.544
(Penny:)So handsome.
00:15.712 --> 00:16.837
(Penny:)Like James Bond.

What was Leonard wearing when he walked into the room with Amy , Penny and Sheldon ?
a0 A button down shirt with a vest
a1 A pinstriped suit
a2 A track suit
a3 A tuxedo √
a4 A clown costume

Fig. 3. Examples from the TVQA dataset. The question had to be written with a [*when/before/after*] clause so that it is localized to a specific moment in the clip. 54% of questions were about *what*, followed by 15% about *who*.

ities were taken from the Charades human-activity dataset [6]. Each video was handed to a pair of annotators. The "Questioner" was tasked to ask questions about activities and events in the video clip, given only 3 video frames. The "Answerer" has access to the video and script, and answers the "Questioner" over a sequence of 10 questions. Once the conversation is complete, the "Questioner" is tasked to summarize the video.



Q. What is the adorable boy looking at ?
He was looking outside the window
Q. Can you tell what he is looking at?
I can't tell what he is looking at
Q. How does the video start?
The boy is sitting on the chair and eating a bread
Q. Is he alone in the video?
Yes he is alone in the video
Q. Is there any audio?
Yes there were some noise on the video
Q. How old does he appear to be?
I think he appear to be 11 years old
Q. How does the video end?
The boy looked out the window
Q. Are there any pets in the video?
There is no pet in the video
Q. Do you know what nationality he is?
I can't tell what nationality he is
Q. Is there anything else I should know?
That is all that happened in the video

Questioner Summary: A boy is in the kitchen, sitting on a chair. He gets up to look out the window.

Q. What is going on when the video starts ?
The man is entering the house
Q. Is he carrying anything?
He is not carrying anything
Q. What does he have over his shoulder?
It looks like he has a jacket
Q. Does he have that when he walks in the door?
He does have that
Q. What does he do when he walks in?
He starts to undress himself
Q. Does he close the door?
He does not close the door
Q. What happens after that?
The man just stands by the door
Q. Does anyone see him?
Nobody else seems to notice
Q. Does he talk at all?
No, he does not talk at all
Q. How does that video end?
The man just stands by the door

Questioner Summary: A man walks through the door and starts to undress

Fig. 4. Examples from the AVSD dataset. The "Questioner" asks a series of questions about the input video and the "Answerer" gives an answer at each round.

Embodied QA [2]. In Embodied QA, an agent is spawned at a random location in a simulated home environment and asked a question about the colour/location of an object. The objective of the agent is to navigate the environment through atomic actions (move forward, turn, *etc.*) and gather visual information to answer the question (Figure 5). The dataset is built on a subset of House3D environments, and the questions were written in specific formats to ensure they were answerable and unambiguous. For example, *location* questions were written as *What room is the `<OBJ>` located in?*, where `<OBJ>` is an unambiguous object that is query-able.

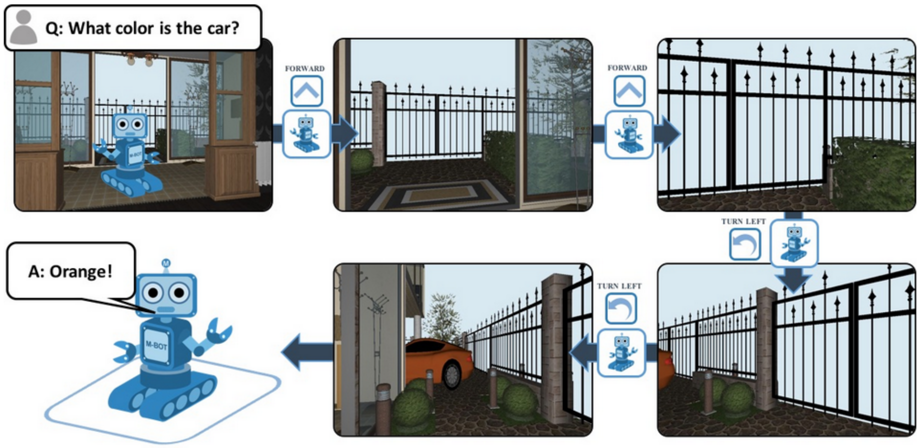


Fig. 5. Example from the EQA dataset. The agent is spawned at a random location in a simulated home environment and navigates the environment to answer a question about the location/colour of an unambiguous object.

5 Data Sample Visualisation

We attach an annotated example and an animated video to showcase the the intended use-case of AssistQ in AI assistants. The video shows 2 examples with egocentric instructional videos, and an example with a third-person instructional video.

Annotated example. We provide 1 annotated example from AssistQ to illustrate the format and information encapsulated in our annotations. The example contains the instructional video (video.mp4), question-answering pairs (qa.json), bounding box coordinates (buttons.csv), narration transcript (script.txt) as well as the image folder containing front-view image(s) of the device.

Examples with egocentric videos. We showcase 2 examples from our AssistQ dataset, namely the EF stove and Bosch washing machine, in an animated video (Figure 6). Through these examples, we demonstrate the use of AssistQ in our proposed Affordance-centric Question-driven Task Completion (AQTC) task. The AI assistant learns offline from an egocentric instructional video so that when the user asks a question (e.g. *stove:* How do I switch on the left stove at maximum heat), the AI assistant provides the user with a series of steps to perform (e.g. *stove:* 1. Press `[here]`, 2. Press `[here]` and 3. Select `[here]` mode 9). Each step is grounded to bounding boxes in the user's image, *i.e.* `[here]` refers to a labelled button/knob on the device. The bounding box is shown on the user's view of the device (*e.g.* through AR glasses) so the user knows the exact action to perform.

Example with third-person video. In the animated video, we also include a Shark vacuum example that had its instructional video recorded in third-person (Figure 7). Compared to the egocentric examples, the third-person example is

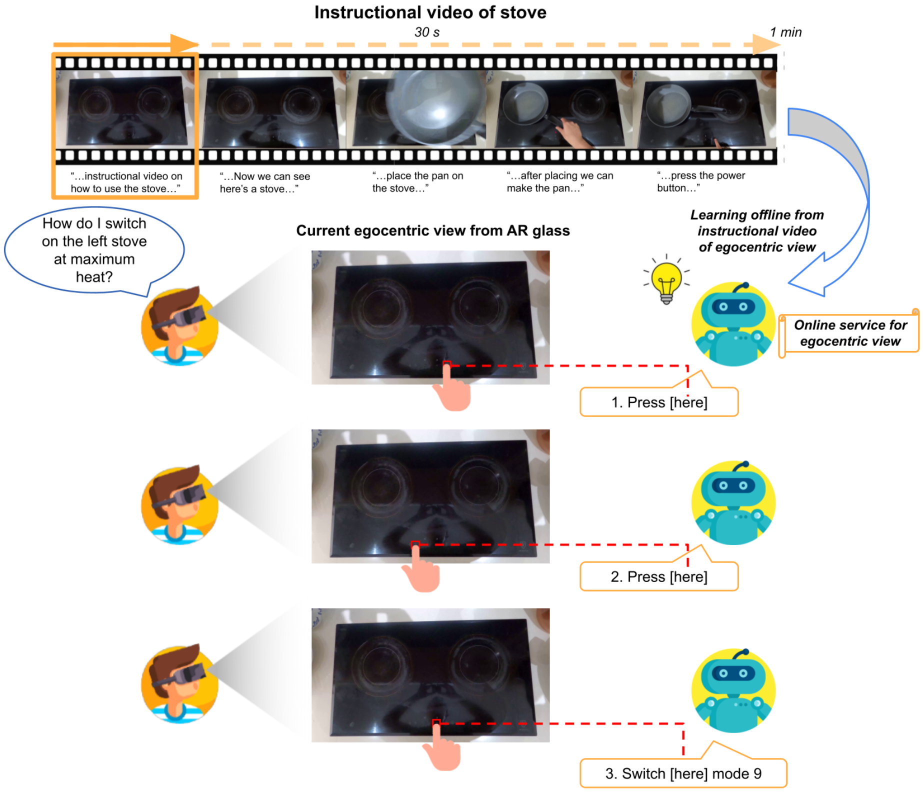

Instructional video of stove
30 s
1 min
"…instructional video on how to use the stove…"
"…Now we can see here's a stove…"
"…place the pan on the stove…"
"…after placing we can make the pan…"
"…press the power button…"
How do I switch on the left stove at maximum heat?
Current egocentric view from AR glass
Learning offline from instructional video of egocentric view
Online service for egocentric view
1. Press [here]
2. Press [here]
3. Switch [here] mode 9

Fig. 6. Illustration of AQTC task with an example from AssistQ. The AI assistant learns offline from a narrated instructional video. The user wonders how to use the appliance and shows the AI assistant their view. The AI assistant predicts the sequence of actions that the user should perform to accomplish the task, and guides the user through each step with bounding boxes over their view of the appliance.

more challenging as it requires the AI assistant to resolve visual information from the third-person video with that of the user's egocentric visual query. The example demonstrates that the AQTC task need not be confined to egocentric videos. We intend to extend the AssistQ dataset to include more examples with third-person videos, as it is more common for product videos/demos to be recorded in third-person. This will allow further research and development in AI assistants to benefit from our AQTC task and AssistQ dataset.

References

1. AlAmri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T.K., Hori, C., Anderson, P., Lee, S., Parikh, D.: Audio visual scene-aware dialog. In: CVPR. pp. 7558–7567 (2019)

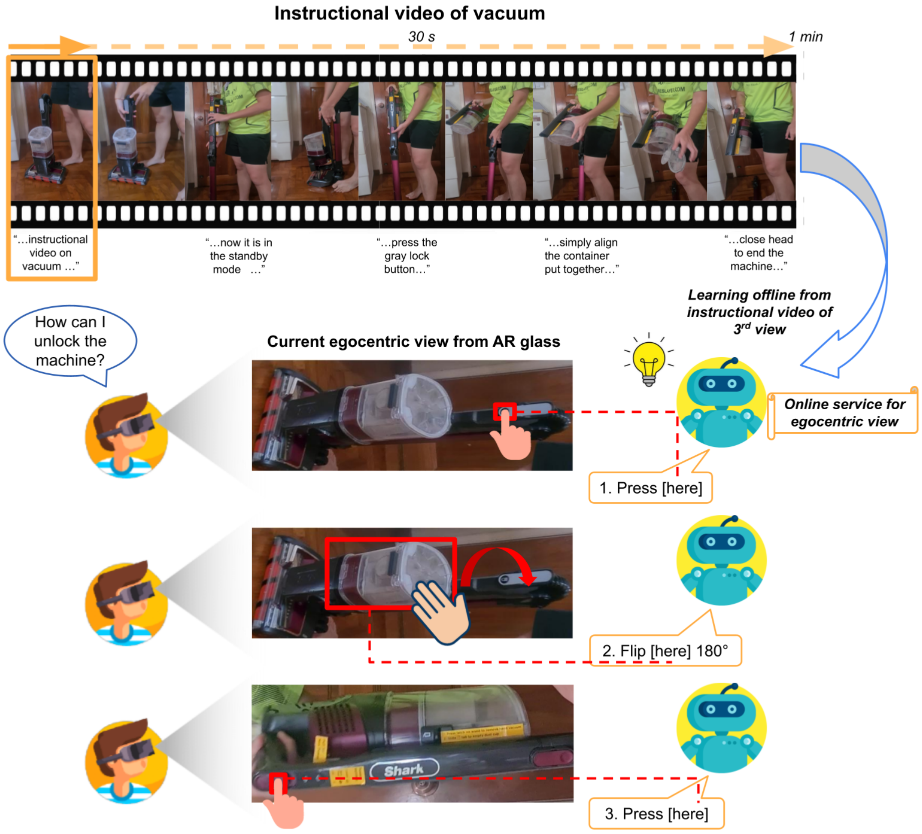

Instructional video of vacuum
30 s
1 min
"…instructional video on vacuum …"
"…now it is in the standby mode …"
"…press the gray lock button…"
"…simply align the container put together…"
"…close head to end the machine…"
Learning offline from instructional video of 3rd view
How can I unlock the machine?
Current egocentric view from AR glass
Online service for egocentric view
1. Press [here]
2. Flip [here] 180°
Shark
3. Press [here]

Fig. 7. Illustration of AQTC task with instructional video recorded from the third-person perspective. The task remains largely unchanged, except that the model has to resolve visual signals from the third-person to the first-person image from the user. This presents new challenges for the AQTC task and we are expanding the AssistQ dataset to include more of such scenarios.

2. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Embodied question answering. In: CVPR. pp. 1–10 (2018)
3. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: TGIF-QA: toward spatio-temporal reasoning in visual question answering. In: CVPR. pp. 1359–1367 (2017)
4. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A., Farhadi, A.: AI2-THOR: An Interactive 3D Environment for Visual AI. arXiv:1712.05474 (2017)
5. Lei, J., Yu, L., Bansal, M., Berg, T.L.: TVQA: localized, compositional video question answering. In: EMNLP. pp. 1369–1379 (2018)
6. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV. pp. 510–526 (2016)
7. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D., Maksymets, O., Gokaslan, A., Vondrus, V., Dharur, S.,

Meier, F., Galuba, W., Chang, A., Kira, Z., Koltun, V., Malik, J., Savva, M., Batra, D.: Habitat 2.0: Training home assistants to rearrange their habitat. In: NeurIPS. pp. 251–266 (2021)

8. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: CVPR. pp. 4631–4640 (2016)