Ego-EXTRA: video-language Egocentric Dataset for EXpert-TRAinee assistance Supplementary Material

Francesco Ragusa[*1,2], Michele Mazzamuto[*1,2], Rosario Forte[1], Irene D'Ambra[1], James Fort[3], Jakob Engel[3], Antonino Furnari[1,2], Giovanni Maria Farinella[1,2]

[1]Department of Mathematics and Computer Science - University of Catania, Italy
[2]Next Vision s.r.l. - Spinoff of the University of Catania, Italy
[3]Meta Reality Labs Research, USA

Abstract

This supplementary document provides additional details on the acuisition and annotation of the Ego-EXTRA dataset, including additional statistics, hardware details and acquisition protocol. We further report additional details on the two-steps human validation of the QA sets and additional qualitative examples not included in the main paper. This material is intended to complement the main manuscript titled: Ego-EXTRA: video-language Egocentric Dataset for EXpert-TRAinee assistance. All data and code are available at https://fpv-iplab.github.io/Ego-EXTRA/.

1. Ego-EXTRA Dataset

1.1. Subjects

Data collection was carried out with the participation of 33 trainees and 4 experts aged between 18 and 52 years. All participants are volunteers who provided their privacy consent and authorization to acquire data in the considered environments using the described protocols, transcribe audio conversations, and publicly release the resulting data for research purposes. Table 1 reports the list of trainee with information about Gender, Age, and Profession.

1.2. Data Acquisition

The Aria glasses worn by trainees for data acquisition are equipped with the visual sensors such as two monochrome scene/SLAM cameras, one RGB camera, and two eye-tracking cameras as well as with non-visual sensors like two inertial measurement units (IMUs), seven-channel spatial microphone array, a magnetometer, a barometer, a GPS receiver, and both Bluetooth and WiFi beacons. For each acquisition session, Aria glasses are connected to a mobile phone using the ARIA mobile companion app [**?**], allowing the user to manage the data capture process by selecting an acquisition profile. In particular, as shown in Figure 1 we used a custom profile with the following characteristics:

- RGB camera with a resolution of 1408x1408 at 15 FPS;
- SLAM camera at 30 FPS;
- Eye-tracking cameras at 30 FPS;

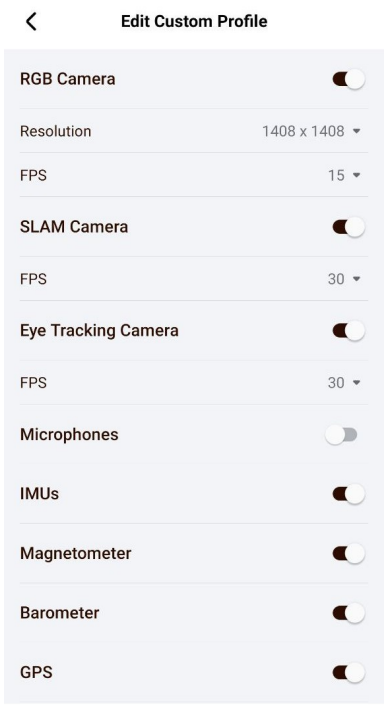

Figure 1. Screenshot of the custom profile used for the acquisition.

Pro-Active:

E: Perfect, and the butter will start to melt, and you need to avoid making lumps with the flour, so you need to stir it. I advise you to lower the butter.
T: Okay?
E: Compared to the flour, so make it adhere to the surface of the pan. Okay. Perfect. Wait a moment for it to melt a bit, and set it to four, too. So, the pan doesn't come, doesn't come. Read this signal it's a signal. Move it to the right, move it to the next position, yes.
T: Here.
E: Yes. Put to Four there, or K, perfect. Now, let the butter melt; it will start to melt and we need to mix it with the flour, avoiding any lumps from forming.
T: Okay.
E: Nothing, that little flame was the other burner that turned off. Everything is normal, right? If you think it's too low and the butter is not melting and you want to speed things up, instead of four, turn it to five, you decide, okay?
T: Ok I'm setting it to five.
E: Perfect. It seems that the butter is starting to melt.

On-Demand:

T: Since it's already melted for a few seconds, can I leave it? The bechamel sauce.
E: I advise you to always keep stirring.
T: Makes it turn, and then I'll do it, I'll do it with.
E: You should do, if necessary.
T: OK, okay.
E: Lower the temperature, set it to one, set the bechamel sauce to one and you can leave it. If it's very low, it shouldn't form volumes.
T: The cooktop occasionally turns off, so I avoid that by positioning better the pan, right? In the meantime, let's press this.
E: I see that, that's perfect, good job.
T: We help the spinach too?
E: Wait.
T: OK.
E: Do you know how you can help? By adding a finger of water to that spinach.
T: A glass?
E: Yes, and raise the temperature of the bechamel sauce again if it seems soft.
T: Yes, yes, OK, OK. Another minute, precisely, I'll recover the bechamel sauce. Can we drain the spinach?
E: Yes, but be careful not to burn yourself.
T: Is this strainer okay?
E: You can go.
T: Perfect.

Figure 2. Example of trainee/expert conversation acquired with our pro-active (left) and on-demand (right) protocols.

- IMUs;
- Magnetometer;
- Barometer;
- GPS;
- WiFi and Bluetooth.

The collected data were then exported in VRS (Visual Record Stream) format, which provides standardized methods to store images, audio, and discrete sensor data in compact, evolution-resilient records that are already synchronized. VRS files are then processed using the ARIA SDK[1] to extract the trainee's RGB egocentric video. Synchronized eye gaze and SLAM are obtained using the Project Aria Machine Perception Services[2] as shown in Figure 3. Audio conversations have been transcribed using a commercial software. An example of trainee/expert dialogue obtained with both acquisition protocols is reported in Figure 2.

[1] https : / / facebookresearch . github . io / projectaria_tools/docs/ARK/sdk

[2] https : / / facebookresearch . github . io / projectaria_tools/docs/ARK/mps

1.3. Synchronization and Raw Data Processing

1.3.1. Gaze Projection

To allow a spatial alignment between the egocentric video streams coming from the ARIA device and the smartphone, the trainee was instructed to observe a QR code placed in the environment before starting the acquisition session. The QR code is used to estimate a rigid transformation, allowing the expert's gaze to be projected onto the trainee's viewpoint. The Expert's video stream is recorded together with the gaze signals collected through the Tobii pro device. To allow temporal synchronization between the egocentric video stream and the two-way audio conversation, the trainee and the expert begin each collection by following a countdown to provide a signal useful for temporal synchronization. Based on the recorded countdown, the video pairs are manually synchronized. We detect the QR codes on both the expert's and trainee's videos in the first 60 seconds of each video and compute a 3×3 homography matrix H, which stays the same for the duration of the video, that maps the expert's frame to the corresponding trainee's

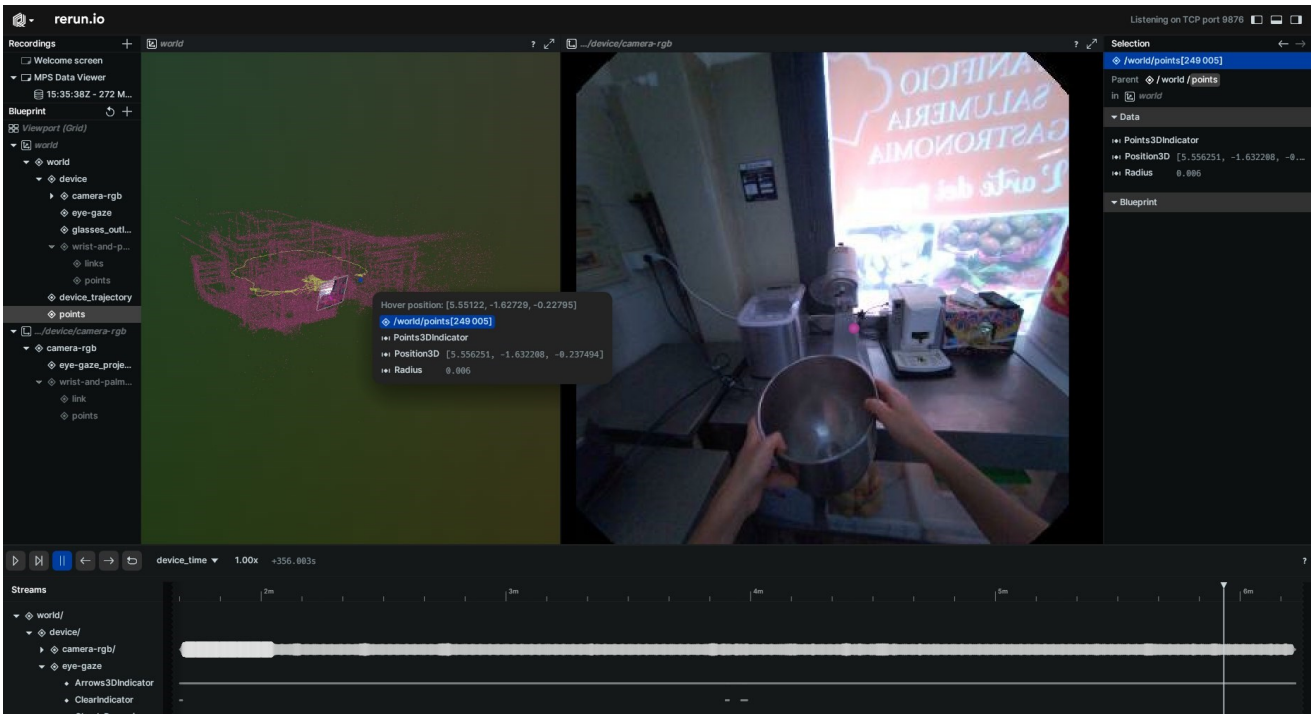

Figure 3. Example of SLAM and eye gaze obtained from the MPS services.

frame. The expert's gaze is therefore projected to the reference frame of the RGB video collected with ARIA, so that both the expert's and trainee's gaze signals are mapped to the same reference system (see Figure 4 and 5).

1.3.2. Translation and correction

Due to privacy issues, the acquired audio conversations cannot be shared. Therefore, we transcribed all conversations using professional software. We then prompted a LLama 3.1 model to translate the transcriptions into English, correcting any grammar or spelling errors. Each phrase was assigned a timestamp derived from the audio and a unique ID. Table 2 reports some examples of corrected transcriptions.

2. Ego-EXTRA VQA Benchmark

2.0.1. QAs Extraction

Transforming trainee-expert conversations turns into question-answer sets is challenging. To overcome this issue, we used conversation transcripts to prompt a Llama 3.1 405B model [**?**] to generate multiple-choice question answer pairs based on the conversations using a specifically designed prompt reported in the following:

I will provide you with a transcript of a video. Simulate watching the video and generate questions that can only be answered well if you are watching the video. For each question, generate one correct answer and four incorrect answers (so a total of 5 options). The incorrect answers should be plausible mistakes that could occur during the execution of that action. Avoid trivial questions. Act as a domain expert and generate multiple-choice questions based on the questions asked by (T:) during

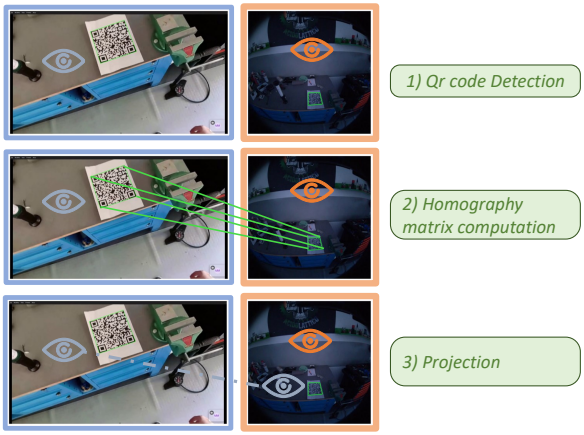

Figure 4. Projection of the expert gaze into the reference point of view of the trainee. In the end of the process, both expert's and trainee's gaze are in the same coordinate system as ARIA's RGB video.

Subject Type	Gender	Age	Profession
Trainee	M	29	PhD Student
Trainee	M	23	PhD Student
Trainee	F	45	Grant Researcher
Trainee	F	36	Unemployed
Trainee	M	25	Master Student
Trainee	F	25	Master Student
Trainee	M	23	Master Student
Trainee	M	46	Teacher
Trainee	F	30	PostDoc
Trainee	F	24	City Councilor
Trainee	F	24	Bachelor Student
Trainee	M	18	High School Student
Trainee	M	25	Waiter
Trainee	M	52	Nurse
Trainee	M	23	Bachelor Student
Trainee	F	23	Master Student
Trainee	M	33	Researcher
Trainee	M	25	PhD Student
Trainee	F	22	Master Student
Trainee	F	45	Housewife
Trainee	F	47	Artist
Trainee	M	22	Bachelor Student
Trainee	M	28	PostDoc
Trainee	M	23	Master Student
Trainee	F	23	Master Student
Trainee	F	23	Bachelor Student
Trainee	F	24	Bachelor Student
Trainee	F	24	Bachelor Student
Trainee	M	29	Psychologist
Trainee	M	28	PostDoc
Trainee	F	24	Bachelor Student
Trainee	F	24	Master Student
Trainee	F	22	Waiter
Expert	M	41	Assembly
Expert	F	29	Bakery manager
Expert	M	47	Bike Shop Manager
Expert	F	45	House Cook

Table 1. We reported the list of people engaged in the data acquisition process highlighting their gender, age, and profession.

the provided transcript. Create as many questions as you think are necessary, judging by the length of the transcript and how many questions the apprentice asks (do like from 7 to 15 questions). Each question should include the subject. Never mention the expert or the trainee.

With this prompt the model generates a question, the correct answer and four plausible but incorrect answers as reported in Table 3.

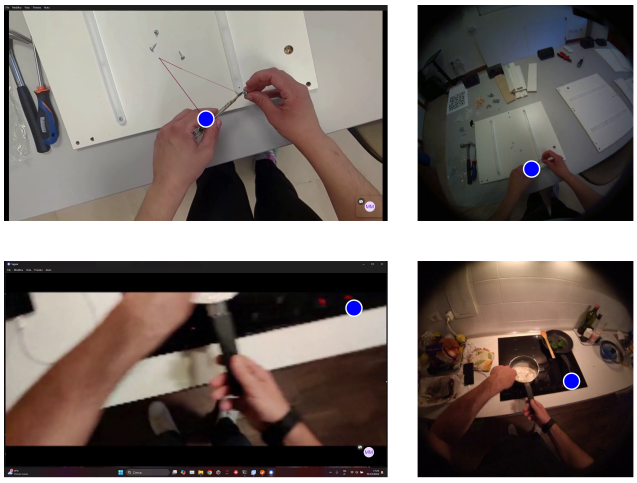

Figure 5. On the left, the trainee's video is streamed to the expert's laptop. On the right, the expert's gaze is reprojected onto the video acquired with the ARIA glasses.

2.0.2. Human Validation

In the initial validation phase, six human annotators reviewed the Question-Answering (QA) candidates. Using a dedicated web interface, each annotator was presented with the video clip, conversation transcript, the correct answer, and a set of distractors for each question. Their task was to flag potential issues via checkboxes, including transcription errors, semantically flawed questions, or excessive similarity between the correct answer and the distractors.

To scale up the validation process, we used the results from this initial phase to create a qualification test for Amazon Mechanical Turk (AMT). The test comprised questions that achieved high inter-annotator agreement among our internal team. For the large-scale validation, we selected only AMT workers with a historical approval rate of at least 90

Examples of questions discarded during this process are reported in Table 4. The web interface used by the annotators is shown in Figure 8.

2.0.3. Grounding

After the textual validation step, also in this case we perform a two phase grounding validation to ensure that each QA candidate is semantically and visually anchored to the video content. This step is crucial to verify that the question is not only well-formed, but also contextually supported by the video segment and the corresponding dialogue.

Each QA item is manually labeled as *GROUNDED*, *NOT GROUNDED*, or *DISCARD* using a dedicated annotation interface by our six internal annotators. Annotators are presented with the video clip and its transcript, where the current conversation turn is highlighted in color and the surrounding turns are shown in grey for context.

A question is marked as *GROUNDED* if it is clearly sup-

Original Transcription	Corrected Transcription
I have to puttthewater inthe spin acid	I have to put the water in the spinach
I mean I don't see lamps	I mean I don't see lumps
Is there a specific order in which I have to crew or is it indifferent?	Is there a specific order in which I have to screw or is it indifferent?
No, you can pass it withthe clock	No, you can pass it with the cloth
That one, thatsilver, this biexactly, thisexam the thisallen key, yes,ok	That one, that silver, this big exactly, this exagonal, this allen key, yes, ok

Table 2. Some examples of transcription errors that have been corrected.

ported by the video and coherent with the dialogue. It is labeled as *NOT GROUNDED* if it is unrelated to the visual or textual context, and as *DISCARD* if it is of low quality or not relevant. Annotators also indicate whether the video contains the correct answer to the question. Figure 6 shows the interface used for this task. The distribution of annotations for Scenario 3 across the labeling categories is summarized in Figure 7. On average, each annotator spent approximately 498.44 seconds completing this task.

Once we obtained labels of grounding for one video per scenario we used them as a qualification set (Similarly to the Human Validation step) to select AMT workers who have an acceptance score above 90% and a perfect score on the qualification set.

2.0.4. Baselines

The following prompt was used for both language-only and video-language models:
You are an expert guiding the procedure shown in the video. The question is: '{question}'. "Choose the correct answer by selecting one number from the following options:" + "".join([f"{i+1}) {q}" for i, q in enumerate(options)]) + "Reply with ONLY the number of the correct answer (1, 2, 3, 4, or 5). Do not explain or justify. Reply with a SINGLE number."

In the case of LLMs, the video is not provided, and they rely solely on the input textual prompt.

Sample Human Baseline We provide a human baseline to compare the discrepancy in understanding between humans and state-of-the-art video-language models. We sampled an average of 54.25 questions per scenario, obtaining a total of 217 questions. We designed a web tool to allow experts to answer the questions while observing the related video clips. The experts involved in answering the questions are the same who participated in the data acquisition process. We collected all the answers and computed the human baseline. Example of the web tool interface is shown in Figure 9.

3. Experiments

Figure 10-11 show qualitative results obtained by the adopted baselines in our VQA benchmark.

3.1. Qualitative Results

Qualitative results are shown in Figures 10 and 11.

Transcript

ID 9: E: Now, let's focus on the next steps.
ID 10: T: Alright, which of the two wheels should I remove first?
ID 11: T: OK, I see.
ID 12: E: You should remove the front wheel.
ID 13: T: Great. Is the angle of the bike okay, or should I adjust it?
ID 14: T: OK, understood.

QA


```
{
"id": 1,
"text": "Which wheel should be removed first?",
"question_involved_ids": "10-13",
"options": [
    "The front wheel",
    "The rear wheel",
    "Both wheels",
    "Only the left wheel",
    "Only the right wheel"
],
"correct_answer": "The front wheel",
"answer_involved_ids": "13",
"question_start_time": "00:00:09,000",
"question_end_time": "00:00:15,000",
"answer_start_time": "00:00:16,000"
}
```


Table 3. An example of QA generation from the transcript of the trainee/expert conversation.

Acknowledgements

This research is supported by Meta Reality Labs, Next Vision s.r.l. and by the project Future Artificial Intelligence Research (FAIR) – PNRR MUR Cod. PE0000013 - CUP: E63C22001940006.

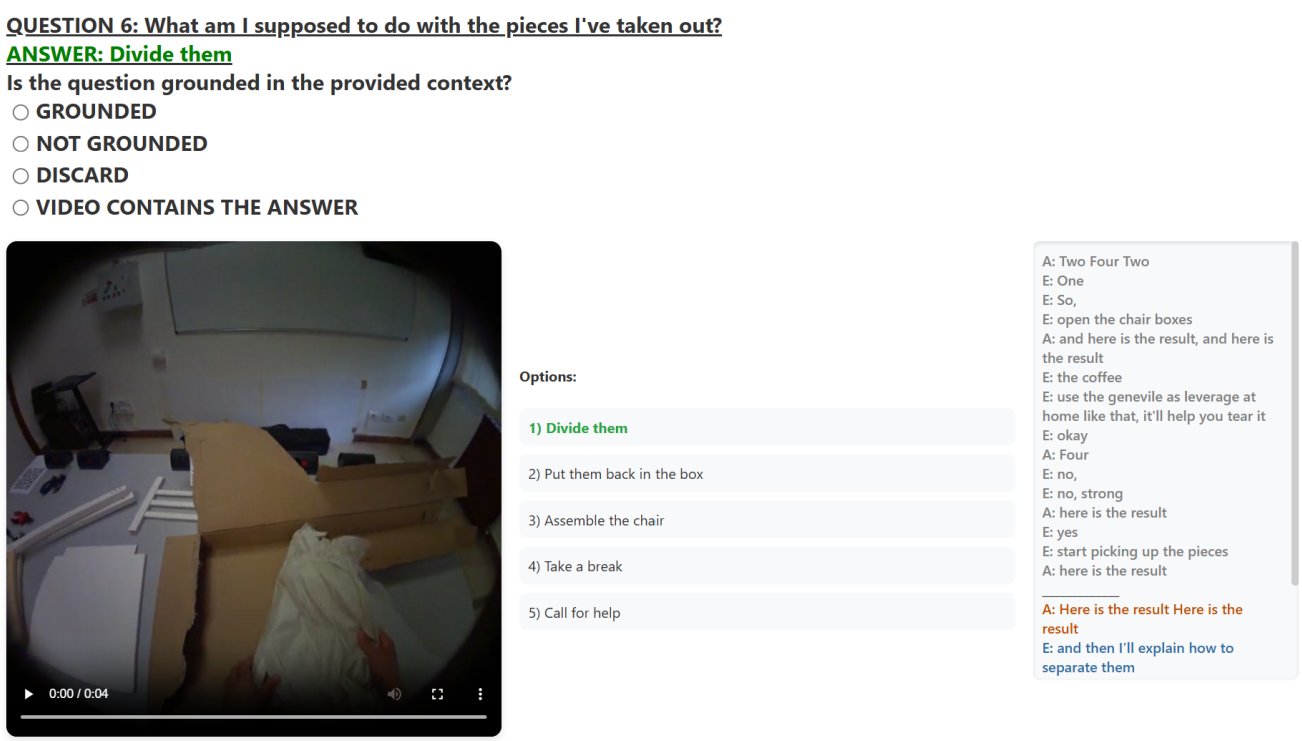

Figure 6. Web tool interface used for grounding validation

QA 1	QA 2
"question": "What is the correct way to insert the wheel?", "options": ["Insert the wheel from here", "Insert the wheel from there", "Do not insert the wheel", "Insert the wheel with the patches", "Insert the wheel without the patches"], "correct_answer": "Insert the wheel from here"	"question": "What is the final state of the chair after following the instructions?", "options": ["Assembled", "Partially disassembled", "Fully disassembled", "Broken", "Reassembled"], "correct_answer": "Fully disassembled"
QA 3	**QA 4**
"question": "What is the purpose of the tare function in the stand mixer?", "options": ["To measure the weight of the ingredients", "To mix the ingredients together", "To adjust the speed of the mixer", "To reset the mixer to zero", "To prepare the mixer for baking"], "correct_answer": "To reset the mixer to zero"	"question": "What is the purpose of crushing the spinach with a fork?", "options": ["To make the spinach more tender", "To make the spinach more flavorful", "To help cook the spinach faster", "To make the spinach more crunchy", "To separate the spinach leaves"], "correct_answer": "To separate the spinach leaves"

Table 4. Examples of discarded questions by human validation.

Question Text	**Options**
What is the first action to take when disassembling the drawer?	1. Pull out the drawer 2. Remove the plastic clips 3. Remove the wooden dowels 4. Unscrew the screws 5. Use the pliers
What is the first action to take when disassembling the drawer?	1. Grab the pliers 2. Pull out the drawer 3. Remove the screws 4. Extract the wooden pegs 5. Remove the plastic clips
What is the initial step in taking apart the drawer?	1. Pull out the drawer 2. Loosen the fasteners 3. Pick up the pliers 4. Remove the plastic clips 5. Take out the wooden rods
What is the initial step in taking apart the drawer?	1. Loosen the fasteners 2. Pick up the pliers 3. Remove the plastic clips 4. Remove the wooden dowels 5. Open the drawer
How do you begin disassembling the drawer?	1. Extract the wooden pegs 2. Remove the plastic clips 3. Pull out the drawer 4. Pick up the pliers 5. Unscrew the screws
How do you begin disassembling the drawer?	1. Unscrew the screws 2. Pick up the pliers 3. Remove the plastic clips 4. Take out the wooden rods 5. Pull out the drawer

Table 5. Examples of obtained multiple-choice question answers considering their variants.

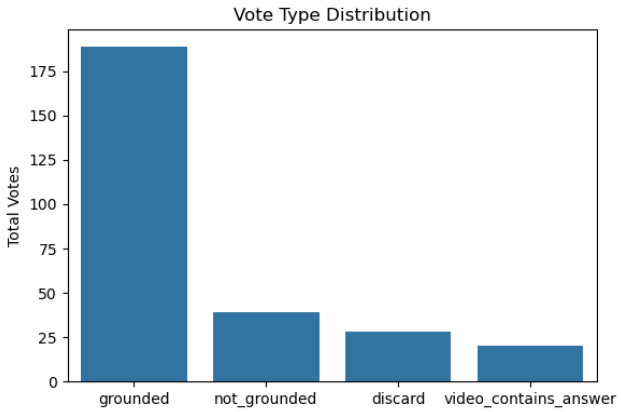

Figure 7. Distribution of grounding labels and average annotation duration.

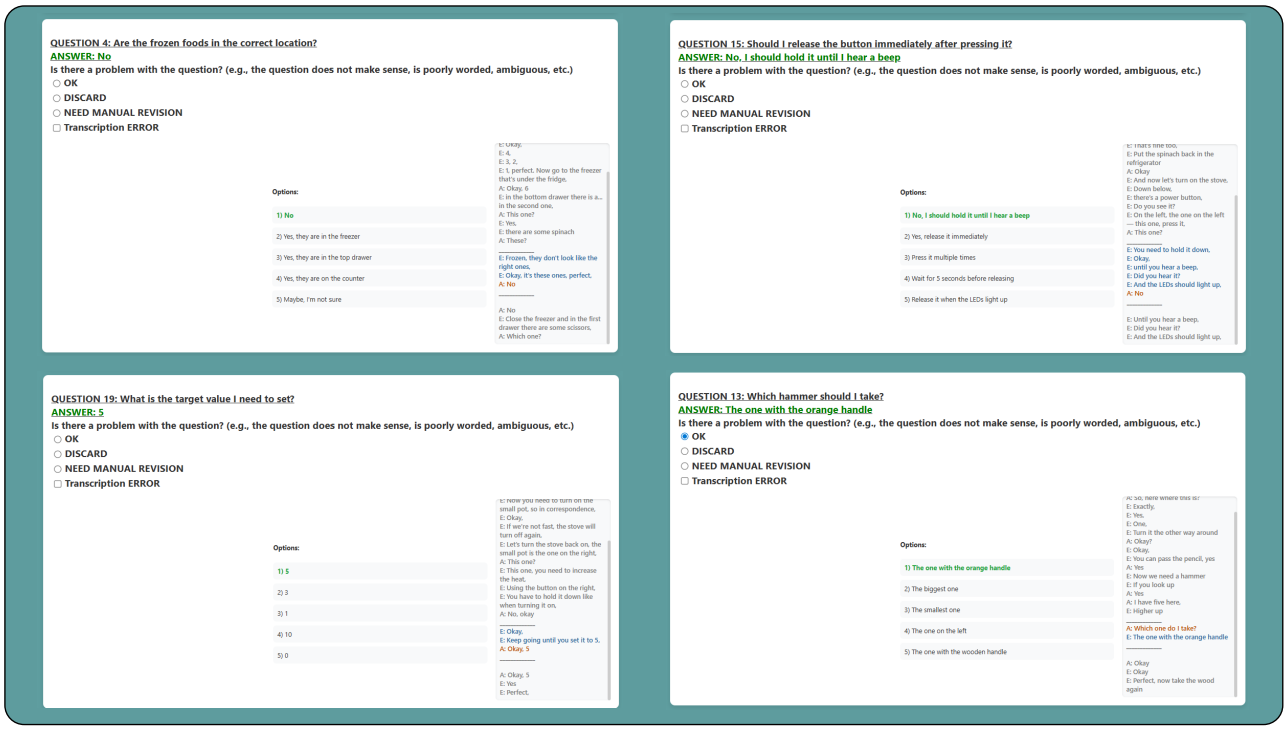

Figure 8. Tool used for human validation. Each question and option is provided with a checkbox.

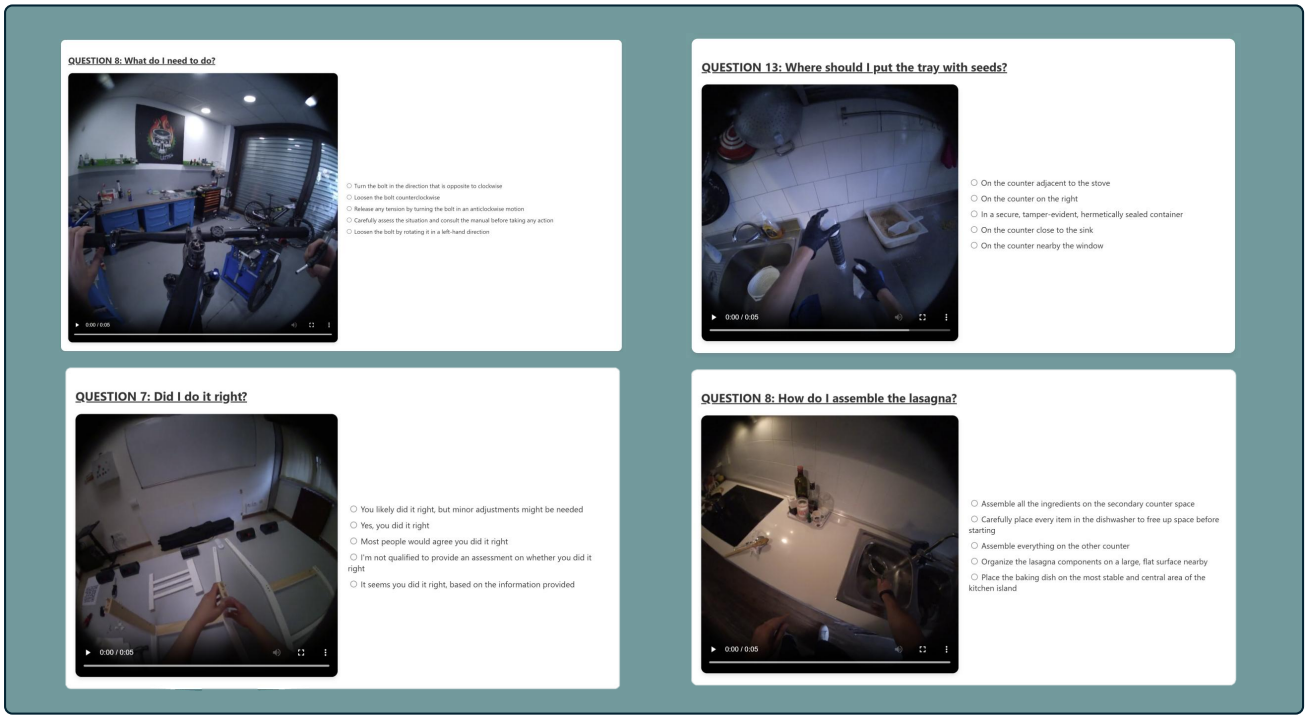

Figure 9. Web tool interface used to obtain the sample human baseline.

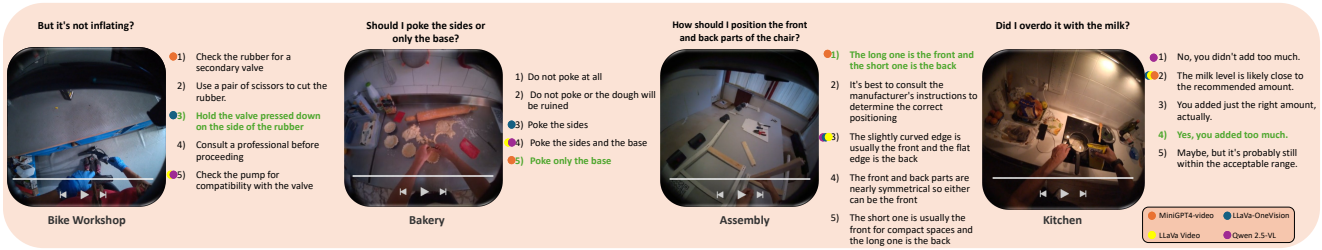

Figure 10. Qualitave results of the proposed VQA benchmark. Correct answer in green, baselines predictions marked with colors.

Figure 11. Qualitative results of the proposed VQA benchmark. Correct answer in green, baselines predictions marked with colors.