Supplementary Materials for "Visual Intention Grounding for Egocentric Assistants"

1. Visualization

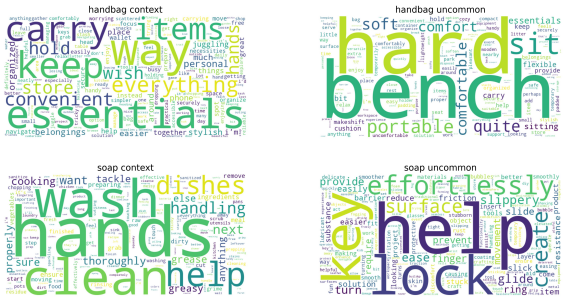


Figure 1. Word cloud visualizations of object affordances under different intention contexts. (a) shows word clouds for the object "handbag", where context-aware intentions highlight its primary function of carrying essentials, while uncommon intentions repurpose it as a cushion for a hard bench. (b) presents word clouds for "soap", commonly associated with handwashing or dish cleaning, but also serving an uncommon purpose, making a stuck lock slippery for easier unlocking.

To quantitatively analyze the linguistic characteristics of these two intention categories, we visualize their distinctive vocabulary distributions through word cloud representations (Figure 1).

To better understand the challenges of visual intention grounding, we visualize grounding results for both context intention sentences and uncommon intention sentences (Figure 3).

2. Dataset Collection Process

2.1. Constructing the Object Primary Affordance List

Inspired by [2], which classifies uncommon affordances as secondary or tertiary, we construct an object primary affordance list (Table 1). This list serves as the foundation for defining both contextual and uncommon object uses in our dataset. A context-aware intention sentence expresses the need for an object's primary function, incorporating environmental cues, whereas an uncommon intention sentence describes an alternative use beyond its primary function.

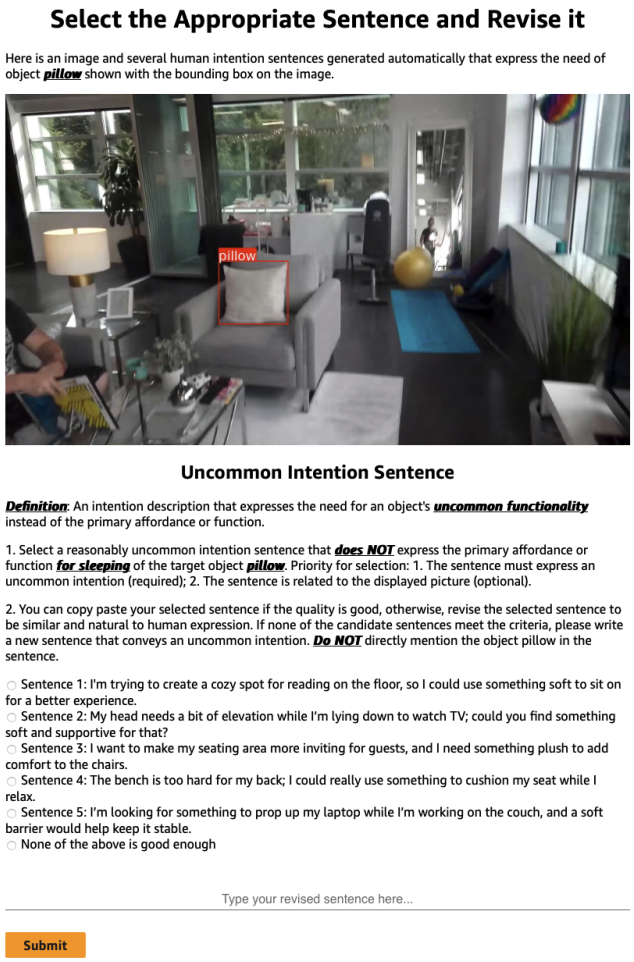
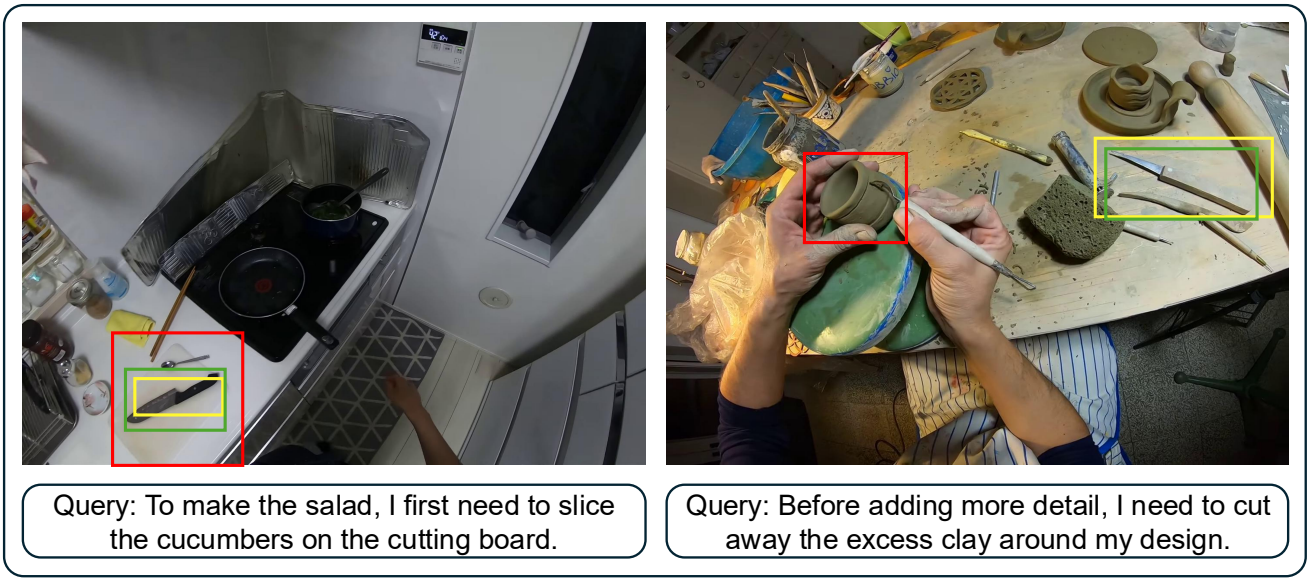


Figure 2. Example of the MTurk annotation interface for refining and validating generated intention sentences. Annotators assess the quality and coherence of each sentence.

2.2. Collecting Intention Sentences

After establishing each object's primary function, we collected five context-aware intention sentences and five uncommon intention sentences per object (Table 2). These sentences were carefully designed to cover diverse real-world scenarios, serving as in-context learning [1] exam-



(a) Context intention sentence grounding.

(b) Uncommon intention sentence grounding.

Figure 3. Visualization of grounding results for different intention types. (a) Context intention sentence grounding. (b) Uncommon intention sentence grounding. In both figures, Naive SFT MiniGPTv2 outputs are shown in red, our RoG SFT MiniGPTv2 outputs are shown in yellow, and ground truth bounding boxes are shown in green.

ples for GPT-4 generation. The prompts used to generate the candidate sentence set for each sample in EgoIntention are provided in Table 3.

2.3. Crowdsourced Annotation via MTurk

To refine and validate the dataset, we employed Amazon Mechanical Turk (MTurk). The MTurk interface (Figure 2) was designed to collect user feedback on generated sentences, ensuring linguistic clarity and coherence.

Additionally, we used a separate MTurk interface (Figure 4) to collect alternative objects and bounding box annotations, leveraging human judgment to ensure high-quality annotations of valid object alternatives and accurate object locations.

Table 1. Common objects and their primary functions, forming the basis for defining context-aware and uncommon affordances in our dataset.

Object	**Primary Function**
Book	For reading
Pillow	For sleeping
Telephone	For talking
Drum (musical instrument)	For playing music
Drill	For punching
Television Set	For watching TV programs
Knife	For cutting
Table	For laying items
Handbag	Holding daily necessities
Chair	For sitting
Soap	For cleaning

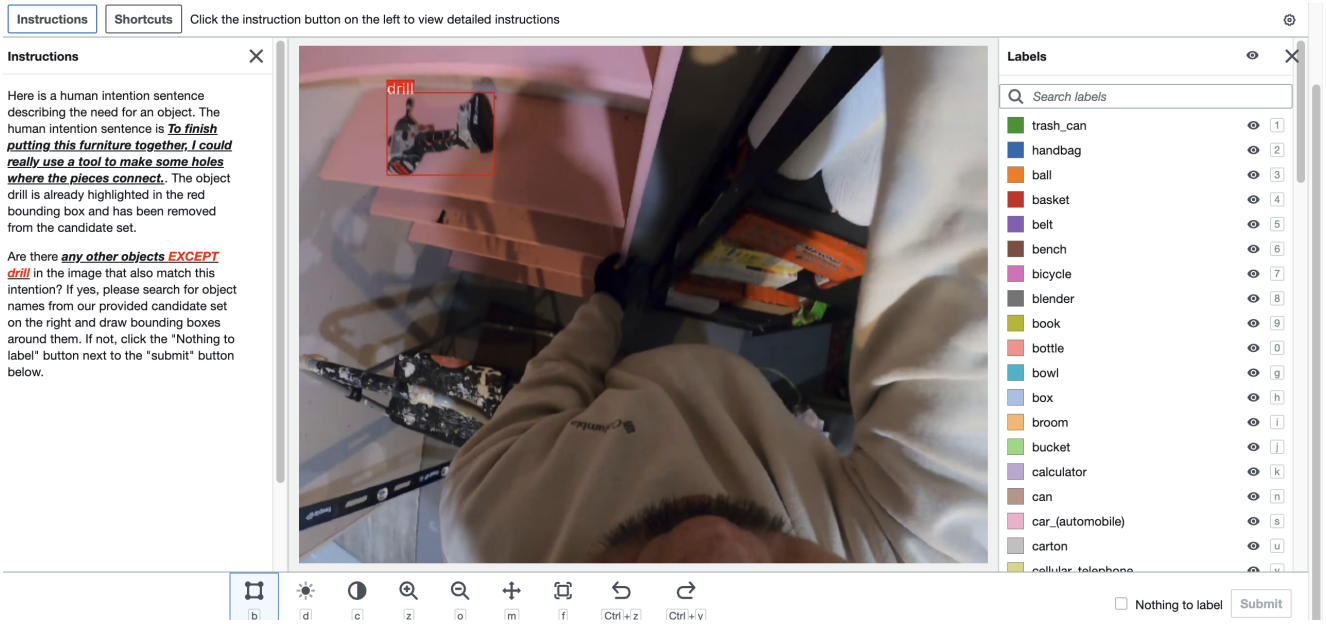


Figure 4. Example of the MTurk annotation interface for collecting alternative object annotations and bounding boxes. This process ensures diverse and high-quality grounding annotations.

Table 2. Examples of context-aware and uncommon intention sentences for selected objects. These sentences serve as in-context learning examples when prompting ChatGPT to generate intention queries for each sample in the EgoIntention dataset. Context-aware sentences emphasize the object's primary function within its environment, while uncommon sentences explore alternative, less conventional uses.

Object	**Example Intention Sentences**
Chair	**Context Intention Sentences:**
	After a long meeting, I need somewhere to sit and respond to emails at my desk.
	We have one extra guest for dinner and need to make sure everyone has a spot at the table.
	I need to sit down comfortably to study for my exams, but there's no place to rest.
	It would be perfect to have a place here where I can relax and read a book next to my coffee table.
	While getting ready, it would be helpful if I had somewhere to sit and do my makeup.
	Uncommon Intention Sentences:
	There are too many small tools scattered around, and I need a platform to keep them organized and within reach while I work.
	A sudden wind has picked up, and I need to anchor the picnic blanket down to prevent it from blowing away.
	I need to reach the ceiling to replace the burnt-out bulb, but nothing seems to be quite tall enough around here.
	I need to get that book from the top shelf, but I can't quite reach it from the floor.
	The soap and sink are too high for the kid. Can you grab something to stand on so they can reach?
Soap	**Context Intention Sentences:**
	My hands are really greasy after working on the car, and I need to clean them before lunch.
	Before I touch anything in the living room, I need to wash up thoroughly.
	I need to ensure everything is sanitary after handling this raw chicken.
	After my workout, I need something to help freshen up in the shower.
	I need something to help get these dishes sparkling clean.
	Uncommon Intention Sentences:
	I need something that can help lubricate this stubborn zipper so it moves smoothly again.
	I need something that can create a fog-resistant coating on this mirror to keep it clear.
	I need something that can provide a good lather for a smooth shave since I don't have any shaving cream.
	I need something that can help this ring slide off my finger more easily.
	I need something that can lubricate this key so it turns smoothly in the lock.

Table 3. Comparison of context-aware and uncommon intention sentence prompts used in GPT-based sentence generation. The prompts guide the model to generate implicit queries requiring contextual reasoning or uncommon object affordances.

Prompt Type	**Example Prompt**
Context Intention Sentence Prompt	**[Role system]** You are in a real-world situation as shown in the input image and use objects in your environment to achieve a specific goal. You need to communicate your intention to an AI robot using natural language, without explicitly naming the object you want. Generate a sentence that requires the use of the target object as shown in the input image. Incorporate contextual information, relationships with other objects, or subtle visual cues. The sentence should be natural and require the AI to understand the context and your implied goals. **Key Points:** 1) Understand the image scenario by visually perceiving the input egocentric image. 2) The generated sentence should emphasize contextual reasoning and implied needs, sounding as natural as everyday human language. 3) Example sentences suggest a goal without directly naming the required object. 4) Sentences should involve relationships between multiple objects or subtle visual cues. **[Role user]** Here are some examples for generating sentences that express the need of target object {OBJECT}. **[Role user and assistant]** {Five IN-CONTEXT LEARNING examples}. **[Role user]** Image scenarios are described in examples, however to finish the following query you need to understand the image scenario from the input image. Now please generate five sentences for target object: {OBJECT} (As shown in the input image, referred by the red bounding box).
Uncommon Intention Sentence Prompt	**[Role system]** You are in a real-world situation as shown in the input image and use objects in your environment to achieve a specific goal. You need to communicate your intention to an AI robot using natural language, without explicitly naming the object you want. When you need something that is not currently available, can anything nearby be a substantial substitute? Generate a sentence that describes a scenario where the target object's uncommon use is required. **Key Points:** 1) The generated sentence should emphasize the uncommon need for the target object. 2) Example sentences suggest a goal without directly naming the required object. 3) You should focus on the uncommon affordance of the target object. Sentences should imply the target object can be used as a substitute for other objects. **[Role user]** Here are some examples for generating sentences that express the uncommon need of target object {OBJECT}. **[Role user and assistant]** {Five IN-CONTEXT LEARNING examples} **[Role user]** Now please generate five sentences for target object: {OBJECT} (As shown in the cropped image).

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[2] Zihang Lai, Senthil Purushwalkam, and Abhinav Gupta. The functional correspondence problem. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15772–15781, 2021. 1