"Is This It?": Towards Ecologically Valid Benchmarks for Situated Collaboration

Dan Bohus
Microsoft Research
Redmond, United States
dbohus@microsoft.com

Sean Andrist
Microsoft Research
Redmond, United States
sandrist@microsoft.com

Yuwei Bao*
University of Michigan
Ann Arbor, United States
yuweibao@umich.edu

Eric Horvitz
Microsoft
Redmond, United States
horvitz@microsoft.com

Ann Paradiso
Microsoft Research
Redmond, United States
annpar@microsoft.com

Abstract

We report initial work towards constructing ecologically valid benchmarks to assess the capabilities of large multimodal models for engaging in situated collaboration. In contrast to existing benchmarks, in which question-answer pairs are generated post hoc over preexisting or synthetic datasets via templates, human annotators, or large language models (LLMs), we propose and investigate an interactive system-driven approach, where the questions are generated by users in context, during their interactions with an end-to-end situated AI system. We illustrate how the questions that arise are different in form and content from questions typically found in existing embodied question answering (EQA) benchmarks and discuss new real-world challenge problems brought to the fore.

Keywords

Embodied Question Answering; Benchmark Datasets; Mixed-Reality Task Assistance; Situated Collaboration

ACM Reference Format:
Dan Bohus, Sean Andrist, Yuwei Bao, Eric Horvitz, and Ann Paradiso. 2024. "Is This It?": Towards Ecologically Valid Benchmarks for Situated Collaboration. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI Companion '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3686215.3690152

1 Introduction

By coupling the generalist abilities of large language models with the capacity to "see," large multimodal models (LMMs) [1, 10, 14, 16] promise to enable a wide array of AI-powered assistance scenarios in the physical world. Numerous applications can be envisioned, including collaborative robots, intelligent monitoring of factory floors, and mixed-reality glasses providing assistance on the fly.

There are still significant gaps between demonstrations of capabilities and the requirements for practical deployment in real-world scenarios. To track the performance of emerging models and understand their capabilities, the research community has developed a variety of benchmarks for video- and embodied- question answering [7, 9, 11, 13, 19, 21, 22]. These benchmarks are typically constructed by identifying a preexisting multimodal dataset (or creating a synthetic one via a virtual environment), and then generating question-answer pairs from templates, human annotators, or via LLMs. The questions are designed to probe model capabilities along various dimensions, such as spatial understanding, episodic memory, and the recognition of objects and their attributes. While these benchmarks provide useful probes for model competency, we argue that they do not accurately capture the types of questions users ask when engaged in a real-time task, and thus do not reflect the full range of challenges that arise during situated collaboration.

To address this issue, we investigate an *interactive system-driven* approach for constructing benchmarks and assessing capabilities of large multimodal models in situated assistance tasks. Rather than starting with a preexisting dataset and creating questions post hoc, we collect data with an end-to-end interactive AI system, and use it to identify challenges and define benchmarks. As we discuss below, this approach leads to different types of questions, and helps highlight novel embodied interaction challenges that go well beyond question answering. We argue that this approach provides a more ecologically valid and useful assessment of LMM capabilities, and that it can serve as a vehicle for real-world, continuous testing of models against new data.

2 Related Work

Numerous visual question answering (VQA) datasets have been created to probe models' multimodal capabilities and to track their performance. Recent datasets extend beyond static image-based tasks and start to incorporate video, 3D, embodiment, and simulated environments for spatial and temporal reasoning. Some of the datasets [9, 19, 21, 22] adopt post hoc question-answer (QA) generation from a third-person perspective as a way to probe the model's understanding of visual inputs. Other datasets [7, 11, 13] are generated in an embodied, situated environment to enable episodic and environment-specific reasoning. Several of these benchmark datasets were constructed by using LLMs [7, 9, 19] and crowd workers [11, 22]. In others, the authors themselves [13] generate hypothetical questions in simulated situations, along with the expected

*Work conducted during an internship at Microsoft Research.

ICMI Companion '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0463-5/24/11
https://doi.org/10.1145/3686215.3690152

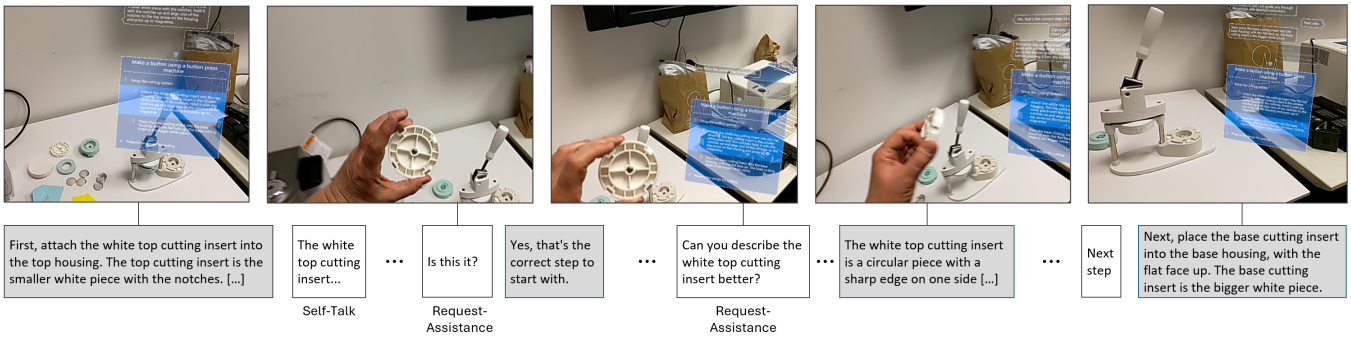

Figure 1: Segment from a participant interacting with Sigma. Sigma utterances have gray background; participant utterances have white background. The temporal axis is not to scale.

ground truth answers. Most of these datasets include only single-turn QAs, and they do not depict the realistic QAs that surface in a natural manner during real-time situated interaction, sometimes in the midst of long chains of multi-turn dialogues with a shared context and common ground that accrues over time.

Several research projects have focused on the challenging task of using LLMs and LMMs to provide people with interactive guidance for procedural tasks. Some of these efforts [5, 6, 12, 17] focus on recognizing visual cues to help people to better understand task procedures. Other projects [8, 15, 18] emphasize recognizing steps, mistakes, objects, states, and gestures from egocentric videos. A few studies have proposed challenges to evaluate both visual and language capabilities for task assistance when considering synchronized multimodal streaming data in mixed reality [2, 20]. The tasks considered in most benchmarks to date focus on the classification of dialog acts, in combination with high-level assessments of end-user satisfaction. To our knowledge, no studies have analyzed individual sentences to assess whether each question is answered correctly or addressed at the appropriate time and level of granularity. In distinction to prior research, we emphasize the importance of evaluating each question and answer in the context of the full situated collaborative history to that point, as captured during an interactive task guidance session with a real system.

3 Interactive System-Driven Benchmarking

Although constructing benchmarks by generating question-answer pairs post hoc on existing data can be useful for systematically probing the capabilities of LMMs with a battery of organized tests, we argue that the questions posed in such benchmarks lack ecological validity. That is, they do not reflect the types of questions end users pose in realistic scenarios, leading to a mismatch between measured model performance and the actual quality of user experiences.

EQA datasets can also be constructed from human-human interactions, where a participant takes on the role of an instructor or expert assistant (perhaps disguised as an AI system via Wizard-of-Oz paradigm) while another person carries out various tasks, asking questions along the way [2, 20] These questions capture the highest degree of ecological validity and might represent a gold standard for embodied question answering tasks. However, they are highly unstructured, contextual, and challenging to organize into a coherent benchmark.

Between the extremes of scenario-independent, artificial questions on one end and unconstrained human-to-human questions on the other, we propose an interactive system-driven approach to constructing ecologically valid situated assistance benchmarks. By building or leveraging an existing application, human-AI interactions can be collected that are situated in space and time within a task context. As we will illustrate below, the challenges and questions that emerge from such interactions are distinct from those posed by existing benchmarks. These questions can be organized into a dataset, and models can be trained to more accurately answer questions from that dataset in the usual static benchmark paradigm. Importantly, these models can be further evaluated by integrating them back into the original application, and testing their abilities to improve user experience by answering questions in new, live interactions. These new interactions can surface new questions and challenges, within cycles of refinement.

4 Data Collection

We used Sigma [3, 4], an open-source mixed-reality task assistance system as the experimental test-bed for data collection. Sigma leverages a HoloLens 2 headset and uses speech recognition, speech synthesis, and LLMs to guide users step-by-step through procedural tasks in the physical world. Throughout the interaction, the system displays a virtual panel with instructions and reads the steps aloud. Users can navigate among steps and they can ask questions, which are in turn answered by prompting an LLM along with the current step context. A sample interaction snippet is illustrated in Figure 1 and more system details are available in [3, 4].

Our long-term goal is to build a situated-assistance benchmark from data collected from users interacting with Sigma on a wide variety of tasks. We report here on a pilot data collection, with the short-term goals of refining the experimental protocol and identifying an initial set of salient challenges and phenomena. We plan to surface these challenges, along with others yet to be discovered, in the eventual benchmark.

For the pilot experiment, we authored and used five task recipes: resetting a Vertuo coffee machine to factory defaults, making coffee with a Nespresso coffee machine, changing the hard drive in a computer, creating craft buttons with a button making machine, and creating a notebook with a notebook binding machine. The experiment was reviewed and approved by our organization's IRB.

After obtaining informed consent, each participant was briefed on how to use Sigma, carried out a short eye calibration sequence, and was presented with final instructions and safety warnings. Each participant carried out 2-4 tasks in a controlled lab space, then was debriefed and compensated $50 for each hour of participation. Below, we report results from 26 interaction sessions collected, spanning the five defined tasks and nine participants.

5 From EQA to Situated Assistance

We present a qualitative analysis of 284 user utterances obtained from the data (after excluding easily parsable explicit navigation commands like *"next step"*, and responses to when the system asks *"Are you ready for the next step?"*)

5.1 Determining Response Obligation

A first observation is that in situated-assistance settings, not all user utterances, even those phrased as questions, create an obligation for the system to respond. In fact, while the dataset contains many questions addressed to the system, a large proportion of utterances (42%) reflect self-talk that users engage in as they perform steps (e.g., in Figure 1). While some self-talk utterances may provide an opportunity for the system to interject in an appropriate manner, they do not necessarily create an obligation to respond. At other times, users make statements that, while not in interrogative form, have the force of a question and release the conversational floor towards the system as a request for help, such as when reporting an unexpected error condition or an unexpected state, e.g., *"I don't see a maximum fill line"* and *"This is not going in."*

To better understand the space of user utterances, we manually annotated them with one of four possible dialog acts: **Request-Assistance** (38%)—which creates an obligation for the system to respond; **Acknowledgment** (7%) to a system utterance; **Self-Talk** (42%); and **Step-Transition** (13%)—in which the user reports that they are ready to transition to the next step or wish to hear a previous step again.

The data highlights that an important capability for situated assistive systems is knowing not just how to answer, but *whether* a given user utterance should be answered in the first place. This task is not as simple as it may appear, and a first challenge can be formulated around it. In many instances, this problem cannot be resolved from text alone, but rather requires leveraging audio and visual information. As one example, there are 33 utterances for which the speech recognition result was *"Ok"*. Of these, 12 denote the completion of the step, 11 are self-talk, and 10 are acknowledgments. Each usage requires a different response (or no response), and telling them apart requires careful consideration of the timing, prosody, and history of actions performed.

5.2 Situated Question Answering

An analysis of the **Request-Assistance** utterances in the collected data reveals that they differ in content and construction from the questions typically encountered in EQA challenge benchmarks. The differences stem largely from the fact that in an interactive setting, the participant and the system have shared attention, memory, and goals. They continuously interact to ground with each other and build mutual understanding over time. As a result, questions arise from real needs, are frequently specific, and are deeply enmeshed in the moment-by-moment flow of the collaboration.

Table 1: Examples of Request-Assistance Utterances by Type

Type	Utterance
Grounding States	"Should the handle bar be up or down at this point?" "But the ca- capsule is coming out when I try to close it." "Mm-hmm. But the blue capsule is too big to fit in."
Grounding Objects	"What is the base housing again?" "Is this it?" "How sharp is it supposed to be?" "Is the white base the ... the thing I just put on?"
Grounding Actions	"How much should I fill?" "Do I place, uh, the loop inside the holes?" "I understand rotating the ... base insert into place. What is rotating the top?"
General Help	"I need help, I'm lost."
Conv. Mgmt.	"Can you say that last part again?"

5.2.1 Situated Questions. We found that most of the user's questions aim to resolve grounding problems between the user and system regarding states, objects, and actions. A few other questions serve as general requests for help and conversation management. Several examples are shown in Table 1.

Often, questions center on resolving language-grounding issues, where the participant does not yet have an understanding of a particular term used by the system, e.g., *"What is the base housing again?"*, or *"What is the pin line?"*. In other cases, the participants understand the language, but need more details to fully ground the system's guidance to a concrete physical object, state or action, e.g., *"How sharp is it supposed to be?"* or *"How much should I fill?"*. Sometimes they simply elicit more information, e.g., *"What does the capsule look like?"* or they aim to verify a specific grounding via a situated reference, e.g., *"Is the white base the, the thing I just put on?"* Yet other times they aim to disambiguate between multiple possible groundings, e.g., *"Is it the blue capsule or the brown capsule?"*.

Many of the questions are anchored in the physical context and contain deictic expressions. Referring expressions for states, objects, and actions often contain personal and demonstrative pronouns such as ***"it"***, ***"this"***, or ***"that"*** as in *"How sharp is* ***it*** *supposed to be?"* or ***"This*** *is not going in."* More complex referring expressions also arise; e.g., in *"Is the white base the ...* ***the thing I just put on****?"* where an object is referred to indirectly by indicating it as an object of a past action. The questions contain both *endophoric* and *exophoric* references; in the former the referent is available in previous utterances, whereas in the latter it is only available in the physical, multimodal context. Sometimes, these references are even combined. In one case, the participant held an object up as he asked the question *"Is* ***this it****?"*. In this context, ***"it"*** is an endophoric reference to the top cutting insert object that the system named in a previous utterance (*"Attach* ***the top cutting insert*** *[...]"*) whereas ***"this"*** is an exophoric reference to the object the user is actually holding in his hand. We emphasize the critical importance and rich

set of research challenges captured by this single example. Situated assistive systems must handle such questions by reasoning about the full, multimodal context, including language and visual history over time. In turn, informative benchmarks should include these phenomena, as they naturally arise in the stream of interaction.

Beyond text and vision, the prosodical contours and audio features of the questions can be informative, and sometimes essential for parsing meaning (they are typically missing in EQA benchmarks where the questions are usually given as text). As we have already discussed, not all requests for help are interrogatives. Furthermore, many questions produced while interacting contain natural disfluencies, hesitations, false-starts, restarts (see again Table 1). These phenomena exacerbate challenges with segmenting and identifying requests for help and understanding how and when to best respond. Prosodical features sometimes dictate sentence semantics, as we have seen in the many facets of *"Ok"* discussed in the previous subsection. Situated assistive systems need to understand these nuances in the acoustic channel to respond appropriately, and informative benchmarks need to capture these phenomena.

Overall, our initial data collection highlights that the questions a situated-assistive system needs to respond to are markedly different in both content and form from questions typically observed in existing EQA benchmarks. Consider the contrast to OpenEQA [13], a recently proposed EQA benchmark in which questions were constructed by asking humans to watch video traces through homes and generate question-answer pairs while imagining themselves to be users of an assistive system. A cursory inspection of the questions in the two datasets reveals many important differences: for example, the Sigma data contains more deictic pronoun constructions, i.e., *"it"* appears in 21% of utterances vs 3% in OpenEQA. The *"How?"* questions in OpenEQA elicit general action plans, like *"How can I clean the room?"* or *"How can I dry my towel?"*, whereas more fine-grained and targeted questions eliciting missing information about actions, objects and states arise in Sigma interactions, e.g., *"How much force should I use on the handle?"* or *"Should I use the small or the big capsule?"* Object-related questions have a different structure in the two datasets: OpenEQA contains probing computer vision questions for recognizing objects and understanding spatial references, e.g., *"What is on the left side of the bed?"*. In contrast, the Sigma dataset contains many questions aimed at grounding a language referent, e.g., *"What is the base housing again?"*

5.2.2 Situated Answers. Throughout the data collection process, the Sigma system used a large language model prompt to identify questions and generate answers. Constructing ecologically valid benchmarks for situated assistance requires not only providing good coverage for the space of questions, but also understanding what makes for a good answer. For instance, we have noticed anecdotally that regardless of the correctness of answers from the LLM, the generations are often generic and too long-winded for the interactive setting. Like questions, good answers should leverage the grounding between the participants, make use of references, and be efficient and to the point.

5.3 Proactive Situated Assistance

Seamless assistance goes well beyond answering questions. As frequently observed in human-human interactive datasets [2, 20], a good collaborative partner not only responds to requests for help, but continuously monitors the situation at hand and proactively intervenes to correct or prevent mistakes, provide clarifications and confirmations, or offer advice and encouragement as needed.

To our knowledge, there are no existing benchmarks that aim to assess capabilities for useful, proactive, system-initiative interventions. The development of such benchmarks raises significant challenges. Such an effort will require novel metrics and evaluation techniques, as both timing, content, form, and their alignment play an important role in the quality of a system-initiated intervention.

The grounding issues that are communicated by participant questions are often visible prior to the participant actually asking the question. Various *streaming* inference tasks for tracking these issues in real-time should be defined as waypoints towards enabling proactive intervention capabilities. Examples include detecting in-stream when a state, object, or action grounding issue arises, detecting when a participant starts or stops performing an action specified in natural language, detecting whether an action is being performed correctly, etc. In addition to understanding what happens in the physical world, understanding cognitive states such as user confusion, frustration, confidence, and focus also plays an important role in generating useful interventions. Human collaborative partners can often track this information in stream by leveraging non-verbal behaviors such as gaze and hand movement patterns, as well as information leaked through self-talk. We plan to develop benchmarks that assess whether and how well large multimodal models can perform on these types of streaming inference tasks, on the road towards enabling proactive interventions.

6 Conclusion and Future Work

In pursuit of constructing ecologically valid benchmarks for assessing the capabilities of LMMs for situated collaboration tasks, we propose and investigate an interactive system-driven approach. An initial analysis of data from a pilot study shows that questions arising through interaction differ in form and content from questions typically found in current EQA benchmarks. New challenges arise around determining when the system has an obligation to respond, and around developing tasks and metrics for benchmarking capabilities for in-stream, proactive interventions.

Based on the lessons learned, we plan to conduct a larger scale data collection with the Sigma system, and to define and publicly release a set of benchmarks and metrics around these challenges. In conjunction with the data, problem formulations and metrics, we will conduct experiments and release results from an initial assessment of existing LMM models on these challenges, coupled with a subsequent error analysis.

Future work includes investigating more natural collaborative settings. While Sigma and other datasets like HoloAssist [20] and WTaG [2] provide a rich source of naturalistic questions, they are still collected in controlled lab settings. We expect that questions asked by people as they coordinate with AI systems and other people *in the wild* will surface additional challenges with the real-time interpretation of language and activities.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219 [cs.CL] https://arxiv.org/abs/2404.14219

[2] Yuwei Bao, Keunwoo Yu, Yichi Zhang, Shane Storks, Itamar Bar-Yossef, Alex de la Iglesia, Megan Su, Xiao Zheng, and Joyce Chai. 2023. Can Foundation Models Watch, Talk and Guide You Step by Step to Make a Cake?. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12325–12341. https://doi.org/10.18653/v1/2023.findings-emnlp.824

[3] Dan Bohus, Sean Andrist, Nick Saw, Ann Paradiso, Ishani Chakraborty, and Mahdi Rad. 2024. SIGMA: An Open-Source Interactive System for Mixed-Reality Task Assistance Research. arXiv:2405.13035 [cs.HC] https://arxiv.org/abs/2405.13035

[4] Dan Bohus, Sean Andrist, Nick Saw, Ann Paradiso, Ishani Chakraborty, and Mahdi Rad. 2024. SIGMA: An Open-Source Interactive System for Mixed-Reality Task Assistance Research–Extended Abstract. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 889–890.

[5] Sonia Castelo, Joao Rulff, Erin McGowan, Bea Steers, Guande Wu, Shaoyu Chen, Iran Roman, Roque Lopez, Ethan Brewer, Chen Zhao, Jing Qian, Kyunghyun Cho, He He, Qi Sun, Huy Vo, Juan Bello, Michael Krone, and Claudio Silva. 2024. ARGUS: Visualization of AI-Assisted Task Guidance in AR. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (2024), 1313–1323. https://doi.org/10.1109/TVCG.2023.3327396

[6] Qiyu Chen, Richa Mishra, Dina El-Zanfaly, and Kris Kitani. 2023. Origami Sensei: Mixed Reality AI-Assistant for Creative Tasks Using Hands. In *Companion Publication of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 147–151. https://doi.org/10.1145/3563703.3596625

[7] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. 2024. An Embodied Generalist Agent in 3D World. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[8] Shih-Po Lee, Zijia Lu, Zekun Zhang, Minh Hoai, and Ehsan Elhamifar. 2024. Error Detection in Egocentric Procedural Task Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18655–18666.

[9] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125* (2023).

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *NeurIPS*.

[11] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2023. SQA3D: Situated Question Answering in 3D Scenes. In *International Conference on Learning Representations*. https://openreview.net/forum?id=IDJx97BC38

[12] A. Maffei, Michela Dalle Mura, Fabio Marco Monetti, and Eleonora Boffa. 2023. Dynamic Mixed Reality Assembly Guidance Using Optical Recognition Methods. *Applied Sciences* (2023). https://api.semanticscholar.org/CorpusID:256451054

[13] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. 2024. OpenEQA: Embodied Question Answering in the Era of Foundation Models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] OpenAI. 2023. GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf

[15] Rohith Peddi, Shivvrat Arya, Bharath Challa, Likhitha Pallapothula, Akshay Vyas, Jikai Wang, Qifan Zhang, Vasundhara Komaragiri, Eric Ragan, Nicholas Ruozzi, Yu Xiang, and Vibhav Gogate. 2023. CaptainCook4D: A dataset for understanding errors in procedural activities. arXiv:2312.14556 [cs.CV]

[16] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. arXiv:2306.14824 [cs.CL] https://arxiv.org/abs/2306.14824

[17] Arvin Christopher C. Reyes, Neil Patrick A. Del Gallego, and Jordan Aiko P. Deja. 2020. Mixed Reality Guidance System for Motherboard Assembly Using Tangible Augmented Reality. In *Proceedings of the 2020 4th International Conference on Virtual and Augmented Reality Simulations* (Sydney, NSW, Australia) *(ICVARS '20)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3385378.3385379

[18] Tim J. Schoonbeek, Tim Houben, Hans Onvlee, Peter H.N. de With, and Fons van der Sommen. 2024. IndustReal: A Dataset for Procedure Step Recognition Handling Execution Errors in Egocentric Videos in an Industrial-Like Setting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 4365–4374.

[19] Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, and Chuang Gan. 2024. SOK-Bench: A Situated Video Reasoning Benchmark with Aligned Open-World Knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13384–13394.

[20] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. 2023. HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 20270–20281.

[21] Bo Wu, Shoubin Yu, Tenenbaum Joshua B Chen, Zhenfang, and Chuang Gan. 2021. STAR: A Benchmark for Situated Reasoning in Real-World Videos. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*.

[22] Yuanhan Zhang Bo Li Songyang Zhang Wangbo Zhao Yike Yuan Jiaqi Wang Conghui He Ziwei Liu Kai Chen Dahua Lin Yuan Liu, Haodong Duan. 2023. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv:2307.06281* (2023).