(LLM) as the backbone for guidance generation, and explored three different mutimodal methods to extract visual and dialog context. Our empirical results have shown promising results with foundation models, and highlighted some challenges and exciting areas of improvement for future research. The dataset and code are available at https://github.com/sled-group/Watch-Talk-and-Guide.

2 Related Work

2.1 Task Guidance Systems

Traditional task guidance systems (Ockerman and Pritchett, 1998, 2000) focused on providing the user with pre-loaded task-specific information without tracking the current state of the environment, or being able to generalize to new tasks (Leelasawassuk et al., 2017; Reyes et al., 2020; Lu and Mayol-Cuevas, 2019; Wang et al., 2016). The complexity of the problem comes from various aspects, such as environment understanding, object and action recognition, user's preference and mental state detection, real-time inference, etc (Manuvinakurike et al., 2018; Kim et al., 2022). In this work, we collected a multimodal dataset with real human-human task guidance interactions to better study the depth and breadth of the problem. We established a strong zero-shot baseline without prior exposure of the given tasks and develop a task guidance system with the help of the latest AR advancement to incorporates both users' perception and dialog.

2.2 Language and Multimodal Foundation Models

Large language foundation models (LLMs) such as ChatGPT,[2] GPT-4 (OpenAI, 2023a), and Bard[3] have demonstrated a wide range of language generation and reasoning capabilities. These models are not only equipped with huge knowledge bases through training on web-scale datasets, but also the in-context learning ability that allows them to learn new tasks from a few examples without any parameter updates (Brown et al., 2020).

Meanwhile, multimodal foundation models such as GPT-4v (OpenAI, 2023b), CLIP (Radford et al., 2021a) are also on the rise to incorporate large scale vision (Betker et al., 2023; Peng et al., 2023; Zhang et al., 2022; Li et al., 2022, 2023), audio (OpenAI, 2022), embodied (Brohan et al., 2023) and other input modalities (Zellers et al., 2021, 2022; Radford et al., 2021b; Li et al., 2022, 2023; Alayrac et al., 2022; Moon et al., 2020) that can reason and generate one modality type given another. While some proprietary models can be difficult or expensive to access (e.g. GPT-4v), other open source multimodal foundation models have been adapted to problems related to task guidance systems, e.g., action success detection (Du et al., 2023). In this work, we leverage the large-pretrained world knowledge embedded in these foundation models, and LLMs' in context learning ability, to build a generalizable situated task guidance system.

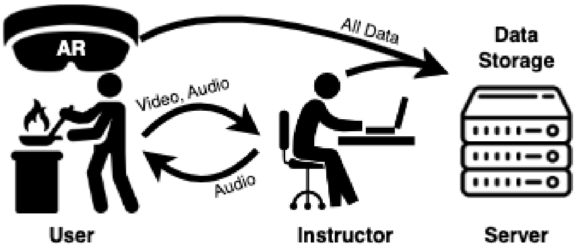

Figure 1: Data collection setup for WTaG.

3 A Dataset for Situated Task Guidance

In this work, we introduce **Watch, Talk, and Guide (WTaG)**, a new dataset for situated task guidance. WTaG includes nearly 10 hours of egocentric videos of human users performing cooking tasks while guided by human instructors through live, natural interaction. Synchronized videos and audio transcripts in WTaG present a variety of challenging phenomena, including perceptual understanding, communications, natural mistakes, and much more. We hope this dataset can serve as the starting point to dive into this complex problem.

Egocentric video datasets (Table 1) have garnered much attention in the past decade thanks to their potential application in interesting research areas such as embodied AI and task guidance systems. Most of the large egocentric video datasets contain unscripted activities (Damen et al., 2018, 2020; Lee et al., 2012; Su and Grauman, 2016; Pirsiavash and Ramanan, 2012; Fathi et al., 2012; Grauman et al., 2022), while others have collected (semi-)scripted activities where camera wearers are asked to follow certain instructions (Sigurdsson et al., 2018; Li et al., 2018; Sener et al., 2022). Classical video understanding datasets such as YouCook2 (Zhou et al., 2017) offered temporal boundaries with task procedural descriptions. Meanwhile, related works

[2] https://openai.com/blog/chatgpt

[3] https://ai.google/static/documents/google-about-bard.pdf