Can Foundation Models Watch, Talk and Guide You Step by Step to Make a Cake?

Yuwei Bao[†], **Keunwoo Peter Yu**[†], **Yichi Zhang**[†], **Shane Storks**[†], **Itamar Bar-Yossef**[†], **Alexander De La Iglesia**[†], **Megan Su**[†], **Xiao Lin Zheng**[§,*], **Joyce Chai**[†]
[†]University of Michigan, [§]Syracuse University

Abstract

Despite tremendous advances in AI, it remains a significant challenge to develop interactive task guidance systems that can offer situated, personalized guidance and assist humans in various tasks. These systems need to have a sophisticated understanding of the user as well as the environment, and make timely accurate decisions on when and what to say. To address this issue, we created a new multimodal benchmark dataset, **Watch, Talk and Guide (WTaG)** based on natural interaction between a human user and a human instructor. We further proposed two tasks: *User and Environment Understanding*, and *Instructor Decision Making*. We leveraged several foundation models to study to what extent these models can be quickly adapted to perceptually enabled task guidance. Our quantitative, qualitative, and human evaluation results show that these models can demonstrate fair performances in some cases with no task-specific training, but a fast and reliable adaptation remains a significant challenge. Our benchmark and baselines will provide a stepping stone for future work on situated task guidance.

1 Introduction

You have probably watched a lot of YouTube videos on how to bake a cheesecake, or how to change the car windshield wipers, but something always goes wrong and you just wish there is an expert right there to guide you through. Can we design an artificial intelligent system to watch, talk and guide humans step by step to complete a given task?

Task guidance for human users is a challenging problem, as it requires an interactive system to have a sophisticated understanding of what the user is doing, under the environment setup, and providing appropriate timely guidance. What is more challenging is if we could design a model that can be easily generalized to any arbitrary task without prior exposure, given a task manual or one demonstration. This requires the model to have a robust knowledge base and in-context learning abilities to easily pick up a new task to guide the human through.

Traditional approaches to develop AI agent for interactive task guidance like this require a large amount of task-specific training or rules to recognize object states (Gao et al., 2016), mistakes in actions (Du et al., 2023), and to interact with humans (Wu et al., 2021), thus limiting their ability to generalize. However, the recent rise of foundation models trained on a large amount of multimodal data from the web (Brown et al., 2020; Radford et al., 2021b; Alayrac et al., 2022; Li et al., 2022, 2023) creates new opportunities for developing robust open-domain AI agents for this problem. These works have undergone a paradigm shift and begun to explore the zero- and few-shot adaptation of these models to various embodied AI problems (Khandelwal et al., 2022; Huang et al., 2022; Ahn et al., 2022; Kapelyukh et al., 2023).

In this work, we extend this paradigm shift by examining the application of recent state-of-the-art foundation models[1](Bommasani et al., 2021) to situated interactive task guidance. We created **Watch, Talk and Guide (WTaG)**, a new multimodal benchmark dataset which includes richly annotated human-human dialog interactions, dialog intentions, steps, and mistakes to support this effort. We define two tasks: (1) *User and Environment Understanding*, and (2) *Instructor Decision Making* to quantitatively and qualitatively evaluate models' task guidance performance. With the inherent complexity of the problem itself, this dataset can help researchers understand the various nuances of the problem in the most realistic human-human interaction setting. We used a large language model

[*]Work done during a summer internship at the University of Michigan.

[1]We use this term to broadly refer to large pre-trained models.