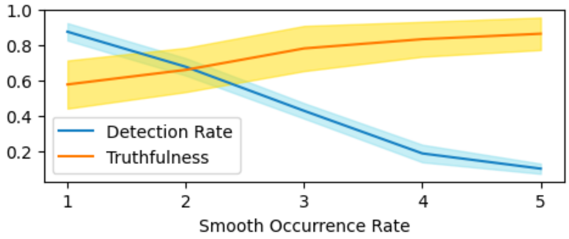
Prompt	Truthfulness
None	26.2±11.9
"Question: What is the user doing? Answer:"	25.4±13.2
"This is a picture of"	25.3±12.1

(a) BLIP2 Scene Description Evaluation



(b) EgoHOS Object and State Detection Evaluation

Figure 8: Vision to Language Translation Performance. (a) BLIP2 scene descriptions are reported to be only about 25% truthful. (b) A smooth occurrence rate of 2 is chosen for the experiments and offers about 70% truthfulness of the extracted visual context.

with the this method. It is much higher than the BLIP-2 Scene Description method, but still noisy enough to confuse the language model.

7 Discussions and Conclusions

In this work, we explored how to leverage foundation models to guide human users step by step to complete a task in a zero-shot setting given an arbitrary task recipe. We created a new benchmark dataset, Watch, Talk and Guide (WTaG), with natural human user and instructor interactions, and mistake guidance. Two tasks were proposed to evaluate model's ability on user and environment understanding, as well as instructor decision making. We used a large language model as the backbone for guidance generation, and compared three configurations of inputs incorporating language and visual context through dialog history, scene description, and object and state detection with multimodal foundation models. We conducted quantitative, and human evaluations of the three methods on our dataset, and discovered several challenges for future work.

First of all, vision to language translation is challenging. Having an accurate and relevant description of what the user is doing and the environment setup is non-trivial. Some of the methods we've experimented with often describe untruthful scenes or objects, or offer true but too generic or irrelevant information, e.g., "the user is preparing food." For future work, it would be helpful if these vision to language models could leverage recipe information, and user's attention (e.g., eye gaze), and other sensory inputs (e.g., sound) to achieve richer and more relevant user and environment descriptions.

Second, the overall performance of each prediction task from the large language model is fairly above random guessing, especially in a zero-shot setting with no task-specific finetuning, but are realistically too low to be useful in practice. There is much room for improvement, but also a great potential to be generalized to more tasks, multi-user or multi-tasking environments.

Third, there is a big difference between hands-on task guidance versus watching a "How-To" YouTube video. If we look closely, most of the model generated responses are instructions, instead of more situation-aware and personalized responses such as question answering and confirmations. It is easy to repeat recipe instructions or perhaps offer a little more detail leveraging the knowledge base, but harder to communicate with the user in a more situation-relevant way. Sometimes, a simple confirmation can boost the user's confidence, and worth a lot more than repeating what the user should do.

Furthermore, besides user intent and hesitation, there are other types of user states that can be helpful but are not currently modeled by our system. Are they familiar with the recipe? Do they look confused? Are they getting annoyed by all the instructions? Are they emotionally stable? All of these can change the way the instructor should talk. It is hard to categorize and collect users' mental states at each time point throughout the task. Sometimes experienced human instructors can infer through user actions and utterances, but the large language models have not demonstrated a strong ability in that way in our experiments.

Lastly, on the model architecture side, we leveraged some of the best performing multimodal and language models currently available, translated the rich audiovisual context collected in WTaG to language, and used LLM as the only reasoning engine to generate guidance. Language is too concise of a medium to offer context, and sometimes a picture is worth a thousand words. The current multimodal foundation models have limited reasoning capabilities to answer complicated questions and offer accurate timely guidance. This opens a wide variety of future opportunities for improving multimodal foundation models for situated task guidance.