Limitations

As exciting as this work is, there are several limitations we would like to acknowledge and hope to improve for future works.

One of the biggest challenges we faced was evaluation. There isn't a great method to comprehensively evaluate generative language models in a scalable way. There are hardly many quantitative ways to capture the syntax, semantics, informativeness, relevancy, etc of the outputs, and human evaluation is costly and subjective. In this work, we tried to include both aspects, by evaluating models' performance on several classification tasks, as well as human evaluation on the guidance context. Ultimately, we need a perceptual agent that can communicate with humans, and guide us step by step to make a cake. But to get there, we believe these are some of the milestones that the models should be able to achieve.

Another challenge related to evaluation was how to evaluate an interactive system in a real world setting. Every perception understanding can change models' decision making, and every decision prediction could completely change what would happen next. In our work, we had to focus on single step decision prediction given the ground truth interaction history. Our recorded dataset serves as resource to train/fine-tune/in-context learn/evaluate models on these decision points. Once a decent performing model is ready, we would love to conduct real-time human-bot interaction evaluations.

The third challenge is scale and resources. Our dataset contains 10 hours of recordings and 3 recipe tasks. A more robust baseline would ideally include a more diverse pool of users and instructors, more tasks, different environment setups, and etc. This would take collected effort and we are planning on expending our work in the future.

On the computation side, due to limited resources, we only ran the full experiments on the latest version of the GPT3.5-turbo-0301 model instead of GPT4. Querying large language models can be costly especially when we are querying at a high frequency (every few seconds per video) with a growing sized prompt. Meanwhile, there will always be another model version update in a few months that may achieve higher performance. However, we believe a lot of our observations still stand as the visual to language translation is a tough bottleneck to offer appropriate context, hallucination is still a big challenge for generative models, and situated personalized guidance is yet to be ideal.

About the experiment results, there is a fair amount of randomness depending on what prompt was used, which version of the model it is, how we set all the hyperparameters for each model, etc. We reduced the temperature to be zero to minimize the randomness, and yet the exact same prompt can lead to different responses due to the inherent randomness of the LLM. An exhaustive prompt tuning and hyperparameter search is almost impossible and most of them can hardly be measured quantitatively. We reported several prompt and hyperparameter tuning results in Section 6, and ran the same prompts through ChatGPT three times. We would try to conduct larger scale tuning and evaluations in future works when resources are available.

Last but not the least, we experimented two vision processing methods, one for frame image captioning, and the other for object segmentation and object state detection. We chose these models as they've been reported to demonstrate state of the art performance recently in related tasks, also with a minor consideration of the processing speed, since the task in nature is ideally real-time. Nevertheless, there are a lot more models out there that can extract more fine-grained information, such as hand gestures, object segmentation, object tracking, or can take advantage of temporal video information instead of still frame inputs. We are planning on exploring more vision or other modality processing models for our task.

Ethics Statement

The Institutional Review Board (IRB) of our institution approved this human subjects research before the start of the study. The WTaG dataset contains identifiable data (audio), but no facial identifiable data as all videos are first person views of the task environment. Members of the study team served as the human instructor, and recruited human subjects as the users. The human subjects were prepared the experiment setup, how to use the augmented reality headset, and potential risks before the experiment, and were debriefed after the recording. The consent to video and audio data for publication is optional and we will share the consented de-identified subset of the data (video and dialog ASR) when the paper is published. All data are stored in password protected internal servers with restricted access to only the study team. The human evaluations were