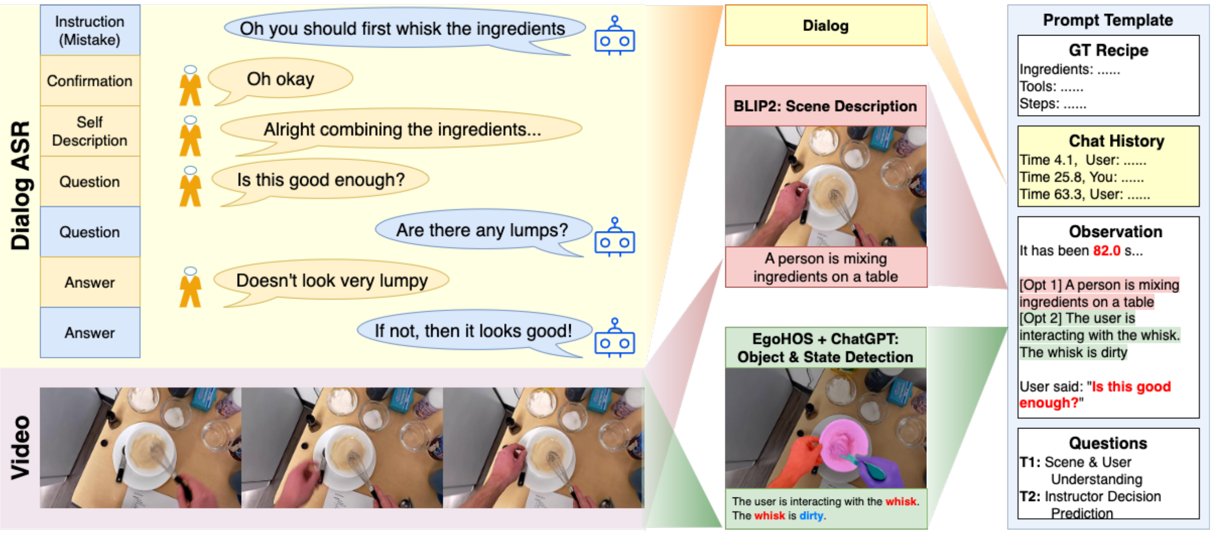
Dialog ASR
Instruction (Mistake)
Oh you should first whisk the ingredients
Confirmation
Oh okay
Self Description
Alright combining the ingredients...
Question
Is this good enough?
Question
Are there any lumps?
Answer
Doesn't look very lumpy
Answer
If not, then it looks good!
Video
Dialog
BLIP2: Scene Description
A person is mixing ingredients on a table
EgoHOS + ChatGPT: Object & State Detection
The user is interacting with the whisk. The whisk is dirty.
Prompt Template
GT Recipe
Ingredients: ......
Tools: ......
Steps: ......
Chat History
Time 4.1, User: ......
Time 25.8, You: ......
Time 63.3, User: ......
Observation
It has been 82.0 s...
[Opt 1] A person is mixing ingredients on a table
[Opt 2] The user is interacting with the whisk. The whisk is dirty
User said: "Is this good enough?"
Questions
T1: Scene & User Understanding
T2: Instructor Decision Prediction

Figure 4: Interactive Task Guidance Pipeline: Synchronized video and dialog transcripts are inputs to the system. We annotated each utterance to reflect dialog intent. The dialog history is inserted into the template, and the latest utterance is part of the observation. To process the videos, each queried frame either goes through BLIP2 for a scene description, or goes through EgoHOS for object and state detection. Zero, or one of the two video extraction output is inserted into the prompt template. The prompts are sent to ChatGPT for instruction predictions.

Instructor Decision Making

1. When to Talk: Should the instructor talk at time t (yes/no).
2. Instructor Intent: If yes to 1, instructor's dialog intention (options).
3. Instruction Type: If yes to 1 and intent in 2 is "Instruction", what type (options).
4. Guidance generation: If yes to 1, what to say in natural language.

5 Methods

Given the WTaG dataset and the two tasks defined above, we explore the application of pre-trained large language and vision foundation models on this problem ***without task-specific training***.

For each recording in WTaG, we feed the video frame by frame to our system, together with the synchronized dialog transcripts, and only query ChatGPT at the three query conditions (Section 4). The visual frames go through optional multimodal information extraction process detailed below, to translate the relevant visual context into natural language. The dialog transcripts are extracted at the first frame each utterance appears, and offer conversational context for model predictions. We designed a prompt template (Figure 4) that includes the ground truth recipe template, user-instructor chat history up to time point t, observations of the user and environment, as well as a list of questions listed in Section 4. We then send the prompts to ChatGPT[7] to answer questions related to the proposed tasks in each query time point. To enrich the prompts and offer more context, we explored the following three methods that translate multimodal precepts from the egocentric video into language:

Language Only (Lan): As a baseline for the LLM, we extracted the ground truth user and instructor dialog up until time t. All the past utterances are added to the prompt as part of the interaction history, and only the most recent user utterance is added to the observation to avoid model cheating. To enable temporal reasoning, observations include how long the user has been following the recipe so far. This method does not offer any visually-dependent information to the LLM backbone, and challenges the model to infer the context purely based on the conversation.

Scene Description (Sce): In this method, we generate a free-text scene description by applying BLIP-2 (Li et al., 2023) to the latest frame image, and insert it into the prompt as part of the observations. Depending on the prompts to BLIP-2, we ask generic questions such as "What is the user doing" or "This is a picture of" to get the descriptions. While this approach is flexible and open-domain, there is no control of how much situation-specific the descriptions would be. Together with the time

[7] `https://openai.com/blog/chatgpt`; used Azure endpoint for GPT3.5-turbo-0301, which is trained on data from up to September 2021.