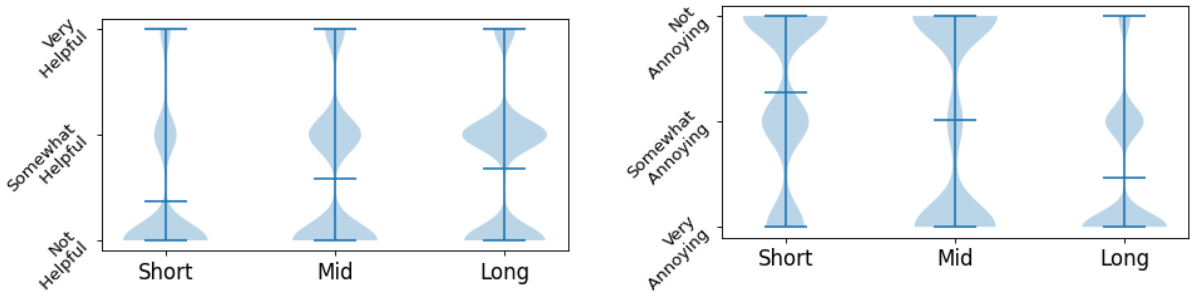

(a) Instruction Helpfulness

(b) Instruction Annoyance

Figure 7: Human Evaluation on Model Generated Guidance: Overall on average, most guidance are not considered as very helpful, and somewhat to very annoying by human evaluators. The violin plot shows the probabilistic distribution of each rating per video category, where three horizontal lines indicate the min, max and mean ratings respectively. In general, task guidance can be a very personalized experience.

This shows that situated natural language task guidance still has a long way to go to be applicable even though they've shown above random chance level performances across most of the tasks evaluated. Among the three groups, although not statistically significant, more guidance in the short videos were considered as "Not Helpful" by humans, whereas more guidance in the long videos were thought as "Very Annoying".

We also measured the inner annotator agreement on these results, and calculated the Cohen's κ (Cohen, 1960) value for helpfulness (0.14) and annoyance (0.02) ratings. This shows that the task guidance can be a very personal experience and that different users have different preferences and tolerance on the guidance. While the methods we experimented with did not inject user preferences into the prompts, and the LLM was unable to detect nuanced user mental preferences through the dialog, future work may begin to study how large pretrained models can be used for these challenges.

6.3 Perceptual Input Extraction

In order for the LLM backbone model to have an accurate assessment of the user and the environment, a high quality visual perception extraction module is a necessity. In this experiment, we did an ablation study on how truthful the Scene Description (Sce) and the Object and State Detection (Obj) modules are while translating the perceptual inputs to language. We conducted a small scale human evaluation on 6 recordings (2 of each recipe) and evaluated the truthfulness of the vision outputs. The truthfulness is defined as whether the vision output of each scene contains information that are not present or completely irrelevant to the scene. For example, "a person is putting ketchup on a plate" for the pinwheel recipe in the WTaG dataset is not truthful, as this action is not part of any recording. Each entry receives a binary value where 0 is 'Not Truthful' and 1 is 'Truthful'. This metric is more factual based than subjective, and all results below were evaluated by one person.

For the Scene Description evaluation (Figure 8a), we tried out 3 different prompts to the BLIP-2 model: no prompt (only the image), "Question: What is the user doing? Answer:", and "This is a picture of". There wasn't a significant difference in truthfulness among the three prompts, and we went with no prompt for all the experiments. However, overall, it was observed that the truthfulness of BLIP-2 is below 30%, which means most of the scene descriptions are actually hallucinating. This could cause huge confusions to the LLMs about exactly what is happening in the scene.

For the Object and State Detection evaluation (Figure 8b), since EgoHOS outputs unstable predictions from frame to frame, we conducted an experiment on detected object smoothing strategies. For a sliding window of 10 frames, we tried out if the same object needs to occur from 1 to 5 times in order to be included in the LLM prompts. For each prompt query, we evaluate 1) if any object and state were detected at all from the module, and 2) if the detection results were truthful. Y axis is the percentage of the detection outputs that are either truthful ($\{0, 1\}$) or detected ($\{0, 1\}$). According to Figure 8b, it was observed that as the required smooth occurrence goes higher, the same object needs to be detected more frequently to be included in the output; while at the same time, fewer objects were detected throughout the recipe overall. We picked smooth rate of 2 for all the experiments. As a result, the visual detection is about 70% accurate