Technical Summary of “Can Foundation Models Watch, Talk and Guide You Step by Step to Make a Cake?”

Part I: Problem Formulation and Experimental Design

Kai-Yu Lu

October 19, 2025

Abstract

This document summarizes the problem formulation, dataset design, experimental methodology, and empirical findings from a study on using foundation models for situated task guidance. The authors introduce WTaG (Watch, Talk and Guide), a benchmark that evaluates whether large language models can observe egocentric video and dialog to provide real-time verbal instructions for procedural tasks. I focus on understanding the technical setup, evaluation protocol, and quantitative results that characterize system performance across different vision-to-language translation strategies.

1 Research Problem and Motivation

The paper addresses situated task guidance: helping users complete procedural tasks through real-time verbal instructions while observing their actions via egocentric video. The core research question asks whether large foundation models can handle this problem in a zero-shot setting without task-specific training, using only a recipe specification and live observations.

The system operates at defined query points where it receives dialog history, recipe text, elapsed time, and optionally a text description of the current video frame. Two capabilities are evaluated: understanding what the user is doing (intent recognition, step detection, mistake identification) and deciding when and what to say (timing, response type, guidance generation).

This problem requires bridging perception and reasoning. Visual observations must be converted into appropriate verbal interventions based on procedural context rather than simple pattern matching. The challenge lies in maintaining reliability when visual evidence is ambiguous and when deciding whether intervention is necessary.

2 Dataset Construction

WTaG contains 56 recordings totaling approximately 10 hours, collected from 17 users and 3 instructors across three cooking recipes: pour-over coffee, peanut butter and jelly pinwheels, and microwave mug cake.

The data collection setup involved users wearing AR headsets that streamed their first-person view to remote instructors. Users received simplified recipes with high-level directions and minimal details, while instructors had complete, detailed recipes. This asymmetry was designed to elicit natural mistakes and questions, creating realistic intervention scenarios.

The recordings captured 4,233 English utterances with time-aligned transcripts. Dialog was automatically transcribed using Microsoft Azure speech recognition and manually corrected. Videos range from 5 to

18 minutes with a median of 10 minutes. Each recording contains a median of 31 instructor utterances and 35 user utterances.

Annotations include multiple layers:

User utterances are labeled with six intent types: Question, Answer, Confirmation, Self Description, Hesitation, and Other.

User mistakes are categorized into three classes: Wrong Action (incorrect motion or procedure), Wrong Object (using the wrong tool or ingredient), and Wrong State (incorrect measurement, temperature, or intensity).

Instructor utterances receive five intent labels: Instruction, Question, Answer, Confirmation, and Other. When the intent is Instruction, it is further classified into four subtypes: Mistake Correction, Current Step (guidance about what the user is doing now), Next Step (guidance about what comes next), and Details (additional information or clarification).

Step boundaries mark the temporal span of each recipe step within each video, enabling step detection evaluation.

The annotation scheme reflects careful task decomposition. Separating mistake types into Wrong Object, Wrong Action, and Wrong State enables analysis of which aspects of procedural understanding the system can and cannot capture.

3 Query Protocol and Task Definitions

Rather than querying at every frame, the authors defined three triggering conditions to generate evaluation points:

- When the ground truth user speaks
- When the ground truth instructor speaks
- When no one has spoken for 10 seconds

This protocol generated 5,921 query points across the dataset. It creates realistic scenarios where staying silent is sometimes the appropriate response, rather than forcing the system to produce output at every opportunity.

At each query point, the system receives:

- The latest video frame
- Dialog history up to that point
- The complete recipe specification
- Elapsed time since the task began

The evaluation defines two main task groups:

3.1 User and Environment Understanding

This group tests whether the system can accurately perceive and interpret the current state:

1. **User Intent Prediction**: Classify the dialog intent of the user's most recent utterance into one of six categories (Question, Answer, Confirmation, Self Description, Hesitation, Other).

2. **Step Detection**: Identify which recipe step the user is currently performing from the list of steps in the recipe.
3. **Mistake Existence**: Binary decision on whether the user has made a mistake at the current time point.
4. **Mistake Type Classification**: If a mistake exists, classify it as Wrong Action, Wrong Object, or Wrong State.

3.2 Instructor Decision Making

This group tests whether the system can make appropriate intervention decisions:

1. **When to Talk**: Binary decision on whether the instructor should speak at this time point. The timing threshold is calibrated based on median instructor speaking speed and words per utterance from the training set.
2. **Instructor Intent**: If the decision is to speak, classify the dialog intent into one of five categories (Instruction, Question, Answer, Confirmation, Other).
3. **Instruction Type**: If the intent is Instruction, further classify it into one of four subtypes (Mistake Correction, Current Step, Next Step, Details).
4. **Guidance Generation**: Generate the actual text of what the instructor should say.

Classification tasks report micro F1 scores computed across all query points. For guidance generation, human evaluation is conducted on a subset of 936 time points across 6 videos.

4 Modeling Approach

All experimental configurations share the same backbone language model (GPT-3.5-turbo-0301) and prompt template structure. They differ only in how visual information is converted to text before being fed to the language model. The API configuration uses temperature equals zero, a maximum token limit of 100, and fixed stop sequences.

Six recordings (two per recipe) were reserved for prompt engineering and hyperparameter tuning. The remaining 50 recordings were used for evaluation. Each configuration was evaluated three times to account for any residual randomness.

4.1 Configuration 1: Language Only

This baseline configuration provides only dialog history and elapsed time, with no visual information. It tests whether conversation alone contains sufficient information to infer user state and make guidance decisions.

The prompt includes:

- The complete recipe specification
- Time-stamped dialog history
- Elapsed time since task start
- The user's most recent utterance (if any)

4.2 Configuration 2: Scene Description

This configuration adds visual context by generating a free-text caption for the current frame using BLIP-2, a vision-language model.

Three prompt variants were tested for BLIP-2:

- No prompt (image only)
- “Question: What is the user doing? Answer:”
- “This is a picture of”

The generated caption is inserted into the main prompt as part of the observation. For example: “Scene description: a person is preparing food on a table.”

4.3 Configuration 3: Object and State Detection

This configuration takes a more structured approach to visual understanding:

Step 1: EgoHOS (Egocentric Hand-Object Segmentation) segments the user’s hands and any objects they are touching in the current frame.

Step 2: For each detected object segment, CLIP (Contrastive Language-Image Pre-training) predicts the object name and its state.

Step 3: To prevent hallucinations, CLIP selects from a candidate list rather than generating free-form predictions. The candidate list is constructed by prompting the language model to extract likely objects and states from the recipe, followed by light manual cleanup.

Step 4: A temporal smoothing rule is applied. An object is only included in the prompt if it appears in at least 2 frames within a sliding window of 10 frames. This reduces noise from unstable frame-by-frame detections.

The resulting structured information is formatted as natural language and inserted into the prompt. For example: “The user is interacting with measuring cup. The measuring cup is filled with water.”

This design prioritizes interpretability by decomposing visual understanding into discrete, verifiable components rather than relying on end-to-end learned representations.

5 Empirical Results

5.1 User and Environment Understanding Performance

All three methods substantially exceeded random chance baseline on user intent prediction and step detection, but struggled severely with mistake detection.

User Intent Prediction: All methods achieved approximately 40 percent micro F1 score, compared to a random baseline of approximately 17 percent (one sixth chance given six categories). Language Only and Object Detection performed similarly at around 40 to 42 percent, while Scene Description was slightly lower and more variable at around 35 percent.

Step Detection: Performance ranged from 60 to 68 percent F1 across methods, compared to a random baseline of approximately 10 percent. Object Detection achieved the highest score at roughly 68 percent, followed by Language Only at 65 percent, and Scene Description at 60 percent. Performance varied by recipe, with all methods performing better on coffee and cake compared to pinwheels.

Mistake Existence Detection: All three methods achieved only approximately 5 percent F1, barely above the random baseline of approximately 2 percent. This represents near-complete failure on this critical task.

Mistake Type Classification: Among the small number of cases where the model correctly detected that a mistake occurred, classification into Wrong Action, Wrong Object, or Wrong State was near chance at approximately 33 percent (one third for three categories).

5.2 Instructor Decision Making Performance

When to Talk: All methods performed around 50 percent, essentially equivalent to random guessing. This indicates the model has difficulty determining when intervention is appropriate. Qualitative analysis revealed that models tend to talk more frequently than human instructors.

Instructor Intent: Once the model decides to speak, intent classification achieved approximately 45 to 48 percent F1 (Language Only highest at 48 percent), compared to a random baseline of 20 percent. This is above chance but still indicates substantial confusion.

Instruction Type: When the intent is Instruction, subtype classification achieved approximately 38 to 42 percent F1, compared to a random baseline of 25 percent.

Distribution analysis revealed systematic bias: models heavily favor producing Instruction intent (approximately 90 percent of responses) and within instructions, overwhelmingly select Current Step (approximately 60 percent). In contrast, human instructors distribute responses more evenly across Answer, Confirmation, Question, and different instruction types.

5.3 Human Evaluation of Generated Guidance

Three human evaluators rated 936 system outputs across 6 test videos on two dimensions:

Helpfulness (1 equals not helpful, 3 equals very helpful): Mean ratings ranged from 1.5 to 2.0 across all methods and video categories, indicating performance between "not helpful" and "somewhat helpful." Videos were stratified by interaction length (short, medium, long based on number of utterances). Guidance in short videos received slightly lower helpfulness ratings.

Annoyance (1 equals very annoying, 3 equals not annoying): Mean ratings similarly ranged from 1.5 to 2.0, indicating responses were perceived as "somewhat annoying" to "tolerable." Guidance in long videos (high interaction) received slightly higher annoyance ratings.

Inter-rater agreement was very low: Cohen's kappa approximately 0.14 for helpfulness and 0.02 for annoyance. This indicates that guidance quality preferences are highly subjective and personalized.

5.4 Vision-to-Language Quality Analysis

A separate perceptual study evaluated the truthfulness of vision module outputs on a subset of 6 videos:

BLIP-2 Scene Descriptions: Across all three prompt variants tested, truthfulness was below 30 percent. Approximately 70 percent of generated captions contained information that was not present in the scene or was completely irrelevant. Different prompts showed no significant difference in quality.

Object and State Detection: With a smoothing rate of 2 (object must appear in at least 2 of 10 frames), the pipeline achieved approximately 70 percent truthfulness. This represents the fraction of detection outputs that accurately described objects and states present in the scene.

The detection rate (fraction of time points where any object was detected) decreased as the smoothing threshold increased, trading coverage for accuracy. The authors selected smoothing rate 2 as a reasonable balance, achieving 70 percent truthfulness while maintaining approximately 85 percent detection rate.

5.5 Comparative Analysis Across Methods

Language Only remained competitive across most tasks, demonstrating that dialog history carries substantial information about task state. It matched or slightly underperformed Object Detection on user intent and step detection by 2 to 3 percentage points, but this difference is relatively small given the added complexity of vision processing.

Object and State Detection produced small but consistent improvements over Language Only on user understanding tasks. The structured approach of detecting hands, objects, and states provided more reliable visual context than free-form captions.

Scene Description often performed worse than Language Only and showed higher variance across trials. The high hallucination rate (70 percent false information) in BLIP-2 captions likely confused the language model rather than helping it.

All methods failed similarly on mistake detection regardless of visual input quality, suggesting that the failure stems from factors beyond visual grounding, such as insufficient temporal context or limitations in how the language model reasons about procedural violations.

Summary

This study demonstrates that foundation models can track procedural progress and understand user communication in situated task guidance without task-specific training. However, they fail to detect errors and over-produce instructions compared to human behavior. The competitive performance of Language Only reveals that dialog carries rich procedural information, but current vision-to-language translation introduces substantial noise (30 to 70 percent error rates) that limits the value of visual modules. The work provides a realistic benchmark and highlights fundamental gaps between classification accuracy and reliable interactive decision-making.