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

Part II: Strengths, Limitations, and Research Implications

Kai-Yu Lu

October 19, 2025

Abstract

This document provides critical analysis of the WTaG study, examining its methodological strengths, identifying key limitations, and discussing implications for multimodal reasoning research. I evaluate what the empirical findings reveal about the gap between visual pattern recognition and contextual understanding in interactive systems, explore technical bottlenecks in vision-to-language translation, and suggest potential research directions that could address the identified challenges.

1 Methodological Strengths

1.1 Realistic Interaction Design

The dataset captures genuinely natural human behavior rather than scripted or synthetic interactions. The asymmetry between user recipes (simplified, high-level) and instructor recipes (complete, detailed) creates authentic challenges. Users naturally make mistakes, hesitate, and ask questions because they lack complete information. Instructors must infer state from limited visual evidence and user comments.

This design is more realistic than alternatives such as:

- Having users follow complete step-by-step instructions (which eliminates natural errors)
- Using simulated environments (which lack the complexity of real-world variability)
- Collecting third-person videos (which miss the perspective constraints of first-person view)

The naturalistic setting means the benchmark reflects actual difficulties that task guidance systems will face in deployment.

1.2 Well-Designed Query Protocol

The three-condition triggering mechanism (user speech, instructor speech, 10-second silence) creates diverse evaluation contexts. Some query points occur during active conversation, where responding is clearly appropriate. Others occur during silent work periods, where staying quiet may be correct.

This is substantially better than querying at fixed time intervals, which would:

- Force evaluation at arbitrary moments unrelated to interaction flow
- Miss critical decision points when someone has just spoken
- Fail to test whether the system knows when not to interrupt

The protocol also generates a large number of evaluation points (5,921) from relatively limited video duration (10 hours), enabling statistically meaningful performance measurement.

1.3 Fine-Grained Annotation Scheme

The multi-layer annotation enables analysis far beyond simple accuracy metrics. Researchers can determine:

- Which types of user communication the system understands well versus poorly
- Whether step detection failures cluster in specific recipes or procedural phases
- Which mistake categories are most difficult to recognize
- How instructor response distributions differ between models and humans

The separation of mistake types (Wrong Object, Wrong Action, Wrong State) is particularly valuable because it connects to different perceptual and reasoning requirements. Wrong Object might be detectable from visual appearance, Wrong Action from motion patterns, and Wrong State from quantitative measurements.

1.4 Systematic Vision Module Comparison

The study does not simply report that "vision helps" or "vision does not help," but rather investigates why different vision-to-language strategies succeed or fail.

Testing multiple BLIP-2 prompts and quantifying caption truthfulness reveals that the problem is not prompt engineering but fundamental limitations in how the model generates scene descriptions.

Varying the EgoHOS smoothing rate and measuring the truthfulness versus coverage tradeoff provides actionable guidance: researchers know that requiring detection in 2 of 10 frames achieves 70 percent accuracy with 85 percent coverage, and can make informed decisions about this tradeoff for their applications.

This kind of module-level analysis identifies specific bottlenecks rather than treating vision as a black box.

1.5 Transparent Evaluation Methodology

Using zero-shot prompting with temperature zero and documenting three independent runs per condition makes results reproducible. The authors do not tune prompts extensively on the test set or cherry-pick successful examples.

Reporting both classification metrics and human evaluation provides complementary perspectives. Automatic metrics reveal performance on well-defined subtasks, while human ratings capture subjective quality dimensions that resist quantification.

Acknowledging low inter-rater agreement (kappa near zero for annoyance) demonstrates intellectual honesty. Many studies would hide or downplay such findings, but here the authors recognize that guidance quality is inherently personal.

2 Key Limitations

2.1 Limited Dataset Scale and Diversity

The dataset contains only three recipes totaling 10 hours from 20 participants. This constrains generalization in several ways:

Task diversity: All three recipes are short cooking tasks (5 to 18 minutes). The findings may not extend to longer tasks, assembly procedures, repair operations, or other procedural domains.

Procedural structure: Cooking recipes have relatively linear structure with clear step boundaries. Tasks with conditional branching, parallel subtasks, or less rigid sequencing might reveal different strengths and weaknesses.

Error patterns: The types of mistakes users make likely depend on task familiarity and complexity. A small set of tasks provides limited coverage of possible error modes.

User population: College students may differ from other user populations in learning speed, communication style, and comfort with technology.

Expanding the dataset to include more tasks, longer sessions, and more diverse users would strengthen confidence that the findings generalize.

2.2 Lack of Temporal Modeling

Each frame is processed independently, which creates fundamental limitations for mistake detection. Many errors cannot be recognized from single snapshots:

Wrong Action: A user might reach for the correct object but then use it incorrectly. The reaching motion looks appropriate in individual frames; only the complete sequence reveals the error.

Wrong State: Determining whether liquid has been poured to the correct level requires tracking the pouring motion over time, not just observing the final state.

Procedural violations: A user might perform a valid action at the wrong time. Single frames cannot distinguish "too early" from "correct timing."

The study acknowledges this limitation but does not explore temporal modeling approaches. Adding even simple temporal features (such as object contact sequences over the past 1 to 2 seconds) might significantly improve mistake detection.

2.3 Insufficient Analysis of Perception Error Propagation

The vision quality study shows that BLIP-2 hallucinates 70 percent of the time and EgoHOS plus CLIP achieves only 70 percent accuracy. However, the paper does not trace how these errors affect downstream tasks.

Important questions remain unanswered:

- Do hallucinated object names mainly hurt mistake detection, or do they also degrade step identification?
- Are state misidentifications more damaging than object misidentifications?
- Can the language model compensate for some visual errors by cross-referencing with dialog?
- Which types of scenes (cluttered, occluded, similar-looking objects) cause the most vision failures?

A systematic ablation study that introduces controlled perception errors (swapping object labels, corrupting state predictions) would reveal which vision improvements matter most for each task.

2.4 Unexplained Mistake Detection Failure

The near-complete failure on mistake detection (5 percent F1) is the most concerning result, yet the paper offers limited explanation. Possible contributing factors include:

Visual limitations: Single frames lack temporal context, and vision modules have high error rates.

Prompt design: The prompt may not clearly specify what constitutes each mistake type, or may not provide enough context about expected versus observed states.

Language model reasoning: GPT-3.5 may lack the reasoning capability to compare observed states against procedural constraints and identify violations.

Data distribution: Mistakes may be relatively rare in the dataset, making this task particularly difficult for zero-shot learning.

The paper does not investigate which of these factors dominates, making it unclear what interventions would most improve performance.

2.5 Limited Human Evaluation Scope

Human evaluation covers only 936 of 5,921 query points (approximately 16 percent) across 6 of 50 test videos. This raises questions:

- Were these 6 videos representative, or might they have unusual characteristics?
- How does guidance quality vary by recipe, user skill level, or session progress?
- Would ratings improve if evaluators were actual task novices versus external judges?

The very low inter-rater agreement (kappa 0.02 for annoyance) suggests that three raters may be insufficient to capture the range of user preferences. Stratifying evaluation by user characteristics (prior cooking experience, chattiness, error rate) might reveal when guidance is more versus less helpful.

2.6 No Computational Cost Analysis

The paper does not report inference time, memory usage, API costs, or energy consumption. For practical deployment:

- Querying GPT-3.5 at 5,921 time points across 10 hours requires approximately one query every 6 seconds, which would be expensive at API pricing
- Running EgoHOS and CLIP on every frame requires substantial GPU resources
- Real-time deployment would require careful optimization to meet latency constraints

Understanding the cost-performance tradeoff is essential for determining when this approach is practical versus when simpler alternatives are preferable.

3 Technical Bottlenecks

3.1 Vision-to-Language Translation Quality

The 30 to 70 percent error rate in visual descriptions represents a fundamental challenge. Two distinct failure modes appear:

BLIP-2 hallucination: Free-form caption generation produces fluent but false descriptions. The model generates generic statements like “a person is preparing food on a table” that are true but uninformative, or specific claims about objects that are not present.

Detection noise: Even with structured object detection, EgoHOS plus CLIP produces 30 percent false positives after temporal smoothing. Hand-object contacts are difficult to segment reliably, and CLIP sometimes misclassifies visually similar objects.

Both problems stem from attempting to compress rich visual information into language. Potential solutions include:

Structured representations: Instead of free-text captions, output a table listing visible objects, their states, hand contacts, and confidence scores. This reduces ambiguity and makes errors more detectable.

Uncertainty estimation: Rather than forcing the vision module to commit to specific object names, allow it to express uncertainty ("possibly a measuring cup or small bowl"). The language model could then request clarification rather than reasoning over false assertions.

Task-specific grounding: Fine-tune vision modules on the specific recipes and objects in the dataset. This would improve accuracy but sacrifice generalization to new tasks.

3.2 Absence of Temporal Context

Single-frame observations cannot capture:

- Motion trajectories (where the hand is moving, not just where it is)
- Action sequences (which objects were touched and in what order)
- State changes (whether quantities are increasing or decreasing)
- Timing information (whether actions are too fast, too slow, or appropriately paced)

Short-horizon temporal features could encode recent history. For example:

- Last 1 to 2 seconds of hand-object contacts as an ordered sequence
- Recent changes in visible object states ("sugar container moved from left to center")
- Duration spent on the current step compared to recipe expectations

This temporal context could be formatted as additional prompt fields without requiring architectural changes to the language model.

3.3 Over-Instruction and Lack of Self-Regulation

Models default to producing instructions about the current step regardless of context. They do not modulate behavior based on:

- User pace (working quickly and confidently versus slowly and hesitantly)
- Recent intervention history (just spoke 5 seconds ago versus have been silent for 2 minutes)
- Evidence quality (clear visual observation versus ambiguous detection)
- User request (explicitly asked for help versus working independently)

Human instructors naturally calibrate intervention frequency based on these cues. Computational approaches could include:

Cost-based triggering: Estimate the cost of speaking (interruption, cognitive load) versus the benefit (preventing errors, building confidence) and speak only when expected benefit exceeds cost.

Policy learning: Rather than zero-shot prompting, train a lightweight policy network on human instructor decisions to predict when to speak.

Explicit uncertainty modeling: When visual evidence is ambiguous, prefer questions ("Are you using the tablespoon or teaspoon?") over assertions ("Use the tablespoon").

4 Research Implications

4.1 Classification Accuracy Versus Reliable Decision-Making

The WTaG results illustrate a critical gap: models achieve 60 to 65 percent F1 on step detection but only 5 percent on mistake detection. Both tasks involve similar visual and linguistic inputs, yet performance differs dramatically.

This suggests that step detection can be solved through pattern matching (recognizing typical object configurations for each step), while mistake detection requires deeper reasoning about what should versus should not occur given procedural constraints.

Similar gaps appear in other domains:

- Visual question answering models that answer correctly while producing inconsistent reasoning
- Medical diagnosis systems that classify accurately but cannot explain their predictions
- Autonomous vehicles that recognize objects reliably but misjudge their intentions

Addressing this gap requires moving beyond maximizing classification accuracy toward building models that maintain internal consistency and ground their decisions in verifiable reasoning.

4.2 The Competitive Performance of Dialog-Only Models

Language Only performs nearly as well as vision-augmented methods on most tasks. This reveals that:

- Users naturally describe what they are doing through questions and confirmations
- Instructors can infer substantial state information from user utterances alone
- Current vision modules add limited value given their high error rates

This finding has practical implications. In deployment scenarios where vision processing is expensive or unreliable, dialog-based approaches may provide acceptable performance at lower cost. However, this also means that vision modules must substantially outperform the Language Only baseline to justify their complexity.

The result also suggests that better integration between vision and language could improve performance. Rather than treating them as independent information sources, models could use language to guide visual attention ("look for the measuring cup the user just mentioned") or use vision to disambiguate language ("which bowl" when multiple bowls are visible).

4.3 Temporal Reasoning as an Open Challenge

The mistake detection failure highlights that static observations are insufficient for understanding procedural tasks. This connects to broader challenges in video understanding:

- Action recognition requires distinguishing actions with similar poses but different motions
- Anomaly detection in surveillance requires recognizing unusual sequences
- Sports analysis requires understanding strategic plans that unfold over extended periods

Most vision-language models process individual frames or short clips independently. Architectures that maintain persistent state representations across longer temporal horizons remain an active research area.

4.4 Personalization and Adaptive Communication

The very low inter-rater agreement (kappa near zero) demonstrates that guidance preferences are highly personal. Some users prefer detailed step-by-step instructions, while others want minimal intervention. Some appreciate confirmations, while others find them annoying.

Current foundation models provide one-size-fits-all responses. Adaptive systems would need to:

- Infer user preferences from early interactions (do they ask many questions or work independently?)
- Adjust intervention frequency based on observed user pace and confidence
- Vary language formality and directness based on user response patterns
- Learn individual user models over repeated sessions

This personalization challenge extends beyond task guidance to any interactive AI system: chatbots, tutoring systems, healthcare assistants, and customer service agents all face similar needs to adapt communication style to individual users.

5 Potential Research Directions

5.1 Structured Visual Representations

Rather than converting vision to free-form language, represent visual information as structured data that language models can more reliably process:

Object-State Table: List recipe-relevant objects, their current states, confidence scores, and spatial relationships. Example:

```
Object          State           Confidence  Location
measuring_cup   filled_water    0.85        user_hand
sugar_jar       lid_removed     0.72        table_center
spoon           not_detected    0.00        unknown
```

Action Sequence: Encode recent hand-object interactions as timestamped events:

```
t-2.1s: right_hand contacted sugar_jar
t-1.8s: right_hand grasped spoon
t-0.5s: spoon entered measuring_cup
```

These representations reduce ambiguity and make errors more detectable. The language model can check whether observations are consistent with recipe expectations using symbolic reasoning rather than pattern matching over noisy text.

5.2 Verification-Augmented Generation

Require each generated guidance sentence to cite specific supporting evidence:

```
Assertion: "Add two tablespoons of sugar"
Evidence:
  - Step requirement: sugar=2_tbsp
  - Observation: sugar_jar visible, confidence=0.72
  - Action: user holding spoon
```

```
Assertion: "The mixture should be smooth"
Evidence:
  - Step requirement: consistency=smooth
  - Observation: [MISSING visual evidence]
  - Fallback: Rephrase as question
  - Output: "Does the mixture look smooth?"
```

Sentences lacking observational support would trigger rewrites as questions or acknowledgments. This constraint could reduce hallucination while keeping advice grounded.

5.3 Cost-Aware Intervention Policies

Model the decision to speak as optimizing a cost function:

$$J = \alpha \cdot P(\text{user error}) + \beta \cdot \text{InterruptionCost} - \gamma \cdot E[\text{helpfulness}]$$

Error probability estimated from: object state consistency with current step, unusual action sequences, user hesitation markers in speech.

Interruption cost estimated from: time since last instructor utterance, user working speed (fast suggests confidence, slow suggests confusion), user question frequency.

Expected helpfulness estimated from: confidence in visual observations, relevance of available guidance to detected issues, user responsiveness to past guidance.

When uncertainty is high, default to confirmations or clarifying questions rather than directives. This approach could reduce over-talking while maintaining effectiveness on critical interventions.

5.4 Perturbation-Based Consistency Evaluation

Test whether models maintain stable reasoning under semantically equivalent variations:

Dialog perturbations: Rephrase user questions (“Is this enough?” versus “Did I add the right amount?”) and verify the instructor response remains consistent.

Visual perturbations: Apply style transfer, adjust brightness, or add background clutter without changing task-relevant content. Verify step detection and mistake recognition remain stable.

Temporal perturbations: Sample frames at different rates or use different temporal windows. Verify that conclusions about user state do not change arbitrarily.

Models that maintain high accuracy while reasoning collapses under these perturbations are unreliable. Consistency metrics could complement accuracy metrics in benchmark evaluation.

5.5 Targeted Robustness Studies

Systematically measure how different perception errors propagate to task performance:

Object swapping: Replace objects in videos with visually similar alternatives (tablespoon swapped for teaspoon). Does mistake detection improve if the vision module is perfect?

State corruption: Deliberately corrupt state labels (report empty when actually full). Which tasks degrade most?

Occlusion simulation: Mask portions of frames to simulate hand occlusions or camera obstructions. How does performance degrade with partial visibility?

This analysis would quantify the value of improving each vision component and guide research priorities.

6 Conclusion

The WTaG study provides a realistic benchmark for situated task guidance and reveals fundamental challenges in applying foundation models to interactive systems. While models can track procedural progress and understand user communication, they fail to detect errors, over-produce instructions, and cannot self-regulate intervention frequency.

The competitive performance of dialog-only approaches demonstrates that conversation carries rich procedural information, but poor vision-to-language translation quality (30 to 70 percent error rates) limits the value of current visual modules. The mistake detection failure highlights the need for temporal reasoning beyond single-frame observations.

These findings illustrate the gap between classification accuracy and reliable decision-making in interactive contexts. Addressing this gap will require structured visual representations, temporal modeling, verification mechanisms, and adaptive communication strategies that go beyond maximizing benchmark scores toward building systems that reason transparently and adapt to individual users.