Towards Dynamic and Realistic Evaluation of Multi-modal Large Language Model

Yijiang Li*[1], Huiqi Zou*[2], Ziang Xiao[2]

[1]UC San Deigo, [2]Johns Hopkins University

*Equal contribution

Introduction

- **Hallucination:** A multimodal model provides information that contradicts with the content of an image.
- **Problems:**
  - Benchmark questions don't match the tasks models face in application scenarios.
  - Single-turn question-answer pairs in existing benchmarks fail to capture the complexity of real-world user interactions.
  - Static benchmarks risk data contamination, affecting their reliability.
- **Solution:** A dynamic evaluation method that emulates user interactions through multi-turn, interrogative conversations.

Experiments

- **Dataset**: Testing samples sourced from the COCO dataset, each annotated with captions, instance bounding boxes and coordinates.
- **Baseline**: Caption-based evaluation method (i.e., Please provide a detailed description).
- **Metric**: CHAIR-i, CHAIR-s, and coverage.

$$\text{coverage} = \frac{|\{\text{all objects mentioned by the VLM}\}|}{|\{\text{all objects mentioned in the ground truth}\}|}$$

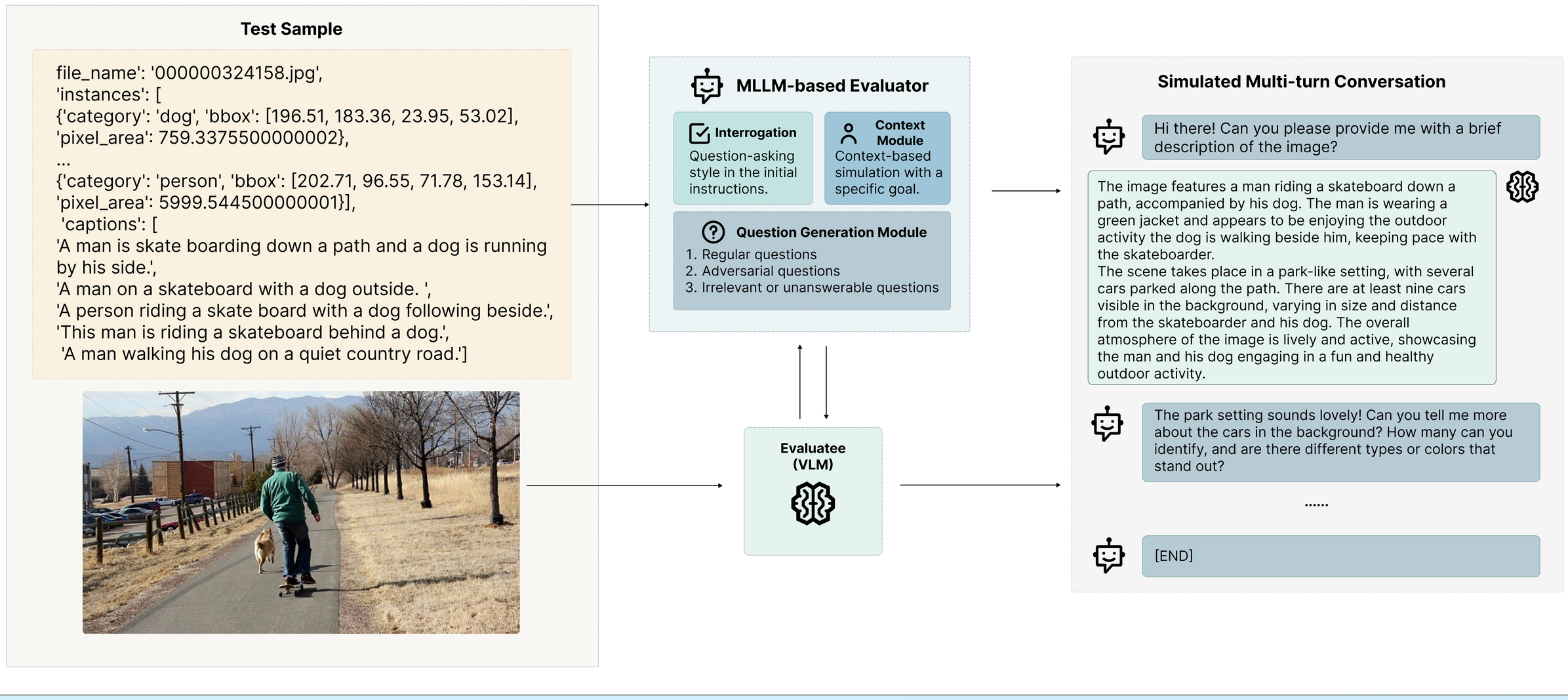
$$\text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|} \quad \text{CHAIR}_s = \frac{|\{\text{sentences with hallucinated object}\}|}{|\{\text{all sentences}\}|}$$

- **Human evaluation/annotation**:
  - Annotate hallucinations in the generated conversational transcripts at the sentence level.
  - Focus on object existence and quantity, attributes (e.g., color, position, actions),relations, and scene text.

Methodology

Multimodal large language models (MLLM)-based evaluation framework:

- **Context module** dynamically assigns the evaluator a context and user role with a specific image-querying objective and an initial question.
- **Question generation module** generates the unanswerable that cannot be resolved using the given image information, as well as adversarial questions that probe hallucinations.



Results

Compared to the caption-based method, model exhibits a **higher hallucination rate**, and **more objects are mentioned** in conversations when using our dynamic evaluator.

Evaluator	CHAIRs	CHAIRi	Coverage	Human
Caption	46	22.9	83.4%	72%
Dynamic	77.5	31.4	96.1%	76%

Table: The performance of LLaVA with different evaluators.

Conclusion

- We introduce a dynamic evaluation framework for MLLM hallucination evaluation.
- Our approach closes the realism gap by accounting for downstream task scenarios and user interactions.
- In the next phase, we will evaluate our framework's validity and reliability on multiple tasks.