Supplementary Material

Paper ID: 15750 ICCV 2025

Contents

A. Datasets and Evaluation Metrics

- **POPE (Polling-based Object Probing Evaluation)** is a framework designed to diagnose hallucination in large vision-language models (LVLMs) by probing the presence of objects through binary classification tasks. It avoids dependence on complex natural language inference by focusing on straightforward object existence verification. POPE introduces three sampling settings—*random*, *popular*, and *adversarial*—to ensure diverse and challenging evaluation scenarios. It samples 500 images each from three datasets: MSCOCO, A-OKVQA, and GQA. For each image, six probing questions are asked, leading to a total of 27,000 evaluation samples.
- **MME (MLLM Evaluation)** is a comprehensive multimodal benchmark that evaluates vision-language models in both perceptual and cognitive dimensions. It consists of 14 subtasks, including 10 focused on perception (e.g., object recognition and attribute detection) and four that address cognitive reasoning. In line with previous works such as [**? ?**], we utilize the *existence* and *counting* subsets to analyze object-level hallucinations. The remaining 12 subsets are used to assess generalization capabilities beyond direct object grounding.
- **ROPE (Recognition-based Object Probing Evaluation)** is an automated evaluation framework that targets multi-object hallucinations in LVLM. Unlike binary-based probing, ROPE uses visual referring expressions to specify object class distributions in images and examines whether models hallucinate non-existent objects or mis-recognize visual entities. In addition, it studies the influence of object salience, frequency in the training distribution, and internal biases of the models. ROPE thus provides fine-grained insights into model behavior in multi-object scenarios.

B. Qualitative Results

Figures 1 and 2 illustrate examples from the Whoops dataset, each containing an element that is ecologically or contextually mismatched. For example, a panda appears below the aurora borealis, despite pandas typically inhabiting bamboo forests in temperate regions. Another example shows a seagull in a dense tropical rainforest, far from its usual coastal habitat. Our method accurately pinpoints these incongruities by integrating visual cues with environmental context, whereas baseline models often hallucinate extraneous details or overlook the mismatch entirely. Another illustrative example (Figure 1 right) shows Abraham Lincoln, who died in 1865, shaking hands with Barack Obama, born in 1961, thus creating a clear temporal incongruity. Baseline models may note superficial details, such as both individuals wearing formal attire, but overlook the deeper anachronism that arises from merging historical figures separated by nearly a century. In contrast, our method leverages contextual reasoning and domain knowledge to identify that these two individuals could never have co-existed, thereby recognizing the core mismatch at a more semantically grounded level.

C. Discussion

C.1. Failure Case

Figure 3 illustrates a representative failure case within the category of“cognition” of the MME benchmark. The baseline model consistently misinterprets the textual or conceptual cues embedded in the images, leading to incorrect answers. For instance, it fails to recognize that a basketball is commonly played with the feet in certain sports contexts (e.g., soccer-style variants), overlooks basic code snippets that clearly indicate a C++ program, and mistakes a desk fan for a suitable cooling solution despite the query context. These failure cases reveal that our proposed method is not immune to the challenges of integrating visual information with real-world and domain-specific knowledge, particularly in highly exceptional scenarios that require an extremely complex level of contextual understanding and inference. In most routine cases, our model performs robustly; however, the observed shortcomings in these demanding cognition tasks suggest that there is still room for refinement. We contend that these instances represent edge cases, and future work will focus on incorporating a richer knowledge base and more sophisticated reasoning mechanisms to address these high-level contextual challenges.

C.2. Future Works

In the future, we plan to integrate a richer and more comprehensive knowledge base into our model, leveraging advanced multi-modal fusion strategies and sophisticated reasoning mechanisms to further enhance contextual understanding. In addition, our goal is to expand our evaluation framework by incorporating a diverse range of datasets that capture various domains and complex scenarios. This expanded analysis will allow for in-depth investigations of failure modes, thereby guiding targeted improvements in both model architecture and training protocols. By addressing these challenges, we expect to significantly boost the overall robustness and performance of our approach in complex cognitive tasks.

Figure 1. Examples of LLava 1.5 results on the Whoops dataset.

Figure 2. Examples of LLava-NeXT (LLaVA1.6) results on the Whoops dataset.

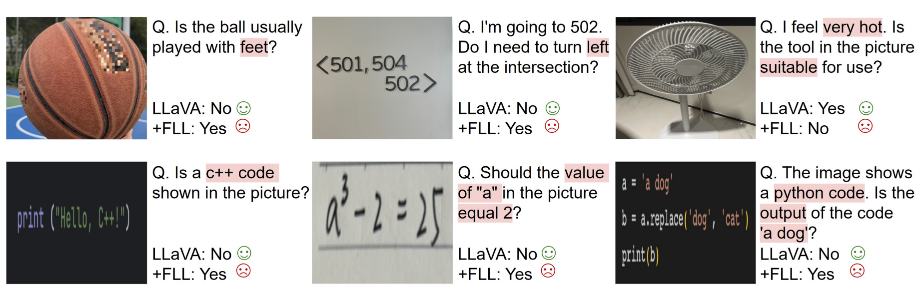

Figure 3. A failure case within the cognition category of MME