Supplementary for Evolutionary Eye-Hand Coordination with Large Language Models

6. Experiment Details

6.1. Reasoning Segmentation

Our study employed the LISA[9] model, utilizing LLaVa-1.0[13] and SAM[8] with a ViT-Huge[4] architecture, trained on eight A800 GPUs on both the train and validation sets of ReasonSeg. Testing was conducted on a single A800 GPU using a subset of 57 images from the ReasonSeg test set. This limitation was due to the GPT-4V[18] API"s (gpt-4-vision-preview) constraint of 100 images per day for each user. Owing to this significant restriction, we did not conduct further multi-round optimization and evolutionary strategy experiments on LISA-EHCO. Ideally, for GPT-4V results, providing both the original images and masked results would enhance the model"s understanding of highlighted areas. However, due to the same constraint, we could only offer a single image overlaid with a mask. This led to the exclusion of certain images where GPT-4V failed to recognize the masked areas from the selected 57 images. Furthermore, we observed that the GPT-4V API (gpt-4-vision-preview) tends to crash and become non-responsive when processing more than four images, potentially impacting the retention of history in multi-round experiments. Figure 6 illustrates additional experimental results on the test set.

6.2. Image Generation

The Recognizer was equipped with LLaVA-1.5. Due to GPT-4V"s limitations, extensive testing on GPT-4V was not feasible. The model selection for GPT-4[17] was GPT-4-Turbo (gpt-4-1106-preview), which showed significantly better performance in refining prompts compared to GPT-4. Figure 7 showcases more experimental results. These results also show some cases between the single loop strategy and the evolutionary strategy.

6.3. Reasoning Detection

Since relevant data was lacking, only a limited number of case studies were conducted on DetGPT[19], revealing improvements in bounding box and instruction interpretations. The text provided to GroundingDINO[14] often contained redundancies, leading to cluttered results. For instance, in cases pointing to the cue ball, the text given to GroundingDINO included the word "ball," resulting in the identification of all remaining pool balls. Direct optimization of GroundingDINO's input text by the Recognizer could potentially yield better results.

7. User Study Details

Our user study achieved a final participation count of 120 individuals, with a total view count of 440, translating to a submission rate of 27%. Among the participants, 22% used desktop devices, while 78% accessed the study using mobile devices. Within the mobile device category, 48% were Android users, and 30% were iOS users. The detailed content of the User Study questionnaire is attached.

7.1. Mitigating Hallucination

Hallucination is a complex open problem. The Evolutionary Eye-Hand Coordination may helps mitigate hallucination caused by unreasonable user inputs in reasoning tasks. As illustrated in Figure 8, when LLMs are paired with tools and confronted with irrational or inappropriate inputs, they tend to produce these hallucinatory outputs. For instance, in scenarios involving images of pool tables, when LLMs are tasked with segmenting objects like bananas, cats, or dogs – which are clearly not present in the image – the models may erroneously identify and segment unrelated parts of the image, such as mistaking a section of the pool table for a banana. This misinterpretation not only demonstrates the limitations of current LLMs in understanding and processing visual data but also highlights the challenges in aligning the model's outputs with realistic and logical expectations.

The introduction of the Evolutionary Eye-Hand Coordination, equipped with a "Recognizer" – essentially acting as the "eyes" of the system – significantly enhances the model's ability to discern and rectify such errors. Upon receiving an input, the Recognizer assesses the visual data and provides feedback to the LLM. This feedback includes information about the visual context, clarifications on the feasibility of the task, and corrections for any inaccuracies in the initial output.

When the Recognizer detects inputs leading to potential hallucinations, it promptly intervenes, signaling to the LLM that the input is unreasonable or illogical. This interaction not only prevents the generation of non-sensical outputs, but also guides the LLM in understanding the nature of plausible and implausible tasks in visual contexts. Consequently, the Evolutionary Eye-Hand Coordination reduces the occurrence of hallucinatory responses, thereby enhancing the reliability and accuracy of LLMs in reasoning tasks.

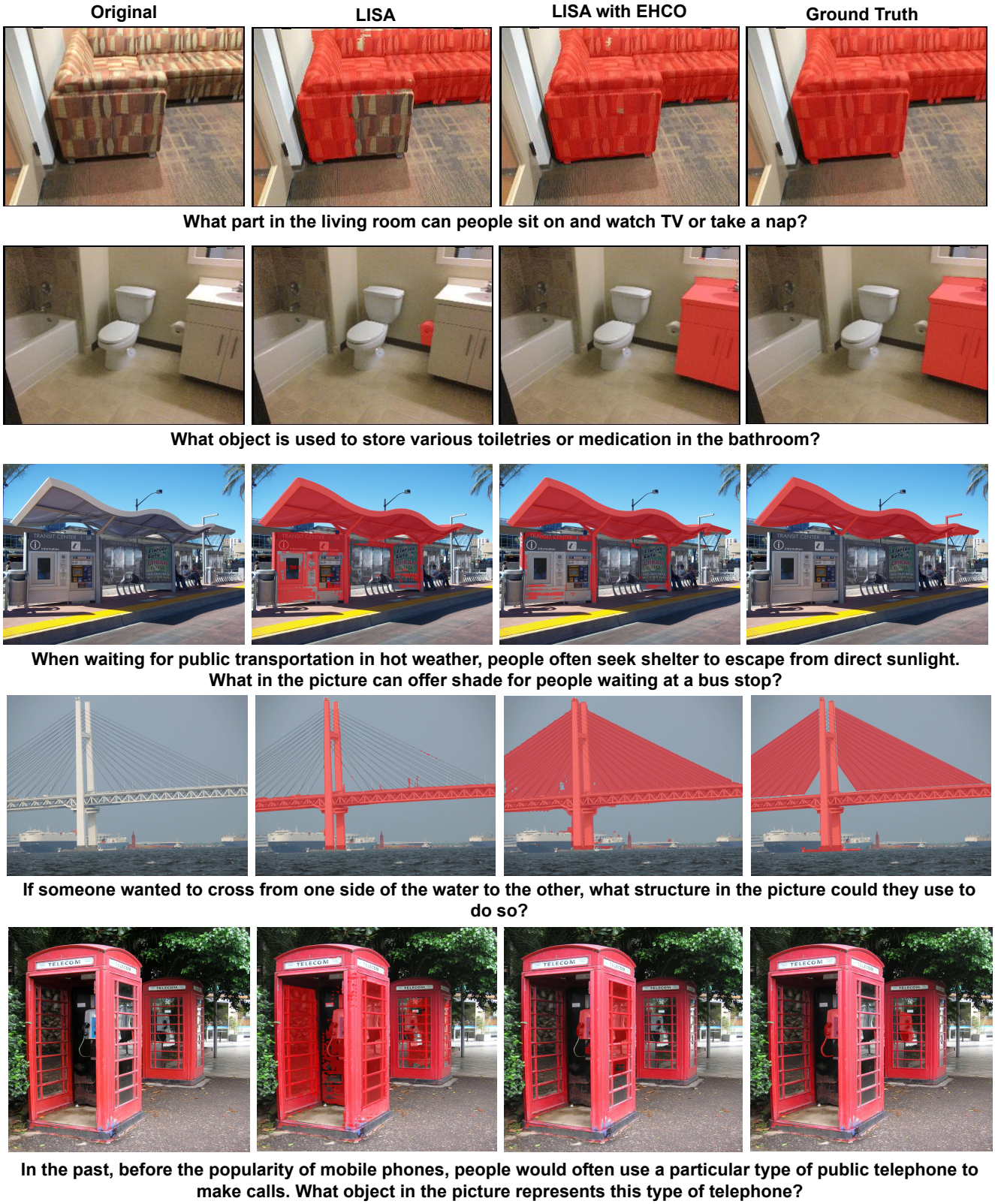

Figure 6. More examples for reasoning segmentation task. LISA in EHCO is performed better than LISA.

Figure 7. Comparative Performance of DallE-3, GPT-4-Assisted DallE-3, and DallE-3 in EHCO with single loop strategy and DallE-3 in EHCO with evolutionary strategy across Four Challenging Imaginative Generation Scenarios. This figure illustrates in "Cat is walking a dog" and "Book is reading a person" scenarios, evolutionary strategy shows better performance than single loop strategy. In the "Cat is walking a dog" scenario, the evolutionary strategy better demonstrates the Cat"s status as the owner of "walking". In the "Moon is shining on the sun" scenario, except for the poorer performance of the GPT-4-Assisted Dalle-3 result, all others successfully express the phenomenon of "Moon shining". Also, in these cases, it is easy to find the consistency of DallE-3 in EHCO outperforms GPT-4-Assisted Dalle-3.

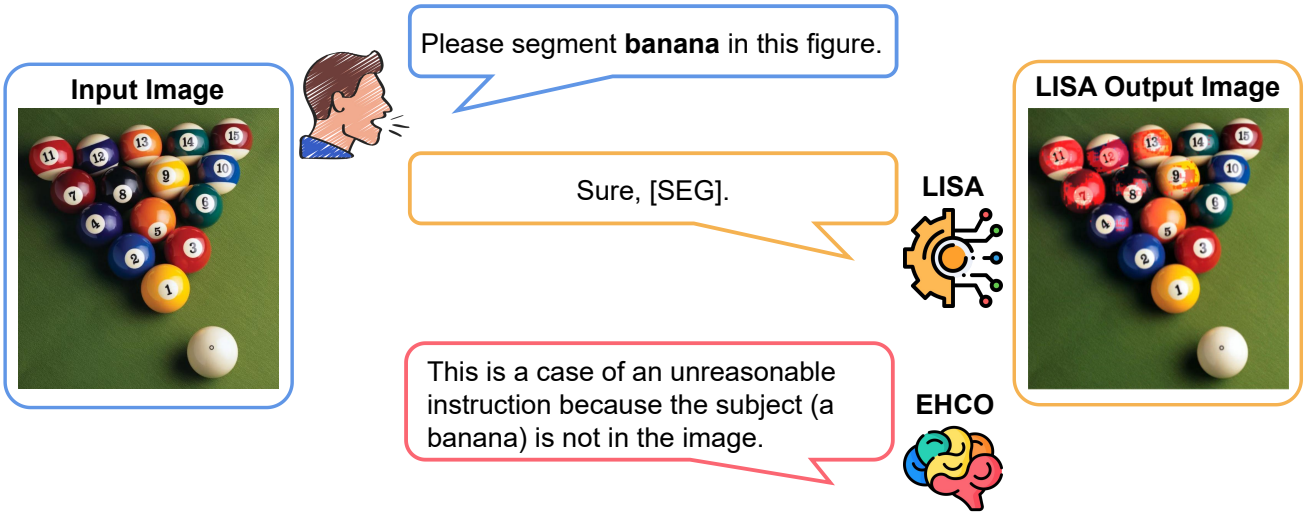

Figure 8. Performance of LISA as a Representative of Visual Tools and Evolutionary Eye-Hand Coordination (ECO) in Response to Unreasonable Input. The figure illustrates a scenario where an unreasonable instruction is given: segmenting a banana from an image of a billiard table. Visual Tools, due to the hallucination phenomenon, often provide incorrect responses, as shown by the erroneous segmentation of the top-left part of the billiard table. In contrast, within the ECO framework, the Recognizer effectively provides feedback, and the Brain component accurately identifies the input as unreasonable, refraining from generating an incorrect response. This comparison demonstrates that ECO can mitigate hallucination issues caused by unreasonable input.

7.2. Prompt Engineering

Prompt engineering plays a crucial role in establishing the Evolutionary Eye-Hand Coordination. The right prompts can significantly enhance the efficiency and optimization ceiling of the evolution, while inappropriate prompts may lead to operational failures. Each tool within the evolution requires specific prompts, typically including instructions that define the scope of tool operations. For instance, in Reasoning Detection tasks, the prompts can adjust the threshold values of Bounding Boxes.

When the Recognizer provides feedback, it includes details about the processed task, such as Mask and Bounding Box cues, along with the results from the current tool output and the user's original instruction. This feedback is presented to the Brain LLM in a natural language text format. The Brain LLM's critical decision is to determine the accuracy of the response and decide whether to halt or continue the evolution. If the evolution continues, the new set of optimized instructions for the next iteration is typically outputted in JSON format. Thanks to the new feature, Json-mode, of GPT-4-Turbo, the stability of the evolution has seen significant improvements.

In optimizing the instructions, particularly for tasks with special requirements like image generation, more conventional, template-based rules are provided as reminders. This approach helps in maintaining a structured and efficient flow of information, ensuring that each iteration of the evolution contributes to progressively optimizing the output in alignment with the user's intent. By carefully engineering these prompts, we can fine-tune the interaction between the Brain LLM and the Recognizer, leading to more accurate and relevant results from the Evolutionary Eye-Hand Coordination.