LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Thesis proposal

Topic: **Look Again: Prompt-Level Visual Repetition for Robust Reasoning in Vision–Language Models**

Supervisor: Ali Modarressi

Examiner: Hinrich Schütze

Level: MSc

Summary:

This project studies a specific reliability issue in Vision–Language Models (VLMs): **incorrect early visual perception cascading into wrong reasoning**. In typical VLMs the image is encoded into `<image>` token embeddings that the language model uses to answer multimodal questions. However, recent analyses of multimodal hallucination and visual grounding failures in VLMs show that models can drift from the actual visual content and produce inconsistent answers, suggesting that early misinterpretations may persist through the reasoning process [1].

The first objective is to quantify how severe this issue is by constructing a simple diagnostic dataset of controlled images (e.g., colored objects with explicit counts) and associated questions. We will deliberately perturb the model's internal reasoning trace — for example, by injecting an incorrect visual observation into a chain-of-thought prompt or a reasoning-capable VLM — and measure to what extent these erroneous visual states lead to incorrect final answers. This causal perturbation approach differs from much of the existing evaluation literature, which typically characterizes hallucination errors at the output level without isolating the impact of specific early visual mistakes.

The second objective is to explore **simple inference-time heuristics** that repeat or represent the image input within the prompt to see if the model can correct a bad initial perception. The idea takes inspiration from re-reading strategies in text reasoning, where re-exposure to the input improves comprehension, as well as from multimodal work that "looks twice" at visual tokens during hidden-state processing [2]. However, our approach is distinct in that the repetition is done **purely at the prompt/trace level**, without modifying model internals. We will compare single-pass reasoning to variants with image repetition or two-stage visual checks to evaluate whether such heuristics reduce error propagation and improve accuracy.

Models that could be used for this task could be any VLM where we prompt for CoT reasoning or already reasoning capable models such as Qwen3-VL.

Requirements: SGLang, HuggingFace, A good understanding of VLMs, reasoning-based models and/or CoT prompting.

References:

1. Hanchao Liu et al. (2024). *A Survey on Hallucination in Large Vision-Language Models*. arXiv: 2402.00253 [cs.CV]. URL: https://arxiv.org/abs/2402.00253

2. Xin Zou et al. (2025). *Look Twice Before You Answer: Memory-Space Visual Retracing for Hallucination Mitigation in Multimodal Large Language Models*. arXiv: 2410.03577 [cs.CV]. URL: https://arxiv.org/abs/2410.03577