Master Thesis Proposal Outline

Thesis Title

Diagnosing and Calibrating VLMs in Quantitative Reasoning

1. Problem Description

While Vision-Language Models (VLMs) have shown remarkable progress, their ability to perform quantitative physical reasoning (QPR)—such as kinematic physics analysis—remains fundamentally limited. Current evaluation paradigms predominantly rely on monolithic accuracy metrics (e.g., Mean Relative Accuracy), which provide a holistic score but offer no insight into the cognitive mechanisms of where or why models fail. Without a fundamental understanding of how VLMs perceive the physical world, comprehend physical laws, and handle temporal grounding, efficiently enhancing VLMs remains elusive. Moreover, the reliability of model confidence in continuous numerical spaces is largely underexplored. This thesis aims to addresses these critical gaps by explicitly decomposing quantitative reasoning into interpretable dimensions and calibrate confidence reliability in English-based kinematic physics tasks.

2. Research Questions (RQs)

- **RQ1 (Mechanistic Diagnostic Decomposition):** Do VLMs identify, comprehend, and locate the minimal necessary cognitive dimensions—namely visual fidelity, physical law comprehension, and temporal grounding—as humans do during quantitative physical reasoning?
- **RQ2 (Confidence Reliability and Calibration):** Are VLMs reliably confident in their numeric answers with an awareness of ground-truth boundaries, and how can neighborhood-informed calibration effectively bridge the calibration gap?

3. Research Objectives

1. **Cognitive Bottleneck Attribution:** Adapt and apply multi-modal diagnostic probes to systematically isolate reasoning failures into distinct cognitive dimensions (visual perception, formula selection, and temporal indexing).
2. **Neighborhood-Informed Calibration:** Evaluate model uncertainty in open-ended generative spaces and propose neighborhood-aware calibration techniques to enhance confidence reliability.
3. **Actionable Roadmap for Trustworthy VLMs:** Provide structural insights into cumulative hallucination effects and guide future architecture design for physically-grounded VLMs.

4. Methodology & Scope

- **Models:** Evaluate state-of-the-art open-source VLMs (e.g., Qwen3-VL, Gemma3, InternVL) in a zero-shot English reasoning setting.

- **Diagnostic Framework:** Employ FACT (decomposing visual fidelity, physical law comprehension, and temporal grounding via targeted probes) and NICE (neighborhood-informed confidence evaluation and calibration).
- **Evaluation Metrics:** Macro F_1, F1 score for probing tasks, Mean Relative Accuracy (MRA) for numerical reasoning, and Neighborhood Confidence Index (NCI) / Neighborhood Calibration Error (NCE) for uncertainty estimation.

5. Expected Contributions

- Shifting the evaluation paradigm from black-box monolithic accuracy to mechanistic, interpretable failure attribution in physical reasoning.
- Establishing novel calibration metrics and neighborhood-informed techniques that effectively mitigate probability over-concentration and improve model trustworthiness.

References

1. Li Puyin, Tiange Xiang, Ella Mao, Shirley Wei, Xinye Chen, Adnan Masood, Li Fei-fei, and Ehsan Adeli. (2025). Quantiphy: A quantitative benchmark evaluating physical reasoning abilities of vision-language models.
2. Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. (2025). Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *ICLR 2025*
3. Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. (2017). On calibration of modern neural networks. *International Conference on Machine Learning (ICML)*, pp. 1321–1330.

If you are interested in the proposal, please contact Jian Lan via lan@dbs.ifi.lmu.de