Master Thesis Topic:

Leveraging Structured Knowledge Grounding to Enhance Physical Reasoning in VLMs

1 Introduction and Background

Current Vision-Language Models (VLMs) demonstrate impressive performance in static scene understanding but struggle with the dynamic and causal logic of the physical world. As highlighted by the PhysBench [1] benchmark, models frequently fail to predict the outcomes of simple physical interactions (e.g., collision, stability, or fluid flow) due to a lack of grounded physical common sense.

To address this, this work proposes to explore a training-centric approach that moves beyond standard Supervised Fine-Tuning (SFT). The core idea is Structured Knowledge Grounding (SKG): a strategy that forces the model to decompose a physical scene into a structured representation (objects, states, and constraints) before generating a reasoning path. By fine-tuning the model to align visual sequences with explicit physical principles, we aim to transform the model from a probabilistic predictor into a causal reasoner capable of zero-shot generalization in complex physical environments.

2 Research Questions

- Q1: Causal Alignment: What are effective ways to fine-tune a VLM to output intermediate Physical State Representations (e.g., mass, friction, velocity labels) which significantly reduce hallucinations in qualitative outcome prediction?
- Q2: Strategy Effectiveness: How does a Chain-of-Thought (CoT) fine-tuning strategy—which explicitly supervises the reasoning steps (Observation → Law Discovery → Prediction)—compare against vanilla SFT on PhysBench's diverse tasks?
- Q3: Cross-Task Robustness: Does a model trained on basic mechanics (e.g., rigid body dynamics) show emergent reasoning capabilities in other PhysBench domains like fluid dynamics or thermodynamics?

3 Tasks & Goals (Proposed Training Strategies)

- **Diagnostic Benchmarking**: Perform an in-depth analysis of baseline VLM performance on **PhysBench** to identify specific failure modes in qualitative reasoning (e.g., collision logic vs. stability prediction).
- **Structured Data Synthesis**: Develop an automated pipeline to augment the dataset with **Reasoning Traces**. This involves pairing visual scenarios with structured "Rationale" (Observation → Physical Law → Prediction) to move beyond simple label-matching.

- **Structured Knowledge Fine-Tuning (SKFT)**: Implement a fine-tuning strategy that supervises the model's internal reasoning path. This forces the VLM to align its visual perception with explicit physical principles during the training phase.
- **Causal Robustness Training**: Integrate **Counterfactual Scenarios** into the training loop (e.g., modifying gravity or mass) to ensure the model learns causal physical laws rather than memorizing statistical visual patterns.
- **Cross-Domain Evaluation**: Validate the generalizability of the proposed training strategy by evaluating the model on unseen physical categories within PhysBench, ensuring the reasoning capability is task-agnostic.

4 Requirements

- Strong proficiency in Python and Deep Learning frameworks (PyTorch, Transformers).
- Experience with Fine-tuning techniques (LoRA, QLoRA, or Full Fine-tuning).
- Ability to handle and augment large-scale multi-modal datasets.

Contact

Please contact only if you are very interested in the topic and believe you are a good match. Please send your CV / self-introduction and transcripts to lan@dbs.ifi.lmu.de

References

- [1] Deitke, M., et al. (2023). *PhysBench: Benchmarking and Analyzing the Physical Reasoning Capabilities of VLMs.* (The primary benchmark for the thesis).
- [2] Wei, J., et al. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.* NeurIPS. (Basis for the reasoning-step supervision).
- [3] Radford, A., et al. (2021). *Learning Transferable Visual Models from Natural Language Supervision.* In Proceedings of ICML.