LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Thesis proposal

Topic: **Explore the Spatial-Temporal Reasoning Capability of Foundation Models in Vision-Language Navigation**

Supervisor: Shengqiang Zhang

Examiner: Prof. Hinrich Schütze

Level: BSc / MSc

Summary: Vision-language navigation (VLN) is a task where an agent navigates in a 3D environment based on natural language instructions. The agent needs to understand the environment and the instructions to make decisions. In this task, it's important for the agent to have a good spatial-temporal understanding and reasoning capability.
In this project, we will explore such capability with some popular foundation models. More specifically, we will collect a dataset which contains a sequence of observation images during the agent's navigation along with the ground-truth trajectories. Then we want to explore whether the tested models can predict the temporal order of these images based on the model's understanding of the environment and the instructions. We will try to test some popular vision-language models in this task and analyze their performance.

Requirements: Good Python programming background, basic knowledge and experience of machine learning.

References: Ku, Alexander, et al. "Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding." arXiv preprint arXiv:2010.07954 (2020).