Statement of Research Interests - Joint PhD Program: ETH Zürich x Microsoft

This Research Statement outlines my technical background and academic objectives in the fields of computer vision, generative modeling, and multimodal reasoning. My primary research focus lies at the intersection of these disciplines, specifically exploring how intelligent systems can move beyond simple pattern recognition toward structured reasoning and world understanding.

My previous research at the Hasso Plattner Institute involved investigating the synergy between identifiable Variational Autoencoders (iVAE) and denoising diffusion probabilistic models. By leveraging the diffusion model's ability to capture data distributions across scales alongside the iVAE's capacity for identifiable latent representations, our team was able to learn disentangled representations more effectively than traditional methods. This architecture specifically enabled the generation of more detailed and nuanced interpolations based on those learned representations. This can be used to create world models that disentangle different parts of the generated environments that do not interact together and therefore should be independent in their representation.

Additionally, in a lecture and seminar on reinforcement learning, I learned the fundamentals of the discipline and did research on reinforcement learning for learning search trees for database indexing. By framing the construction of a search tree as a Tree-Markov Decision Process, we enabled an agent to learn the optimal split points tailored to specific, non-uniform query distributions. Utilizing an Advantage Actor-Critic architecture, we achieved a 40% reduction in average lookup time for skewed access patterns. This research improved my understanding of reward shaping and policy optimization in complex, multi-step state spaces.

Currently, I am beginning my Master's thesis, which focuses on detecting semantically complex interacting groups of objects using Vision-Language Models (VLMs). The objective is to enable models to reason about object relations and group dynamics at a pixel-wise granularity (e.g. detecting, describing and reasoning about socially affiliated human groups in street view imagery), addressing a specific challenge where current state-of-the-art models struggle.

My long-term research interest is to evolve visual representation learning toward deep reasoning and semantically grounded representations. I am particularly interested in how causal and multimodal principles, such as integrating depth estimation, touch, or sensory signals in general, can help models generalize to new domains. By extending reasoning-enabled vision systems with world understanding and the ability to take actions based on it, we can train models which make decisions grounded in their environments. I therefore want to contribute to the development of a unified foundation model that treats "Action" as a first-class modality alongside Vision and Language. Concretely, I am interested in developing action-conditioned world models that allow agents to reason about object interactions, anticipate the consequences of interventions, and plan accordingly.

For this, I plan to build on my diverse background in reinforcement learning, generative image models, and vision-language models to develop systems trained with disentangled representations and improved semantic comprehension of visual input. Here, I believe ETH Zürich and Microsoft provide an excellent, well-resourced academic environment to achieve this, through access to world-class foundational models and researchers.