Research Statement

Berk Çiçek
cicekberk8@gmail.com — berk-cicek.github.io

My research bridges foundation models and embodied AI, targeting contact-rich robotic manipulation through Vision-Language-Action architectures enhanced with adaptive world models. I address a fundamental question: *How can we leverage semantic reasoning of large-scale foundation models while maintaining physical grounding for precise, force-aware control?*

Previous Work: Neuro-Symbolic Embodied AI

My master's thesis, **Contact-VLA**, addresses a critical limitation of current VLA models—their inability to perform zero-shot planning in contact-rich scenarios requiring explicit physical reasoning. While end-to-end policies like OpenVLA excel at demonstrated tasks, they struggle with dynamic contact interactions such as pivoting objects against walls or maintaining constant forces.

To overcome this, I developed **CoRAL** (Contact-Rich Adaptive LLM-based Control), a neuro-symbolic framework that decouples semantic reasoning from reactive control. Rather than using foundation models as direct action predictors, CoRAL employs them as *world model designers* and *cost function synthesizers* for a Model Predictive Path Integral (MPPI) controller. The architecture features two nested feedback loops: a high-frequency inner loop where MPPI performs reactive control with real-time force feedback, and a low-frequency outer loop where an LLM performs online system identification to refine physical parameters (mass, friction) and adapt strategies mid-execution.

This explicitly addresses *physical parameter ambiguity* in vision-based world models. A VLM provides semantic priors (e.g., "this wooden board is light") that initialize the physics simulator for MPPI rollouts. When robot interaction reveals model mismatch—such as unexpected slippage from underestimated friction—the LLM diagnostically updates both world model parameters and cost structure, enabling task recovery. Validated on simulation and real Franka Emika Panda hardware, CoRAL achieves over 50% higher success rates than VLA baselines on novel contact-critical tasks while maintaining explainability through natural language reasoning. This work is under review at RSS 2026.

Complementing this, my **H-MaP** work (RA-L 2025, ICRA 2026) tackles sequential manipulation by combining discrete symbolic planning with continuous motion optimization for multi-stage tasks, while **FViT-Grasp** explores efficient vision transformers for real-time 6-DoF grasp detection. I am also developing **PuzzleWorld-3D**, an API-based benchmark for evaluating reasoning in embodied AI agents through programmatic interfaces designed for foundation models and VLA architectures.

Future Research Vision

I am excited to explore three interconnected directions: (1) **Scalable World Models**—investigating how self-supervised learning on large-scale robotic datasets can produce generalizable physical priors for zero-shot transfer, (2) **Multimodal VLA Architectures**—extending beyond RGB-D to incorporate force/torque and tactile sensing, with pre-training objectives that capture contact dynamics, and (3) **Embodied Reasoning at Scale**—developing memory architectures that enable agents to accumulate and reuse manipulation strategies across diverse contexts.

Why ETH Zürich and Microsoft

The joint PhD program uniquely combines ETH Zürich's academic rigor with Microsoft's industrial-scale resources and access to state-of-the-art foundation models—essential for training and evaluating embodied AI systems at scale with diverse real-world data. Working with Prof. Marc Pollefeys and Dr. Oier Mees, whose expertise in 3D vision and robot learning aligns perfectly with my research interests, would provide an ideal environment to push the boundaries of what foundation models can achieve in physical world interactions.