

GuardVLA

GuardVLA: Hierarchical Safety Reasoning for Trustworthy Vision-Language-Action Systems

Programm / Ausschreibung	FORPA, Dissertaionen 2024, Industrienae Dissertationen 2026	Status	laufend
Projektstart	01.09.2026	Projektende	31.08.2029
Zeitraum	2026 - 2029	Projektlaufzeit	36 Monate
Barwert der Projektförderung	€ 109.950		
Keywords	Safety-Aware Embodied AI, Vision-Language-Action Models, Continuous Risk Modelling		

Projektbeschreibung

Vision-Language-Action-Modelle (VLA) entwickeln sich zunehmend zu einer Schlüsseltechnologie im Bereich der embodied AI. Sie ermöglichen es Robotern, visuelle Wahrnehmung, Sprachverständnis und Handlungsgenerierung in einer gemeinsamen Steuerungsarchitektur zu verbinden. Derzeitige VLA-Systeme sind jedoch noch nicht ausreichend für den Einsatz in sicherheitskritischen realen Umgebungen geeignet. Zu den zentralen Herausforderungen zählen ihr hoher Rechenaufwand, eine eingeschränkte Echtzeitfähigkeit, stark vereinfachte binäre Sicherheitsrepräsentationen sowie das Fehlen strukturierter Mechanismen zur Sicherheitsüberwachung, Fehlerdiagnose und Wiederherstellung während der Ausführung.

Ziel dieses Projekts ist die Entwicklung eines sicherheitsbewussten VLA-Frameworks, das kontinuierliche Risikomodellierung und hierarchisches Sicherheitsreasoning in die robotische Steuerung integriert. Sicherheit soll dabei nicht lediglich als binäre Unterscheidung zwischen „sicher“ und „unsicher“ verstanden werden. Stattdessen wird Risiko als abgestufter, kontextabhängiger Sicherheitswert modelliert. Das Framework umfasst zwei komplementäre Ebenen: eine hochfrequente physische Sicherheitsebene, die unmittelbare Risiken erkennt und reaktive Eingriffe ermöglicht, sowie eine langsamere Reasoning-Ebene, die Fehlerursachen diagnostiziert, Neuplanung unterstützt und korrigierende Eingaben für die Roboter-Policy generiert.

Als Ergebnis wird eine modulare und übertragbare Sicherheitsarchitektur für VLA- und VLN-basierte Robotersysteme erwartet. Im Rahmen des Projekts werden sicherheitsrelevante Grenzen bestehender VLA- und VLN-Architekturen identifiziert, kontinuierliche und schweregradabhängige Sicherheitsmodelle entwickelt, ein zweischichtiger Sicherheitsmechanismus implementiert und der Ansatz in repräsentativen robotischen Szenarien validiert. Die erwarteten Ergebnisse sollen die Robustheit, Reaktionsfähigkeit, Wiederherstellungsfähigkeit und Vertrauenswürdigkeit von embodied-AI-Systemen verbessern und damit ihren zukünftigen Einsatz in Bereichen wie Servicerobotik, Logistik im Gesundheitswesen, autonomen Systemen und industrieller Automatisierung unterstützen.

Abstract

Vision-Language-Action (VLA) models are becoming a key technology for embodied AI, enabling robots to combine visual perception, language understanding, and action generation within a unified control architecture. However, current VLA systems are not yet suitable for safety-critical real-world deployment. Major limitations include high computational demands, limited real-time responsiveness, simplified binary safety representations, and the lack of structured mechanisms for safety monitoring, failure diagnosis, and recovery during execution.

This project aims to develop a safety-aware VLA framework that integrates continuous risk modelling and hierarchical safety reasoning into robotic control. Instead of treating safety only as a binary safe/unsafe decision, the proposed approach models risk as a graded, context-dependent safety cost. The framework will consist of two complementary layers: a high-frequency physical safety layer for immediate risk detection and reactive intervention, and a slower reasoning layer for diagnosing failure causes, replanning, and generating corrective inputs for the robot policy.

The expected outcome is a modular and transferable safety architecture for VLA/VLN-based robotic systems. The project will identify safety-relevant limitations of current VLA/VLN architectures, develop continuous and severity-aware safety models, implement a dual-layer safety mechanism, and validate the approach in representative robotic scenarios. The results are expected to improve robustness, responsiveness, recovery capability, and trustworthiness of embodied AI systems, supporting their future deployment in domains such as service robotics, healthcare logistics, autonomous systems, and industrial automation.

Projektpartner

- Virtual Vehicle Research GmbH