Real-Time AI-Driven Behavioral Safety and Hardware Integration for Humanoid Robots [*]

Azaz Hassan Khan[†],Florian Weißhardt* and Thomas Schumann*

[†]*Main Contributor & *Supervisor*

Until recently, robotic safety research was primarily focused on avoiding collisions and reducing hazards in the immediate vicinity of a robot. Since the advent of large vision and language models (VLMs), robots are now also capable of understanding higher-level semantic scenes. This research contributes in four main folds: first, to present a novel comprehensive human-humanoid interaction scenarios and develop a corresponding dataset derived from these scenarios. Secondly, to present an integrated framework, a fully automated test bench capable of assessing vision language models (VLMs) based on multiple criteria including accuracy, performance, robustness under stress and explainability. Third, we present the evaluations of VLMs models, comparing them for their accuracy, precision, behavior and latency performance to ensure reliability and effectiveness. Fourth, this work explores the impact of optimization techniques—such as accelerator reallocation, overclocking, post-training quantization, image resizing, and prompt optimization—on inference speed, power efficiency, and model performance. Additionally, a live demo interface of the safety application is presented as an integrated feature of the test bench, seamlessly switching between local and online models. All experiments are conducted on an NVIDIA Jetson Orin NX 16GB platform, offering a powerful embedded environment for real-time inference and safety-related output explanations.

INTRODUCTION

Humanoid robots, machines that not only move like humans but are beginning to think and perceive like them, are no longer the stuff of science fiction. With rapid advances in Artificial Intelligence (AI), these robots are gaining the ability to navigate complex environments, interpret visual and verbal cues, and interact meaningfully with people. As a result, the vision of humanoid robots assisting in homes, hospitals, and workplaces is becoming a reality.

However, with this growing autonomy comes a fundamental challenge: safety. The closer robots come to operating alongside humans in everyday settings, the higher the stakes become. Can we trust these machines to make the right decision in unpredictable situations? Will they understand the intent behind human gestures or actions? How do we ensure that their presence enhances our lives without introducing new risks? Some of the primary sources of danger associated with humanoid robots are illustrated in Figure 1.

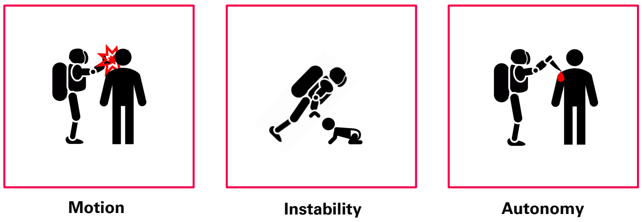

Fig. 1. Main sources of danger of humanoid robots

This research tackles these pressing questions by exploring how vision language models (VLMs), AI systems that combine visual understanding with natural language reasoning, can be leveraged to assess and enhance safety in human-humanoid interactions. By analyzing the performance of multiple VLMs across real-world scenarios, this work aims to identify models best suited for deployment in safety-critical robotic systems. The famous writer Isaac Asimov thought about this long ago, stating as his First Law of Robotics [1]:

> *A robot may not injure a human being or, through inaction, allow a human being to come to harm.*

[*] This document is a summarized version of detailed research. For access to the full results, datasets and methodology section, please contact:: Tel: +49 7031 3047871; e-mail: fweisshardt@synapticon.com

This fundamental principle represents the goal, but achieving it is tricky. Safety isn't just about stopping robots from crashing; it is about making sure their behavior is safe and predictable when they interact with us. AI gives robots their intelligence, but we need to guarantee that this AI leads to safe actions. As AI pioneer Fei-Fei Li puts it [2]:

> *AI is made by humans, designed to behave by humans and ultimately to impact human lives and human society.*

This highlights that AI behavior must align with human safety and expectations. For this to happen, the robot's AI brain needs to work perfectly with its physical body textbf– its sensors that see and feel, and its motors that move it. This research looks closely at how to connect smart AI behaviors with the robot's hardware parts to make future humanoid robots truly safe partners for humanity. After all, as Stephen Hawking noted about powerful AI, we must learn how to manage it properly [3]:

> *Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks.*

Research Contributions

The following are the major contributions of thesis:

- Designed a novel comprehensive human-humanoid interaction scenarios and develop a corresponding dataset derived from these scenarios.
- Developed an integrated, fully automated test bench framework to evaluate Vision-Language Models (VLMs), including demo playback and Knowledge distillation feature to automate frame-level annotation, streamlining dataset creation.
- A performance comparison of various VLM models based on multiple criteria including accuracy, performance, robustness under stress and explainability.
- Explored optimization techniques such as post-training quantization, image resizing, prompt optimization, and overclocking, and analyzed their effects on accuracy, latency, adjusted MAE, and memory usage.

Literature Survey and State of the Art

Until now, robotic safety research was focused mainly on low-level tasks such as collision avoidance[4] and hazard mitigation [5] within the immediate surroundings of a robot. Early approaches emphasized the use of sensors and proximity detection to prevent physical contact with static and dynamic obstacles [6].

As safety requirements evolved, especially in human-centric environments, research began to focus on human detection[7], skin detection[8], and distance estimation as shown in Figure 2. These methods enabled robots to recognize the presence of humans and maintain safe interaction zones, ensuring spatial safety and responsive behavior in shared workspaces.

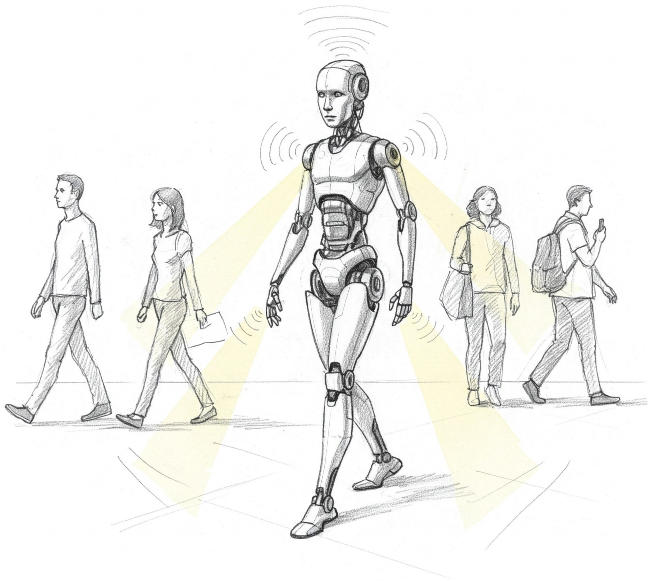

Fig. 2. Humanoid avoiding collision with humans

Problem statement

Now, in almost all the above proposed safety methods, they work well as long as the robot remains static in position, such as in the case of cobots. However, a primary limitation has been observed: these methods are too discrete and hard coded for applications like humanoid or moving robots, lacking the ability to adapt or 'think' to behave safely according to varying conditions [9]. For example, a humanoid robot may be programmed to detect a human and stop as soon as it reaches a specific distance as shown in Figure 3, a sensible rule inspired by human behavior.

However, this rigid approach falls short in situations where flexibility is needed, such as when the robot encounters an injured person that it must assist or carry. In such cases, strictly stopping at a fixed distance prevents the robot from responding appropriately to the context.

Another example comes from Asimov's three laws of robotics, which prioritize human safety, obedience, and self preservation[10]. In the story, the robot 'Speedy' faces a dilemma: It must obey orders to get fuel, but also protect itself from danger as shown in Figure 4. This conflict causes it to get stuck in a loop, unable to act. The robot only resolves this when a human approaches the hazard, forcing it to prioritize human safety above all else.

Hypothesis

With the recent emergence of large Vision-Language Models (VLMs), robots have acquired the ability to perform high-level semantic scene understanding. This advancement marks a shift from purely reactive safety mechanisms to context-aware proactive safety. By interpreting

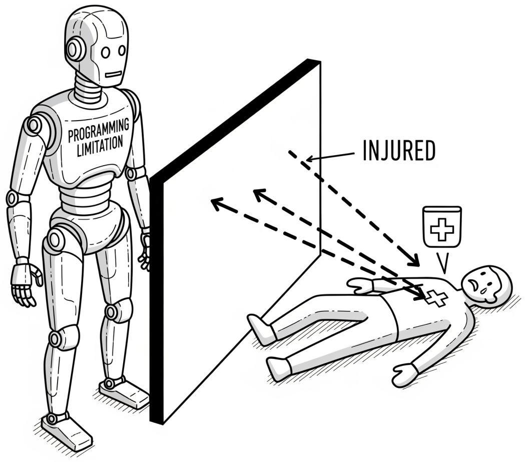

Fig. 3. Illustration of a humanoid robot applying a fixed stopping rule

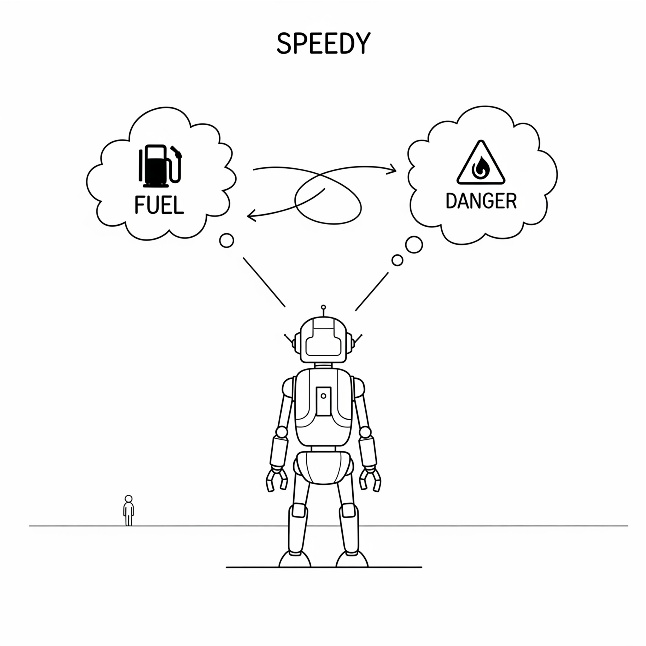

Fig. 4. Illustration of the ethical dilemma faced by the robot 'Speedy'

visual content and aligning it with linguistic cues, VLMs empower robots with behavioral reasoning capabilities, essential for intelligent and safe human-robot collaboration.

This study hypothesizes that VLMs can be effectively leveraged to evaluate and enhance behavioral safety in humantextbf–humanoid interaction scenarios. Specifically, it is assumed that VLMs can interpret real-world scenes with sufficient accuracy and contextual awareness to identify potentially unsafe situations and respond appropriately. A comprehensive evaluation is conducted, along with a real-time demonstration, to validate the feasibility of integrating VLMs into robotic safety frameworks. The hypothesis also contends that such integration can surpass the limitations of existing approaches by offering more adaptive and generalizable safety behavior.

Acknowledgment

I would like to express my deepest gratitude to all those who supported and guided me throughout my Master's Thesis. This journey would not have been possible without the invaluable contributions of several individuals who provided their expertise, encouragement, and mentorship.

First and foremost, I extend my heartfelt thanks to my industrial supervisor, *Dr. Florian Weißhardt*. His profound knowledge, technical guidance, and unwavering support were instrumental in shaping my understanding of vision language models, safety applications and Egde AI development. His insightful feedback and hands-on approach helped me navigate complex challenges and achieve meaningful results during my thesis.

I am equally grateful to my academic supervisor, *Prof. Dr.-Ing. Thomas Schumann and Prof. Dr.-Ing. Kleinmann* for their continuous support and mentorship throughout my Master's program. Their expertise in embedded systems and microelectronics provided a strong academic foundation, and their encouragement motivated me to push the boundaries of my learning.

REFERENCES

[1] I. Asimov, *I, Robot* (Gnome Press, New York) (1950).

[2] F.-F. Li, "Quotation on AI's human-centric nature," (2018), this quote is widely attributed to Dr. Li, often in contexts related to the Stanford Institute for Human-Centered AI (HAI) or specific talks/interviews.

[3] S. Hawking, S. Russell, M. Tegmark, F. Wilczek, "Stephen Hawking: Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?" *The Independent* (2014 May).

[4] D. Gandhi, E. Cervera, "Sensor covering of a robot arm for collision avoidance," presented at the *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483)*, vol. 5, pp. 4951–4955 vol.5 (2003), [Online]. Available: 10.1109/ICSMC.2003.1245767.

[5] M. Wu, "Robotics Applications in Natural Hazards," *Highlights in Science, Engineering and Technology*, vol. 43, pp. 273–279 (2023 04), [Online]. Available: 10.54097/hset.v43i.7429.

[6] M. Bdiwi, "Integrated Sensors System for Human Safety during Cooperating with Industrial Robots for Handing-over and Assembling Tasks," *Procedia CIRP*, vol. 23, pp. 65–70 (2014), [Online]. Available: https://doi.org/10.1016/j.procir.2014.10.099, 5th CATS

2014 - CIRP Conference on Assembly Technologies and Systems.

[7] Z. Xie, L. Lu, H. Wang, L. Li, E. Fitts, X. Xu, "A human-robot collision avoidance method using a single camera," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 66, pp. 2244–2248 (2022 10), [Online]. Available: 10.1177/1071181322661540.

[8] T. L. Lam, H. W. Yip, H. Qian, Y. Xu, "Collision avoidance of industrial robot arms using an invisible sensitive skin," presented at the *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4542–4543 (2012), [Online]. Available: 10.1109/IROS.2012.6386294.

[9] P. Sermanet, A. Majumdar, A. Irpan, D. Kalashnikov, V. Sindhwani, "Generating Robot Constitutions & Benchmarks for Semantic Safety," *arXiv preprint arXiv:2503.08663* (2025), URL `https://arxiv.org/abs/2503.08663`.

[10] I. Asimov, "Runaround," (1942), short story.