Friends of CHEDDAR

Vision-Language Models on the Edge for Real-Time Robotic Perception

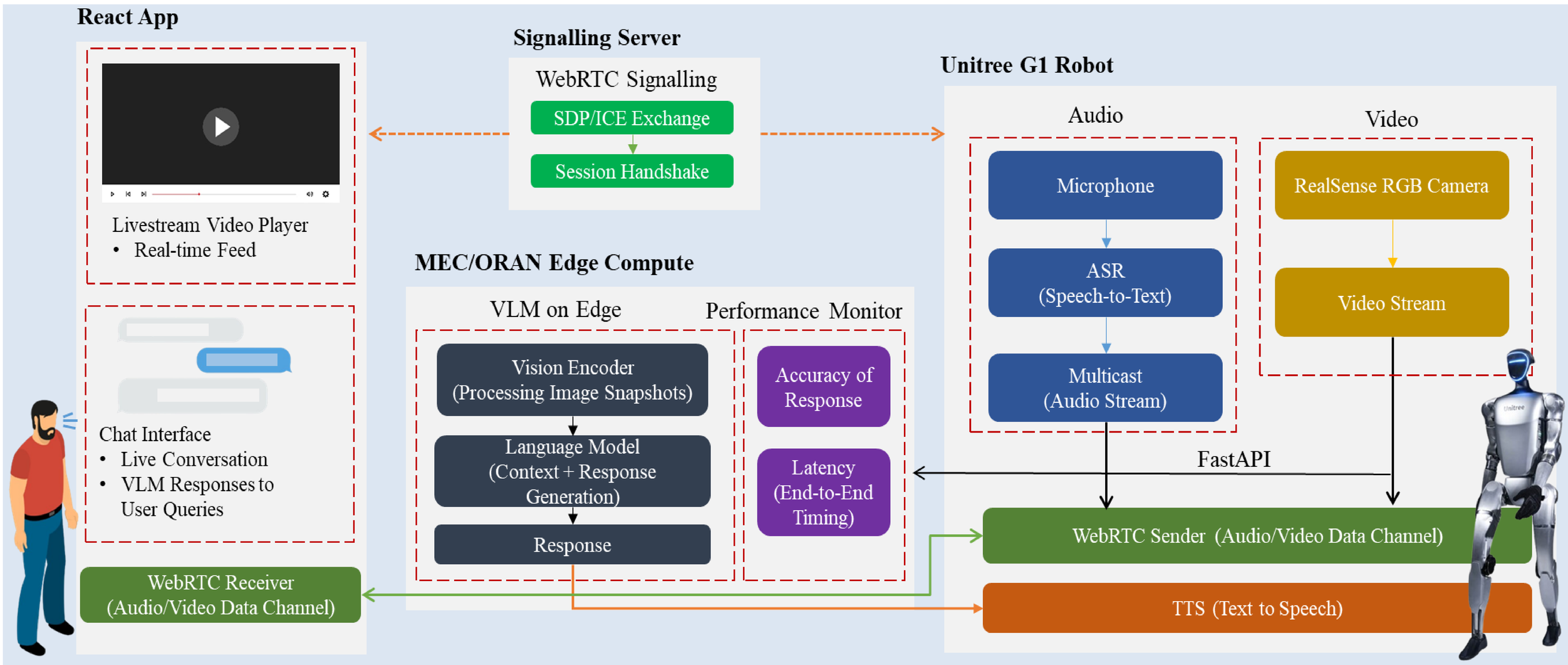
Sarat Ahmad, Maryam Hafeez, Syed A. Zaidi
School of Electronic and Electrical Engineering, University of Leeds, Leeds, UK

Motivation

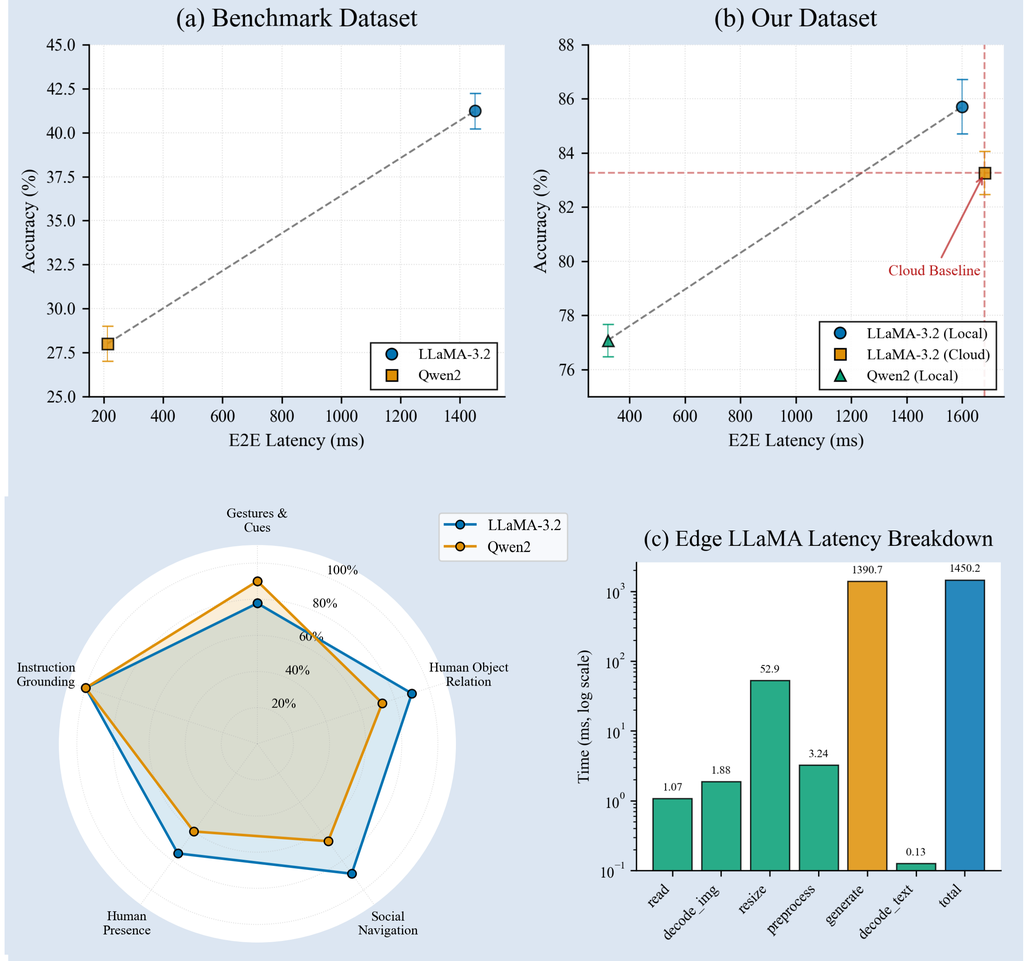
Problem:
VLMs can enable humanoid robots to understand and respond to the world using both vision and language. However, real-time deployment of such large models in robotics faces: high latency from cloud-based inference, limited onboard compute and memory on robots, privacy and reliability concerns when streaming sensitive data to remote servers.

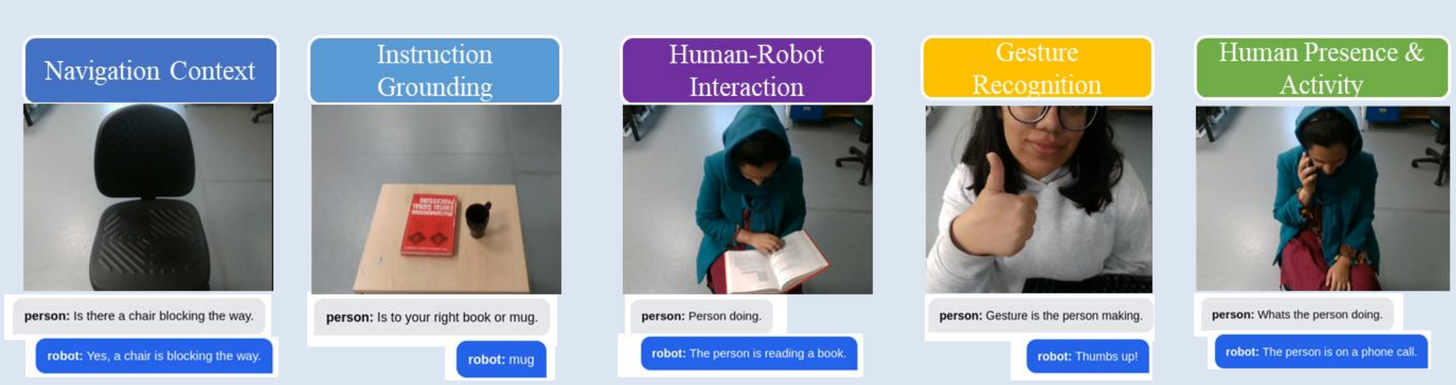
Our Solution:
- Edge Intelligence Framework using ORAN + MEC to bring VLM inference closer to the robot
- Real-time perception pipeline with Unitree G1 humanoid robot as testbed
- Edge deployment of LLaMA-3.2-11B and Qwen2-VL-2B
- Benchmark evaluation on Robo2VLM and robot-collected datasets comparing edge vs. cloud performance

Methodology



Results





- ❑ Edge LLaMA: near-cloud accuracy, 5% lower latency
- ❑ Qwen2: sub-second latency (↓50%)
- ❑ LLaMA shows balanced accuracy across all HRI categories; Qwen2 strong on perception tasks, weaker on reasoning tasks
- ❑ Large model - better reasoning; small model - faster response
- ❑ Autoregressive decoding is the main bottleneck (>85% of latency)
- ❑ Edge gives lower bandwidth, better privacy
- ❑ Achieved real-time human–robot interaction