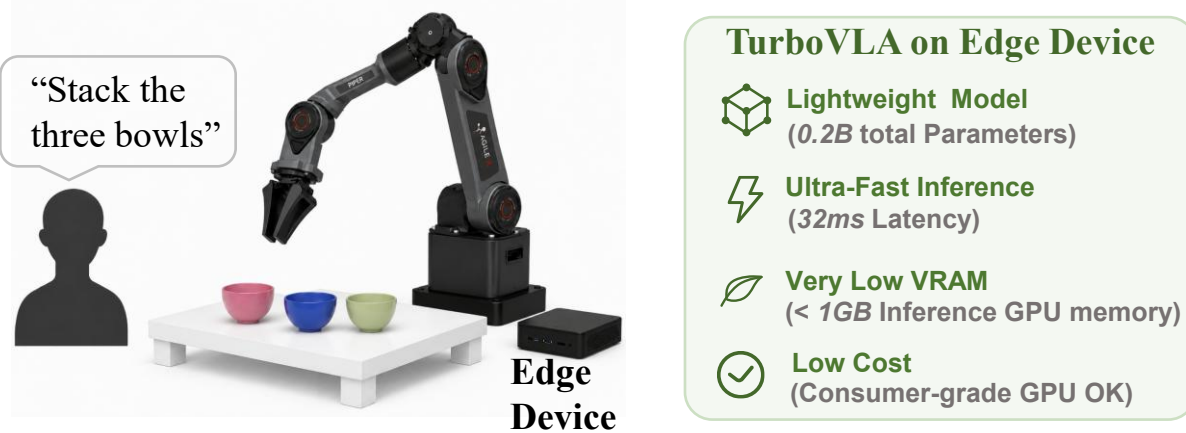


LLM-centric VLA Deployment



☹️ Relies on network, high latency, high cost, limited by environment

Ours TurboVLA Deployment



😄 Direct edge inference, low latency, low cost and more reliable

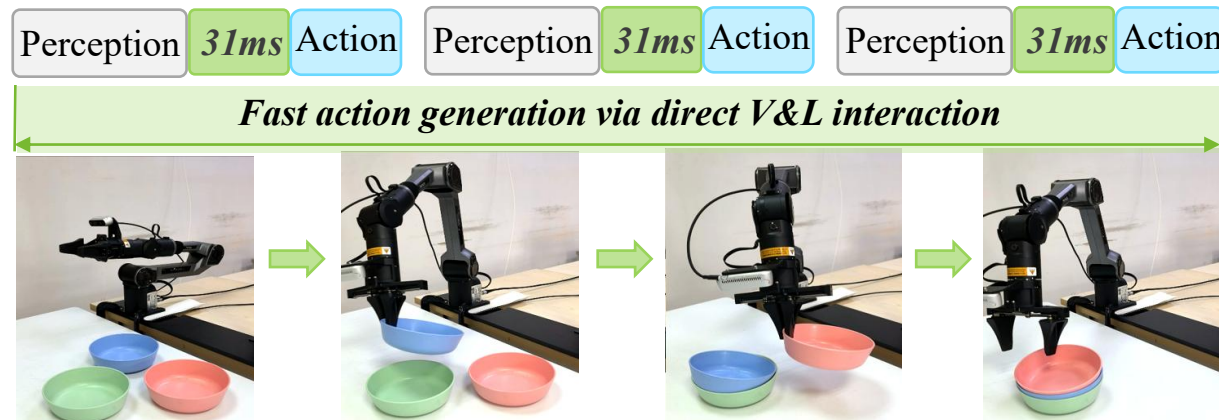
$\pi 0.5$ Execution



LLM-centric inference limits action updates to 11 Hz

☹️ Unfinished

TurboVLA Execution



TurboVLA enables action updates at 32 Hz on RTX 4090

😄 Finished