

Offload or Overload: A Platform Measurement Study of Mobile Robotic Manipulation Workloads

Sara Pohland¹, Xenofon Foukas², Ganesh Ananthanarayanan², Andrey Kolobov²,
Sanjeev Mehrotra², Bozidar Radunovic², Ankit Verma²

¹UC Berkeley, ²Microsoft Corp.

Abstract

Mobile robotic manipulation—the ability of robots to navigate spaces and interact with objects—is a core capability of physical AI. Foundation models have led to breakthroughs in their performance, but at a significant computational cost. We present the first measurement study of mobile robotic manipulation workloads across onboard, edge, and cloud GPU platforms. We find that the full workload stack is infeasible to run on smaller onboard GPUs, while larger onboard GPUs drain robot batteries several hours faster. Offloading alleviates these constraints but introduces its own challenges, as additional network latency degrades task accuracy, and the bandwidth requirement makes naive cloud offloading impractical. Finally, we quantify opportunities and pitfalls of sharing compute across robot fleets. We believe our measurement study will be crucial to designing inference systems for mobile robots.

1 Introduction

While AI-based systems, such as ChatGPT and Claude, have transformed digital workflows, they remain largely disconnected from the physical world, leaving much of AI’s potential untapped. We believe the next wave of AI will center on *physical AI*: embodied robotic assistants deployed as coordinated fleets in homes, enterprises, and industrial settings. Companies like Tesla, Google, and Figure AI are already pursuing general physical intelligence, signaling an industry-wide push.

A core capability underlying physical AI is *mobile robotic manipulation*—the ability of robots to navigate spaces and interact with objects. This capability is the building block for multi-step tasks that combine navigation (moving to a target) with object manipulation (moving objects). Mobile manipulation unlocks real-world applications across homes (e.g., clearing a countertop), enterprises (e.g., delivering mail), and industrial settings (e.g., moving tools). Recent advances in foundation models have driven breakthroughs in mobile robotic manipulation, enabling generalization across tasks, objects, environments, and robotic embodiments.

However, these advances come at a *high computational cost*. Robotics foundation models are orders of magnitude larger than the classical models and expert system policies traditionally used in robotics. Modern robots can be equipped with

powerful onboard CPUs and GPUs, but these compute platforms can lead to a significant increase in the robot’s cost and a decrease in its battery lifetime. While lightweight onboard GPUs are an option, they present a hard tradeoff by increasing the execution time of the foundation models, thus reducing the robot’s practical utility. Lighter models can reduce the execution time but at the expense of accuracy.

Objectives: We present the *first measurement study of workloads involved in mobile robotic manipulation* on onboard, edge, and cloud GPU platforms, with both wireless and wired connectivity, to address the following key questions:

- How severe are the tradeoffs between memory usage, task performance, and power consumption for onboard GPUs?
- If workloads were to be offloaded, what are the effects of the added network constraints (latency and bandwidth) on the execution time and accuracy of the workloads?
- When offloading, what are the implications of having multiple robots share the network and GPU compute?

We argue that answering these questions will enable the design of more effective mobile robotic manipulation systems.

Our measurement study considers different robot form factors, workloads for mobile robotic manipulation, GPU computes and wireless connectivity. The SO-101, TurtleBot 4, and Stretch 3 robots represent distinct points in the design space of robots in cost and capabilities; §3. We consider a wide variety of GPUs for onboard compute (Jetson Orin and Thor), edge compute (L4 and DGX Spark), and cloud compute (A100 GPU), as well as wireless connectivity for the robots; §3. Finally, we use state-of-the-art robotic workloads for perception, planning, navigation, and manipulation; §4. Our extensive setup gives us confidence that our results can broadly inform decisions in designing robotic inference systems.

1) Onboard GPU robotics: With onboard robot GPUs, we observe that smaller GPUs, such as the Jetson Orin with 32GB of memory, cannot even fit the full stack of models for mobile robotic manipulation. On GPUs with sufficient memory, mapping and planning can be slower by up to 383% compared to an A100, thus limiting the robot’s abilities in dynamic spaces. Navigation sees a 30% drop in its timely detection of obstacles with lighter GPUs. While the manipulation VLA models do not dramatically slow down with smaller GPUs, the slowdown is still sufficient to drop their accuracies by 50%. Finally, the

larger onboard GPUs such as Jetson Thor drain robot batteries considerably faster, by up to 160% (or a few hours), for even the larger robots. Our study in §5 concludes that the impact on accuracy, execution time, and battery lifetime of onboard GPUs offers a case for offloading model inference.

2) Offloading considerations: Offloading robotics AI inference, nonetheless, presents a non-trivial design space. We quantify that inference times on cloud-class GPUs are considerably smaller than on edge-class GPUs, but the network connectivity to reach the GPU for offloading presents a challenge. Even a few tens of milliseconds of additional latency can drop the accuracy of manipulation by over 10%. Further, offloading requires continuously transmitting image streams from the robot, which can saturate the network. Attempts to compress the streams can drop accuracy of manipulation and semantic mapping by nearly 20%. Thus, offloading is not a panacea and requires straddling a complex tradeoff of GPUs, connectivity, and video compression, as we explain in §6.

3) Multiplexing robotic offloads: Offloading inference of a *fleet* of robots to an edge compute (or cloud) presents both challenges and opportunities. We show that offloading compute to a shared platform provides opportunity to batch requests across multiple robots and to leverage statistical multiplexing across time-varying workloads. This means that we can use less compute to serve the same workload. However, the inference time and network latencies rise under contention from multiple robots, thus an offloading solution requires the use of admission control or careful resource management with performance guarantees for high-priority robotics tasks.

In summary, we make the following contributions: (i) We identify the emerging importance of mobile robotics manipulation and design a suite of representative workloads for benchmarking. (ii) We analyze the complex tradeoff in latency, accuracy, and power between equipping robots with onboard GPUs and offloading their inference to the edge or cloud. (iii) We present the opportunities and challenges in compute and communication with multiplexing inference of large robot fleets. We believe that our extensive study will help guide further work on inference systems for mobile robots.

2 Background & Motivation

In this section, we describe the recent work in robotics and compute infrastructure that motivated our studies.

2.1 Mobile robotic manipulation

Mobile manipulation is a fundamental capability in robotics that promises to significantly alter the future of robotics [13], with potential impacts across important areas of society, ranging from reindustrialization [44] to household assistance and elderly care [7]. Recent advances in foundation models are beginning to enable mobile manipulation to move beyond the

Task type	Focus	Examples
Perception	Sense & understand the environment	SLAM, object detection, semantic mapping
Planning	Interpret high-level goals & plan actions	Command interpretation, task sequencing
Navigation	Move safely & efficiently to a target location	Path following, obstacle avoidance
Manipulation	Interact physically with objects	Grasping, lifting, opening, placing

Table 1: Examples of typical tasks required to enable a mobile robotic manipulation application.

highly structured factory environments into less structured, open-world environments [13, 45]. Mobile robotic manipulation integrates four key competencies: perception, planning, navigation, and manipulation [56] (summarized in Table 1).

1) Perception—i.e., using onboard sensors to collect data and generate a representation of the robot’s surroundings—is tightly intertwined with the other competencies and often involves simultaneous localization and mapping (SLAM), whereby the robot tries to estimate its own location while constructing a map of the environment [51]. In contrast to geometric grid maps, robots increasingly integrate open-vocabulary object detection and segmentation [57] to obtain semantically rich representations for detailed reasoning [40].

2) After perceiving its environment, the robot generates a **plan** to accomplish its assigned task, which involves task decomposition. For example, to collect all dirty dishes, the robot must search the office for dish-like objects, then navigate to each dish and grasp it, and finally transport the collected dishes to the kitchen sink. In mobile manipulation, planning involves grounding natural language instructions into actions based on the robot’s environmental perception [12, 46].

3) The steps planned by the robot typically involve navigation and manipulation. In **navigation**, the goal is to move from one location to another while avoiding collisions with obstacles and other agents and possibly following a low-cost or socially compliant path [28]. Navigation requires both global planning on a map [38] and local planning to react to dynamic or unforeseen obstacles [50].

4) Finally, **manipulation** is the physical interaction with objects via, e.g., grasping, lifting, opening, or placing [27]. Successful manipulation requires implicit or explicit comprehension of geometry, pose uncertainty, contact dynamics, and task-specific constraints, all of which become substantially more challenging in unstructured human environments [24], which is the focus of recent work [5].

2.2 Advances in robotics capabilities

With the advent of *foundation models* containing billions of parameters, the ability of robots to engage in mobile manipulation has dramatically changed. Combining large language models (LLMs) (e.g., GPT [36]) with vision-language embedding models (e.g., CLIP [39]) has led to vision language models

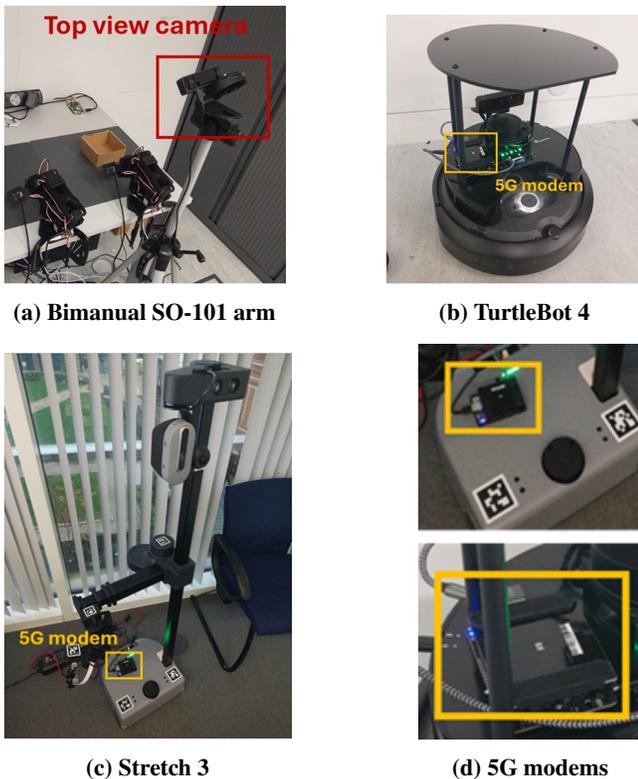


Figure 1: Robotic platforms used for evaluating workloads.

(VLMs) such as LLaVA [30] and GPT-4V [36]. VLMs provide the backbone for open-vocabulary object detection models, semantic segmentation models (e.g., SAM [25]), 3D world modeling systems (e.g., VLMaps [18]), vision language action (VLA) models for robot manipulation (e.g., $\pi_{0.5}$ [20]), and planning capability for embodied question answering (EQA) (e.g., GraphEQA [43]). Additionally, the emergence of video models has led to the development of World Action Models (WAMs) (e.g., DreamZero [55]). Alongside advancements in foundation models, improvements in SLAM have made perception and navigation more robust in diverse and dynamic environments, while reducing navigation times, avoiding expensive 3D lidars, and enabling faster reaction times for collision avoidance through GPU acceleration. The NVIDIA Isaac ROS toolkit offers a low-latency visual 3D SLAM package [26] and uses GPU acceleration with nvblox [31], thus dramatically improving real-time robot navigation. Together, these models have fundamentally transformed the capabilities of perception, planning, navigation, and manipulation.

3 Experimental Setup

Our setup can be broken down into the robotic platforms (§3.1), and the compute and network infrastructure (§3.2).

3.1 Robotic platforms

Our goal in selecting robots for our experiments was to span key points in the design space, including robotic functionalities (e.g., manipulation, locomotion), hardware capabilities (e.g., sensors) and cost. We focused on robots with open software stacks and full programmability, allowing us to build on community-developed models. We selected three robotic platforms: Stretch 3, TurtleBot 4, and SO-101 (Figure 1). Together, these platforms cover a range of costs, onboard compute, battery power, and functional abilities.

1) Stretch 3: Stretch 3 is an open-source mobile manipulator (Figure 1c), designed for operation in homes and workplaces. It is equipped with a 7-DoF gripper, two RGB-D cameras (Intel RealSense D435i mounted on the head and Intel RealSense D405 on the gripper), a differential-drive base, and a 2D LiDAR (SLAMTEC RPLIDAR A1). It comes with an Intel NUC 12 compute and two 12V batteries providing ~ 216 Wh combined. Stretch provides an open-source software stack, Stretch AI [41], with models for perception, planning, navigation, and manipulation. The cost of the robot is \$25K.

2) TurtleBot 4: TurtleBot 4 is an open-source mobile robotics platform built on the iRobot Create 3 base (Figure 1b). It includes a Luxonis OAK-D Pro RGB-D camera and a SLAMTEC RPLIDAR A1, both connected to a Raspberry Pi 4. In contrast to Stretch 3, it does not include a gripper, and is therefore suited for search and surveillance tasks but not manipulation. The platform includes a 26 Wh battery (14.4 V nominal). With widespread community support, it supports many navigation packages [42], and also allows open-source models and algorithms to be easily adapted. Its cost is \$2K.

3) Bimanual SO-101: The SO-101 is a robotic arm developed by RobotStudio and Hugging Face. It is suited for manipulation tasks, such as pick-and-place, and can be mounted on a mobile base, such as the LeKiwi platform [48]. Each arm offers 6 degrees of freedom. The arm includes a Raspberry Pi-based wrist camera and connects to external compute, such as Raspberry Pi 5, through USB-C. For our experiments, we used a static bimanual configuration consisting of two camera-equipped SO-101 arms (Figure 1a), along with an additional Luxonis OAK-D camera mounted on a tripod to provide a top-down view. The arms are integrated with the LeRobot SDK [11] for data collection, training, and evaluation using state-of-the-art VLA models. Each arm costs \sim \$300.

3.2 Compute & network infrastructure

The architecture of our testbed (Figure 2), consists of robot, edge, and cloud components, as well as their connectivity.

3.2.1 Robot onboard compute: Driven by recent trends of small form-factor and energy efficient options for robotics GPU acceleration, we leveraged (1) an Nvidia Jetson Orin AGX with 32GB of unified memory and (2) an Nvidia Jetson

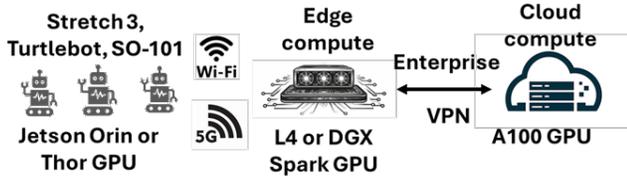


Figure 2: Experimental setup of robots, edge compute with Wi-Fi and 5G connectivity, and cloud compute.

Thor with 128GB of unified memory as the primary onboard compute options. We also experimented with (3) an Nvidia Jetson Orin Nano as a lightweight alternative, but, with only 8GB of unified memory, few of the models could fit. We extended the robots to include Nvidia Jetsons as additional onboard compute. To measure the power consumption of the onboard compute boards, we used a smart plug (Shelly Plug Gen 4), that allows us to export power measurements in real time over the network using a network API [1].

3.2.2 Edge compute: To offload to the edge, we considered: (1) an HPE server with an Intel Xeon 6338N CPU, an Nvidia L4 GPU with 24GB of memory, and an Intel E810 NIC at 25 Gbps; (2) an Nvidia DGX Spark workstation with 128GB of *unified* memory and ConnectX-7 NIC at 200 Gbps.

The aforementioned options for compute onboard and at the edge capture different representative points in the design space of cost, power, and performance.

3.2.3 Cloud compute: To offload to the cloud, we used an Azure VM with an AMD EPYC 7V13 CPU, an A100 GPU with 80GB of memory, and a Mellanox NIC at 100 Gbps.

3.2.4 Connectivity: The robots have wireless connectivity using both Wi-Fi and 5G. We deployed a Netgear Business WAX602 Wi-Fi 6 access point, and we equipped the robots with Wi-Fi 6 AX900 USB adapters. For 5G, we deployed a private 5G network using srsRAN [49], open5GS, and a Foxconn RPQN 7800 radio unit, and equipped the robots with 5G modems (Figure 1d). The edge compute was connected to the cloud VM via a FortiGate firewall and VPN.

4 Workloads Under Study and Methodology

We construct a suite of evaluation benchmarks consisting of key building blocks of mobile robotic manipulation (Table 2). (i) semantic understanding of the environment and long-term planning (§4.1), (ii) safe (i.e., collision-free) navigation (§4.2), and (iii) multi-step manipulation involving sub-tasks (§4.3). For the canonical task of “finding all the dirty dishes in the office and bringing them to the kitchen sink”, the robot needs to understand the environment it is in and make a plan (operation (i)), safely navigate to key locations (operation (ii)), and then manipulate the objects (operation (iii)).

4.1 Semantic mapping and planning

Semantic mapping provides understanding of the robot’s environment, which in turn is used by a planner to devise the steps to execute mobile robotic manipulation tasks.

4.1.1 VLMaps: As described in §2.2, foundation models have enabled rich 3D scene modeling. VLMaps [18] is one such popular approach that allows for continuously updating an open-vocabulary semantic map. Like most modern semantic mappers, it builds a 3D point cloud, where each point is a vector in a joint image-text embedding space. We improved the VLMaps implementation to use the SAM segmentation model [25], CLIP [39] to generate image embeddings, and Depth Anything [54] to improve depth estimates. These modifications together substantially improved accuracy of the 3D map when we deployed it on the TurtleBot robot (§3.1). The map generated by VLMaps is semantically rich and is used in conjunction with a language model for open-vocabulary navigation tasks. VLMaps requires high bandwidth image streams and high compute to generate its map. Map updates are expected with medium latency (seconds to minutes).

4.1.2 GraphEQA: Building upon 3D mapping, GraphEQA [43] combines mapping *as well as* planning. While similar in spirit to VLMaps, GraphEQA’s mapping has a key differentiator: it uses a 3D semantic scene graph, in which nodes correspond to spatial elements such as rooms and objects, and edges capture their geometric and semantic relationships. The scene graph allows GraphEQA to answer questions, such as “is the microwave in the kitchen left open?”. To generate its answer, GraphEQA uses a multimodal LLM planner (such as Qwen-VL-2.5-3B [4]) that interprets natural language queries and plans the steps needed to answer the above question, such as navigating to the kitchen, identifying the microwave, and understanding the state of its door. GraphEQA is supported by the Stretch AI framework of the Stretch robot (§3.1). It requires low bandwidth because it only processes videos occasionally, but requires high compute when a question is asked. It tolerates medium response latency (seconds).

4.2 Collision-free navigation

Navigation in mobile robotic manipulation applications allows robots to move to the locations of objects of interest.

4.2.1 RTAB-Map & nvblox: We evaluate SOTA SLAM methods, which consist of two parts: localization and mapping (§2.1). For localization, we employ RTAB-Map [29]—a widely used SLAM method that fuses RGB-D, Stereo, and LiDAR sensor inputs. Although RTAB-Map can also generate a map of the environment, map updates can take many seconds, which is not sufficient in highly dynamic environments or even some static environments, depending on the operating speed of the robot. In our task, the TurtleBot has a partially

Class	Workload	Task description	Operations involved	Embodiment	BW	Lat	Cmp
Semantic understanding & long-term planning	VLMaps [18]	The robot updates a semantic map with new information as it moves from point A to point B of the lab. The robot is asked to check if the door of the lab has been left open.	Perception & navigation	TurtleBot	H	M	H
	GraphEQA [43]		Perception, planning, & navigation	Stretch 3	L	M	H
Multi-step mobile manipulation	$\pi_{0.5}$ [20]	The robot is tasked with handing over a tape from one arm to the other and to place it in a box within a time window of 70s.	Perception, planning, & manipulation	SO-101	H	L	H
	DreamZero [55]	The robot is tasked with grabbing a frying pan from the hob.	Perception, planning, & manipulation	Simulation (video rec.)	H	L	H
Collision-free navigation	RTAB-Map [29] & nvblox [31]	The robot is tasked to navigate to some location, which requires it to turn around a corner and avoid a previously unmapped obstacle.	Perception & navigation	TurtleBot	H	L	M

Table 2: Robotic workloads considered for the experimental evaluation characterized by bandwidth (BW), latency (Lat), and compute (Cmp) requirements as low (L), medium (M), or high (H).

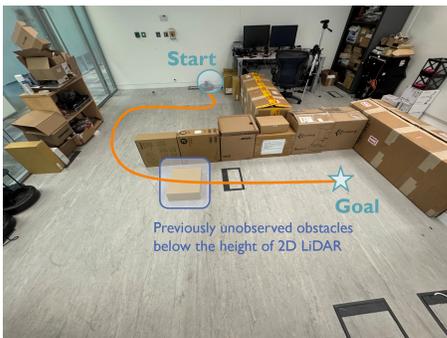


Figure 3: Trajectory of TurtleBot for navigation task and unmapped hidden obstacle.

known map and is tasked with turning around a corner with an unmapped obstacle (Figure 3), which is not visible to the robot until after it has turned, so it must quickly detect the obstacle and update the map. RTAB-Map does not update its map quickly enough to avoid collision, so we use a GPU-based occupancy mapping method, nvblox [31], which can update the map an order of magnitude faster.

4.3 Mobile manipulation

Finally, manipulation models allow robots to interact with objects by picking and placing them.

4.3.1 $\pi_{0.5}$: Foundation models have been used to design VLAs to enable manipulation skills generalizing across tasks and robot embodiments (§2.2). The most widely used of these models are π_0 [19] and its successor, $\pi_{0.5}$ [20]. These VLAs combine a VLM backbone, PaliGemma [8], with a smaller action expert diffusion model that processes robot states and actions. This architecture allows a single model to generalize across robots with differing action spaces to perform a

wide range of tasks. Using the LeRobot framework [11], we fine-tune the $\pi_{0.5}$ model on the bimanual SO-101 robotic arm (§3.1; Figure 1c) to *hand over a tape from one arm to the other, and then place it in a box*. Such coordinated manipulation is representative of real-world scenarios in manufacturing and enterprises, for moving tools, hazardous objects, or cups and cans. We deploy the fine-tuned $\pi_{0.5}$ model using action chunks of 50 actions and an execution horizon of 20 actions. We also use the state-of-the-art real-time chunking (RTC) algorithm for tolerating high inference and network latencies [9], but we still expect an end-to-end inference latency of 100–200 ms.

4.3.2 DreamZero: Another emerging class of robot foundation models, called World Action Models (WAMs), tightly couple action generation with world modeling. DreamZero [55] is a representative WAM that builds on a large pretrained video diffusion backbone to jointly predict future visual observations and robot actions conditioned on language and robot state. By aligning actions with predicted video futures, DreamZero leverages spatiotemporal priors from web-scale videos, enabling generalization to unseen tasks, spaces, and robots without explicit planning. Due the lack of support of DreamZero for the onboard GPUs that we studied, we use it in simulation, using video feeds of a robot tasked with grabbing a frying pan from the hob. Our evaluation of this workload focused on its compute and performance requirements.

5 Analysis of Onboard Compute Options

We leverage our setup (§3) and workloads (§4) to investigate the resource demands of robotic applications. We run each workload on three onboard GPUs: Jetson Orin AGX, Thor, and Nano. We report *i*) memory consumption (§5.1), *ii*) end-to-end task execution time and/or accuracy (§5.2), and *iii*) power

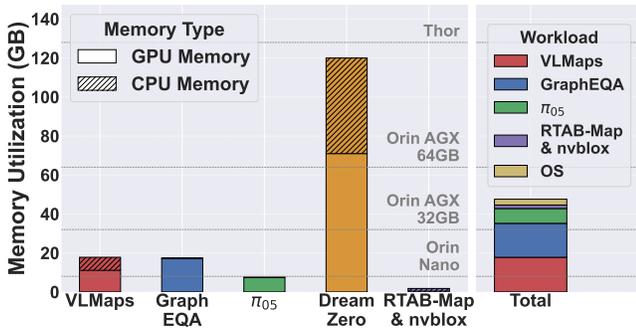


Figure 4: GPU and CPU memory consumption of workloads.

consumption (§5.3). We find that the demanding requirements of robotics workloads make onboard GPUs inadequate.

5.1 Memory usage

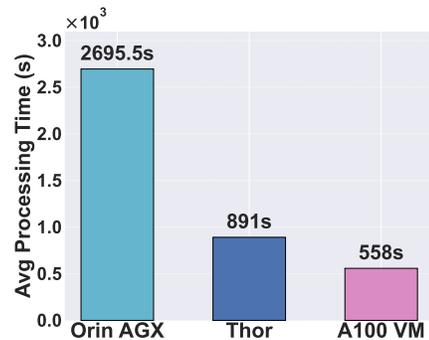
Given the unified memory architecture of platforms like the Orin and the Thor, the CPU and GPU tasks have to contend for the same shared memory. We therefore measure both the CPU and GPU memory for all workloads using the monitoring tools provided by Nvidia (jetson-stats [33] for the Jetson Orin AGX, and nvidia-smi [35] for the Jetson Thor), as well as the Nsight Systems tool [34], for GPU and CPU behavior.

As depicted in Figure 4, we observe that all the workloads under study (§4) require several gigabytes of memory, ranging from approximately 2GB for the RTAB-Map & nvblox workload, to more than 120GB for DreamZero.

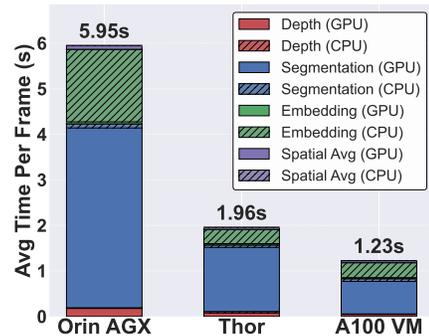
As we can see in Figure 4, even without taking DreamZero into account, the full stack of workloads, including semantic mapping and planning, mobile manipulation, and collision-free navigation, plus the OS, requires a minimum of approximately 50GB of memory. This makes it challenging to simultaneously deploy several workloads for the robot on memory-constrained platforms, like the Nano or the 32GB Orin. For reference, Figure 4 includes the memory capacities of the GPUs typical for onboard robotics, showing that the demands are stringent *even for one workload*. While one could swap models into memory on demand, the overheads of many seconds will be prohibitive for real-time mobile manipulation.

5.2 Task performance

We next evaluate the accuracy and execution time of the workloads of Table 2 on the onboard GPU options of Jetson Thor, Orin AGX, and Nano (where applicable). For *comparative reference*, we use the accuracy and execution time achieved when offloading the workloads to the A100 GPU VM, with an ideal maximum round-trip network latency of less than 15ms. We repeat each task described in Table 2 twenty times to report our results (Figures 5–8).



(a) Average processing time for 450 frames.



(b) CPU and GPU time for processing a single frame.

Figure 5: VLMaps execution time across compute platforms.

5.2.1 VLMaps: Recall from §4 that VLMaps is used to generate a rich 3D map of a robot’s environment. We measure the time for map generation with 450 frames (30s of video at 15 FPS). Note that the lower the generation time, the more current the robot’s understanding of its environment, which is vital in many dynamic situations. As Figure 5a shows, compared to our reference execution time on an A100, the execution time jumps by 59% with the Thor and 383% with the Jetson Orin AGX—a slowdown that can degrade the utility of the semantic map and be detrimental to the robot’s performance in frequently changing environments. Upon deeper inspection of the time and task breakdown bars in Figure 5b, the frame processing is primarily a GPU-bound operation, with the majority of the time spent on the GPU task of segmentation using the SAM model (see §4.1). Interestingly, the second biggest contributor to the total frame processing time is the CPU part of the embedding stage, which involves large data copying. Nonetheless, this CPU task performs similarly on the Thor and the A100 VM, making the GPU task of segmentation the main contribution to the spike in the overall processing time.

5.2.2 GraphEQA: Recall from §4.1 that GraphEQA uses its 3D map to construct and execute a workflow of steps for the robot, in order to answer questions about the environment.

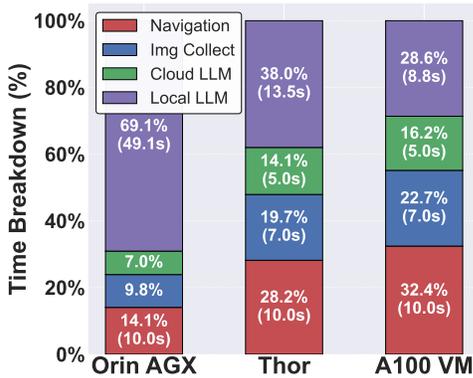
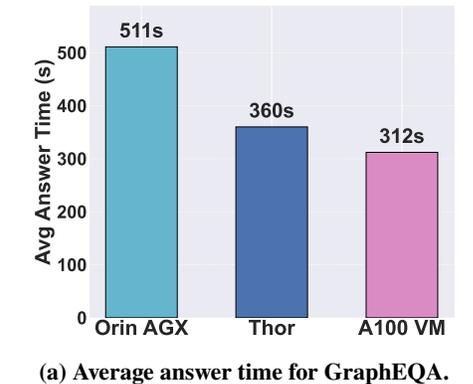


Figure 6: GraphEQA execution time across compute platforms.

As seen in Figure 6a, the end-to-end execution time of the task varies significantly across GPU platforms, with the A100 VM requiring on average 312s, the Thor 360s, and the Orin AGX 511s. For spatial exploration, interactivity and small user waiting times are important for the utility of the application, and the inflation in Figure 6a can be sizable. As shown in the breakdown in Figure 6b, the main contributor to this time increase is the GPU inference time for the local VLM, which is used to identify featured object names from the collected image observations. With VLM model sizes predicted to increase, the GPU compute will be the expected bottleneck.

5.2.3 RTAB-Map & nvblox: Given the small memory footprint of the nvblox component, we evaluate its performance on the more lightweight Jetson Nano, in addition to the Thor and Orin AGX. As observed in Figure 7a, the task success rate is 100% on all platforms except the Nano, for which it drops to 70%. This indicates that, despite the lightweight nature of the navigation workload, higher-end onboard compute might still be required to ensure high reliability, considering that collision-free navigation is a safety-critical operation. Inspection of the failed runs reveals that the accuracy reduction is due to delays in detecting and updating unmapped occupancies

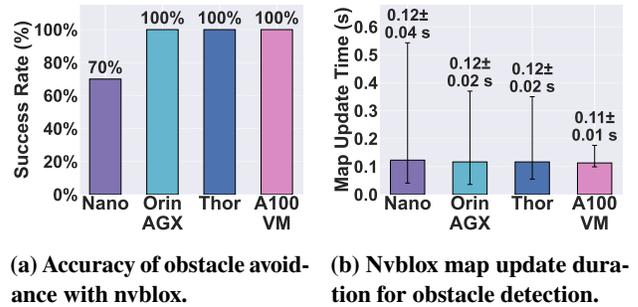


Figure 7: Nvblox success rates across compute platforms.

(§4.2), shown in Figure 7b. While the average latency is approximately the same, the Nano sees significantly higher standard deviation, and even higher max latencies that can reach up to half a second. One could mitigate this by slowing the robot, but in many scenarios (e.g., factory floors, warehouses), reduced speed directly translates to reduced productivity.

5.2.4 $\pi_{0.5}$: For the handover task (§4.3), Figure 8a shows the average end-to-end task execution time of successful attempts. We observe that it is the same on the A100 and the Thor (~34s), and increases by ~23% on the Orin AGX (~42s). Crucially, the task accuracy (Figure 8b) on the A100 VM and the Thor is 80%, while it drops to 30% on the Orin. This drop is a direct consequence of the latency for generation of action chunks: approximately 60ms for the A100, 144ms for the Thor and 440ms for the Orin AGX (see Figure 11 in Section 6.1). High latencies lead to the depletion of the action queue before a new action chunk is generated, leading to stop-and-go behavior and movement jerkiness that affects the accuracy of the task [10, 52]. We visually observed such jerkiness for the Orin AGX. Also, we note that the latency affected the accuracy of the sub-tasks (i.e., pick up tape, handover, put in the box) differently. As shown in Figure 8c, failures manifested in the sub-tasks of picking up the tape and handing it over to the other arm. From our visual observations, this is due to the higher precision required by those sub-tasks, where mistakes of even a few centimeters can lead to dropping or failing to grab the tape.

5.2.5 DreamZero: We evaluated DreamZero on simulations and focus on its execution time. We only consider the Thor and the A100, since the model does not fit on the Orin AGX. The average inference time of DreamZero is 6.22s on the A100 and 19.27s on the Thor (see Figure 11 in §6.1). Such latencies would also reduce task accuracy, as DreamZero requires sub-second inference, currently achievable only with high-end multi-GPU setups (2×GB200) [55], further underscoring the trend toward compute that exceeds onboard capabilities.

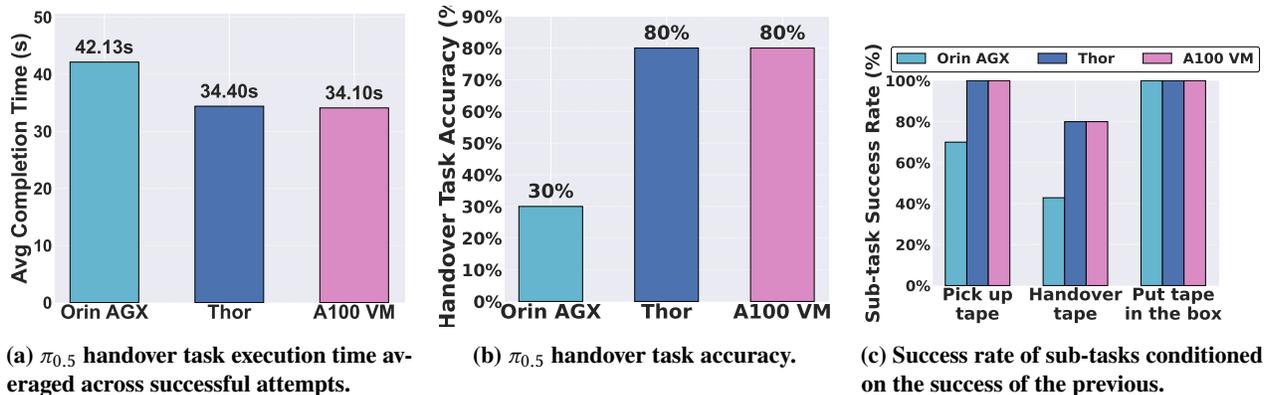


Figure 8: $\pi_{0.5}$ execution times and success rates across compute platforms.

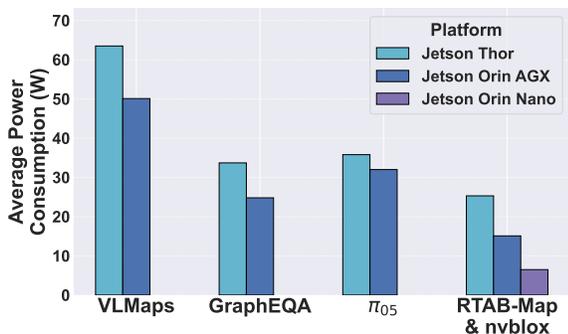


Figure 9: Power consumption of onboard compute when running the workloads of Table 2.

Takeaway #1: Modern robotics workloads require GPUs beyond what is typically fitted on board robots today, leading to expensive on-robot upgrades or offloading.

5.3 Power consumption

Another important dimension is the onboard compute power consumption and its effect on the battery lifetime of the robot. To this end, we measured the power consumption of our onboard compute platforms (Thor, Orin AGX, Nano), while running the various workloads of Table 2, excluding DreamZero, which can only be hosted in server-grade GPUs. The results are summarized in Figure 9 and capture the average power consumption during the execution of each workload individually, including both busy and idle periods. In the case of the Nano, we only report the power consumption for the RTAB-Map & nvblox workload, as it is the only workload that can run on this platform. From Figure 9 we observe that the VLMaps workload is the most demanding, followed by $\pi_{0.5}$, GraphEQA, and RTAB-Map & nvblox. It is noteworthy that the Thor consumes only marginally more power compared to the Orin AGX, while exhibiting significantly better compute capabilities (§5.2).

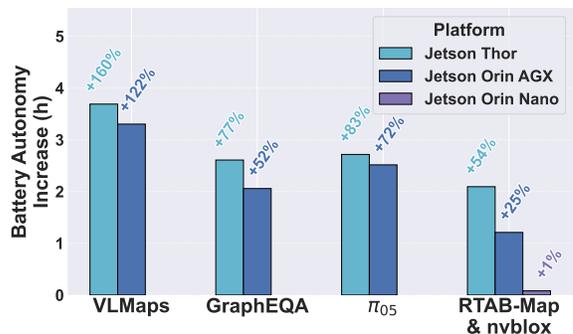


Figure 10: Stretch 3 battery lifetime increase from offloading, by replacing high-end onboard compute with a Raspberry Pi 5.

To put the power numbers into perspective, we estimate the increase in battery lifetime for the Stretch 3 robot if all computations were offloaded to a remote location and the onboard GPUs were replaced with a Raspberry Pi 5 used only for data transmission (~6W). As shown in Figure 10, given the robot’s battery capacity of 216Wh and the 30W baseline power consumption of the robot’s sensors and motors, such an offloading strategy could yield substantial battery savings, increasing the robot’s operational lifetime by several hours.

Takeaway #2: Onboard GPUs significantly increase robot power consumption. Offloading compute can therefore substantially extend a robot’s battery lifetime.

6 Analysis of Offloading Options

Following the observations of §5, we examine the choices of offloading to a GPU server on the edge and in the cloud. Two factors determine this choice: (i) end-to-end task latency (§6.1), and (ii) bandwidth to transmit data to the server (§6.2).

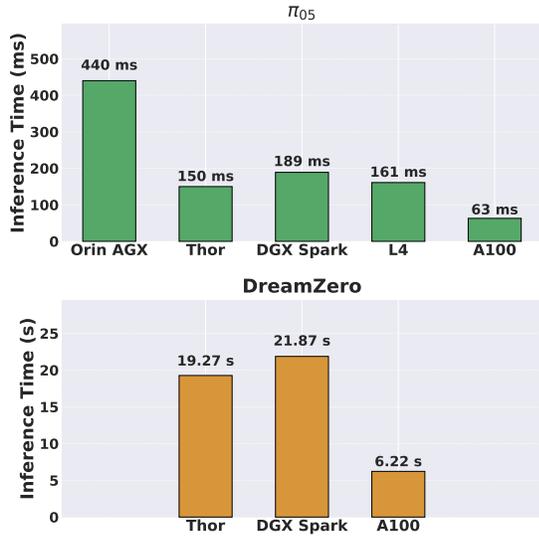


Figure 11: Inference time for $\pi_{0.5}$ and DreamZero across different compute platforms.



Figure 12: Example 5G cell configurations with tradeoffs in terms of latency and throughput, as measured in our testbed.

6.1 Latency considerations

Latency components: The end-to-end task latency when offloading depends on three factors: the compute latency, latency of the wireless link to reach the base station or edge compute, and the wide-area latency (if the server is in the cloud). All three latency components can vary depending on the infrastructure characteristics. As shown in Figure 11, the compute latency of $\pi_{0.5}$ and DreamZero (§4.3) vary with the compute platform: edge platforms, such as the DGX Spark and the L4 server, introduce higher inference latencies compared to server-grade GPUs, like the A100. Wired latency is affected by the choice of the offloading location and link type [22]. Similarly, wireless link latencies can be affected by several parameters, such as the wireless technology (e.g., Wi-Fi or 5G), its configurations, and the number of robots sending data. For example, Wi-Fi latency can be from a few ms in an idle network to tens of ms in a busy one (Figure 17). 5G latency additionally depends on the cell configuration (Figure 12).

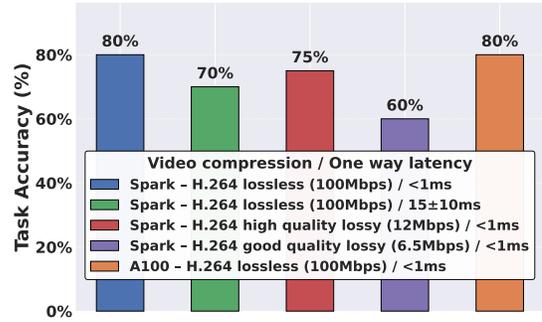


Figure 13: $\pi_{0.5}$ accuracy with different latency and bandwidth.

Config.	15 FPS	10 FPS	5 FPS	1 FPS
640x480	20.3 / -	13.7 / 91.5%	7.2 / 89.8%	1.8 / 71.2%
320x240	5.4 / 84.7%	3.8 / 84.7%	2.1 / 79.7%	0.7 / 69.5%
160x120	1.7 / 81.4%	1.2 / 79.7%	0.8 / 78.0%	0.5 / 72.9%

Table 3: Average bandwidth (Mbps) / recall with VLMaps across image resolutions and frame rates.

Latency vs. Accuracy: We measure the impact of increased network latency on the accuracy of robotics workloads. As shown in Figure 13, while the accuracy of the $\pi_{0.5}$ workload is 80% on the DGX Spark when the network connectivity is ideal, it drops to 70% once we introduce additional network latency (normal distribution with 10ms mean and 15ms standard deviation one-way [22]). The loss of accuracy with added network latency can be compensated by higher-end GPUs. For example, by replacing the DGX Spark with the A100 VM, accuracy goes back to 80%, albeit at increased cost. We have observed similar trends for RTAB-Map & nvblox for the same hardware and latency setups, where the accuracy drops to 80%, compared to the 100% under ideal network conditions.

6.2 Bandwidth considerations

When offloading workloads, the robot must continuously transmit sensor data over the network. The feasibility of such offloading therefore depends not only on latency, but also on sustained network bandwidth. We explore the impact of bandwidth using $\pi_{0.5}$ and VLMaps as representative examples.

In the case of $\pi_{0.5}$, we stream lossless video from three cameras mounted on the bimanual SO-101 setup with a resolution of 640×480 at 30 FPS. This requires an average uplink bandwidth of ~ 100 Mbps and results in a task accuracy of 80%, as shown in Figure 13. In multi-robot settings, bandwidth demands scale with the number of robots, which can quickly saturate the network capacity, even with tens of robots.

Bandwidth vs. Accuracy: To alleviate the stress on the network bandwidth, we experiment with lossy H.264 video compression with two different quality configurations, “high” and

Model	C1. Batch Runtime (s)			C2. Speedup vs. Seq.			C3. Latency vs. Thor B=1			C4. GPU Mem (GB)			C5. Mem. Savings		
	B=1	B=2	B=4	B=1	B=2	B=4	B=1	B=2	B=4	B=1	B=2	B=4	B=1	B=2	B=4
$\pi_{0.5}$ [20]	0.0633	0.0956	0.158	1.00	1.32	1.60	+56.1%	+33.7%	-9.5%	7.90	8.25	8.70	0%	47.8%	45.6%
SAM [25]	0.366	0.725	1.448	1.00	1.01	1.01	+12.9%	-45.5%	-79.3%	2.52	2.53	2.56	0%	49.3%	71.5%
CLIP [39]	0.0758	0.1398	0.263	1.00	1.08	1.16	-10.6%	-8.7%	-31.7%	3.90	3.96	4.08	0%	49.7%	47.9%
Depth															
Anything [54]	0.0722	0.133	0.273	1.00	1.08	1.06	+4.8%	-14.5%	-27.2%	1.28	1.29	1.29	0%	49.7%	74.8%
Qwen2.5 [4]	1.217	1.262	1.368	1.00	1.93	3.55	+27.2%	+24.4%	+18.1%	7.52	7.77	8.25	0%	48.4%	72.6%

Table 4: Batching of robotic models on A100 GPU. (C1.) Batch Runtime is the time to process all B requests as a batch. (C2.) Speedup relative to running B requests one after another. (C3.) Latency vs. Thor is the latency relative to a single Thor request (+ = faster, - = slower). (C4.) Memory utilizations. (C5.) Mem. Savings vs. B Inst. is the memory reduction compared to loading B separate instances.

“good”, which result in bandwidth requirements of 12 and 6.5 Mbps, respectively. We ensure that no packets are lost, which could lead to corrupt frames. As we can observe in Figure 13, the bandwidth reduction comes at the cost of a drop in the accuracy of the handover and pick-and-place task.

Similar observations can be made about the performance and bandwidth tradeoffs of VLMaps. Here, we reduce the bandwidth for streaming the robot’s RGB images by reducing the frame rate and resolution of the video (Table 3). We treat the map generated by VLMaps with the 640×480 , 15 FPS configuration as a reference, which identifies 59 unique objects by matching map embeddings against a large set of indoor text labels. For the alternative resolutions and frame rates, we compute recall as the percentage of these 59 objects that are also identified (Table 3). Reducing frame rate has a considerable effect: at 640×480 , recall drops to 91.5% at 10 FPS and 71.2% at 1 FPS. Resolution reductions similarly degrade recall. At 15 FPS, lowering resolution from 640×480 to 320×240 yields 84.7% recall, and further reducing to 160×120 yields 81.4%.

Takeaway #3: *Obtaining the potential benefits of offloading robotics workloads requires carefully considering the complex tradeoff involving workload performance, network latency, bandwidth, and available GPU resources.*

7 Multiplexing Robotic Workloads

Offloading unlocks opportunities for multiplexing robotic workloads for reducing the deployment cost. In this section, we explore the opportunities and challenges to such sharing.

7.1 Sharing opportunities

7.1.1 Batching: To understand the extent to which sharing of edge/cloud compute resources is possible, we first consider several robots simultaneously performing the same type of task and explore the extent to which robotic workloads can benefit from batching requests to the corresponding models. We focus on VLMaps, GraphEQA, and $\pi_{0.5}$, given that batching does not apply to nvblox. In the case of VLMaps, we decompose

the process to its core underlying models (i.e., CLIP, SAM and Depth Anything v2). GraphEQA utilizes Qwen-VL-2.5-3B, which we study using generation responses of 32 tokens. Table 4 shows the results for the A100 GPU for batches ($B = 1, 2, 4$). As we can observe (column C5), batching brings significant memory savings for all models, ranging from 45.6% for $\pi_{0.5}$ to 74.8% for DepthAnything (when $B=4$).

Inference latency improves significantly compared to sequential execution of the inference requests (column C2), with $\pi_{0.5}$ and Qwen models speeding up by 1.6 and 3.55, respectively. As we see in column C3, by serving all four requests in parallel on the more powerful A100 GPU, inference latency increases only marginally for $\pi_{0.5}$ while decreasing significantly for Qwen, relative to a single request on the onboard Thor GPU. This means the A100 can serve 4 robots simultaneously with latency comparable to *each* robot having its own dedicated Thor. In the case of SAM, CLIP, and Depth Anything, the inference latency increases roughly linearly with the batch size. The vision models (SAM, CLIP, and Depth Anything) do not benefit from sharing the prefill phase as in the case of LLM-based workloads ($\pi_{0.5}$ and Qwen) [21, 37].

Takeaway #4: *Robotic workloads involving VLAs and VLMs present significant opportunities for sharing GPUs through real-time batching, more so than pure vision models that are typically used in perception tasks.*

7.1.2 Statistical multiplexing opportunities: We use the Nvidia nsys profiling tool to measure the GPU Streaming Multiprocessors (SM) utilization for the various workloads under study. As illustrated in Figures 14 and 15 for the DGX Spark and the A100 VM, all the workloads present patterns of periodic activity, followed by idle GPU periods, exposing ample opportunities for the multiplexing of workloads. The periodicity of active and idle periods depends on the task, with the idle gaps being larger in the case of the A100 GPU. In the case of the $\pi_{0.5}$ model and GraphEQA, the active GPU periods correspond to the generation of actions for the robot to execute (move the arms for $\pi_{0.5}$ and explore a new area for

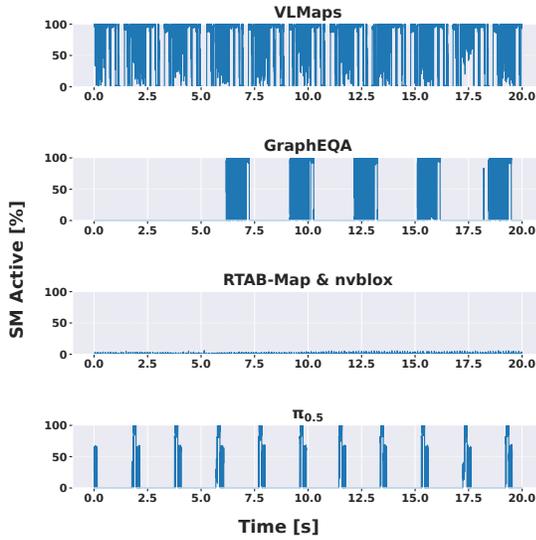


Figure 14: SM utilization for studied workloads on DGX Spark.

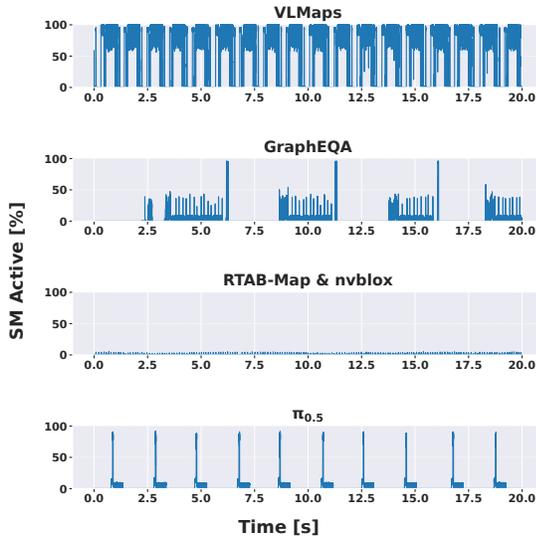


Figure 15: SM utilization for studied workloads on A100 VM.

GraphEQA), while the idle periods correspond to the time of executing those actions or sitting idle. On the other hand, VLMaps and RTAB-Map & nvblox are constantly processing frames, with idle GPU periods corresponding to CPU activity (e.g., to transfer frames to and from the GPU).

All the workloads exhibit low absolute utilization or reach their peaks only briefly (e.g., the peaks of GraphEQA and $\pi_{0.5}$ only last for a few tens of milliseconds). This means that significant sharing opportunities exist even during active periods. It should be noted that while VLMaps is more computationally demanding in the average case, its best-effort nature allows

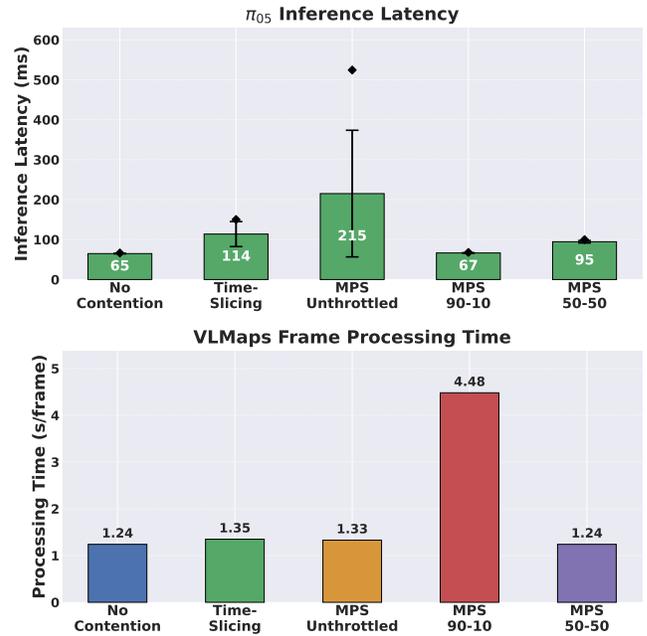


Figure 16: Impact of $\pi_{0.5}$ and VLMaps sharing the compute.

the delay of its execution, to prioritize other more latency-sensitive workloads, such as $\pi_{0.5}$. Furthermore, the semantic map indexing of VLMaps could also be paused while the robot is not moving around, since no new information is added from steady frames, freeing up even more compute resources.

Takeaway #5: *Robotic workloads have alternating idle and busy periods that present opportunities for statistical multiplexing, which become more ample with higher-grade GPUs.*

7.2 Sharing challenges

7.2.1 Compute sharing: Here we focus on the performance implications of statistical multiplexing. We present one concrete example of sharing the A100 VM between $\pi_{0.5}$ and VLMaps, but note that our findings extend to the other workloads under study. For this experiment, we separate the two workloads to run on different CPU sets, using taskset, and consider four different widely-used policies for sharing the GPU: i) default GPU scheduling with time-slicing, ii) MPS scheduling, where both workloads are not throttled by the GPU scheduler, iii) MPS scheduling, where the latency critical $\pi_{0.5}$ workload is given 90% of the GPU SMs, leaving 10% for VLMaps, and iv) MPS scheduling with 50-50 assignment of SMs between the two workloads. In all four cases, we measure the inference latency of $\pi_{0.5}$, as well as the time required to process a single frame with VLMaps.

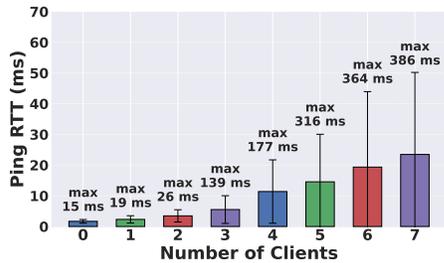


Figure 17: Ping RTT of idle client with varying additional connected clients, each generating 50Mbps of UL TCP traffic.

In Figure 16, we observe that sharing the GPU using time slicing (i) and unthrottled MPS (ii) results in an inference latency increase of 75% and 230%, respectively. This is because, as shown in Figures 14 and 15, the VLMaps workload is more computationally intensive, resulting in long scheduling delays for $\pi_{0.5}$. By introducing throttles, and guaranteeing higher SM availability for $\pi_{0.5}$, its inference latency reduces. In the optimal case, when restricting VLMaps to 10% of the SMs (iii), the latency increases by just 3%. However, this static throttling results in a significant increase of 261% in the frame processing time of VLMaps, since the SMs allocated to $\pi_{0.5}$ can no longer be reallocated to VLMaps even during idle periods.

7.2.2 Wireless spectrum sharing: As with the scheduling of compute, high latency can be a critical issue when sharing radio resources for offloading the workloads of several robots. To demonstrate, we capture a trace of the traffic that our Stretch 3 robot sends to the offloading server for the GraphEQA model (~50 Mbps uplink traffic), and we replay this traffic over our Wi-Fi 6 network for a number of robot clients, while measuring the ping RTT of another (idle) client.

As Figure 17 shows, the RTT of the idle client significantly increases with the number of active clients, and can reach close to 25ms in the mean and up to 50ms in the standard deviation. We note that this latency increase was observed, despite the several optimizations in our Wi-Fi network to ensure low latency, including using OFDMA and centralized scheduling, as well as ensuring no interference from neighboring access points. As with unthrottled compute sharing, the scheduling delays of Wi-Fi increases with the contending users [16, 17].

Takeaway #6: *In sharing compute and wireless networks between robots, it is challenging to balance real-time and high-throughput workloads, necessitating dynamic QoS-aware scheduling, traffic shaping, and network slicing.*

8 Need for a Distributed Robotics Platform

Our measurement study highlights the growing need for powerful GPUs to support emerging mobile robotic manipulation workloads, as well as the potential for cost-effective shared

infrastructure. Furthermore, we demonstrate that no single deployment strategy—onboard, edge, or cloud—is optimal for all deployment scenarios and use cases due to inherent tradeoffs across workload performance, network latency, bandwidth, power consumption, and monetary cost.

These characteristics motivate the need for a *distributed robotics inference platform* that spans robots, edge, and cloud resources. Such a platform should support distributed orchestration of workloads based on bandwidth, latency, and compute requirements, and enable dynamic, QoS-aware scheduling of GPU, CPU, and network resources. Although similar ideas have been explored in domains such as mobile [15] and telecommunications [3], robotic systems introduce new constraints. In particular, closed perception–action loops impose strict real-time requirements, creating new challenges for system design and implementation to ensure safety and reliable performance in physical-world tasks.

9 Related Work

While foundation models have significantly advanced the state of mobile robotic manipulation, most existing research emphasizes algorithmic accuracy and generalization, often overlooking the system-level constraints.

1) System constraints: Current mobile robotic manipulation literature primarily evaluates task performance under idealized or simulated conditions [56]. While these works demonstrate impressive capabilities, they often assume that substantial compute resources are readily available. As a result, the impact of heterogeneous onboard compute platforms remains under-explored. In parallel, the robotics community has studied energy-efficient motion planning and control [53]. However, little work has examined how modern AI workloads affect the energy consumption and battery longevity of robots. Our work addresses this gap by systematically measuring the energy, memory, and performance trade-offs of core robotics workloads across several onboard compute platforms (§5). To our knowledge, this is the first study to quantify how task accuracy and execution time of robotics foundation models vary as a function of system-level resource constraints.

2) Offloading considerations: A natural strategy for handling resource-intensive workloads is offloading to edge or cloud devices. Offloading has been extensively studied in mobile and edge computing [15], particularly for deep neural network inference on resource-constrained devices [47] and large-scale video analytics [6]. However, such systems generally operate in soft real-time or open-loop settings, where occasional latency fluctuations are tolerable. In contrast, robotic systems operate within closed perception–action loops, where delays or dropped computations can directly degrade control performance or compromise safety. Thus, offloading considerations in robotic systems differ from those in mobile computing.

Within robotics, offloading has received comparatively limited empirical attention. Early surveys highlighted the potential for cloud infrastructure to accelerate workloads like SLAM and motion planning while noting risks introduced by network latency and connectivity variability [23]. More recent studies have quantified these effects, measuring how fixed communication latencies influence trajectory tracking and gesture recognition [32]. Despite this work, there is little measurement-based understanding of how the modern robotics workloads behave under offloading regimes. Prior studies largely focused on classical robotics pipelines or relatively small neural networks, which do not capture the memory footprint and bandwidth demands of contemporary multimodal models. To address this gap, we provide a quantitative analysis of offloading constraints for robotics workloads (§6).

3) Resource sharing in multi-robot settings: While multi-robot systems are well studied in the context of coordinated control [14] and collaborative mapping [2], the systems implications of shared compute infrastructure remain largely unexplored. This paper provides the first comprehensive measurement study of these multiplexing effects for robotics workloads, characterizing both the opportunities and limitations of shared infrastructure in multi-robot deployments (§7).

10 Conclusions

In this work, we performed the first measurement study of mobile robotic manipulation workloads across onboard, edge, and cloud GPU platforms. Our findings reveal that offloading such workloads can be beneficial in terms of performance and battery savings for the robots, but introduces its own challenges, due to the effects of network latency and bandwidth on task accuracy. Finally, the study highlights the opportunities and challenges of sharing compute across robot fleets. We hope that this study will help guide future research in the space of inference systems for mobile robots.

References

- [1] Allterco Robotics. [n.d.]. Shelly Plug Gen 4 Technical Documentation. <https://shelly-api-docs.shelly.cloud/gen2/Devices/Gen4/ShellyPlugUSG4/>. Accessed March 12, 2026.
- [2] Randa Almadhoun, Tarek Taha, Lakmal Seneviratne, and Yahya Zweiri. 2019. A survey on multi-robot coverage path planning for model reconstruction and mapping. *SN Applied Sciences* 1, 8 (July 2019), 847. <https://doi.org/10.1007/s42452-019-0872-y>
- [3] Ganesh Ananthanarayanan, Matthew Balkwill, Xenofon Foukas, Zhihua Lai, Bozidar Radunovic, Connor Settle, and Yongguang Zhang. 2025. Distributed AI Platform for the 6G RAN. In *Proceedings of the 2nd ACM Workshop on Open and AI RAN*. 15–21.
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu and Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report.
- [5] Shuanghao Bai, Wenxuan Song, Jiayi Chen, Yuheng Ji, Zhide Zhong, Jin Yang, Han Zhao, Wanqi Zhou, Wei Zhao, Zhe Li, Pengxiang Ding, Cheng Chi, Haoang Li, Chang Xu, Xiaolong Zheng, Donglin Wang, Shanghang Zhang, and Badong Chen. 2025. Towards a Unified Understanding of Robot Manipulation: A Comprehensive Survey. arXiv:2510.10903 [cs.RO] <https://arxiv.org/abs/2510.10903>
- [6] Tong Bai, Haoran Zhao, Lei Huang, Zhipeng Wang, Dong In Kim, and Arumugam Nallanathan. 2026. A Decade of Video Analytics at Edge: Training, Deployment, Orchestration, and Platforms. *IEEE Communications Surveys & Tutorials* 28 (2026), 2127–2162. <https://doi.org/10.1109/COMST.2025.3563377>
- [7] BBC News. n.d.. These robots can clean, exercise - and care for you in old age. Would you trust them to? <https://www.bbc.co.uk/news/articles/c9wdzzyyglq5o>. Accessed March 13, 2026.
- [8] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Al-abdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. 2024. PaliGemma: A versatile 3B VLM for transfer. <https://arxiv.org/abs/2407.07726>
- [9] Kevin Black, Manuel Y Galliker, and Sergey Levine. 2025. Real-Time Execution of Action Chunking Flow Policies. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=UKR2zO5uww>
- [10] Kevin Black, Manuel Y Galliker, and Sergey Levine. 2025. Real-time execution of action chunking flow policies. *arXiv preprint arXiv:2506.07339* (2025).
- [11] Remi Cadene, Simon Alibert, Alexander Soare, Quentin Gallouedec, Adil Zouitine, Steven Palma, Pepijn Kooijmans, Michel Aractingi, Mustafa Shukor, Dana Aubakirova, Martino Russi, Francesco Capuano, Caroline Pascal, Jade Choghari, Jess Moss, and Thomas Wolf. 2024. LeRobot: State-of-the-art Machine Learning for Real-World Robotics in Pytorch. <https://github.com/huggingface/lerobot>.
- [12] Vanya Cohen, Jason Xinyu Liu, Raymond Mooney, Stefanie Tellex, and David Watkins. 2024. A survey of robotic language grounding: tradeoffs between symbols and embeddings. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (Jeju, Korea) (IJCAI '24)*. Article 885, 11 pages. <https://doi.org/10.24963/ijcai.2024/885>
- [13] Colossus. n.d.. When Will Robots Go Mainstream? <https://colossus.com/article/when-will-robots-go-mainstream/>. Accessed March 13, 2026.
- [14] Jorge Cortes and Magnus Egerstedt. 2017. Coordinated Control of Multi-Robot Systems: A Survey. *SICE Journal of Control, Measurement, and System Integration* 10, 6 (2017), 495–503. <https://doi.org/10.9746/jcmsi.10.495>
- [15] Eduardo Cuervo, Aruna Balasubramanian, Dae-ki Cho, Alec Wolman, Stefan Saroiu, Ranveer Chandra, and Paramvir Bahl. 2010. MAUI: making smartphones last longer with code offload. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services (San Francisco, California, USA) (MobiSys '10)*. Association for Computing Machinery, New York, NY, USA, 49–62. <https://doi.org/10.1145/1814433.1814441>
- [16] Xenofon Foukas and Bozidar Radunovic. 2025. The future of the industrial AI edge is cellular. In *Proceedings of the 26th International Workshop on Mobile Computing Systems and Applications*. 61–66.

- [17] Ushasi Ghosh, Shivang Aggarwal, Mohamed Ibrahim, Puneet Sharma, and Dinesh Bharadia. 2026. Beyond Wi-Fi: Private 5G for Better Multi-User Quality-of-Experience. In *Proceedings of the 27th International Workshop on Mobile Computing Systems and Applications*. 7–12.
- [18] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. 2023. Visual Language Maps for Robot Navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. London, UK.
- [19] Physical Intelligence. 2024. π_0 : A Vision-Language-Action Flow Model for General Robot Control. <https://arxiv.org/abs/2410.24164>
- [20] Physical Intelligence. 2025. $\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization. <https://arxiv.org/abs/2504.16054>
- [21] Wenqi Jiang, Jason Clemons, Karu Sankaralingam, and Christos Kozyrakis. 2026. How Fast Can I Run My VLA? Demystifying VLA Inference Performance with VLA-Perf. *arXiv preprint arXiv:2602.18397* (2026).
- [22] Yuchen Jin, Sundararajan Renganathan, Ganesh Ananthanarayanan, Junchen Jiang, Venkata N. Padmanabhan, Manuel Schroder, Matt Calder, and Arvind Krishnamurthy. 2019. Zooming in on wide-area latencies to a global cloud provider. In *Proceedings of the ACM Special Interest Group on Data Communication (Beijing, China) (SIGCOMM '19)*. Association for Computing Machinery, New York, NY, USA, 104–116. <https://doi.org/10.1145/3341302.3342073>
- [23] Ben Kehoe, Sachin Patil, Pieter Abbeel, and Ken Goldberg. 2015. A Survey of Research on Cloud Robotics and Automation. *IEEE Transactions on Automation Science and Engineering* 12, 2 (2015), 398–409. <https://doi.org/10.1109/TASE.2014.2376492>
- [24] Charles C. Kemp, Aaron Edsinger, and Eduardo Torres-Jara. 2007. Challenges for robot manipulation in human environments [Grand Challenges of Robotics]. *IEEE Robotics & Automation Magazine* 14, 1 (2007), 20–29. <https://doi.org/10.1109/MRA.2007.339604>
- [25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything.
- [26] Alexander Korovko, Dmitry Slepichev, Alexander Efitov, Aigul Dzhumuratova, Viktor Kuznetsov, Hesam Rabeti, Joydeep Biswas, and Soha Pouya. 2025. cuVSLAM: CUDA accelerated visual odometry and mapping. *arXiv:2506.04359 [cs.RO]* <https://arxiv.org/abs/2506.04359>
- [27] Oliver Kroemer, Scott Niekum, and George Konidaris. 2021. A review of robot learning for manipulation: challenges, representations, and algorithms. *J. Mach. Learn. Res.* 22, 1, Article 30 (January 2021), 82 pages.
- [28] Thibault Kruse, Amit Kumar Pandey, Rachid Alami, and Alexandra Kirsch. 2013. Human-aware robot navigation: A survey. *Robotics and Autonomous Systems* 61, 12 (2013), 1726–1743. <https://doi.org/10.1016/j.robot.2013.05.007>
- [29] Mathieu Labbé and François Michaud. 2019. RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation. *Journal of Field Robotics* 36, 2 (2019), 416–446. <https://doi.org/10.1002/rob.21831> [arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.21831](https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.21831)
- [30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.
- [31] Alexander Millane, Helen Oleynikova, Emilie Wirbel, Remo Steiner, Vikram Ramasamy, David Tingdahl, and Roland Siegwart. 2024. nvbox: GPU-Accelerated Incremental Signed Distance Field Mapping. *arXiv:2311.00626* <https://arxiv.org/abs/2311.00626>
- [32] Luca Muratore and Nikos Tsagarakis. 2023. XBot2D: towards a robotics hybrid cloud architecture for field robotics. *Frontiers in Robotics and AI* Volume 10 - 2023 (2023). <https://doi.org/10.3389/frobt.2023.1168694>
- [33] NVIDIA. [n. d.]. Jetson Stats. https://developer.nvidia.com/embedded/community/jetson-projects/jetson_stats. Accessed: 24 February 2026.
- [34] NVIDIA. [n. d.]. NVIDIA Nsight Systems. <https://developer.nvidia.com/nsight-systems>. Accessed: 24 February 2026.
- [35] NVIDIA. [n. d.]. NVIDIA System Management Interface (nvidia-smi). <https://docs.nvidia.com/deploy/nvidia-smi/>. Accessed: 24 February 2026.
- [36] OpenAI. 2023. GPT-4 Technical Report.
- [37] Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, ĩrigo Goiri, Saeed Maleki, and Ricardo Bianchini. 2024. Splitwise: Efficient generative llm inference using phase splitting. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 118–132.
- [38] B.K. Patle, Ganesh Babu L, Anish Pandey, D.R.K. Parhi, and A. Jagadeesh. 2019. A review: On path planning strategies for navigation of mobile robot. *Defence Technology* 15, 4 (2019), 582–606. <https://doi.org/10.1016/j.dt.2019.04.011>
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html>
- [40] Sonia Raychaudhuri and Angel X Chang. 2025. Semantic Mapping in Indoor Embodied AI - A Survey on Advances, Challenges, and Future Directions. *Transactions on Machine Learning Research* (2025). <https://openreview.net/forum?id=USgQ38RG6G>
- [41] Hello Robot. 2025. *Stretch AI*. <https://hello-robot.com/stretch-ai>
- [42] Clearpath Robotics. [n. d.]. *Indoor Autonomy Software*. <https://clearpathrobotics.com/indoornav/>
- [43] Saumya Saxena, Blake Buchanan, Chris Paxton, Peiqi Liu, Bingqing Chen, Narunas Vaskevicius, Luigi Palmieri, Jonathan Francis, and Oliver Kroemer. 2025. GraphEQA: Using 3D Semantic Scene Graphs for Real-time Embodied Question Answering. In *Conference on Robot Learning (CoRL)*.
- [44] SemiAnalysis. n.d.. America Is Missing the New Labor Economy: Robotics (Part 1). <https://newsletter.semianalysis.com/p/america-is-missing-the-new-labor-economy-robotics-part-1>. Accessed March 13, 2026.
- [45] SemiAnalysis. n.d.. Robotics: Levels of Autonomy. <https://newsletter.semianalysis.com/p/robotics-levels-of-autonomy>. Accessed March 13, 2026.
- [46] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. <https://arxiv.org/abs/1912.01734>
- [47] Md. Maruf Hossain Shuvo, Syed Kamrul Islam, Jianlin Cheng, and Bashir I. Morshed. 2023. Efficient Acceleration of Deep Learning Inference on Resource-Constrained Edge Devices: A Review. *Proc. IEEE* 111, 1 (2023), 42–91. <https://doi.org/10.1109/JPROC.2022.3226481>
- [48] SIGRobotics-UIUC. 2025. LeKiwi: Low-Cost Mobile Manipulator. <https://github.com/SIGRobotics-UIUC/LeKiwi>
- [49] Software Radio Systems. [n. d.]. srsRAN Project. <https://www.srslte.com/>. Accessed: March 12, 2026.
- [50] Sevag Tafnakaji, Hadi Hajieghrary, Quentin Teixeira, and Yasemin Bekiroglu. 2023. Benchmarking local motion planners for navigation of mobile manipulators. In *2023 IEEE/SICE International Symposium on System Integration (SII)*. 1–6. <https://doi.org/10.1109/SII55687.2023>

10039133

- [51] Hamid Taheri and Zhao Chun Xia. 2021. SLAM; definition and evolution. *Engineering Applications of Artificial Intelligence* 97 (2021), 104032. <https://doi.org/10.1016/j.engappai.2020.104032>
- [52] Jiaming Tang, Yufei Sun, Yilong Zhao, Shang Yang, Yujun Lin, Zhuoyang Zhang, James Hou, Yao Lu, Zhijian Liu, and Song Han. 2025. Vlash: Real-time vlas via future-state-aware asynchronous inference. *arXiv preprint arXiv:2512.01031* (2025).
- [53] Mingyu Wu, Che Fai Yeong, Eileen Lee Ming Su, William Holderbaum, and Chenguang Yang. 2023. A review on energy efficiency in autonomous mobile robots. *Robotic Intelligence and Automation* 43, 6 (09 2023), 648–668. <https://doi.org/10.1108/RIA-05-2023-0060> arXiv:<https://www.emerald.com/ria/article-pdf/43/6/648/2346033/ria-05-2023-0060.pdf>
- [54] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *CVPR*.
- [55] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. 2026. World Action Models are Zero-shot Policies. *arXiv preprint arXiv:2602.15922* (2026).
- [56] Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, Zsolt Kira, Manolis Savva, Angel Chang, Devendra Singh Chaplot, Dhruv Batra, Roozbeh Mottaghi, Yonatan Bisk, and Chris Paxton. 2024. HomeRobot: Open-Vocabulary Mobile Manipulation. <https://arxiv.org/abs/2306.11565>
- [57] Chaoyang Zhu and Long Chen. 2024. A Survey on Open-Vocabulary Detection and Segmentation: Past, Present, and Future. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 8954–8975. <https://doi.org/10.1109/TPAMI.2024.3413013>