Embodied AI with Foundation Models for Mobile Service Robots: A Systematic Review

Matthew Lisondra, Beno Benhabib and Goldie Nejat, *IEEE Member*

Abstract— **Rapid advancements in foundation models, including Large Language Models, Vision-Language Models, Multimodal Large Language Models, and Vision-Language-Action Models have opened new avenues for embodied AI in mobile service robotics. By combining foundation models with the principles of embodied AI, where intelligent systems perceive, reason, and act through physical interactions, robots can improve understanding, adapt to, and execute complex tasks in dynamic real-world environments. However, embodied AI in mobile service robots continues to face key challenges, including multimodal sensor fusion, real-time decision-making under uncertainty, task generalization, and effective human-robot interactions (HRI).**

In this paper, we present the first systematic review of the integration of foundation models in mobile service robotics, identifying key open challenges in embodied AI and examining how foundation models can address them. Namely, we explore the role of such models in enabling real-time sensor fusion, language-conditioned control, and adaptive task execution. Furthermore, we discuss real-world applications in the domestic assistance, healthcare, and service automation sectors, demonstrating the transformative impact of foundation models on service robotics. We also include potential future research directions, emphasizing the need for predictive scaling laws, autonomous long-term adaptation, and cross-embodiment generalization to enable scalable, efficient, and robust deployment of foundation models in human-centric robotic systems.

Index Terms— Mobile Service Robots, Foundation Models, AI-Enabled Robotics

I. INTRODUCTION

In recent years, there have been rapid advancements in foundation models including Large Language Models (LLMs) [1]-[6], Vision Language Models (VLMs) [7]-[9], Multimodal Large Language Models (MLLMs) [10], [11] and Vision Language Action Models (VLAs) [12]-[14]. These models are significantly changing the overall field of robotics and its numerous applications. In particular, there have been notable improvements to personalized task management and context-aware smart assistant capabilities for domestic assistance [15]-[17], healthcare [18]-[21], and service automation [9]-[11]. Foundation models are trained on vast datasets to understand and generate human-like text/speech, as well as interpret visual information and reason. For example, language-guided mobile

Figure 1. A mobile service robot performs a domestic assistance task by responding to a language command from a user with limited mobility. The robot uses a VLM to detect candidate objects in the environment, while a MLLM interprets the urgency and semantic intent of the user's query. A VLA model grounds the instruction to executable actions, enabling the robot to search drawers, visually identify the correct medicine, and plan a fetch trajectory for delivery, demonstrating embodied language-conditioned control in a home. Image obtained from the Habitat 2.0 Simulator [60].

service robots utilizing foundation models can carry-out complex instructions [25]-[27], perform fetch and carry tasks [28]-[31] and/or socially interact [32]-[34] with human users.

As robots are deployed in real-world environments to increasingly interact with humans, the integration of language-based intelligence has become crucial for intuitive and effective communications. In particular, mobile service robots can provide adaptable, context-aware assistance in dynamic and unpredictable environments to achieve various goals, from navigating person-centered environments to object or person detection and identification. Their ability to autonomously interpret human instructions, reason about tasks, and execute actions has enabled them to be deployed in numerous environments, from warehouses [35], [36] to hospitals [12], [37] and personal homes [27], [38], [39].

By leveraging the advancements in foundation models, service robots can bridge the gap between AI-driven perception and human-like decision-making, leading to natural human-robot interactions (HRI) and collaborations in everyday

scenarios. However, to fully integrate the promising capabilities of foundation models, the open challenges in the development and deployment of embodied AI mobile service robots must be addressed. These open challenges include:

1) *Multi-modal Perception*, where sensor noise, occlusions, dynamic lighting conditions, and changing spatial configurations can hinder the successful integration and fusion of multiple sensory modes [40]-[44];

2) *Translation of Natural Language Instructions into Executable Robot Actions*, where an inherent gap exists between the abstract linguistic representations and the precise, context-dependent executable commands required by robots [45]-[49]. This is further complicated by language ambiguity, lack of physical commonsense, and the need for long-horizon planning in dynamic human-centric environments [50]-[52];

3) *Uncertainty Estimation*, where mobile service robots need to operate in unpredictable environments, requiring real-time, context-aware decisions despite the presence of partial observations, and ambiguous user commands [53]-[55]. However, current embodied AI methods struggle with uncertainty estimation, leading to overconfident predictions or failures in high-risk scenarios [56]-[59]; and,

4) *Computational Capabilities*, where mobile service robots are limited by their onboard computational capabilities. This makes real-time inference with large-scale deep learning models challenging, as these require powerful GPU clusters.

In general, foundation models can address the above challenges. They can bridge the gap between high-level language understanding and low-level robotic control to enable more intelligent, adaptable, and interactive robots for diverse real-world applications, as shown in Fig. 1.

To-date, existing surveys have primarily examined either broad, non-task-specific applications in general-purpose robotics [61], [62] or language-conditioned manipulation tasks for stationary robotic arms [63]. These have not yet investigated the role of mobility in enabling robots to assist with human-centered tasks. In this paper, we present the first systematic review on the integration of foundation models in mobile service robotics, examining their role in advancing embodied AI. We identify research challenges, and discuss how foundation models can enable task generalization, dynamic scene understanding, and social compliance. In particular, we focus on the promising applications of domestic assistance, healthcare and service automation for mobile service robots.

II. OPEN CHALLENGES OF EMBODIED AI FOR MOBILE SERVICE ROBOTS

In this section, we outline four challenges that need to be addressed to leverage the potential of foundation models in service robotics.

A. Challenge #1: Multimodal Perception

Multimodal perception is essential for mobile service robots to interpret complex environments. These modalities often differ in spatial resolution, sampling frequency, and noise characteristics, making real-time sensor fusion and alignment difficult in human-centered environments. We discuss four core limitations below.

1) Cross-Modal Representation: Sensory data from various robotic sensors need to be fused into a shared latent space to support real-time reasoning in mobile service robots. Fusion architectures are commonly classified into [64]: (*i*) early fusion, where raw inputs are concatenated or merged at the pixel level, (*ii*) late fusion, which aggregates decisions from different modalities, and/or (*iii*) intermediate fusion, where features from each modality are encoded and combined using attention mechanisms, joint embeddings, or cross-modal transformers. However, each fusion type poses challenges: early fusion is sensitive to noise and modality-specific distortions, late fusion can result in delayed or inconsistent decisions when modalities disagree, and intermediate fusion requires learning aligned feature spaces across sensors with different noise, sampling rates, and spatial resolutions. These challenges are compounded by spatial misalignment, temporal desynchronization, and representational inconsistency [41], [42]. In mobile service robotics, such inconsistencies can hinder object association, scene parsing, and HRI, particularly in tasks that require fine-grained context interpretations, such as locating household items in cluttered spaces, or monitoring patient conditions using multi-sensor data.

2) Latency Issues: Sensor streams often operate at different rates, leading to temporal desynchronization [65], [66]. Without accurate spatiotemporal calibration, perception pipelines can produce blurred depth estimates, and erroneous visual-inertial pose predictions, especially, for robot localization [65]. Furthermore, misaligned fused representations such as feature maps, point clusters, and IMU-integrated trajectories can lead to incorrect obstacle boundaries and delayed semantic segmentation [67]. For mobile service robots, these delays can impair situational awareness.

3) Uncertainty Propagation Across Modalities: Sensor quality may degrade due to lighting changes, occlusions, or environmental noise [68]-[70]. Uncertainty propagation occurs when sensor noise or degradation in one modality affects the reliability of fused estimates across modalities in joint estimation pipelines. Yet, many fusion methods assume stationary sensor noise models and fixed modality trust, which can result in overconfident predictions or action outputs, [71], [72]. This is problematic for mobile service robots, as overconfidence can lead to unsafe behaviors. Adaptive confidence weighting ,which assigns modality-specific weights based on online reliability metrics, exists in research [73], [74], but remains underdeveloped for real-world deployment due to key limitations, such as: (*i*) being computationally expensive, (*ii*) assuming static environments, and (*iii*) lacking robustness to abrupt sensor failures and scene changes. These constraints make them difficult to implement for mobile service robots.

4) Domain Adaptation and Transferability Issues: Early, late, and intermediate fusion models are often trained in static, well-lit indoor datasets and degrade in deployment across cluttered households, dim hospital corridors, or dynamic outdoor venues [75], [76]. Domain adaptation methods such as domain randomization [77], self-supervised online learning [78], and adversarial style transfer [79] face limitations in mobile service robot deployments. Transferability is further hindered when models are overfit to specific training conditions. Mobile

service robots cannot retrain models in real time during operation, making them vulnerable to perceptual drift.

B. Challenge #2: Translation of Natural Language Instructions into Executable Robot Actions

Enabling language-guided mobile service robots to follow high-level natural language commands provided by humans remains a challenge. Key limitations include:

1) Symbolic-to-Embodied Mapping and Instruction Ambiguity: Natural language, typically, expresses tasks at a high level such as "*clean the room*" or "*get the medicine*", whereas robot controllers require low-level action sequences. Bridging this gap is difficult when instructions are ambiguous or underspecified, lacking object labels, spatial context, or temporal constraints. Classical planners (*e.g.*, STRIPS [45] and (SHOP2) [46], [47]), or logic-based (PDDL) [48], [49], decompose commands into discrete subgoals, however, rely on hand-engineered operators that are rule-based and discretized, and too domain-specific to generalize [80]-[83] across unseen environments. This challenge is critical for mobile service robots, which must translate diverse, user-issued commands into continuous, context-aware actions while adapting to unpredictable spatial layouts, and real-time task demands.

2) Lack of Embodied Commonsense and Physical Constraints Awareness: Many symbolic planners, grounding models and PDDL-based models lack awareness of physical constraints, [84], [85]. Language-only systems, limitation-based models, or semantic parsers trained on demonstrations may incorporate basic kinematics or affordances, but lack generalizable reasoning about physical feasibility, robot limitations, and environment variability, [50]-[52]. As a result, they can produce physically impossible plans, unsafe actions, or failed HRI.

3) Failure in Long-Horizon Task Planning: Many mobile service tasks, such as medication delivery or multi-room navigation, require reasoning across multiple steps, maintaining memory, and adapting plans as conditions change [86]. Classical planners such as Simple Task Network (STN) Planning [87] or Fast Forward (FF) [88]) can efficiently decompose tasks in static domains, but are unable to handle unexpected subgoals, branching, or replanning in dynamic environments [89], [90]. Learning-based models such as Neural Task Graphs (NTG) [91] and Option-Critic [92] can offer more flexible sequencing, but often suffer from policy drift, error accumulation, and limited memory [93]-[95]. Mobile service robots need to execute real-world, multi-step tasks which require robust, adaptive, and context-aware task progression.

C. Challenge #3: Uncertainty Estimation

Uncertainty is inherent in language-guided mobile service robotics, arising from sensor noise, human ambiguity, partial observability, and/or dynamic environments [56]-[59], [96], [97]. Uncertainty estimation in mobile service robots have four core limitations:

1) Lack of Explicit Uncertainty Quantification: Mobile service robots often lack the ability to estimate confidence in their decisions. While probabilistic methods consider modeling confidence [56], [96], they are computationally intensive for real-time robotic applications. Thus, many rule-based models

[57] and standard deep learning models [58], [59] lack built-in uncertainty estimation mechanisms. Calibrated uncertainty [98] has not been widely implemented in mobile service robotics. In human-centric environments, this can compromise both human safety and trust.

2) Failure in Long-Horizon Uncertainty Estimation: Robots must reason about the consequences of their actions across extended time horizons. However, traditional state estimation and planning methods, such as Kalman Filters [99], [100], belief-space planners and sampling-based approaches [101], do not model how uncertainty accumulates [102]-[104], which may lead to underestimating uncertainty as prediction horizons increase [105], [106]. Belief-space motion planners [107] and sampling-based approaches such as Partially Observable Sparse Samplers [101] can become computationally intractable as the planning horizon increases [93]-[95]. This can limit the ability of mobile service robots to anticipate risk, recover from uncertain states, or adapt plans over time.

3) Uncertainty in HRI: Service robots must manage uncertainty in user-facing decisions, including interpreting ambiguous verbal and nonverbal instructions. To-date, the majority of robots used in HRI cannot communicate uncertainty or request clarification [108]-[111]. This can lead to social errors such as interrupting users, performing unintended actions, or failing to respond appropriately in a conversation. These failures often stem from misaligned intent interpretation. Plan-based dialog management frameworks [53], behavior-tree-based HRI controllers [54], and Partially Observable Markov Decision Process (POMDP)-based dialog managers [55] often assume fixed interaction protocols and lack the ability to model, quantify, or express uncertainty in real time, [112]. This challenge for mobile service robots can lead to safety risks.

D. Challenge #4: Computational Capabilities

Mobile service robots have limited onboard computational power, making real-time inference with large-scale symbolic and deep learning models challenging. Key limitations include:

1) Perception and Planning Computational Overhead: Large-scale deep learning models often require significant memory [113], [114], and computational resources [115]. For example, models with convolutional backbones [142] or 3D segmentation networks [143] can exceed the real-time processing capacity of edge computing units commonly used in mobile service robots. This results in delayed perception updates, sluggish motion planning, and unsafe lags during user interactions or obstacle avoidance. Classical planning frameworks [116], and the Dynamic Window Approach (DWA) planner [117] are designed for lightweight CPU-based inference using simplified 2D cost maps. While efficient, these frameworks are not scalable to high-dimensional semantic inputs like multimodal feature maps or RGB-D affordance graphs [118], [119], which are critical for mobile service robots.

2) Lack of Adaptive Resource Allocation for Real-Time Inference: Most classical mobile service robot frameworks such as traditional navigation stacks [120] and behavior-based control architectures [121] process sensor data at fixed input resolutions and allocate static CPU or GPU budgets, regardless of changing scene complexity or task urgency [122], [123]. This

can result in underutilization of onboard computing during low-complexity tasks, and overload of processing units during high-complexity tasks [124], [125]. Only a few classical visual processing methods [126], [127] and ElasticFusion [128], include adaptive techniques. However, these are largely heuristic and do not generalize to full perception, planning, and control pipelines required by mobile service robots [129]-[131].

To address the aforementioned challenges, foundation models can be utilized to extend the capabilities of mobile service robots, enabling them to manage real-time perception, interpret complex multimodal inputs, and reason about actions in human-centric environments.

III. OPPORTUNITIES OF FOUNDATION MODELS FOR MOBILE SERVICE ROBOTS

This section outlines how foundation models address the four core challenges identified in Section II.

A. Addressing Challenge #1: Multimodal Perception

Robust multimodal perception is essential for interpreting complex environments. Particularly, multimodal VLMs and MLLMs can address this challenge by providing pre-trained multimodal representations of visual, textual and spatial features that enable the integration of heterogeneous sensory data through unified representation learning, temporal synchronization, uncertainty-aware fusion, and domain generalization [132], [133]. We discuss below the key advances of leveraging foundation models for this challenge.

1) Cross-Modal Representation Gap: Foundation models address the challenge of aligning heterogeneous sensor streams into shared latent spaces using: 1) cross-modal tokenization [134], [135], 2) spatial encoding [132], and 3) attention-based fusion [136]. Perception and language inputs, robot motion traces, and natural language commands are tokenized into a unified discrete format to allow joint multimodal reasoning. For example, Magma [134] applies cross-modal tokenization by encoding object features, robot motion traces, and User Interface (UI) navigation points into a unified token space for transformer-based action grounding and planning. These mechanisms address representational fragmentation and enable robust, flexible multimodal perception critical for mobile service robots.

2) Latency Issues: Strategies for reducing latency include: 1) learning time-aware representations [137], 2) aligning asynchronous inputs [138], and 3) dynamically adjusting modality-specific update rates [139]. For learning time-aware representations, video foundation models such as VideoJAM [137] and InternVideo2 [138] model visual appearance and optical-flow motion jointly across time, enabling coherent spatiotemporal grounding across asynchronous frames. Object-centric physical modeling frameworks align asynchronous inputs. These advances enable mobile service robots to perform low-latency navigation, manipulation, and monitoring tasks.

3) Uncertainty Propagation Across Modalities: Mobile service robots operating in cluttered or human-centric environments must assess the reliability of multimodal inputs in real time. Foundation models enable uncertainty-aware fusion, where each modality's contribution is dynamically weighted by estimated confidence [140]. This directly addresses the problem of uncertainty propagation across modalities, where sensor degradation in one input can affect fused estimates and downstream decisions. For example, DeepSeek-R1 [141], a reinforcement-trained LLM, and Segment Anything Model (SAM) [142], a VLM foundation model, incorporate reinforcement-aligned and spatial confidence mechanisms to suppress unreliable perceptual inputs and enhance segmentation under ambiguous scenes. These strategies allow service robots to apply lower weights to unreliable sensors and maintain robust perception under occlusion and/or noise.

4) Domain Adaptation and Transferability: Service robots must generalize perception and reasoning in diverse environments, from hospitals and homes to offices and public spaces. This requires self-supervised vision-only foundation models, multimodal VLMs, and MLLMs that can adapt to new domains and transfer knowledge from large-scale pretraining distributions to unseen real-world environments with minimal labeled data or task-specific finetuning [143]. Transferability refers to the ability of these models to be trained on one domain and be able to perform effectively in different domains without retraining. To support this, foundation models leverage: 1) self-supervised pretraining [144]-[145], 2) cross-domain token alignment [133], [146], and 3) few-shot generalization [147], [148] demonstrated by Qwen2-VL [149], a VLM that dynamically adjusts tokenization based on image resolution for domain-flexible perception. These models allow mobile service robots to maintain perception and reasoning performance when encountering unfamiliar layouts, textures, lighting conditions, and semantic content.

B. Addressing Challenge #2: Translation of Natural Language Instructions into Executable Robot Actions

Understanding and executing natural language instructions remains a core challenge for mobile service robots. In response, foundation models leverage language-conditioned policies [12], vision-language alignment [150], and multimodal reasoning [151], [152] to bridge the gap between symbolic language and embodied robotic control. In particular, recent foundation models have advanced instruction-following in mobile service robots by addressing the symbolic-to-embodied gap [153]-[158], incorporating physical commonsense [159], [160], and supporting long-horizon planning and adaptation [161]-[163]. We discuss these key advances below.

1) Symbolic-to-Embodied Mapping and Instruction Ambiguity: Foundation models address the challenge of grounding high-level natural language into executable robot actions by learning shared semantic representations and integrating real-time perceptual context. Unlike classical planners that rely on predefined operators or logical rules [164], foundation models align symbolic instructions with continuous control spaces, enabling more flexible, environment-aware grounding for mobile service robots. For example, LIMO [153], a reasoning-focused language model, introduces a recurrent latent reasoning module to preserve and reuse semantic context across reasoning steps, supporting both short- and long-horizon tasks. VLMs and MLLMs can enhance a mobile robot's ability in interpreting vague language, resolving ambiguities, and translating abstract

commands into actions that respect spatial context, task constraints, and human dialogue.

2) Lack of Embodied Commonsense and Physical Constraints Awareness: In general, foundation models can address the absence of physical commonsense, by incorporating: 1) physics-based priors [165], 2) affordance reasoning [159], and 3) constraint-aware planning [13], [148], [160] into language-grounded policy architectures. These capabilities address the embodied AI gap where robots often lack real-world intuition about object properties, support stability, or task feasibility which is critical in mobile service contexts. Genesis [165] is a generative physics simulation foundation model and encodes differentiable physics into object-centric representations, allowing mobile robots to simulate interactions such as stacking, collision, or deformation before execution. This prevents unsafe actions such as placing heavy items on top of fragile ones or interacting with unreachable or obstructed objects. These capabilities allow mobile service robots to reason about physical constraints to ensure infeasible actions are not taken, and to reduce safety risks and user dissatisfaction.

3) Failure in Long-Horizon Task Planning: Mobile service robots frequently face tasks that span multiple steps, require context retention, and require real-time adaptation to changing environments. To manage this complexity, foundation models 1) decompose high-level language instructions into structured subtasks [161], [162], 2) track prior actions [162], [163], and 3) revise plans based on sensory feedback or environmental changes [166]-[168]. They enhance long-horizon task planning by enabling instruction decomposition, memory retention, and adaptive replanning which are important for mobile service robots operating in dynamic settings. For example, Code-as-

Policies [161], a code-writing LLM, translates language into modular control programs for robust task sequencing, while LLM-Planner [162], an LLM-based planner, performs few-shot grounded planning that adapts instructions to visual context and tracks prior actions through goal-conditioned history encoding. s1 [163], a reasoning-optimized LLM, introduces budget forcing in order to dynamically scale reasoning depth at test time, mitigating error propagation and improving task continuity for ambiguous or multi-step planning. These foundation models assist mobile service robots to execute multi-step, context-sensitive instructions by integrating structured planning, foresight, and adaptability.

C. Addressing Challenge #3: Uncertainty Estimation

Robust uncertainty estimation for mobile service robots enables such robots to work in unpredictable environments, where sensor noise, ambiguous input, and unexpected events can compromise decision-making. Foundation models address this challenge by supporting explicit confidence modeling, prediction long-horizon risk prediction, and uncertainty-aware HRI [169]-[181]. We discuss these below.

1) Lack of Explicit Uncertainty Quantification: Effective decision-making, under ambiguous or noisy conditions, requires service robots to estimate their confidence in perception and actions. Foundation models address this by supporting: 1) reinforcement-guided value modeling [141], 2) attention-based input reweighting [169]-[171], and 3) temporal uncertainty tracking [165], [172]. The LLM DeepSeek-R1 [141] estimates action confidence using reinforcement learning signals. The VLMs Perceiver IO [169] and Uni-Perceiver v2 [170] dynamically down-weight
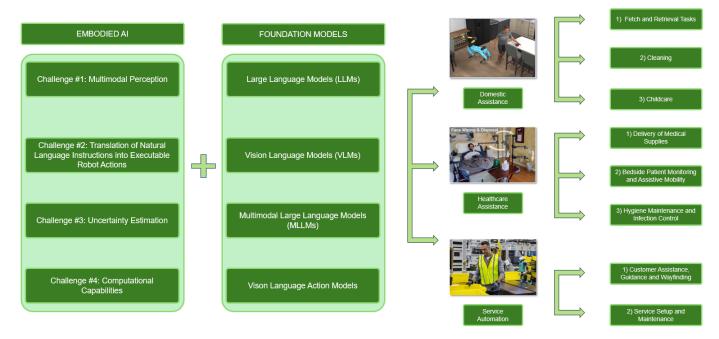

Figure 2. Overview of how foundation models address open challenges in embodied AI for mobile service robots. Embodied AI faces four key challenges: multimodal perception, translating natural language into executable actions, uncertainty estimation, and limited onboard computational capabilities. Foundation models—including LLMs, VLMs, MLLMs, and VLAs—offer scalable solutions to these challenges. Their integration enables real-world deployment of mobile service robots across three primary domains: domestic assistance (object retrieval, cleaning, childcare), healthcare assistance (medical delivery, patient monitoring, hygiene support), and service automation (customer guidance, event setup), thereby enhancing robot autonomy, adaptability, and human-centered task execution.

unreliable sensory inputs through adaptive cross-modal attention layers. The MLLM GPT-4o [171] modulates input contributions by estimating token-level reliability during multimodal reasoning. Temporal uncertainty is addressed by the VLM OmniFusion [165], which performs multimodal temporal fusion with evidential attention that models feature confidence over time. These models allow mobile service robots to maintain reliable performance when navigating cluttered spaces, interpreting noisy inputs, or executing high-risk tasks in uncertain environments.

2) Failure in Long-Horizon Uncertainty Estimation: Planning over extended time horizons presents a major challenge for mobile service robots. To manage this, foundation models use latent world modeling and temporal abstraction [173]-[175] to simulate outcomes in latent space, helping robots anticipate risks such as pose drift, occlusion or localization failure without real world execution. On the other hand, the vision-based self-supervised foundation models MAE [176] and Swin Transformer [177] offer an alternative approach by reconstructing occluded or corrupted spatiotemporal patches from video sequences in order to improve a robot's ability to refine sensory confidence during long, partially observable tasks. These world-model-based and vision-based foundation models enable mobile service robots to forecast degraded visibility, avoid overconfident long-term commitments, and be able to adaptively replan when predictions become unreliable.

3) Uncertainty in Human-Robot Interactions (HRI): Effective HRI requires mobile service robots to detect uncertain or ambiguous human behaviors such as gestures, gaze direction, and to communicate their own uncertainty clearly during task execution. Foundation models address this by integrating behavior forecasting, social intent modeling, and natural language grounding. For example, Lumiere [178], a diffusion-based generative video model, enables motion-consistent human behavior prediction by generating temporally coherent video simulations of potential human trajectories, supporting socially aware navigation. These foundation models enable mobile service robots to perform socially fluent, interpretable, and adaptive interactions in dynamic environments.

D. Addressing Challenge #4: Computational Capabilities

Mobile service robots often operate under constrained budgets, making real-time perception and planning with large deep learning and symbolic models difficult. Foundation models address this challenge by enabling scalable model compression [182]-[184], efficient attention distillation [185], [168], adaptive token processing [186], [187], and on-device execution [188], [189]. These advances allow mobile robots to maintain real-time responsiveness, energy efficiency, and task throughput even under strict power, memory, and latency constraints. We discuss these key advances below.

1) Perception and Planning Computational Overhead: Real-time perception and planning can often exceed the computational limits of embedded systems. This creates latency bottlenecks that impair decision-making during navigation, interaction, or manipulation. Foundation models address computation limits through: 1) architectural streamlining [190], [167], 2) model compression [182]-[184], and 3) efficient attention mechanisms tailored for edge devices [168], [185]. The LLM EdgeFormer [190] uses architectural streamlining by applying dynamic token control to maintain responsiveness during instruction-following. The VLMs DeiT [182], [183] compress vision transformers via attention-based distillation to enable accurate perception with minimal compute. The MLLM VisionLLaMA [184] generalizes LLaMA-style transformers to vision tasks such as object classification, detection, and dense 3D scene understanding using a compact shared backbone with 2D rotary positional encodings. These models reduce the inference bottlenecks of perception and planning pipelines, practical for compute-constrained mobile service robots operating under real-time, energy, and workload constraints.

2) Lack of Adaptive Resource Allocation for Real-Time Inference: Efficient task execution under changing computational loads is essential for mobile service robots operating with strict latency and energy constraints. Foundation models address this by incorporating adaptive computation strategies. For example, the LLMs Switch Transformer [186] and GShard [187] use Mixture-of-Experts (MoE) architectures to selectively activate task-relevant modules, reducing computation during simpler tasks while maintaining full capacity when needed. These methods support adaptive, latency-aware operation critical for reliable real-time mobile service robot navigation, interaction, and manipulation.

Addressing the core challenges of multimodal perception, language grounding, task generalization, uncertainty estimation, and computational efficiency, foundation models offer scalable solutions for mobile service robotics.

IV. REAL-WORLD APPLICATIONS FOR MOBILE SERVICE ROBOTS WITH EMBEDDED FOUNDATION MODELS

The global mobile service robotics market is projected to experience rapid growth across domestic, healthcare, and public service sectors. In particular, the service robotics market is expected to nearly double—growing from $47.10 billion USD in 2024 to $98.65 billion USD by 2029—with an average compound annual growth rate of 15.9% [191]. This surge reflects broader trends toward integrating service robots into homes, healthcare facilities, and public spaces [192], [193]. These trends directly relate to the challenges discussed in Section II and the foundation model solutions explored in Section III. Specifically, foundation models are increasingly being embedded into mobile service robots to enhance domestic assistance, healthcare assistance, and service automation (see Fig. 2). In this section, we examine how these models address real-world needs and enable safer, more adaptive, and socially intelligent robot behaviors across these domains.

A. Domestic Assistance

Domestic assistance refers to everyday tasks to support comfort, safety, and quality of life [194], [195]. These tasks include the fetch and retrieval of household objects [196], [197]-[199], cleaning [38], [200], supportive childcare assistance [201], [202], and cooking activities [203]-[207].

1) Fetch and Retrieval Tasks: These tasks, often triggered by natural language, gestures, or contextual cues, involve

multimodal perception, ambiguous instruction grounding, uncertainty management, and efficient planning. Fetch and retrieval tasks decompose into sub-tasks: 1) object localization, 2) grasp planning and action refinement, 3) navigation and multi-step delivery, and 4) handover and placement reasoning. For example, object localization was achieved via the VLM CLIPort [27] on the Franka Emika Panda Robot. CLIP enables the grounding of natural language prompts to pixel-space visual features, allowing the robot to locate and pick target objects precisely. Grasp planning and action refinement is enabled on the same Franka Emika Panda Robot using Octo [196], which combines a T5 encoder [156] with a diffusion policy to generalize grasp strategies under occlusion and pose uncertainty. Navigation and multi-step delivery can be performed using the Fetch robot in [30], where GPT-4 [182], GPT-4V [179] and SAM [178] are integrated to adapt paths and retrieval plans. Then, CogNav [31], an MLLM using the same Fetch robot, uses GPT-4 [145], and GPT-4V [157] to sequence exploration and mapping in homes. In [197], PARTNR, a benchmark and system for Planning and Reasoning in Embodied Tasks was benchmarked on both the Franka Emika Panda Robot and a Spot robot with a custom mounted 7-DoF arm. It leveraged LLaMA3.1-8B [208] and CodeLLaMA-70B [209] to execute collaborative fetch routines such as retrieving medical kits, improving multi-agent coordination and enabling

the robots to perform socially aware and assistive retrieval tasks.

2) Cleaning: Cleaning tasks involve a robot autonomously organizing, decluttering, and sorting household items into designated locations [210]. Unlike object-specific fetch tasks, cleaning often requires generalizing abstract commands such as "*put everything away*" into sequential, context-aware behaviors across many objects and categories. Cleaning tasks can be decomposed into the following sub-tasks: 1) user preference inference and object categorization, 2) object localization and sorting, and 3) task execution with multimodal grounding and action control. For user-specific preferences and categorization, the LLM in TidyBot [38] which uses GPT-3.5 [148] and can be deployed on a custom holonomic base with a Kinova Gen3 robotic arm, prompts GPT-3.5 [148] to infer placement and manipulation rules such as "*put light clothes in the drawer*" from a few examples, generating personalized tidying policies. These rules are, then, parsed to predict destinations for unseen objects based on learned categories. Object recognition and matching are achieved by TidyBot through the combination of GPT-3.5 [148] with CLIP [135].

3) Childcare: Childcare tasks require a mobile service robot to assist with daily routines, object retrievals, and monitoring of young children [211]. Example tasks include responding to prompts such as "*bring me my blanket*," guiding hygiene routines like teeth brushing after meals, and detecting safety

1) Domestic Assistance	Fetch and Carry Tasks	CLIPort [27], Octo [196] (both Franka Emika Panda Robot); Robi Butler [30], CogNav [31] (Fetch Robot); PARTNR [197] (Franka Emika Panda Robot or Spot + 7-DoF arm), RoboPoint [198], OpenVLA [167]
	Cleaning	TidyBot [38], RT-2 [14], VIMA [200] (all use custom holonomic base + Kinova Gen3 or Franka Emika Panda Robot)
	Childcare	Smart Help [29] (custom mobile base + Kinova Gen3 arm), SIF [201] (Spot + 7-DoF arm), π_0 [202] (ARX-X7; compatible with Kinova Gen3 and Spot Robot)
	Cooking	RING [33] (Unitree Go1 / Locobot / Stretch RE-1), HarmonicMM [203] (Stretch RE-1), Plan-Seq-Learn [204], BOSS [205], [39] (all Franka Emika Panda Robot); ALOHA and Mobile ALOHA [206], [207]
2) Healthcare Assistance	Delivery of Medical Supplies	ProgPrompt [214], LLM-Grounder [215] (both Franka Emika Panda); SayCan [12] (Everyday Robots), VLM-Social-Nav [179] (Clearpath Jackal Robot)
	Bedside of Patient Monitoring and Assistive Mobility	MoMa-LLM [216] (Fetch Robot), OLiVia-Nav [180] (Clearpath Jackal Robot), GSON [217] (custom mobile base with RGB + 2D LiDAR)
	Hygiene Maintenance and Infection Control	AutoRT [218] (Everyday Robots), SayPlan [219] (Franka Emika Panda Robot), RoboGen [220]
3) Service Automation	Customer Assistance, Guidance and Wayfinding	OVSG [225] (Ackermann Steering Robot), LM-Nav [226] (Clearpath Jackal Robot); ViNT [228] (Vizbot, Unitree Go1, Clearpath Jackal, and LoCoBot)
	Service Setup and Maintenance	Hi Robot [229], AgiBot [230] (Agibot Robot), PhysObjects [231] (Franka Emika Panda Robot)

Table 1. Summary of mobile service robots using foundation models for domestic, healthcare, and service automation applications.

risks such as toddlers approaching stairs or handling small objects. Childcare can be decomposed into following sub-tasks: 1) behavioral intent grounding, *2)* dynamic instruction decomposition, and 3) real-time multimodal control and safety adaptation. Behavioral intent grounding is addressed in [29], where Smart Help whose robot platform has an on-board camera and custom mounted Kinova Gen3 arm, uses the LLM GPT-3.5-turbo-instruct [148] to predict user intention from short or underspecified prompts by learning probabilistic mappings between behavioral cues and task goals. This enables the robot to interpret requests like "*help me*" in context-specific ways, such as retrieving a dropped toy or fetching a comfort item. Real-time multimodal control and safety is addressed in π_0 [202] which uses the ARX-X7 robot and applies the VLM PaliGemma [212]. π_0 tokenizes vision-language prompts into continuous motor action commands, enabling a mobile robot to execute environment-aware child supervision behaviors, such as dynamically blocking unsafe areas or re-routing around obstacles during interactive play sessions. These advances in behavioral grounding, instruction decomposition, and multimodal action planning enable service robots to safely and responsively support childcare tasks.

4) Cooking: Cooking tasks involve mobile service robots executing multi-step procedures to assist elderly individuals, busy families, or users with disabilities. Cooking includes such subtasks as: 1) ingredient and utensil localization, 2) appliance interaction and sensing, and 3) bimanual manipulation of cooking items. Ingredient and utensil localization is addressed in [33], using either the Unitree Go1 robot, the Locobot or the Stretch RE-1 robot, which integrates the VLMs SIGLIP-ViT [213] and CLIP [135] to visually locate tools like cutting boards from natural language cues. Sequential task decomposition and planning is addressed in [204] using the Franka Emika Panda robot which employs the LLM GPT-4 [145], and the VLM SAM [142] to transform cooking goals such as "*make a sandwich*" into grounded action sequences—such as "*pick bread*," "*retrieve lettuce*"—linked to visual goals. By integrating perception, planning, sensory feedback, and dexterous control across diverse kitchen tasks, LLM, VLM, MLLM-empowered robots can assist users in cooking routines with greater autonomy, precision, and responsiveness.

B. Healthcare

Healthcare assistance tasks involve supporting clinical workflows, assisting people, and infection control in dynamic hospital environments. These tasks include the delivery of medical supplies [214], [215], bedside patient monitoring and assistive mobility [216], [217], and hygiene maintenance and infection control [218]-[220].

1) Delivery of Medical Supplies: Delivery of medical supplies involves mobile service robots autonomously transporting medications, IV fluids, lab samples, and surgical instruments throughout hospitals and clinics. These deliveries are essential to maintaining smooth clinical workflows, reducing staff workload, and minimizing human error in time-sensitive medical tasks. Robotic delivery tasks can be decomposed into the following sub-tasks: 1) symbolic-to-action mapping and long-horizon planning, 2) spatial grounding of objects and target locations, 3) feasibility evaluation and adaptive plan selection, and 4) social navigation in dynamic clinical settings. Symbolic-to-action mapping and long-horizon delivery planning is addressed in [214] using ProgPrompt, which employs the LLMs Codex [221] and GPT-3 [2] on a custom mobile manipulator base with a Franka-Emika Panda robotic arm. ProgPrompt translates natural language prompts such as "*take these surgical tools to the prep room*" into executable control programs like grab(kit), goto(prep_room), and handover, enabling the robot to autonomously generate structured task plans. It uses an LLM-API and known user-provided prompting primitives of information in a room to generate the task sequence. Spatial grounding of object and target locations is supported in [215], where LLM-Grounder combines GPT-4 [145] and the VLM OpenScene [222] using the same aforementioned mobile robot. LLM-Grounder resolves spatial references like "*on the left tray by the door*" into 3D coordinates for pick-and-place execution, allowing the robot to disambiguate complex clinical instructions and correctly localize supplies and delivery zones even in cluttered hospital environments. Social navigation in dynamic clinical environments is addressed in [179], where VLM-Social-Nav merges GPT-4V [157] and YOLO [223] on a Clearpath Jackal robot, although similar deployments can be made on the Everyday Robots or custom mobile manipulators. VLM-Social-Nav enables the robot to interpret real-time human gestures, access signals, and crowd dynamics. Based on these multimodal cues, the robot adapts its navigation policy in real-time to yield, reroute, or proceed according to human intentions and clinical social norms, ensuring safe and socially compliant delivery of medical supplies in crowded hospital corridors.

2) Bedside Patient Monitoring and Assistive Mobility: Hospitals impose strict constraints on proximity, etiquette, and timely intervention. Bedside monitoring and assistive mobility tasks decompose into sub-tasks: 1) patient intent interpretation and assistive task planning, 2) safe co-navigation and obstacle adaptation, and 3) multi-agent forecasting and proactive intervention. Patient intent interpretation and assistive task planning are performed by MoMa-LLM [216], which uses the Fetch robot as its robotic platform. MoMa-LLM integrates GPT-4 [145] for high-level task reasoning such as deciding between exploration, navigation, object interaction, and GPT-3.5-turbo-1106 [148] for real-time room classification based on detected objects. This enables the robot to interpret verbal prompts such as "*adjust the bed*" or "*bring the IV pole closer*" into structured multi-step action plans tailored to a patient's bedside context and hospital room layout. Safe co-navigation and obstacle adaptation is addressed in OLiVia-Nav [180] on a Clearpath Jackal robot. OLiVia-Nav integrates a custom VLM (like GPT-4V [157]) offline to generate rich social captions describing human-environment interactions, and uses a lightweight distilled VLM, SC-CLIP [224], during navigation. Multi-agent forecasting and proactive intervention is addressed in [217] on a custom 2-wheel differential drive mobile base with RGB cameras and 2D LiDAR; it leverages GPT-4o [171] for visual prompting to identify pedestrian group formations from RGB imagery, supporting the prediction of future multi-agent interactions. These integrated advances in patient intent

understanding, safe navigation, and future forecasting allow mobile service robots to provide responsive, reliable, and proactive bedside support in hospital environments.

3) Hygiene Maintenance and Infection Control: In hospitals, hygiene maintenance tasks involve disinfecting high-touch surfaces, restocking personal protective equipment (PPE), managing biomedical waste, and monitoring air quality in areas such as ICUs, isolation wards, and operating rooms. Hygiene maintenance and infection control robot tasks decompose into sub-tasks: 1) spatial perception and contamination zone identification, 2) hygiene task grounding and low-level action planning, and 3) real-time PPE and tool handling and proactive environmental monitoring. Spatial perception and contamination zone identification is addressed in AutoRT [218] using the Everyday Robots mobile manipulator platform equipped with a mobile base, manipulator arm, and RGB-D camera integrating MLLM [218] and parses environment descriptions and generate safe, goal-directed manipulation instructions under a robot constitution framework. This enables the robot to autonomously explore hospital environments, identify high-risk contamination zones based on ambiguous natural language queries, and prioritize high-touch surface disinfection critical for infection control. Hygiene task grounding and low-level action planning is addressed in SayPlan [219], deployed on a Franka Emika Panda robotic arm mounted on an Omron LD-60 mobile base which uses the LLM GPT-4 [145] to perform semantic search over hierarchical 3D Scene Graph representations of complex hospital rooms, and iteratively refines long-horizon action plans. This enables the robot to ground abstract instructions such as "*sanitize the entire ICU wing*" into feasible, step-by-step disinfection plans that adapt to multi-room hospital layouts, supporting scalable and context-aware hygiene operations. These integrated advances in contamination zone perception, grounded hygiene planning, and real-time adaptive behavior allow mobile service robots to support autonomous, safe, and proactive hygiene maintenance and infection control across high-risk environments.

C. Service Automation

Service automation tasks involve helping customers, managing events, and navigating dynamic public venues such as malls, airports, and museums. These tasks include customer assistance, guidance, and wayfinding [225]-[228] as well as service setup and maintenance [229]-[231].

1) Customer Assistance, Guidance and Wayfinding: Mobile service robots in public venues such as malls, airports, and museums assist customers by interpreting requests, navigating large and dynamic spaces, and supporting multi-stage guidance under social constraints. These tasks decompose into sub-tasks: 1) language-to-goal spatial grounding, 2) multi-stage waypoint sequencing, and 3) symbolic-spatial planning for real-time urgency-based redirection. Language-to-goal spatial grounding is addressed in OVSG [225], which integrates GPT-3.5 [148] or LLaMA [232] with a Detic-CLIP [135] feature extractor on the ROSMASTER R2 Ackermann steering robot equipped with RGB-D cameras. OVSG constructs open-vocabulary 3D scene graphs by combining object embeddings, spatial relation encoding, and abstract relationship features, enabling robots to

localize free-form queries like "*find the nearest pharmacy across from the information desk*" into precise spatial targets. Multi-stage waypoint sequencing is addressed in LM-Nav [226], which integrates GPT-3 [2], CLIP [135], and ViNG [227] on a Clearpath Jackal robot with an RGB camera and odometry sensors. LM-Nav parses instructions into landmark sequences using GPT-3, visually grounds landmarks using CLIP [135], and constructs a topological graph where ViNG predicts temporal traversability. This enables multi-location navigation under sequential or conditional goals such as "*first go to security, then to Gate A12*" in complex outdoor environments. These advances in semantic grounding, waypoint planning, and symbolic exploration empower mobile service robots to autonomously assist customers across dynamic public venues.

2) Service Setup and Maintenance: Event setup and maintenance tasks involve preparing venues, arranging equipment, adjusting layouts, and cleaning up during conferences, trade shows, or public events. Robots must interpret vague and evolving instructions such as "*set up chairs near the left projector,*" "*tidy this booth area,*" or "*build this structure based on the manual.*" These sub-tasks decompose into: 1) adaptive motion skill chaining, 2) generalized layout planning and setup, and 3) spatial safety validation and manipulation. Adaptive motion skill chaining is addressed in [229] using the Mobile ALOHA robot [207] where the MLLM PaliGemma-3B [212], and the VLA π_0 are combined to interpret commands like "*tidy this booth area*" and sequence low-level motion primitives for dynamic event setup with human feedback in the loop, and online robot task correction. Generalized layout planning and setup are supported in [230] on the AgiBot robot, where the VLM InternVL2.5-2B is integrated with the Genie Operator to generalize across diverse event layouts through latent action space planning. Spatial safety validation and manipulation are addressed in [231] which integrates PG-InstructBLIP [233] a fine-tuned VLM, with the Franka Emika Panda robot to reason about object properties such as fragility, mass, and material. By combining skill sequencing, layout planning, and safety validation, mobile service robots can robustly handle complex event setup and maintenance tasks under real-world conditions.

V. FUTURE RESEARCH DIRECTIONS

The integration of foundation models into mobile service robots has demonstrated promising solutions for challenges in multimodal perception, language-to-action translation, uncertainty estimation, and computational efficiency. However, the following technical research gaps need to be addressed in order for the widespread deployment of these robots in homes, hospitals and public environments. Autonomous and adaptable mobile service robots will require advancements in: 1) scaling laws and data efficiency, 2) autonomous long-term adaptation mechanisms, and 3) cross-embodiment skill generalization. We outline these potential research directions that extend advancements in Section III and the applications in Section IV.

A. Establishing Scaling Laws and Data Efficiency

While transferability and domain adaptation of foundation models have been addressed through large-scale pretraining [13], [234] embodied AI currently lacks predictive scaling laws that can forecast generalization from model size, dataset diversity, or training compute resources [14], [235]. Unlike vision or language tasks, which involve fixed input-output mappings such as image classification or sentence completion, service robotics tasks require embodied tasks that involve dynamic physical interactions with environments and human users. Therefore, simply scaling models or data volume is insufficient for robust out-of-distribution adaptation, because real-world robotics introduces non-linear challenges such as 1) contact dynamics, and 2) semantic ambiguities, not present in traditional datasets. Future work should consider scaling issues related to task diversity, such as contact-rich manipulation involving pushing deformable objects, dynamic obstacle negotiation; and environment complexity involving clutter, occlusions and lighting variations when, for example, sensor modalities themselves are corrupted with false information due to perhaps spills in homes, being knocked over in a busy hospital emergency room or in a large crowd as people navigate around a busy mal. Data collection strategies must also shift from maximizing dataset size alone towards capturing semantic diversity and failure-rich scenarios [236].

B. Autonomous Adaptation for Long-term Deployment

Although advances in uncertainty modeling and robustness under occlusions and noise have been achieved [150], [234], LLMs, VLMs, MLLMs, VLAs are brittle under long-term environmental changes, such as evolving lighting, human behaviors, and object configurations. Most models are benchmarked using static datasets, such as COCO for vision or ImageNet for classification, or constrained domains like RoboTHOR [265], and are not evaluated in dynamic, open-world service environments. Future research should develop scalable continual evaluation frameworks that monitor robot robustness across long-term deployments, lengths of up to several months to even years in settings like hospitals where layouts, tasks, and users change over time. Mobile service robots must autonomously adapt through: 1) self-supervised goal discovery based on novel environmental cues [237], 2) autonomous error detection and corrective policy refinement during task failures [238], and 3) real-time human-in-the-loop reinforcement learning from minimal feedback [239]. Unlike classical continual learning, which targets static classification outputs [240], service robotics must support continual semantic memory updates, object-context drift handling, and dynamic policy adjustment during embodied operation.

C. Cross-Embodiment Generalization

Current AI-embedded service robots remain constrained by 1) limited physical skill acquisition beyond pre-trained distributions [14], 2) imprecise fine-grained motion planning for contact-rich tasks [236], and 3) poor generalization across diverse robot embodiments [241]. Future work should develop skill acquisition pipelines that 1) leverage human demonstration videos [14] and 2) integrate precision refinement modules for trajectory control, grasp adjustment, and dynamic stabilization. These modules must support cross-embodiment generalization, enabling transfer of learned policies across robots with varying kinematics, actuation systems, or sensor modalities. Approaches such as multi-embodiment training, where these foundation models (e.g. LLMs, VLMs, MLLMs, VLAs) are trained on a diverse set of robot morphologies, multi-physics domain adaptation (adapting to different physical interaction dynamics across robots), and semantic-motion decoupling (separating high-level task semantics from low-level motion control) will be critical to achieving hardware-agnostic skill transfer. Addressing cross-embodiment generalization will be essential for deploying foundation models on heterogeneous service robot fleets in homes, hospitals and public spaces.

VI. CONCLUSIONS

This systematic review examined the integration of foundation models into mobile service robotics, focusing on how they address four core challenges: multimodal perception, translation of natural language into executable actions, uncertainty estimation across perception and planning, and computational efficiency for real-time operation. We analyzed how recent advances in LLMs, VLMs, MLLMs, and VLAs provide technical solutions to these challenges, such as cross-modal fusion, instruction grounding, confidence-weighted reasoning, and lightweight scalable architectures. We also discussed real-world applications in domestic assistance, healthcare, and service automation, highlighting how foundation models enable context-aware, socially responsive, and generalizable robot behaviors for service robots. While significant progress has been made towards bridging semantic reasoning and embodied robot actions in real world environments, limitations still remain for achieving reliable and scalable deployment in terms of scaling laws and data efficiency, autonomous long-term adaptation, and cross-embodiment generalization. Research in these areas will allow mobile service robots to safely, adaptively, and collaboratively interact in dynamic human environments to become proactive assistants in everyday life.

REFERENCES

[1]   A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," 2018. mikecaptain.com.

[2]   T. Brown *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 1877–1901. Accessed: Dec. 02, 2024. [Online]. Available: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[3]   OpenAI *et al.*, "GPT-4 Technical Report," Mar. 04, 2024, *arXiv*: arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774.

[4]   H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," Jul. 19, 2023, *arXiv*: arXiv:2307.09288. doi: 10.48550/arXiv. 2307.09288.

[5]   R. Taori *et al.*, "Alpaca: A strong, replicable instruction-following model," *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, vol. 3, no. 6, p. 7, 2023.

[6]   A. Chowdhery *et al.*, "PaLM: Scaling Language Modeling with Pathways," https://doi.org/10.48550/arXiv.2204.02311.

[7]   A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International*

Conference on Machine Learning, PMLR, Jul. 2021, pp. 8748–8763. Accessed: Dec. 02, 2024. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[8]	P. Anderson *et al.*, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 6077–6086. doi: 10.1109/CVPR.2018.00636.

[9]	L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A Simple and Performant Baseline for Vision and Language," Aug. 09, 2019, *arXiv*: arXiv:1908.03557. doi: 10.48550/arXiv.1908.03557.

[10]	A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," Apr. 13, 2022, *arXiv*: arXiv:2204.06125. doi: 10.48550/arXiv.2204.06125.

[11]	H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT Pre-Training of Image Transformers," Sep. 03, 2022, *arXiv*: arXiv:2106.08254. doi: 10.48550/arXiv.2106.08254.

[12]	B. Ichter *et al.*, "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances," in *Proceedings of The 6th Conference on Robot Learning*, PMLR, Mar. 2023, pp. 287–318. Accessed: Dec. 03, 2024. [Online]. Available: https://proceedings.mlr.press/v205/ichter23a.html

[13]	D. Driess *et al.*, "PaLM-E: An Embodied Multimodal Language Model," https://doi.org/10.48550/arXiv.2303.03378.

[14]	A. Brohan *et al.*, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," Jul. 28, 2023, *arXiv*: arXiv:2307.15818. doi: 10.48550/arXiv.2307.15818.

[15]	D. Rivkin *et al.*, "AIoT Smart Home via Autonomous LLM Agents," *IEEE Internet of Things Journal*, 2024, Accessed: Jan. 08, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10729865/?casa_token=weIX6nYaShQAAAAA:h9pD7B1BXVrQJdCC6x5pUy59gtZO_uUig2kXULhD4_dDMaKMsSq7VerxASAHA0_ZyGacFNuu7GI

[16]	M. Giudici, L. Padalino, G. Paolino, I. Paratici, A. I. Pascu, and F. Garzotto, "Designing Home Automation Routines Using an LLM-Based Chatbot," *Designs*, vol. 8, no. 3, p. 43, 2024.

[17]	Y. Li *et al.*, "Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security," May 08, 2024, *arXiv*: arXiv:2401.05459. doi: 10.48550/arXiv.2401.05459.

[18]	A. Pandya, "ChatGPT-Enabled daVinci Surgical Robot Prototype: Advancements and Limitations," *Robotics*, vol. 12, no. 4, Art. no. 4, Aug. 2023, doi: 10.3390/robotics12040097.

[19]	M. A. Salichs *et al.*, "Mini: A New Social Robot for the Elderly," *Int J of Soc Robotics*, vol. 12, no. 6, pp. 1231–1249, Dec. 2020, doi: 10.1007/s12369-020-00687-0.

[20]	A. Paiva, I. Leite, H. Boukricha, and I. Wachsmuth, "Empathy in Virtual Agents and Robots: A Survey," *ACM Trans. Interact. Intell. Syst.*, vol. 7, no. 3, p. 11:1-11:40, Sep. 2017, doi: 10.1145/2912150.

[21]	X. Zhao, M. Li, C. Weber, M. B. Hafez, and S. Wermter, "Chat with the Environment: Interactive Multimodal Perception Using Large Language Models," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2023, pp. 3590–3596. doi: 10.1109/IROS55552.2023.10342363.

[22]	Y. Xia, J. Zhang, N. Jazdi, and M. Weyrich, *Incorporating Large Language Models into Production Systems for Enhanced Task Automation and Flexibility*. 2024. doi: 10.51202/9783181024379.

[23]	M. Fakih, R. Dharmaji, Y. Moghaddas, G. Quiros, O. Ogundare, and M. A. Al Faruque, "LLM4PLC: Harnessing Large Language Models for Verifiable Programming of PLCs in Industrial Control Systems," in *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*, Lisbon Portugal: ACM, Apr. 2024, pp. 192–203. doi: 10.1145/3639477.3639743.

[24]	Z. Wang and H. Qin, "Intelligent industrial production process automatic regulation system based on LLM agents," in *2024 5th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)*, IEEE, 2024, pp. 133–137. Accessed: Jan. 08, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10692701/?casa_token=4yvprI8IlIYAAAAA:F_gFkYWq8pfd0Mjk8Czjp0X9QwOUc7b1WSNVQwFjy3lZxoybatiI35ly-uuIq9UaM3n-JbHjgmM

[25]	W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022, pp. 9118–9147. Accessed: Dec. 12, 2024. [Online]. Available: https://proceedings.mlr.press/v162/huang22a.html

[26]	W. Huang *et al.*, "Inner Monologue: Embodied Reasoning through Planning with Language Models," Jul. 12, 2022, *arXiv*: arXiv:2207.05608. doi: 10.48550/arXiv.2207.05608.

[27]	M. Shridhar, L. Manuelli, and D. Fox, "CLIPort: What and Where Pathways for Robotic Manipulation," in *Proceedings of the 5th Conference on Robot Learning*, PMLR, Jan. 2022, pp. 894–906. Accessed: Dec. 03, 2024. [Online]. Available: https://proceedings.mlr.press/v164/shridhar22a.html

[28]	A. Narcomey, N. Tsoi, R. Desai, and M. Vázquez, "Learning Human Preferences Over Robot Behavior as Soft Planning Constraints," Mar. 28, 2024, *arXiv*: arXiv:2403.19795. doi: 10.48550/arXiv.2403.19795.

[29]	Z. Cao, Z. Wang, S. Xie, A. Liu, and L. Fan, "Smart Help: Strategic Opponent Modeling for Proactive and Adaptive Robot Assistance in Households," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2024, pp. 18091–18101. doi: 10.1109/CVPR52733.2024.01713.

[30]	A. Xiao, A. Gupta, Y. Deng, K. Li, and D. Hsu, "Robi Butler: Multimodal Remote Interaction with Household Robotic Assistants," presented at the 2nd Workshop on Mobile Manipulation and Embodied Intelligence at ICRA 2024, May 2024. Accessed: Jan. 22, 2025. [Online]. Available: https://openreview.net/forum?id=FH03VfMqOR

[31]	Y. Cao *et al.*, "CogNav: Cognitive Process Modeling for Object Goal Navigation with LLMs," Dec. 11, 2024, *arXiv*: arXiv:2412.10439. doi: 10.48550/arXiv.2412.10439.

[32]	X. Puig *et al.*, "Habitat 3.0: A Co-Habitat for Humans, Avatars and Robots," Oct. 19, 2023, *arXiv*: arXiv:2310.13724. doi: 10.48550/arXiv.2310.13724.

[33]	A. Eftekhar *et al.*, "The One RING : a Robotic Indoor Navigation Generalist", DOI:10.48550/arXiv.2412.14401.

[34]	J. Hu *et al.*, "FLaRe: Achieving Masterful and Adaptive Robot Policies with Large-Scale Reinforcement Learning Fine-Tuning," Sep. 30, 2024, *arXiv*: arXiv:2409.16578. doi: 10.48550/arXiv.2409.16578.

[35]	H. Fan, X. Liu, J. Y. H. Fuh, W. F. Lu, and B. Li, "Embodied intelligence in manufacturing: leveraging large language models for autonomous industrial robotics," *J Intell Manuf*, vol. 36, no. 2, pp. 1141–1157, Feb. 2025, doi: 10.1007/s10845-023-02294-y.

[36]	G. Salierno, L. Leonardi, and G. Cabri, "Generative AI and Large Language Models in Industry 5.0: Shaping Smarter Sustainable Cities," *Encyclopedia*, vol. 5, no. 1, , Mar. 2025, doi: 10.3390/encyclopedia5010030.

[37]	S. Li *et al.*, "RoboNurse-VLA: Robotic Scrub Nurse System based on Vision-Language-Action Model," Sep. 29, 2024, *arXiv*: arXiv:2409.19590. doi: 10.48550/arXiv.2409.19590.

[38]	J. Wu *et al.*, "TidyBot: personalized robot assistance with large language models," *Auton Robot*, vol. 47, no. 8, pp. 1087–1102, Dec. 2023, doi: 10.1007/s10514-023-10139-z.

[39]	Z. Zhao, H. Tang, and Y. Yan, "Audio-Visual Navigation with Anti-Backtracking," in *Pattern Recognition*, vol. 15318, A. Antonacopoulos, S. Chaudhuri, R. Chellappa, C.-L. Liu, S. Bhattacharya, and U. Pal, Eds., in Lecture Notes in Computer Science, vol. 15318. , Cham: Springer Nature Switzerland, 2025, pp. 358–372. doi: 10.1007/978-3-031-78456-9_23.

[40]	C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture," in *Computer Vision – ACCV 2016*, vol. 10111, S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds., in Lecture Notes in Computer Science, vol. 10111. , Cham: Springer International Publishing, 2017, pp. 213–228. doi: 10.1007/978-3-319-54181-5_14.

[41]	D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 167, pp. 12–23, Sep. 2020, doi: 10.1016/j.isprsjprs.2020.06.014.

[42]	D. Zuñiga-Noël, J.-R. Ruiz-Sarmiento, R. Gomez-Ojeda, and J. Gonzalez-Jimenez, "Automatic Multi-Sensor Extrinsic Calibration For Mobile Robots," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2862–2869, Jul. 2019, doi: 10.1109/LRA.2019.2922618.

[43]	A. Jaegle *et al.*, "Perceiver IO: A General Architecture for Structured Inputs & Outputs," Mar. 15, 2022, *arXiv*: arXiv:2107.14795. doi: 10.48550/arXiv.2107.14795.

[44]	B. Xia, J. Zhou, F. Kong, Y. You, J. Yang, and L. Lin, "Enhancing 3D object detection through multi-modal fusion for cooperative perception," *Alexandria Engineering Journal*, vol. 104, pp. 46–55, Oct. 2024, doi: 10.1016/j.aej.2024.06.025.

[45]	R. E. Fikes and N. J. Nilsson, "STRIPS: A new approach to the application of theorem proving to problem solving," *Artificial intelligence*, vol. 2, no. 3–4, pp. 189–208, 1971.

[46]	D. S. Nau *et al.*, "SHOP2: An HTN planning system," *Journal of artificial intelligence research*, vol. 20, pp. 379–404, 2003.

[47] D. Nau *et al.*, "Applications of SHOP and SHOP2," *IEEE Intelligent Systems*, vol. 20, no. 2, pp. 34–41, Mar. 2005, doi: 10.1109/MIS.2005.20.

[48] M. Fox and D. Long, "PDDL2. 1: An extension to PDDL for expressing temporal planning domains," *Journal of artificial intelligence research*, vol. 20, pp. 61–124, 2003.

[49] M. Ghallab, D. Nau, and P. Traverso, *Automated Planning: theory and practice*. Elsevier, 2004. Accessed: Apr. 08, 2025. [Online]. Available: https://books.google.ca/books?hl=en&lr=&id=uYnpze57MSgC&oi=fnd&pg=PP1&dq=Ghallab,+M.,+Nau,+D.,+%26+Traverso,+P.+%22Automated+Planning:+Theory+and+Practice.%22Elsevier,+2004.&ots=XoMVM3NHTe&sig=3kILRU74sX6R9HDBGa4Si5d6JNo

[50] D. K. Misra, J. Sung, K. Lee, and A. Saxena, "Tell me Dave: Context-sensitive grounding of natural language to manipulation instructions," *The International Journal of Robotics Research*, vol. 35, no. 1–3, pp. 281–300, Jan. 2016, doi: 10.1177/0278364915602060.

[51] J. MacGlashan *et al.*, "Grounding English Commands to Reward Functions," in *Robotics: Science and Systems XI*, Robotics: Science and Systems Foundation, Jul. 2015. doi: 10.15607/RSS.2015.XI.018.

[52] D. Chen and R. Mooney, "Learning to interpret natural language navigation instructions from observations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2011, pp. 859–865. Accessed: Apr. 2025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/7974

[53] D. Bohus and A. I. Rudnicky, "The RavenClaw dialog management framework: Architecture and systems," *Computer Speech & Language*, vol. 23, no. 3, pp. 332–361, 2009.

[54] M. Colledanchise and P. Ögren, "How Behavior Trees Modularize Hybrid Control Systems and Generalize Sequential Behavior Compositions, the Subsumption Architecture, and Decision Trees," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 372–389, Apr. 2017, doi: 10.1109/TRO.2016.2633567.

[55] S. Young, M. Gašić, B. Thomson, and J. D. Williams, "Pomdp-based statistical spoken dialog systems: A review," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013.

[56] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," https://doi.org/10.48550/arXiv.1506.02142.

[57] M. Beetz, M. Tenorth, D. Jain, and J. Bandouch, "Towards automated models of activities of daily life," *Technology and Disability*, vol. 22, no. 1–2, pp. 27–40, Feb. 2010, doi: 10.3233/TAD-2010-0285.

[58] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-End Training of Deep Visuomotor Policies," Journal of Machine Learning Research, vol. 17, no. 39, pp. 1–40, 2016. [Online]. Available: https://www.jmlr.org/papers/volume17/15-522/15-522.pdf

[59] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel, "Benchmarking Deep Reinforcement Learning for Continuous Control," in Proceedings of the 33rd International Conference on Machine Learning (ICML), vol. 48, New York, NY, USA: PMLR, 2016, pp. 1329–1338. [Online]. Available: https://proceedings.mlr.press/v48/duan16.html

[60] A. Szot *et al.*, "Habitat 2.0: Training Home Assistants to Rearrange their Habitat," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, pp. 251–266. Accessed: Apr. 28, 2025. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/021bbc7ee20b71134d53e20206bd6feb-Abstract.html

[61] Y. Hu *et al.*, "Toward General-Purpose Robots via Foundation Models: A Survey and Meta-Analysis," Oct. 01, 2024, *arXiv*: arXiv:2312.08782. doi: 10.48550/arXiv.2312.08782.

[62] H. Zhou *et al.*, "Bridging Language and Action: A Survey of Language-Conditioned Robot Manipulation," Dec. 02, 2024, *arXiv*: arXiv:2312.10807. doi: 10.48550/arXiv.2312.10807.

[63] R. Firoozi *et al.*, "Foundation models in robotics: Applications, challenges, and the future," *The International Journal of Robotics Research*, p. 02783649241281508, Sep. 2024, doi: 10.1177/02783649241281508.

[64] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[65] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "OpenVINS: A Research Platform for Visual-Inertial Estimation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, May 2020, pp. 4666–4672. doi: 10.1109/ICRA40945.2020.9196524.

[66] J. Rehder, R. Siegwart, and P. Furgale, "A General Approach to Spatiotemporal Calibration in Multisensor Systems," *IEEE Transactions on Robotics*, vol. 32, no. 2, pp. 383–398, Apr. 2016, doi: 10.1109/TRO.2016.2529645.

[67] X. Zuo, P. Geneva, W. Lee, Y. Liu, and G. Huang, "LIC-Fusion: LiDAR-Inertial-Camera Odometry," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov. 2019, pp. 5848–5854. doi: 10.1109/IROS40897.2019.8967746.

[68] C. Cadena *et al.*, "Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016, doi: 10.1109/TRO.2016.2624754.

[69] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017, doi: 10.1109/TRO.2017.2705103.

[70] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-Scale Direct Monocular SLAM," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 834–849. doi: 10.1007/978-3-319-10605-2_54.

[71] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-Manifold Preintegration for Real-Time Visual–Inertial Odometry," *IEEE Transactions on Robotics*, vol. 33, no. 1, pp. 1–21, Feb. 2017, doi: 10.1109/TRO.2016.2597321.

[72] T. Lupton and S. Sukkarieh, "Visual-Inertial-Aided Navigation for High-Dynamic Motion in Built Environments Without Initial Conditions," *IEEE Transactions on Robotics*, vol. 28, no. 1, pp. 61–76, Feb. 2012, doi: 10.1109/TRO.2011.2170332.

[73] J. Gawlikowski *et al.*, "A survey of uncertainty in deep neural networks," *Artif Intell Rev*, vol. 56, no. 1, pp. 1513–1589, Oct. 2023, doi: 10.1007/s10462-023-10562-9.

[74] J. H. Jung, Y. Choe, and C. G. Park, "Photometric Visual-Inertial Navigation With Uncertainty-Aware Ensembles," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2039–2052, Aug. 2022..

[75] N. Sünderhauf *et al.*, "The limits and potentials of deep learning for robotics," *The International Journal of Robotics Research*, vol. 37, no. 4–5, pp. 405–420, Apr. 2018, doi: 10.1177/0278364918770733.

[76] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, Jan. 2017, doi: 10.1177/0278364916679498.

[77] J. Tremblay *et al.*, "Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 1082–10828. doi: 10.1109/CVPRW.2018.00143.

[78] B. Sofman, E. Lin, J. A. Bagnell, J. Cole, N. Vandapel, and A. Stentz, "Improving robot navigation through self-supervised online learning," *Journal of Field Robotics*, vol. 23, no. 11–12, pp. 1059–1075, 2006, doi: 10.1002/rob.20169.

[79] H. Porav, W. Maddern, and P. Newman, "Adversarial Training for Adverse Conditions: Robust Metric Localisation Using Appearance Transfer," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 1011–1018. doi: 10.1109/ICRA.2018.8462894.

[80] M. Görner, R. Haschke, H. Ritter, and J. Zhang, "MoveIt! Task Constructor for Task-Level Motion Planning," in *2019 Int. Conf. on Robotics and Automation*, May 2019, pp. 190–196. doi: 10.1109/ICRA.2019.8793898.

[81] S. Tellex *et al.*, "Approaching the symbol grounding problem with probabilistic graphical models," *AI magazine*, vol. 32, no. 4, pp. 64–76, 2011.

[82] T. Williams, R. Cantrell, G. Briggs, P. Schermerhorn, and M. Scheutz, "Grounding natural language references to unvisited and hypothetical locations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2013, pp. 947–953. Accessed: Apr. 08, 2025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/8563

[83] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2010, pp. 259–266. Accessed: Apr., 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5453186/?casa_token=bIJQBwyDGMAAAAAA:4dZ1SSreI2XvNN3dnPQsmD7Vo58XWdsYPYqwFAJ4v0xXP7WaDp0oLNM2x3Xm8v5TLWJHjOB

[84] M. R. Walter *et al.*, "Language Understanding for Field and Service Robots in a Priori Unknown Environments," *Field Robotics*, vol. 2, pp. 1191–1231, Jun. 2022, doi: 10.55417/fr.2022040.

[85] R. Paul, J. Arkin, N. Roy, and T. M. Howard, "Grounding Abstract Spatial Concepts for Language Interaction with Robots," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, Australia: International Joint Conferences on Artificial Intelligence Organization, Aug. 2017, pp. 4929–4933. doi: 10.24963/ijcai.2017/696.

[86] L. P. Kaelbling and T. Lozano-Pérez, "Hierarchical task and motion planning in the now," in *2011 IEEE International Conference on Robotics and Automation*, May 2011, pp. 1470–1477. doi: 10.1109/ICRA.2011.5980391.

[87] F. Mohr, T. Lettmann, and E. Hüllermeier, "Planning with Independent Task Networks," in *KI 2017: Advances in Artificial Intelligence*, vol. 10505, G. Kern-Isberner, J. Fürnkranz, and M. Thimm, Eds., in Lecture Notes in Computer Science, vol. 10505. , Cham: Springer International Publishing, 2017, pp. 193–206. doi: 10.1007/978-3-319-67190-1_15.

[88] J. Hoffmann, "FF: The fast-forward planning system," *AI magazine*, vol. 22, no. 3, pp. 57–57, 2001.

[89] M. Helmert, "The fast downward planning system," *Journal of Artificial Intelligence Research*, vol. 26, pp. 191–246, 2006.

[90] C. Dornhege, P. Eyerich, T. Keller, S. Trüg, M. Brenner, and B. Nebel, "Semantic Attachments for Domain-Independent Planning Systems," in *Towards Service Robots for Everyday Environments: Recent Advances in Designing Service Robots for Complex Tasks in Everyday Environments*, E. Prassler, M. Zöllner, R. Bischoff, W. Burgard, R. Haschke, M. Hägele, G. Lawitzky, B. Nebel, P. Plöger, and U. Reiser, Eds., Berlin, Heidelberg: Springer, 2012, pp. 99–115. doi: 10.1007/978-3-642-25116-0_9.

[91] W. Mao, R. Desai, M. L. Iuzzolino, and N. Kamra, "Action Dynamics Task Graphs for Learning Plannable Representations of Procedural Tasks," Jan. 11, 2023, *arXiv*: arXiv:2302.05330. doi: 10.48550/arXiv.2302.05330.

[92] P.-L. Bacon, J. Harb, and D. Precup, "The option-critic architecture," in *Proceedings of the AAAI conference on artificial intelligence*, 2017. Accessed: Apr. 09, 2025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/10916

[93] H. Kurniawati, Y. Du, D. Hsu, and W. S. Lee, "Motion planning under uncertainty for robotic tasks with long time horizons," *The International Journal of Robotics Research*, vol. 30, no. 3, pp. 308–323, Mar. 2011, doi: 10.1177/0278364910386986.

[94] M. Lauri, D. Hsu, and J. Pajarinen, "Partially Observable Markov Decision Processes in Robotics: A Survey," *IEEE Transactions on Robotics*, vol. 39, no. 1, pp. 21–40, Feb. 2023, doi: 10.1109/TRO.2022.3200138.

[95] D. Silver and J. Veness, "Monte-Carlo Planning in Large POMDPs," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2010. Accessed: Apr. 16, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2010/hash/edfbe1afcf9246bb0d40eb4d8027d90f-Abstract.html

[96] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Apr. 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html

[97] M. G. Mohanan and A. Salgoankar, "A survey of robotic motion planning in dynamic environments," *Robotics and Autonomous Systems*, vol. 100, pp. 171–185, Feb. 2018, doi: 10.1016/j.robot.2017.10.011.

[98] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*, PMLR, 2017, pp. 1321–1330. Accessed: Apr. 24, 2025. [Online]. Available: http://proceedings.mlr.press/v70/guo17a.html

[99] D. Simon, "Kalman filtering," *Embedded systems programming*, vol. 14, no. 6, pp. 72–79, 2001.

[100] M. Khodarahmi and V. Maihami, "A review on Kalman filter models," *Archives of Computational Methods in Engineering*, vol. 30, no. 1, pp. 727–747, 2023.

[101] H. Kurniawati, D. Hsu, and W. S. Lee, "Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces.," in *Robotics: Science and systems*, Citeseer, 2008. Accessed: Apr. 09, 2025. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=dbe2062d6f7603f87c2c5d9e4015e29c0dd19ca6

[102] R. P. Jr, R. Tedrake, L. Kaelbling, and T. Lozano-Perez, "Belief space planning assuming maximum likelihood observations," http://hdl.handle.net/1721.1/62571.

[103] E. A. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373)*, Oct. 2000, pp. 153–158. doi: 10.1109/ASSPCC.2000.882463.

[104] R. Van der Merwe and E. A. Wan, "The square-root unscented Kalman filter for state and parameter-estimation," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, May 2001, pp. 3461–3464 vol.6.

[105] J. D. Brouk and K. J. DeMars, "Kalman Filtering with Uncertain and Asynchronous Measurement Epochs," *NAVIGATION: Journal of the Institute of Navigation*, vol. 71, no. 3, Sep. 2024, doi: 10.33012/navi.652.

[106] C. Naab and Z. Zheng, "Application of the unscented Kalman filter in position estimation a case study on a robot for precise positioning," *Robotics and Autonomous Systems*, vol. 147, p. 103904, Jan. 2022, doi: 10.1016/j.robot.2021.103904.

[107] R. Platt Jr, R. Tedrake, L. Kaelbling, and T. Lozano-Perez, "Belief space planning assuming maximum likelihood observations," 2010, Accessed: Apr. 9, 2025. [Online]. Available: https://dspace.mit.edu/handle/1721.1/62571

[108] J. Leusmann, C. Wang, M. Gienger, A. Schmidt, and S. Mayer, "Understanding the Uncertainty Loop of Human-Robot Interaction," Mar. 14, 2023, *arXiv*: arXiv:2303.07889. doi: 10.48550/arXiv.2303.07889.

[109] R. Cumbal, J. Lopes, and O. Engwall, "Uncertainty in Robot Assisted Second Language Conversation Practice," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, in HRI '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 171–173. doi: 10.1145/3371382.3378306.

[110] J. Hough and D. Schlangen, "It's Not What You Do, It's How You Do It: Grounding Uncertainty for a Simple Robot," in *2017 12th ACM/IEEE Int. Conf. on Human-Robot Interaction*, Mar. 2017, pp. 274–282. Accessed: Apr. 2025. [Online]. Available: https://ieeexplore.ieee.org/document/8534946/

[111] S. Trick, D. Koert, J. Peters, and C. A. Rothkopf, "Multimodal Uncertainty Reduction for Intention Recognition in Human-Robot Interaction," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov. 2019, pp. 7009–7016. doi: 10.1109/IROS40897.2019.8968171.

[112] T. Williams, R. Cantrell, G. Briggs, P. Schermerhorn, and M. Scheutz, "Grounding natural language references to unvisited and hypothetical locations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2013, pp. 947–953. Accessed: Apr. 24, 2025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/8563

[113] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[114] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3D Graph Neural Networks for RGBD Semantic Segmentation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 5209–5218. doi: 10.1109/ICCV.2017.556.

[115] T. Zhou, F. Porikli, D. J. Crandall, L. Van Gool, and W. Wang, "A survey on deep learning technique for video segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 6, pp. 7099–7122, 2022.

[116] S. Quinlan and O. Khatib, "Elastic bands: connecting path planning and control," in *[1993] Proceedings IEEE International Conference on Robotics and Automation*, May 1993, pp. 802–807 vol.2. doi: 10.1109/ROBOT.1993.291936.

[117] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robotics & Automation Magazine*, vol. 4, no. 1, pp. 23–33, Mar. 1997, doi: 10.1109/100.580977.

[118] Y. Zhu *et al.*, "Visual Semantic Planning Using Deep Successor Representations," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 483–492. doi: 10.1109/ICCV.2017.60.

[119] H.-B. Zhang *et al.*, "A Comprehensive Survey of Vision-Based Human Action Recognition Methods," *Sensors*, vol. 19, no. 5, Art. no. 5, Jan. 2019, doi: 10.3390/s19051005.

[120] S. Pütz, J. S. Simón, and J. Hertzberg, "Move base flex a highly flexible navigation framework for mobile robots," in *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, IEEE, 2018, pp. 3416–3421. Accessed: Apr. 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8593829/?casa_token=kovb94KFJ8oAAAAA:dnpBLyjWVwX1qp6pE3e9DycA7hjsbDdGQS1_4mCEOAUV13qWtJC5KV5KTiiRc2sCRL4eZpng

[121] R. C. Arkin, *Behavior-based robotics*. MIT press, 1998. Accessed: Apr. 28, 2025. [Online]. Available: https://books.google.ca/books?hl=en&lr=&id=mRWT6alZt9oC&oi=fnd&pg=PR11&dq=behavior+based+robotics+ronald&ots=47YsjiOcpE&sig=57BnE1kGPyL6dNtoRSESTo5JJ8s

[122] F. Noroozi, M. Daneshmand, and P. Fiorini, "Conventional, Heuristic and Learning-Based Robot Motion Planning: Reviewing Frameworks of Current Practical Significance," *Machines*, vol. 11, no. 7, Art. no. 7, Jul. 2023, doi: 10.3390/machines11070722.

[123] Dr. A. Pandey, "Mobile Robot Navigation and Obstacle Avoidance Techniques: A Review," *International Robotics & Automation Journal*, vol. 2,

pp. 1–12, May 2017, doi: 10.15406/iratj.2017.02.00023.

[124]   R. C. Arkin, *Behavior-based Robotics*. MIT Press, 1998.

[125]   F. L. Lewis and S. S. Ge, *Autonomous Mobile Robots: Sensing, Control, Decision Making and Applications*. CRC Press, 2018.

[126]   X. Guo, M. Lyu, B. Xia, K. Zhang, and L. Zhang, "An Improved Visual SLAM Method with Adaptive Feature Extraction," *Applied Sciences*, vol. 13, no. 18, Art. no. 18, Jan. 2023, doi: 10.3390/app131810038.

[127]   G. Kurz, M. Holoch, and P. Biber, "Geometry-based Graph Pruning for Lifelong SLAM," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2021, pp. 3313–3320. doi: 10.1109/IROS51168.2021.9636530.

[128]   T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison, "ElasticFusion: Dense SLAM without a pose graph.," in *Robotics: science and systems*, Rome, 2015, p. 3. Accessed: Apr. 09, 2025. [Online]. Available: https://roboticsproceedings.org/rss11/p01.pdf

[129]   A. Murali, W. Liu, K. Marino, S. Chernova, and A. Gupta, "Same Object, Different Grasps: Data and Semantic Knowledge for Task-Oriented Grasping," in *Proceedings of the 2020 Conference on Robot Learning*, PMLR, Oct. 2021, pp. 1540–1557. Accessed: Apr. 24, 2025. [Online]. Available: https://proceedings.mlr.press/v155/murali21a.html

[130]   S. Mayya *et al.*, "Adaptive and Risk-Aware Target Tracking for Robot Teams With Heterogeneous Sensors," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5615–5622, Apr. 2022, doi: 10.1109/LRA.2022.3155805.

[131]   J. Wang, S. Lin, and A. Liu, "Bioinspired Perception and Navigation of Service Robots in Indoor Environments: A Review," *Biomimetics*, vol. 8, no. 4, Art. no. 4, Aug. 2023, doi: 10.3390/biomimetics8040350.

[132]   A. Team *et al.*, "Aether: Geometric-Aware Unified World Modeling," Mar. 25, 2025, *arXiv*: arXiv:2503.18945. doi: 10.48550/arXiv.2503.18945.

[133]   J. Li, J. Wu, W. Zhao, S. Bai, and X. Bai, "PartGLEE: A Foundation Model for Recognizing and Parsing Any Objects," Jul. 23, 2024, *arXiv*: arXiv:2407.16696. doi: 10.48550/arXiv.2407.16696.

[134]   J. Yang *et al.*, "Magma: A Foundation Model for Multimodal AI Agents," Feb. 18, 2025, *arXiv*: arXiv:2502.13130. doi: 10.48550/arXiv.2502.13130.

[135]   A. Radford *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *Proceedings of the Int. Conference on Machine Learning*, PMLR, Jul. 2021, pp. 8748–8763. Accessed: Apr. 22, 2025. [Online]. Available: https://proceedings.mlr.press/v139/ radford21a.html

[136]   H. Diao, Y. Cui, X. Li, Y. Wang, H. Lu, and X. Wang, "Unveiling Encoder-Free Vision-Language Models," Oct. 29, 2024, *arXiv*: arXiv:2406.11832. doi: 10.48550/arXiv.2406.11832.

[137]   H. Chefer *et al.*, "VideoJAM: Joint Appearance-Motion Representations for Enhanced Motion Generation in Video Models," Feb. 04, 2025, *arXiv*: arXiv:2502.02492. doi: 10.48550/arXiv.2502.02492.

[138]   Y. Wang *et al.*, "InternVideo2: Scaling Foundation Models for Multimodal Video Understanding," Aug. 14, 2024, *arXiv*: arXiv:2403.15377. doi: 10.48550/arXiv.2403.15377.

[139]   X. Chu *et al.*, "MobileVLM V2: Faster and Stronger Baseline for Vision Language Model," Feb. 06, 2024, *arXiv*: arXiv:2402.03766. doi: 10.48550/arXiv.2402.03766.

[140]   M. Cho *et al.*, "Cocoon: Robust Multi-Modal Perception with Uncertainty-Aware Sensor Fusion," Oct. 16, 2024, *arXiv*: arXiv:2410.12592. doi: 10.48550/arXiv.2410.12592.

[141]   DeepSeek-AI *et al.*, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," Jan. 22, 2025, *arXiv*: arXiv:2501.12948. doi: 10.48550/arXiv.2501.12948.

[142]   A. Kirillov *et al.*, "Segment Anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France: IEEE, Oct. 2023, pp. 3992–4003. doi: 10.1109/ICCV51070.2023.00371.

[143]   M. Oquab *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," Feb. 02, 2024, *arXiv*: arXiv:2304.07193. doi: 10.48550/arXiv.2304.07193.

[144]   T. Ren *et al.*, "DINO-X: A Unified Vision Model for Open-World Object Detection and Understanding," Dec. 06, 2024, *arXiv*: arXiv:2411.14347. doi: 10.48550/arXiv.2411.14347.

[145]   OpenAI *et al.*, "GPT-4 Technical Report," Mar. 04, 2024, *arXiv*: arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774.

[146]   H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2024. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

[147]   S. Tong *et al.*, "Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs," Dec. 04, 2024, *arXiv*: arXiv:2406.16860. doi: 10.48550/arXiv.2406.16860.

[148]   J. Ye *et al.*, "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models," Dec. 23, 2023, *arXiv*: arXiv:2303.10420. doi: 10.48550/arXiv.2303.10420.

[149]   P. Wang *et al.*, "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," Oct. 03, 2024, *arXiv*: arXiv:2409.12191. doi: 10.48550/arXiv.2409.12191.

[150]   A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 8748–8763. Accessed: Apr. 27, 2025. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[151]   Z. Chen *et al.*, "How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites," Apr. 29, 2024, *arXiv*: arXiv:2404.16821. doi: 10.48550/arXiv.2404.16821.

[152]   NVIDIA *et al.*, "Cosmos World Foundation Model Platform for Physical AI," Mar. 18, 2025, *arXiv*: arXiv:2501.03575. doi: 10.48550/arXiv.2501.03575.

[153]   Y. Ye, Z. Huang, Y. Xiao, E. Chern, S. Xia, and P. Liu, "LIMO: Less is More for Reasoning," Feb. 05, 2025, *arXiv*: arXiv:2502.03387. doi: 10.48550/arXiv.2502.03387.

[154]   P. Zhang *et al.*, "InternLM-XComposer: A Vision-Language Large Model for Advanced Text-image Comprehension and Composition," Dec. 14, 2023, *arXiv*: arXiv:2309.15112. doi: 10.48550/arXiv.2309.15112.

[155]   Q. Gu *et al.*, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2024, pp. 5021–5028. Accessed: Dec. 03, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10610243/?casa_token=p-TNeEXvx6IAAAAA:VLXE8nCl4UTIOTyBG9oARwDWE6bouha1QNK1ea90XoEGAn5p95EQTOklctksMizpgRFiU3lr

[156]   C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020. [Online].

[157]   Z. Yang *et al.*, "The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)," Oct. 11, 2023, *arXiv*: arXiv:2309.17421. doi: 10.48550/arXiv.2309.17421.

[158]   A. Défossez *et al.*, "Moshi: a speech-text foundation model for real-time dialogue," Oct. 02, 2024, *arXiv*: arXiv:2410.00037. doi: 10.48550/arXiv.2410.00037.

[159]   G. Zuo, J. Tong, H. Liu, W. Chen, and J. Li, "Graph-based Visual Manipulation Relationship Reasoning in Object-Stacking Scenes," in *2021 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2021, pp. 1–8. doi: 10.1109/IJCNN52387.2021.9534389.

[160]   X. Chen *et al.*, "PaLI-X: On Scaling up a Multilingual Vision and Language Model," May 29, 2023, *arXiv*: arXiv:2305.18565. doi: 10.48550/arXiv.2305.18565.

[161]   J. Liang *et al.*, "Code as Policies: Language Model Programs for Embodied Control," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023, pp. 9493–9500. doi: 10.1109/ICRA48891.2023.10160591.

[162]   C. H. Song, B. M. Sadler, J. Wu, W.-L. Chao, C. Washington, and Y. Su, "LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France: IEEE, Oct. 2023, pp. 2986–2997. doi: 10.1109/ICCV51070.2023.00280.

[163]   N. Muennighoff *et al.*, "s1: Simple test-time scaling," Mar. 01, 2025, *arXiv*: arXiv:2501.19393. doi: 10.48550/arXiv.2501.19393.

[164]   L. P. Kaelbling and T. Lozano-Pérez, "Hierarchical task and motion planning in the now," in *2011 IEEE International Conference on Robotics and Automation*, May 2011, pp. 1470–1477. doi: 10.1109/ICRA.2011.5980391.

[165]   X. Zhou, S. Gandhi, L. Fan, Z. Lin, Y. Du, P. Abbeel, J. Wu, and F. Xia, "GENESIS: A Generative and Universal Physics Engine for Robotics and Beyond," arXiv preprint arXiv:2401.01454, 2024. [Online]. Available: https://genesis-embodied-ai.github.io/

[166]   Q. Li, Y. Lin, Q. Luo, and L. Yu, "DreamerV3 for Traffic Signal Control: Hyperparameter Tuning and Performance," Mar. 04, 2025, *arXiv*: arXiv:2503.02279. doi: 10.48550/arXiv.2503.02279.

[167] M. J. Kim *et al.*, "OpenVLA: An Open-Source Vision-Language-Action Model," Sep. 05, 2024, *arXiv*: arXiv:2406.09246. doi: 10.48550/arXiv.2406.09246.

[168] M. Abdin *et al.*, "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone," Aug. 30, 2024, *arXiv*: arXiv:2404.14219. doi: 10.48550/arXiv.2404.14219.

[169] A. Jaegle *et al.*, "Perceiver IO: A General Architecture for Structured Inputs & Outputs," Mar. 15, 2022, *arXiv*: arXiv:2107.14795. doi: 10.48550/arXiv.2107.14795.

[170] H. Li *et al.*, "Uni-Perceiver v2: A Generalist Model for Large-Scale Vision and Vision-Language Tasks," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 2691–2700. doi: 10.1109/CVPR52729.2023.00264.

[171] OpenAI *et al.*, "GPT-4o System Card," Oct. 25, 2024, *arXiv*: arXiv:2410.21276. doi: 10.48550/arXiv.2410.21276.

[172] A. Awadalla *et al.*, "OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models," Aug. 07, 2023, *arXiv*: arXiv:2308.01390. doi: 10.48550/arXiv.2308.01390.

[173] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to Control: Learning Behaviors by Latent Imagination," Mar. 17, 2020, *arXiv*: arXiv:1912.01603. doi: 10.48550/arXiv.1912.01603.

[174] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba, "Mastering Atari with Discrete World Models," Feb. 12, 2022, *arXiv*: arXiv:2010.02193. doi: 10.48550/arXiv.2010.02193.

[175] Q. Li, Y. Lin, Q. Luo, and L. Yu, "DreamerV3 for Traffic Signal Control: Hyperparameter Tuning and Performance," Mar. 04, 2025, *arXiv*: arXiv:2503.02279. doi: 10.48550/arXiv.2503.02279.

[176] C. Feichtenhofer, H. Fan, Y. Li, and K. He, "Masked Autoencoders As Spatiotemporal Learners," Oct. 21, 2022, *arXiv*: arXiv:2205.09113. doi: 10.48550/arXiv.2205.09113.

[177] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.

[178] O. Bar-Tal *et al.*, "Lumiere: A Space-Time Diffusion Model for Video Generation," in *SIGGRAPH Asia 2024 Conference Papers*, in SA '24. New York, NY, USA: Association for Computing Machinery, Dec. 2024, pp. 1–11. doi: 10.1145/3680528.3687614.

[179] D. Song, J. Liang, A. Payandeh, A. H. Raj, X. Xiao, and D. Manocha, "VLM-Social-Nav: Socially Aware Robot Navigation Through Scoring Using Vision-Language Models," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 508–515, Jan. 2025, doi: 10.1109/LRA.2024.3511409.

[180] S. Narasimhan, A. H. Tan, D. Choi, and G. Nejat, "OLiVia-Nav: An Online Lifelong Vision Language Approach for Mobile Robot Social Navigation," Mar. 08, 2025, *arXiv*: arXiv:2409.13675. doi: 10.48550/arXiv.2409.13675.

[181] Y. Huang, Q. Zhang, P. S. Y, and L. Sun, "TrustGPT: A Benchmark for Trustworthy and Responsible Large Language Models," Jun. 20, 2023, *arXiv*: arXiv:2306.11507. doi: 10.48550/arXiv.2306.11507.

[182] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," Jan. 15, 2021, *arXiv*: arXiv:2012.12877. doi: 10.48550/arXiv.2012.12877.

[183] H. Touvron, M. Cord, and H. Jégou, "DeiT III: Revenge of the ViT," Apr. 14, 2022, *arXiv*: arXiv:2204.07118. doi: 10.48550/arXiv.2204.07118.

[184] X. Chu, J. Su, B. Zhang, and C. Shen, "VisionLLaMA: A Unified LLaMA Backbone for Vision Tasks," Jul. 08, 2024, *arXiv*: arXiv:2403.00522. doi: 10.48550/arXiv.2403.00522.

[185] MiniMax *et al.*, "MiniMax-01: Scaling Foundation Models with Lightning Attention," Jan. 14, 2025, *arXiv*: arXiv:2501.08313. doi: 10.48550/arXiv.2501.08313.

[186] W. Fedus, B. Zoph, and N. Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[187] D. Lepikhin *et al.*, "GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding," Jun. 30, 2020, *arXiv*: arXiv:2006.16668. doi: 10.48550/arXiv.2006.16668.

[188] W. Chen, W. Huang, X. Du, X. Song, Z. Wang, and D. Zhou, "Auto-scaling Vision Transformers without Training," Feb. 27, 2022, *arXiv*: arXiv:2202.11921. doi: 10.48550/arXiv.2202.11921.

[189] G. Qu, Q. Chen, W. Wei, Z. Lin, X. Chen, and K. Huang, "Mobile Edge Intelligence for Large Language Models: A Contemporary Survey," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2025, doi: 10.1109/COMST.2025.3527641.

[190] T. Ge, S.-Q. Chen, and F. Wei, "EdgeFormer: A Parameter-Efficient Transformer for On-Device Seq2seq Generation," Dec. 29, 2022, *arXiv*: arXiv:2202.07959. doi: 10.48550/arXiv.2202.07959.

[191] Service Robotics Market Size, Share and Trends, MarketsandMarkets, Dec. 2024. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/service-robotics-market-681.html

[192] "Agarwal: Four decades of sustainable tourism research:... - Google Scholar." Accessed: Apr. 28, 2025. [Online]. Available: https://scholar.google.com/scholar_lookup?hl=en&volume=26&publication_year=2024&pages=e2643&journal=International+Journal+of+Tourism+Research&issue=2&author=R.+Agarwal&author=A.+Mehrotra&author=A.+Mishra&author=N.+P.+Rana&author=R.+Nunkoo&author=M.+Cho&title=Four+decades+of+sustainable+tourism+research%3A+Trends+and+future+research+directions&doi=https%3A%2F%2Fdoi.org%2F10.1002%2Fjtr.2643

[193] A.-H. Chiang and S. Trimi, "Impacts of service robots on service quality," *Serv Bus*, vol. 14, no. 3, pp. 439–459, Sep. 2020, doi: 10.1007/s11628-020-00423-8.

[194] C. Webster and S. Ivanov, "Robots, Artificial Intelligence and Service Automation in Tourism and Quality of Life," in *Handbook of Tourism and Quality-of-Life Research II: Enhancing the Lives of Tourists, Residents of Host Communities and Service Providers*, M. Uysal and M. J. Sirgy, Eds., Cham: Springer International Publishing, 2023, pp. 533–544. doi: 10.1007/978-3-031-31513-8_36.

[195] A.-H. Chiang and S. Trimi, "Impacts of service robots on service quality," *Serv Bus*, vol. 14, no. 3, pp. 439–459, Sep. 2020, doi: 10.1007/s11628-020-00423-8.

[196] O. M. Team *et al.*, "Octo: An Open-Source Generalist Robot Policy," May 26, 2024, *arXiv*: arXiv:2405.12213. doi: 10.48550/arXiv.2405.12213.

[197] M. Chang *et al.*, "PARTNR: A Benchmark for Planning and Reasoning in Embodied Multi-agent Tasks," Oct. 31, 2024, *arXiv*: arXiv:2411.00081. doi: 10.48550/arXiv.2411.00081.

[198] W. Yuan *et al.*, "RoboPoint: A Vision-Language Model for Spatial Affordance Prediction for Robotics," Jun. 15, 2024, *arXiv*: arXiv:2406.10721. doi: 10.48550/arXiv.2406.10721.

[199] R. Zheng *et al.*, "TraceVLA: Visual Trace Prompting Enhances Spatial-Temporal Awareness for Generalist Robotic Policies," Dec. 25, 2024, *arXiv*: arXiv:2412.10345. doi: 10.48550/arXiv.2412.10345.

[200] Y. Jiang, X. Gao, C. Yu, Y. Zhu, L. Fei-Fei, A. Torralba, A. Garg, and Y. Zhu, "VIMA: General Robot Manipulation with Multimodal Prompts," arXiv preprint arXiv:2210.03094, 2022. [Online]. Available: https://arxiv.org/abs/2210.03094

[201] S. Y. Min *et al.*, "Situated Instruction Following," Jul. 15, 2024, *arXiv*: arXiv:2407.12061. doi: 10.48550/arXiv.2407.12061.

[202] K. Black, T. Xiao, B. Ichter, S. Levine, and K. Goldberg, "π0: A Vision-Language-Action Flow Model for General Robot Control," arXiv preprint arXiv:2410.24164, 2024. [Online]. Available: https://arxiv.org/abs/2410.24164

[203] R. Yang, Y. Kim, R. Hendrix, A. Kembhavi, X. Wang, and K. Ehsani, "Harmonic Mobile Manipulation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2024, pp. 3658–3665. doi: 10.1109/IROS58592.2024.10802201.

[204] M. Dalal, T. Chiruvolu, D. S. Chaplot, and R. Salakhutdinov, "PLAN-SEQ-LEARN: Language Model Guided RL for Solving Long Horizon Robotics Tasks," arXiv preprint arXiv:2405.01534, 2024. [Online]. Available: https://arxiv.org/abs/2405.01534

[205] J. Zhang, L. Xiao, J. Gardner, A. Singh, P. Abbeel, and A. Kumar, "Bootstrap Your Own Skills: Learning to Solve New Tasks with Large Language Model Guidance," arXiv preprint arXiv:2310.10021, 2023. [Online]. Available: https://arxiv.org/abs/2310.10021

[206] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," Apr. 23, 2023, *arXiv*: arXiv:2304.13705. doi: 10.48550/arXiv.2304.13705.

[207] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation," in Proceedings of the Conference on Robot Learning (CoRL), 2024.

[208] A. Grattafiori *et al.*, "The Llama 3 Herd of Models," Nov. 23, 2024, *arXiv*: arXiv:2407.21783. doi: 10.48550/arXiv.2407.21783.

[209] B. Rozière *et al.*, "Code Llama: Open Foundation Models for Code," Jan. 31, 2024, *arXiv*: arXiv:2308.12950. doi: 10.48550/arXiv.2308.12950.

[210] J. Kim *et al.*, "Control strategies for cleaning robots in domestic applications: A comprehensive review," *International Journal of Advanced Robotic Systems*, vol. 16, no. 4, p. 1729881419857432, Jul. 2019, doi: 10.1177/1729881419857432.

[211] H. J. Lym, H. I. Son, D.-Y. Kim, J. Kim, M.-G. Kim, and J. H. Chung, "Child-centered home service design for a family robot companion," *Front. Robot. AI*, vol. 11, p. 1346257, Jul. 2024, doi: 10.3389/frobt.2024.1346257.

[212] L. Beyer *et al.*, "PaliGemma: A versatile 3B VLM for transfer," Oct. 10, 2024, *arXiv*: arXiv:2407.07726. doi: 10.48550/arXiv.2407.07726.

[213] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid Loss for Language Image Pre-Training," Sep. 27, 2023, *arXiv*: arXiv:2303.15343. doi: 10.48550/arXiv.2303.15343.

[214] I. Singh *et al.*, "ProgPrompt: Generating Situated Robot Task Plans using Large Language Models," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023, pp. 11523–11530. doi: 10.1109/ICRA48891.2023.10161317.

[215] J. Yang *et al.*, "Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent," in *International Conference on Robotics and Automation*, IEEE, 2024, pp. 7694–7701. Accessed: Dec. 06, 2024. [Online]. Available: https://ieeexplore.ieee.org/ abstract/document/ 10610443/?casa_token=FBXejBAVc20AAAAA:6gepiF616qP62GXtX5UYM YK0RHxYWB7t0Oa_295N40XANWyjQBzt0zr1Rn0LIdeSGZcbqxvX

[216] D. Honerkamp, M. Büchner, F. Despinoy, T. Welschehold, and A. Valada, "Language-grounded dynamic scene graphs for interactive object search with mobile manipulation," *IEEE Robotics and Automation Letters*, 2024, Accessed: Apr. 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/ abstract/document/10632580/?casa_token=wvLIUG6rg7oAAAAA:ROSpHX 4UZlSsRHIXB_YhYye6b1Vk0rt6QNlZa4v5BfBCL8- uW59IPhcis8UejZZ8pVTSCAyl

[217] S. Luo *et al.*, "GSON: A Group-based Social Navigation Framework with Large Multimodal Model," Apr. 08, 2025, *arXiv*: arXiv:2409.18084. doi: 10.48550/arXiv.2409.18084.

[218] M. Ahn *et al.*, "AutoRT: Embodied Foundation Models for Large Scale Orchestration of Robotic Agents," Jul. 02, 2024, *arXiv*: arXiv:2401.12963. doi: 10.48550/arXiv.2401.12963.

[219] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suenderhauf, "SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Robot Task Planning," Sep. 27, 2023, *arXiv*: arXiv:2307.06135. doi: 10.48550/arXiv.2307.06135.

[220] Y. Wang *et al.*, "RoboGen: Towards Unleashing Infinite Data for Automated Robot Learning via Generative Simulation," Jun. 14, 2024, *arXiv*: arXiv:2311.01455. doi: 10.48550/arXiv.2311.01455.

[221] M. Chen *et al.*, "Evaluating Large Language Models Trained on Code," Jul. 14, 2021, *arXiv*: arXiv:2107.03374. doi: 10.48550/arXiv.2107.03374.

[222] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, "OpenScene: 3D Scene Understanding with Open Vocabularies," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 815–824. doi: 10.1109/CVPR52729.2023.00085.

[223] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," May 09, 2016, *arXiv*: arXiv:1506.02640. doi: 10.48550/arXiv.1506.02640.

[224] S. Bai, Y. Liu, Y. Han, H. Zhang, and Y. Tang, "Self-Calibrated CLIP for Training-Free Open-Vocabulary Segmentation," Mar. 09, 2025, *arXiv*: arXiv:2411.15869. doi: 10.48550/arXiv.2411.15869.

[225] H. Chang *et al.*, "Context-Aware Entity Grounding with Open-Vocabulary 3D Scene Graphs," Sep. 27, 2023, *arXiv*: arXiv:2309.15940. doi: 10.48550/arXiv.2309.15940.

[226] D. Shah, B. Osiński, B. Ichter, and S. Levine, "LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action," in *Proceedings of The 6th Conference on Robot Learning*, PMLR, Mar. 2023, pp. 492–504. Accessed: Apr. 28, 2025. [Online]. Available: https://proceedings.mlr.press/v205/shah23b.html

[227] D. Shah, B. Eysenbach, G. Kahn, N. Rhinehart, and S. Levine, "ViNG: Learning Open-World Navigation with Visual Goals," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, May 2021, pp. 13215–13222. doi: 10.1109/ICRA48506.2021.9561936.

[228] D. Shah *et al.*, "ViNT: A Foundation Model for Visual Navigation," Oct. 24, 2023, *arXiv*: arXiv:2306.14846. doi: 10.48550/arXiv.2306.14846.

[229] L. X. Shi *et al.*, "Hi Robot: Open-Ended Instruction Following with Hierarchical Vision-Language-Action Models," Feb. 26, 2025, *arXiv*: arXiv:2502.19417. doi: 10.48550/arXiv.2502.19417.

[230] AgiBot-World-Contributors *et al.*, "AgiBot World Colosseo: A Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems," Mar. 13, 2025, *arXiv*: arXiv:2503.06669. doi: 10.48550/arXiv.2503.06669.

[231] J. Gao *et al.*, "Physically grounded vision-language models for robotic manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2024, pp. 12462–12469. Accessed: Apr. 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10610090/ ?casa_token=6HQa38yzVtcAAAAA:MjibRIY7VhR1fF8ftUFdYzfMGbqf9ss 2BKxUTEHWBQKvZA7F17P9YIZBcG7Zjrrc-pzZ9cUw

[232] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," Feb. 27, 2023, *arXiv*: arXiv:2302.13971. doi: 10.48550/arXiv.2302.13971.

[233] W. Dai *et al.*, "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning," Jun. 15, 2023, *arXiv*: arXiv:2305.06500. doi: 10.48550/arXiv.2305.06500.

[234] R. Bommasani *et al.*, "On the Opportunities and Risks of Foundation Models," Jul. 12, 2022, *arXiv*: arXiv:2108.07258. doi: 10.48550/arXiv.2108.07258.

[235] X. Xiao *et al.*, "Robot learning in the era of foundation models: a survey," *Neurocomputing*, vol. 638, p. 129963, Jul. 2025, doi: 10.1016/j.neucom.2025.129963.

[236] G. R. Team *et al.*, "Gemini Robotics: Bringing AI into the Physical World," Mar. 25, 2025, *arXiv*: arXiv:2503.20020. doi: 10.48550/arXiv.2503.20020.

[237] A. Zadaianchuk, G. Martius, and F. Yang, "Self-supervised Reinforcement Learning with Independently Controllable Subgoals," in *Proceedings of the 5th Conference on Robot Learning*, PMLR, Jan. 2022, pp. 384–394. Accessed: Apr. 27, 2025. [Online]. Available: https://proceedings.mlr.press/v164/zadaianchuk22a.html

[238] I. Moreira, J. Rivas, F. Cruz, R. Dazeley, A. Ayala, and B. Fernandes, "Deep Reinforcement Learning with Interactive Feedback in a Human–Robot Environment," *Applied Sciences*, vol. 10, no. 16, Art. no. 16, Jan. 2020, doi: 10.3390/app10165574.

[239] D. J. H. Iii and D. Sadigh, "Few-Shot Preference Learning for Human-in-the-Loop RL," in *Proceedings of The 6th Conference on Robot Learning*, PMLR, Mar. 2023, pp. 2014–2025. Accessed: Apr. 27, 2025. [Online]. Available: https://proceedings.mlr.press/v205/iii23a.html

[240] A. Ayub, C. Nehaniv, and K. Dautenhahn, "Interactive Continual Learning Architecture for Long-Term Personalization of Home Service Robots," Mar. 06, 2024, *arXiv*: arXiv:2403.03462. doi: 10.48550/arXiv.2403.03462.

[241] A. O'Neill *et al.*, "Open X-Embodiment: Robotic Learning Datasets and RT-X Models : Open X-Embodiment Collaboration0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, May 2024, pp. 6892–6903. doi: 10.1109/ICRA57147.2024.10611477.