Foundation Models in Robotics: A Comprehensive Review of Methods, Models, Datasets, Challenges and Future Research Directions

Anonymous authors
Paper under double-blind review

Abstract

Over the recent years, the field of robotics has been undergoing a transformative paradigm shift from fixed, single-task, domain-specific solutions towards adaptive, multi-function, general-purpose agents, capable of operating in complex, open-world, dynamic environments. This tremendous advancement is primarily driven by the emergence of Foundation Models (FMs), i.e., large-scale neural-network architectures trained on massive, internet-scale, heterogeneous datasets that provide unprecedented capabilities in multi-modal understanding/reasoning, long-horizon planning, and cross-embodiment generalization. In this context, the current study provides a holistic, thorough, systematic, and in-depth review of the research landscape of FMs in robotics. In particular, the evolution in the field is initially delineated through five distinct research phases, spanning from the early incorporation of native Natural Language Processing (NLP) and Computer Vision (CV) models to the current frontier of multi-sensory generalization and real-world deployment. Subsequently, a highly-granular, multi-criteria, taxonomic investigation of the literature methods is performed, examining the following key aspects: a) The employed foundation model types (i.e., LLMs, VFMs, VLMs, and VLAs), b) The underlying neural network architectures, c) The adopted learning paradigms, d) The different learning stages of knowledge incorporation, e) The most common robotic tasks (including perception, planning, navigation, manipulation, and human-robot interaction), and f) The main real-world application domains. For each defined criterion/aspect, a methodical comparative analysis of the various categories of approaches and critical insights are provided. Moreover, a thorough report on the publicly available datasets, required for model training and evaluation, is provided per considered robotic task. Furthermore, a comprehensive and hierarchical discussion on the current open challenges and promising future research directions in the field is incorporated.

1 Introduction

Over the recent years, the field of robotics has witnessed unprecedented and transformative technological advancements, which have boosted the evolution from fixed, single-task, domain-specific setups to adaptive, general-purpose, open-world solutions (Newbury et al., 2023; Mokssit et al., 2023). Apart from significant developments in hardware and material sciences, a key driving force for this technological rise comprises the progress in the fields of Artificial Intelligence (AI) and Machine Learning (ML) (Soori et al., 2023). In particular, nowadays robotic platforms exhibit increased levels of efficiency, dexterity, autonomy, precision, and adaptation across a wide set of diverse tasks, while operating in complex and dynamic environments (Billard et al., 2025).

Robotic research has so far been dominated by two main (though not mutually exclusive) modeling paradigms, namely automatic control and machine learning approaches (Lee et al., 2024b). Classic automation control relies on the fundamental principle of initially defining a mathematical model of a system for predicting its behavior and, subsequently, designing a controller for enabling it to perform a specific task. Such approaches (often termed model-based) require explicit programming and have been proven to be efficient for implementing

tasks in structured and predictable environments (Lee et al., 2024b). However, they are characterized by low adaptability (reprogramming is needed) and they are typically mathematically complex (Rakhmatillaev et al., 2025). On the other hand, ML methods focus on enabling robots to learn from data and experiences. To this end, ML approaches are shown to exhibit high adaptability (e.g., tackling novel and previously unseen circumstances) and to be efficient in handling tasks in complex, unstructured, and dynamic (or even unknown) environments (Tang et al., 2025). Nevertheless, ML methods are often computationally expensive and typically require large datasets for training purposes (Hu et al., 2023).

Foundation Models (FMs) constitute a recent and, yet, very powerful paradigm in the fields of AI and ML (Awais et al., 2025). In particular, FMs are constructed through training on massive, internet-scale, multi-modal datasets and can be adapted to a wide range of diverse downstream tasks, such as language, vision, and audio processing. In practice, FMs serve as a versatile and reusable basis for efficiently developing specialized or domain-specific (multi-task) applications, avoiding the need for training from scratch and using extensive training datasets. More recently, FMs have also been introduced in the field of robotics, exhibiting the following, among others, advantageous characteristics (that extend the ones of traditional ML-based approaches) (Firoozi et al., 2025; Xiao et al., 2025b): a) Improved transferability across related tasks, environments, and embodiments, b) More reusable and generalizable representations, c) Increased semantic understanding and open-world capabilities, d) Support for sim-to-real transfer and cross-domain adaptation, e) Multi-modal integration and semantic alignment, f) Enhanced versatility in language-conditioned and perception-driven robotic behavior, g) Interpretation of natural language instructions, h) Hierarchical and long-horizon task decomposition and planning, and i) Improved policy generalization. The above favorable attributes, though, are accompanied by critical/unique challenges, including, indicatively (Firoozi et al., 2025; Xiao et al., 2025b): a) Inference latency and high computational cost, b) Limited real-time deployability, c) Lack of semantic and physical grounding, d) Data scarcity and embodiment bias, e) Safety risks and unforeseen failure modes, f) Limited interpretability, transparency, and diagnosability, and g) Ethical, alignment, and regulatory imperatives.

As outlined above, FMs induce transformative effects on and lead to unprecedented performance/capability accomplishments in the field of robotics, fundamentally reforming robot design, learning, programming, and deployment practices. In this context, the current study aims to holistically and comprehensively investigate, map, and analyze in depth the research landscape of robotic FM methods. In particular, the main contributions of this work are:

- Outline of the **evolution in robotic FM research**, focusing on describing the most common FMs proposed in the literature and the main observed phases, which comprise the following ones: a) Phase 1 (2018–2021): Integration of native Natural Language Processing (NLP) and Computer Vision (CV) models; b) Phase 2 (2021–2022): Grounded planning with Vision–Language (VL) representations; c) Phase 3 (2022–2023): Embodied Vision–Language–Action (VLA) policies; d) Phase 4 (2023–2024): Memory, autonomous task composition, and Web-to-robot transfer; and e) Phase 5 (2024–present): Multi-sensory generalization and real-world deployment;
- Holistic, thorough, systematic, highly-granular, multi-criteria, **taxonomic investigation of robotic FM approaches**, taking into account the following main criteria: a) The type of the employed FM with respect to the input-output modalities involved; b) The nature of the underlying Neural Network (NN) architecture; c) The learning paradigm adopted for developing a robotic FM; d) The learning stage at which knowledge is incorporated to a FM; e) The task controlled by a robotic FM; and f) The application domain where a robotic FM is used. For each defined criterion, a methodical comparative analysis of the various categories of approaches and critical insights are provided;
- Thorough report of the **public datasets/benchmarks** used for training/evaluation purposes;
- Extensive discussion of **current challenges** and **future research directions** in the field.

Regarding existing surveys in the field, Table 1 comparatively analyzes the current work with the relevant literature reviews of (Hu et al., 2023; Xu et al., 2024c; Ma et al., 2024c; Jang et al., 2024; Kawaharazuka et al., 2024; Firoozi et al., 2025; Xiao et al., 2025b; Tayyab Khan & Waheed, 2025; Kawaharazuka et al., 2025;

Table 1: Comparative analysis of recent surveys in the field of foundation models in robotics.

Article	Scope	Methodology	Primary contributions	Limitations
Hu et al (2023) (arXiv)	Broad survey and meta-analysis on the use of FMs towards general-purpose robots	• Literature analysis per robotic task • Performance analysis per robotic task • Analysis of 301 papers	• Separation between CV/NLP and native robotic FMs • Identification of trends based on experimental results • Discussion on open challenges and future research directions	• Discussion of early developments in robotic FMs • No systematic and theoretical comparison between different approaches • No discussion on recent developments (e.g., world models, diffusion policies, etc.) • No discussion on application domains
Xu et al (2024c) (arXiv)	Task-oriented survey focusing on robotic manipulation	• Literature categorization to high-level planning and low-level control approaches • Analysis of 64 papers	• Discussion on form and assistant perspectives of planning • Analysis of focused components in the learning process • Discussion on open challenges and future research directions	• Investigation of only robotic manipulation approaches • Limited literature coverage • No systematic and theoretical comparison between different approaches • No discussion on application domains
Ma et al (2024c) (arXiv)	VLA-oriented survey emphasizing on embodied AI aspects	• Taxonomic analysis of VLAs based on individual components, control policies, and high-level task planning • Analysis of 785 papers	• Systematic comparison of methods • Analysis of strengths and limitations per category • Discussion on open challenges and future research directions	• No explicit discussion on other FM types (i.e., LLMs, VFMs, VLMs) • No systematic analysis and comparisons across multiple criteria (i.e., NN architecture, learning paradigm, learning stage, robotic task) • No discussion on application domains
Jang et al (2024) (IJCAS)	Application-oriented survey focusing on robotic autonomy	• Literature categorization based on perception, task planning, and control • Analysis of environmental setups • Analysis of 255 papers	• Discussion of impact of individual FM components on robot autonomy • Analysis of robotic platforms and simulation environments • Discussion on future research directions	• No systematic analysis across multiple criteria (i.e., NN architecture, learning paradigm, learning stage) • No theoretical comparison between different approaches • No discussion on application domains
Kawahara et al (2024) (AR)	Application-oriented survey focusing on component replacement with FMs	• Literature categorization based on perception, planning, and data augmentation • Analysis of 225 papers	• Analysis of input-output relationships in FMs • Discussion on the role of FMs in perception, motion planning, and control • Discussion on future research directions	• No systematic analysis across multiple criteria (i.e., FM type, NN architecture, learning paradigm, learning stage) • No theoretical comparison between different approaches • No discussion on application domains
Firoozi et al (2025) (IJRR)	Broad survey emphasizing on the use of FMs in robotics and embodied AI	• Literature analysis regarding decision-making, planning, and control • Analysis of 233 papers	• Literature investigation regarding FMs that are native and relevant to robotics • Analysis of embodied AI aspects • Discussion on open challenges and future research directions	• No systematic analysis across multiple criteria (i.e., FM type, NN architecture, learning paradigm, learning stage) • No theoretical comparison between different approaches • No discussion on application domains
Xiao et al (2025b) (Neurocomputing)	Robot learning-oriented survey for generalist robots	• Analysis of robot learning in manipulation, navigation, task planning, and reasoning • Analysis of 464 papers	• Systematic and hierarchical analysis of robot learning techniques • Discussion on open challenges and future research directions	• No systematic analysis across multiple criteria (i.e., FM type, NN architecture, learning stage) • No theoretical comparison between different approaches • No discussion on application domains
Tayyab Kh & Waheed (2025) (arXiv)	Integration-oriented survey focusing on real-world deployment	• Literature analysis regarding integrated, system-level strategies • Feasibility analysis in real-world environments • Analysis of 175 papers	• Literature categorization across simulation-driven design, open-world execution, sim-to-real transfer, and adaptable robotics • Discussion on open challenges and future research directions	• No systematic analysis across multiple criteria (i.e., FM type, NN architecture, learning paradigm, learning stage, robotic task) • No theoretical comparison between different approaches • No discussion on application domains
Kawahara et al (2025) (IEEE Access)	VLA-oriented survey emphasizing on full-stack aspects	• VLA literature analysis regarding both software and hardware perspectives • Analysis of 427 papers	• Systematic analysis of architectural designs and learning paradigms • Discussion on practical implementation aspects • Discussion on future research directions	• No explicit discussion on other FM types (i.e., LLMs, VFMs, VLMs) • Coarse categorization of all learning paradigms and robotic tasks • No theoretical comparison between different approaches • No discussion on application domains
Sapkota et al (2025) (arXiv)	VLA-oriented survey detailing research evolution aspects	• Literature analysis covering FM evolution, key progress areas, and application domains • Analysis of 299 papers	• Analysis of research evolution • Discussion on architectural innovations, training strategies, and real-time inference • Analysis of application domains • Discussion on open challenges and future research directions	• No explicit discussion on other FM types (i.e., LLMs, VFMs, VLMs) • No systematic analysis across multiple criteria (i.e., NN architecture, learning paradigm, learning stage) • No theoretical comparison between different approaches
Current survey	Thorough and systematic analysis of the landscape of FMs in robotics	• Structured and systematic literature analysis, querying 6 major databases, using specific inclusion/exclusion criteria, and applying iterative screening • Analysis of 435 papers	• Investigation of research evolution in 5 distinct phases • Highly-granular multi-criteria (6) taxonomic analysis of literature, examining FM types, NN architectures, learning paradigms, learning stages, robotic tasks, and application domains • Per criterion methodical comparative analysis of different approaches and insights reporting • Comprehensive and hierarchical discussion on current challenges and future research directions	–

Sapkota et al., 2025), investigating the following aspects: a) Survey scope, b) Review methodology, c) Primary contributions, and d) Main limitations. Examining Table 1, it can be seen that literature works exhibit the following limitations: a) They often remain relatively specific/narrow in scope (i.e., adopting a task-, model-, application-, learning-, or integration-oriented perspective), leading to a non-thorough investigation of the overall research landscape, b) They consider a single or very few literature analysis criteria, resulting into a non-comprehensive examination of the literature works, and c) They commonly adopt a narrative-style discussion of the literature, avoiding to provide a systematic comparison of the different approaches. On

the contrary, the current survey provides a thorough and in-depth analysis of the research landscape of the use of FMs in robotics, demonstrating the following key advantageous/distinctive characteristics: a) It targets a holistic investigation of the overall field, b) It adopts a structured and systematic literature review methodology, supporting the search in 6 major databases, the use of specific inclusion/exclusion criteria, and the application of an iterative screening process, c) It documents the main 5 distinct research evolution phases, as well as the key trends associated with each of them, d) It supports a highly-granular, multi-criteria (6), taxonomic investigation of the literature, examining the different FM types, NN architectures, learning paradigms, learning stages, robotic tasks, and application domains, e) It incorporates a per criterion methodical comparative analysis of the different approaches and facilitates the reporting of critical insights, and f) It provides a comprehensive and hierarchical discussion on the current challenges and future research directions in the field.

The remainder of the manuscript is organized as follows: Section 2 outlines the adopted methodology for reviewing the relevant literature. Section 3 describes the evolution in robotic FM research and indicates the most widely adopted models. Section 4 delineates the criteria used for analyzing the literature, as well as the resulting categories of robotic FM methods structured in the form of a taxonomy. Sections 5-10 discuss in detail the various categories of robotic FM approaches, taking into account the type of the employed FM, the nature of the underlying NN architecture, the adopted learning paradigm, the learning stage at which knowledge is integrated to the FM, the performed robotic task, and the selected application domain, respectively. Section 11 reviews the publicly available datasets for training and evaluating robotic FM methods. Sections 12 and 13 discuss the current challenges and future research directions in the field, correspondingly, while Section 14 concludes the paper.

2 Literature review methodology

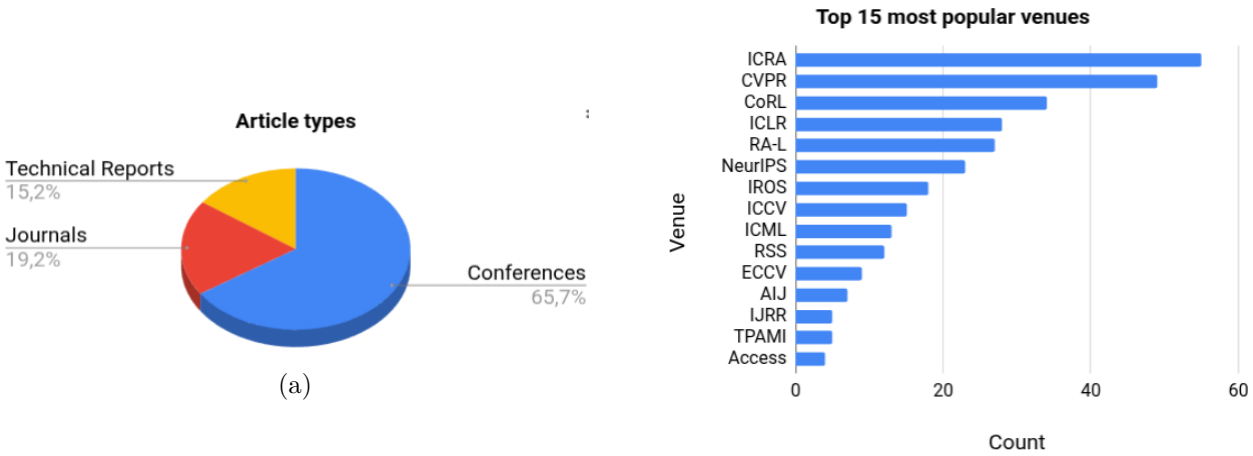
In order to efficiently and thoroughly identify/map the robotic FM literature, while at the same time detecting key concepts and trends, a systematic approach was followed for ensuring comprehensiveness and relevance of the selected research works. In particular, a structured literature review methodology was adopted, consisting of the iterative main steps described below.

2.1 Scope and objectives formulation

The fundamental goal of the performed survey study was to review the robotic FM literature, i.e., approaches for various robotic tasks (namely, perception, reasoning, planning, navigation, manipulation, and human-robot interaction) whose execution relies on the use of an underlying FM. In particular, the focus was on identifying the most recent advancements and works with substantial contribution to the field, emphasizing the following objectives: a) The main categories of methods based on multiple criteria, b) The utilized datasets, c) Current challenges, and d) Future research directions.

2.2 Literature search

In order to ensure broad and thorough coverage of the relevant literature, the search strategy involved querying several major scientific databases, including IEEE Xplore, Google Scholar, Scopus, DBLP, arXiv, and Web of Science. The actual search was performed by combining targeted keywords/terms (e.g., “foundation model”, “robotics”, “vision-language-action”, “large language model”, etc.) with Boolean operators (i.e., “AND”, “OR”, “NOT”). To guarantee contemporary relevance, the search primarily focused on research works published within the last five years; however, certain earlier seminal studies were also included. An example of the query used in the Scopus database is the following:
TITLE-ABS-KEY(“foundation model” OR “vision-language model” OR “visual foundation model” OR “large language model” OR “vision-language-action” OR “robotic foundation model”)
AND TITLE-ABS-KEY(“robot” OR “robotics” OR “manipulation” OR “navigation” OR “human-robot interaction” OR “embodied”)
AND PUBYEAR > 2020
AND PUBYEAR <= 2026



(a) (b)

Figure 1: Key bibliometric analytics regarding robotic FM literature: (a) Article types, and (b) Top-15 most popular venues.

AND (SRCTYPE(j) OR SRCTYPE(p))
AND EXCLUDE(DOCTYPE, "bk")
Multiple search steps were performed iteratively, involving refinements to the employed keywords so as to retrieve more relevant works. Moreover, the list of references of each research article was also analyzed in order to identify additional relevant studies.

2.3 Screening

Initial screening relied on excluding duplicate records, non-English papers, and articles without full-text access, in order to maintain the integrity of the review study. Then, article selection was performed taking into account title and abstract information, so as to eliminate irrelevant works. Subsequently, in-depth full-text review was performed for considering only research studies that: a) Focus on demonstrating approaches for implementing various robotic tasks (i.e., perception, planning, navigation, manipulation, and human-robot interaction), where a FM comprises a key algorithmic component, and b) Exhibit substantial theoretical and/or experimental contributions. Additionally, priority was given to research works originating from prominent robotics and AI/ML publication venues. Eventually, a total of 435 articles were selected for analysis and were included as references in the current manuscript.

Key bibliometric analytics regarding the performed literature review study are illustrated in Fig. 1, while in-depth analysis of the identified robotic FM works is provided in Sections 3–11.

3 Robotic FM research evolution

3.1 Research phases

Although foundation models have only relatively recently been introduced in the field of robotics, they have decisively contributed towards transformative effects, while gradual advancements and emerging research trends can be identified in the literature. In particular, the evolution of robotic FM research can be roughly classified into subsequent and distinct phases, each corresponding to a critical paradigm shift regarding how perception, reasoning, and control procedures are consolidated in a robotic system. The overall research progress concentrates on repositioning from isolated/modular designs to integrated/general-purpose agents (Reed et al., 2022; Driess et al., 2023). The considered research phases are graphically illustrated in Fig. 2, along with key/milestone works associated with each of them, while they are explained in detail in the followings.

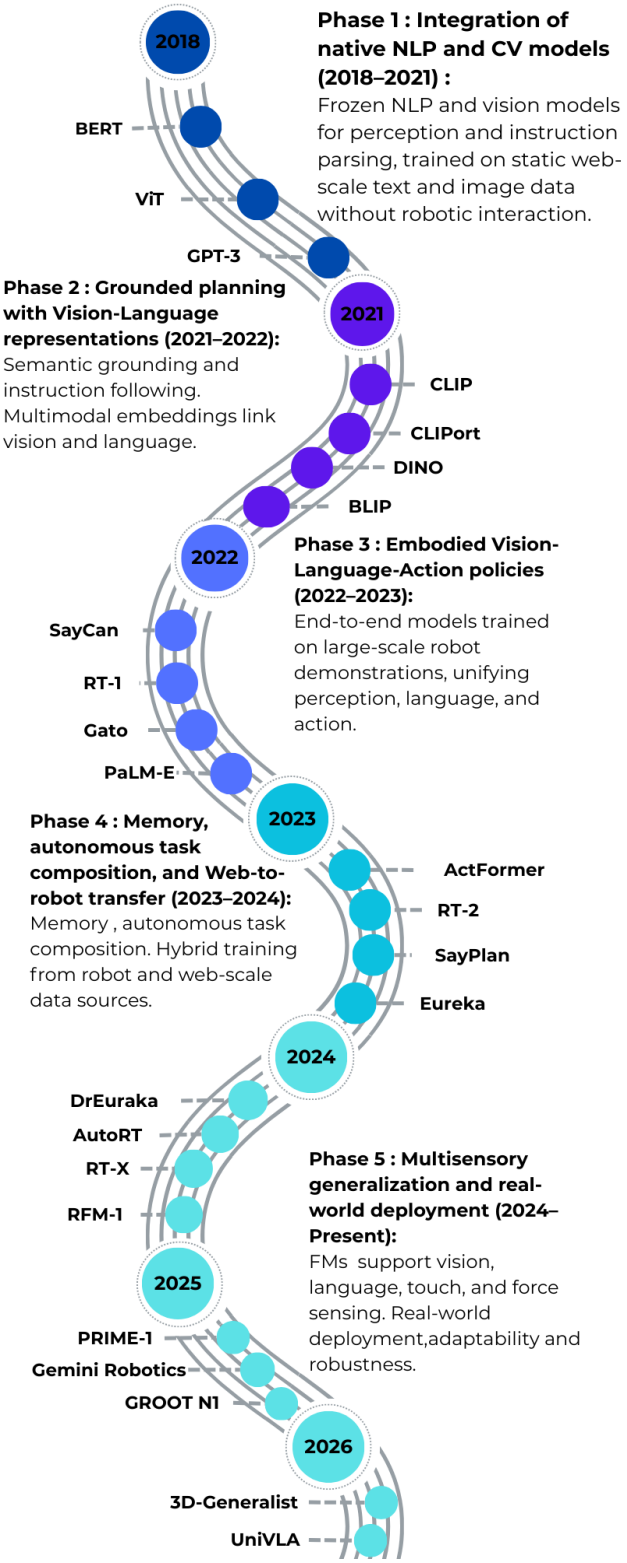


Figure 2: Main phases in robotic FM research and key/milestone works.

Table 2: Most common and widely adopted robotic FMs ('Emb.' denotes whether the model is embodied, 'Param.' indicates the number of parameters, and 'Public' denotes whether the model is publicly available).

Model	Year	Type	Input	Output	Emb.	Param.	Public	Key innovation
BERT (Devlin et al., 2019)	2018	LLM	Text	Text embedding	N	110M	Y	Interpretation of natural-language commands and translation to executable action sequences
ViT (Yuan et al., 2021)	2020	VFM	Image	Class logit	N	86M	Y	Advanced visual perception, especially in capturing long-range dependencies and global contextual information
GPT-3 (Brown et al., 2020)	2020	LLM	Text	Text	N	175B	N	Zero/few-shot language reasoning for translating natural language instructions to executable plans or code
CLIP (Radford et al., 2021)	2021	VLM	Text, image	Embedding	Y	175M	Y	Identification and localization of objects from free-form verbal cues, while allowing zero-shot grounding and task specification
DINO (Zhang et al., 2023)	2021	VFM	Image	Embedding	N	86M	Y	Creation of object attention maps for facilitating robot manipulation tasks
OWL-ViT (Minderer et al., 2022)	2022	VLM	Text, image	Box, label	N	100M	Y	Zero-shot open-vocabulary object detection based on text queries
BLIP (Li et al., 2022a)	2022	VLM	Text, image	Text, embedding	N	480M	Y	Zero-shot multimodal understanding and generation, based on bidirectional vision–language modeling
SayCan (Brohan et al., 2023b)	2022	VLA	Text, environment	Action	Y	–	N	Generation of interpretable and feasible robotic action plans, taking into account object affordance information
RT-1 (Brohan et al., 2023a)	2022	VLA	Text, image	Action	Y	–	Y	Unified, end-to-end policy generator for mobile manipulation, trained using over 130,000 real-world robot state–action trajectories
Gato (Reed et al., 2022)	2022	VLA	Text, image, state	Action	Y	1.18B	N	Multi-modal, multi-task, multi-embodiment generalist policy learning agent, supporting more than 600 different tasks
PaLM-E (Driess et al., 2023)	2023	VLA	Text, image	Text, action	Y	562B	N	Multimodal embodied reasoning, enabling the grounding of open-ended instructions and the generation of long-horizon action plans
SAM (Kirillov et al., 2023)	2023	VFM	Image, prompt	Mask	Y	1.2B	Y	Promptable, zero-shot image segmentation for generating high-quality object masks
ActFormer (Xu et al., 2023)	2023	VFM	Text, image	Action	Y	–	N	Long-horizon action planning, aligning high-level human intent and low-level robot control
RT-2 (Zitkovich et al., 2023)	2023	VLA	Text, image	Action	Y	–	N	Generalization to novel/unseen instructions and objects, by joint training over both robot state–action trajectories and Web-scale vision–language data
SayPlan (Rana et al., 2023)	2023	VLA	Text, scene graph	Plan	Y	–	N	Scalable, long-horizon task planning in large-scale 3D spatial environments
Eureka (Ma et al., 2024a)	2024	LLM	Task description	Reward	Y	–	Y	Direct and iterative generation of Python code implementing robot reward functions
DrEureka (Ma et al., 2024b)	2024	LLM	Simulation, task configuration	Reward	Y	–	Y	Automation of entire sim-to-real robot training pipeline, based on high-level prompts
AutoRT (Ahn et al., 2024)	2024	VLA	Text, image	Action	Y	–	N	Generation of robot actions, while incorporating a set of safety rules
RT-X (O'Neill et al., 2024)	2024	VLA	Text, image	Action	Y	–	Y	Single, general-purpose architecture accounting for different robots, tasks and environments, supporting 527 skills (160,266 tasks) performed by 22 different robotic platforms
RFM-1 (Sohn et al., 2024)	2024	VLA	Text, image, video	Action	Y	8B	N	Execution of complex robot tasks, based on multimodal physics-informed reasoning
PRIME-1 (Inc., 2025)	2025	VLA	Image	3D features, action	Y	–	N	Real-world adaptive control for multi-task operational settings
Gemini robotics (Team et al., 2025)	2025	VLA	Text, image	Action	Y	–	Y	On-device multimodal reasoning for multi-task, dexterous, bi-manual robot operations
GR00T N1 (Bjorck et al., 2025)	2025	VLA	Text, image	Action	Y	2B	Y	Single, multi-task, general-purpose architecture for humanoid robots
3D-GENERALIST (Sun et al., 2026)	2025	VLM	Text, image	Action code, 3D scene	N	–	N	VLM-as-policy framework for iterative generation of simulation-ready 3D environments, jointly refining layout, materials, lighting, and assets via self-improvement fine-tuning
UniVLA (Wang et al., 2026b)	2026	VLA	Text, image, video	Action, image	Y	8.5B	Y	Unified autoregressive token-space modeling of vision, language, and action, enabling world-model post-training from large-scale video data and strong long-horizon policy learning

3.1.1 Phase 1: Integration of native Natural Language Processing (NLP) and Computer Vision (CV) models (2018–2021)

In the early attempts of incorporating FMs in robotic systems, native large-scale off-the-shelf NLP and CV models are used, aiming at enhancing robotic platforms with improved perception capabilities (for example,

identifying and tracking objects in camera video sequences, translating human verbal requests to robot symbolic goals, etc.) (Hong et al., 2021). The data required for training such FMs largely originate from standard Web text and image repositories (information sources that contain no robot action policies, no time-varying sensor streams, and no interactions with the physical world). These networks are integrated following a modular approach, requiring a separate conventional controller for realizing motion and action planning. Due to this fact, NLP and CV FMs serve a supportive role (boosting perception and language grounding tasks), while the robot's behavior remains determined by a hand-engineered downstream pipeline (Shridhar et al., 2022; Hong et al., 2021).

3.1.2 Phase 2: Grounded planning with Vision-Language (VL) representations (2021–2022)

With the emergence of vision-language and scalable language models, increased capabilities for integrating richer semantic priors and more flexible instruction grounding to robots are introduced. In particular, multimodal embeddings are used by robotic systems for linking visual input with natural language commands and generating (or evaluating) task sequences (Sun et al., 2022; Brohan et al., 2023b). Such developments take advantage of broader access to aligned image-text data and pretrained language models (Radford et al., 2021), although robotic data remains relatively limited. The latter constrains robots to utilize a relatively decreased number of own recorded/captured experiences; hence, most training samples originate from third-party datasets (e.g., curated demonstration benchmarks or Web-scraped image–text pairs), forcing models to generalize from proxy sources, rather than direct embodied trials (Shah et al., 2023a).

3.1.3 Phase 3: Embodied Vision-Language-Action (VLA) policies (2022–2023)

This phase marks the introduction of comprehensive and unified robot policies, derived directly from training using large-scale robot-demonstration datasets. In particular, FMs process vision, language, and task context simultaneously, while outputting action sequences using the same architecture (Brohan et al., 2023a; Reed et al., 2022; Driess et al., 2023). This shift towards end-to-end training is enabled by the availability of real-world robotic data at scale (O'Neill et al., 2024), collected considering hundreds of tasks and corresponding variations (e.g., different object types, lighting conditions, camera viewpoints, robot embodiments, etc.). As a consequence, robotic platforms are enhanced by incorporating the capability to generalize across goals and environments without explicit task engineering (Shridhar et al., 2023; Zitkovich et al., 2023).

3.1.4 Phase 4: Memory, autonomous task composition, and Web-to-robot transfer (2023–2024)

Building on prior advances, this phase concerns systems capable of long-horizon planning (Ajay et al., 2023), world-state tracking (Wu et al., 2023b), and autonomous skill discovery (Nam et al., 2023). In this respect, FMs are trained or adapted using both structured robot trajectories and unstructured Web-scale corpora (Kim et al., 2025). Additionally, data diversity increases, combining multimodal internet data with robot-collected experiences (Mees et al., 2024). Moreover, robotic agents synthesize and evaluate their own tasks based on FM reasoning pipelines, boosting semantic autonomy and self-improvement (Parakh et al., 2024).

3.1.5 Phase 5: Multisensory generalization and real-world deployment (2024–Present)

More recently, research advancements have focused on building robust generalist robotic systems, capable of efficiently operating in unstructured real-world environments (Bjorck et al., 2025; Team et al., 2025). The utilized FM-based solutions support multimodal inputs, including vision, language, touch, force, and proprioception (Li et al., 2026b), as well as real-time adaptation (Routray et al., 2026). Additionally, robotic agents are trained and refined using a combination of simulation and real-world interaction data, significantly extending the boundaries in robustness, transferability, safety, and autonomy (Zhao et al., 2026a; Guo et al., 2026). Consequently, FM-based solutions are being widely adopted across industrial, assistive, and mobile platforms (Sohn et al., 2024).

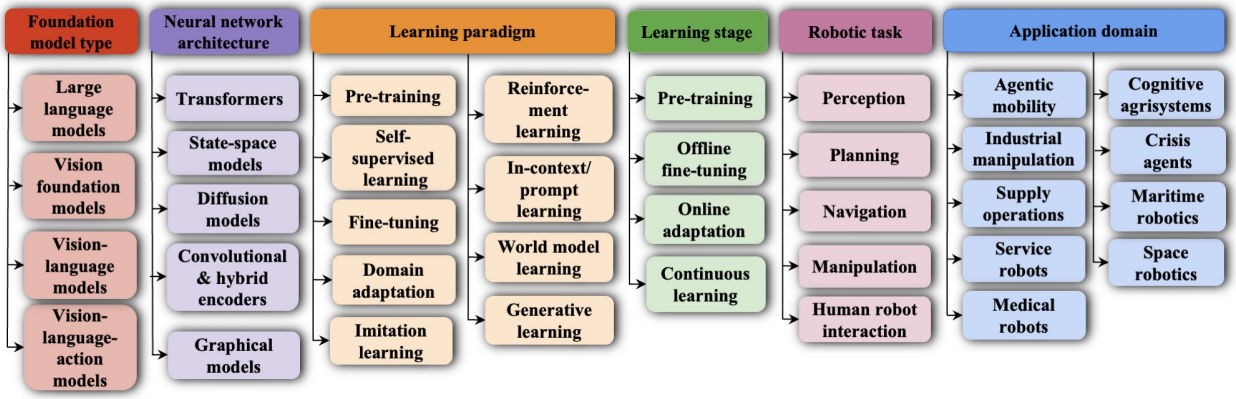


Figure 3: Key criteria and main resulting categories of robotic FM methods.

3.2 Key foundation models

Throughout the different research phases described in Section 3.1, key FM architectures have been introduced, which, on the one hand, have led to significant technological advancements and capabilities, and, on the other hand, have served as the basis for numerous methods in the field. In particular, the most common and widely adopted robotic FMs, along with their main characteristics (namely, type, input/output modalities, adoption of embodied design, number of parameters, availability of publicly available implementation, and key innovation), are briefly summarized in Table 2.

4 Key criteria and main categories of robotic FM methods

This section provides a systematic overview of the landscape of robotic FM methods. For facilitating the analysis, a set of complementary and diverse criteria are defined (each focusing on a specific/key aspect of a robotic FM system), resulting into the classification of the literature works into a corresponding set of main categories. The different criteria used and the resulting categories are graphically illustrated in Fig. 3 and detailed as follows:

- **Foundation model type**: FMs in robotics can be grouped taking into account the number and the nature of the input-output modalities involved, which critically dictates their overall capabilities and how they manage robot perception, reasoning, and action procedures. The main types of FMs are:
  - **Large Language Models (LLMs)**: These receive textual streams of data as input (sometimes also multimodal information) and, subsequently, generate high-level action policies. Their fundamental functionality relies on translating natural language instructions into task goals, programs, or supervision signals (Brohan et al., 2023b; Liang et al., 2023; Driess et al., 2023).
  - **Vision Foundation Models (VFMs)**: These receive as input visual information streams (e.g., RGB, LiDAR, thermal, etc.) and output corresponding delicate representations (e.g., object detection masks, depth maps, dense embeddings, etc.) for facilitating subsequent robotic tasks (Kirillov et al., 2023; Zhang et al., 2023).
  - **Vision-Language Models (VLMs)**: These exploit correlations and inter-dependencies among the input visual and textual data, in order to enable complex robot operations (e.g., visual grounding, language-conditioned mapping, question answering, etc.) (Radford et al., 2021; Liu et al., 2024d).
  - **Vision-Language-Action models (VLAs)**: These comprise native robotic models that map multi-modal inputs directly/end-to-end to generalist action policies, across multiple types of tasks and embodiments (Zitkovich et al., 2023; Kim et al., 2025; Bjorck et al., 2025).

- **Neural Network (NN) architecture**: The type of the underlying neural network architecture that is employed in any FM solution largely defines its capabilities, efficiency, and limitations in robotic applications. The main NN types used in robotic FM methods are:
  - **Transformers**: These rely on self-attention to capture long-range dependencies and to model complex multimodal inputs in a unified way (Vaswani et al., 2017; Driess et al., 2023; Zitkovich et al., 2023).
  - **State-Space Models (SSMs)**: Their fundamental functionality is grounded on the use of a set of first-order differential or difference equations for modeling complex, dynamic operations, typically involving multiple input and output signals (Gu & Dao, 2024; Liu et al., 2024c; Lenz et al., 2025).
  - **Diffusion models**: These employ an iterative denoising process, in order to generate diverse and realistic data samples (e.g., action policies) (Ho et al., 2020; Chi et al., 2025).
  - **Convolutional and hybrid encoders**: Convolutional Neural Network (CNN) encoders are efficient in learning hierarchical feature representations and modeling local patterns in the input data, while they are often combined with transformer or diffusion networks for further enhancing visual perception (Nair et al., 2023; Brohan et al., 2023a; Chi et al., 2025).
  - **Graphical models**: These allow the processing of data that exhibit irregular and/or complex relations, while also enabling the generation of context-aware representations (Rana et al., 2023; Gu et al., 2024; Patel & Song, 2025).
- **Learning paradigm**: In order to develop a robust robotic FM solution, different/diverse learning techniques, principles, and approaches can be adopted. The most commonly met learning paradigms in the literature, which are typically combined in a comprehensive learning methodology, are the following:
  - **Pre-training**: This relies on initially training a FM using vast (internet-scale), broad, and often unlabeled datasets, aiming at modeling rich, general-purpose representations and comprehensive world knowledge (Zitkovich et al., 2023; Driess et al., 2023; Team et al., 2025).
  - **Self-Supervised Learning (SSL)**: This enables ML models to generate their own supervisory signals directly from raw, unlabeled data, thereby circumventing the need for costly external human-provided labels (Gao et al., 2025b; Nazeri et al., 2025).
  - **Fine-tuning**: This allows generalist FMs to efficiently adapt to specific tasks, environments, and robotic platforms, while demonstrating significant efficiency and cost-effectiveness (Mees et al., 2024; Yadav et al., 2026).
  - **Domain Adaptation (DA)**: This aims at adjusting a FM, originally trained on a source domain, to maintain its accuracy and performance when applied to a new target domain (Li et al., 2026a; Zheng et al., 2026).
  - **Imitation Learning (IL)**: Also known as Learning from Demonstration (LfD) or Robot programming by Demonstration (PbD), it relies on the consideration of an autonomous agent learning to execute tasks or acquiring new skills, by observing/emulating demonstrations provided by an expert (Wan et al., 2024; Fu et al., 2025; Cai et al., 2024b).
  - **Reinforcement Learning (RL)**: This is grounded on the use of an agent learning to make sequential decisions and to adjust its behavior through trial-and-error interactions with its surrounding environment, taking into account feedback received in the form of rewards or penalties for its actions (Ma et al., 2024a; Tziafas & Kasaei, 2024; Wang et al., 2024d).
  - **In-context/prompt learning**: This paradigm enables inference-time adaptation of FMs, by guiding model behavior through demonstrations, examples, or task-specific instructions. As such, it supports flexible task adaptation and behavior specification, allowing pre-trained models to generalize to new scenarios through contextual conditioning alone (Huang et al., 2023b; Grigorev et al., 2025; Yin et al., 2025c).
  - **World Model (WM) learning**: WMs enable the grounding of FM knowledge into physically, real-world, plausible predictions, by modeling environmental dynamics and predicting the consequences of robot actions (Gao et al., 2025b; Zhou et al., 2025b).

  - **Generative Learning (GL)**: This facilitates towards reducing the reliance on extensive real-world datasets during training, by artificially synthesizing diverse and high-quality robot experiences (Zhao et al., 2026a; Heppert et al., 2026).

- **Learning stage**: The particular phase, during the overall learning process, at which knowledge is incorporated to a FM, largely defines the type/nature of the acquired knowledge, algorithmic/development details, and key assumptions about the model behavior. In this respect, the main learning stages identified in the literature are summarized as follows:

  - **Pre-training**: This is the first and by-far the most computationally intensive step in the FM development life-cycle, which involves the processing of massive, internet-scale, diverse, and usually multi-modal datasets for learning general-purpose feature representations (Zitkovich et al., 2023; Driess et al., 2023).
  - **Offline fine-tuning**: Following generic pre-training, this step focuses on adjusting the FM knowledge structures to the requirements/nuances of particular application domains or downstream tasks, making use of (minimal) additional training data (Mees et al., 2024; Li et al., 2026a).
  - **Online adaptation**: This corresponds to real-time adjustments of a robot's behavior for maintaining performance during deployment, involving the acquisition of new skills, response to novel tasks, or handling of unforeseen environmental conditions (Wang et al., 2024d; Grigorev et al., 2025).
  - **Continuous learning**: This aims at enabling FMs to continuously acquire new skills, to refine existing ones, and to maintain performance in dynamic, real-world environments in the long term (Wan et al., 2024; Murillo-González & Liu, 2025).

- **Robotic task**: The introduction of FMs has induced transformative effects in the materialization and execution of all core robotic tasks, i.e., specific jobs, actions, or functions that robots perform in order to achieve a goal. The most common, pronounced robotic tasks, where FMs are applied to, are as follows:

  - **Perception**: FMs equip robots with enhanced capabilities to perceive and reason about their surrounding environment, largely relying on visual information processing streams and often combined with additional modalities (e.g., natural language inputs) (Radford et al., 2021; Jiang et al., 2024a; Yamazaki et al., 2024; Nguyen et al., 2024).
  - **Planning**: FMs enable robots to interpret complex, high-level, human-like commands (e.g., in natural language form) and to subsequently translate them into (long-horizon) sequences of low-level, executable, and discrete actions (Brohan et al., 2023b; Liang et al., 2023; Chen et al., 2024c; Singh et al., 2023).
  - **Navigation**: FMs significantly boost robot navigation capabilities, by moving away from traditional, task-specific models and heading towards more generalized, adaptable schemes for efficient operation in complex, unstructured, and dynamic environments (Shah et al., 2023c; Wang et al., 2024a; Huang et al., 2023a; Xu et al., 2024b).
  - **Manipulation**: Often equally termed motor control, it is enhanced by the use of FMs by shifting away from task-specific programming to more generalized, adaptable approaches, also supporting more dexterous and precise manipulation tasks (Brohan et al., 2023a; Zitkovich et al., 2023; Driess et al., 2023; Bjorck et al., 2025).
  - **Human–Robot Interaction (HRI)**: FMs enable robotic platforms to understand and interact/respond with/to humans in a more intuitive, natural, flexible, and human-like way (Izquierdo-Badiola et al., 2024; Liu et al., 2024e; Bärmann et al., 2024; Irfan et al., 2024).

- **Application domain**: FMs have significantly enhanced several aspects of robot capabilities (e.g., autonomy, complex decision-making, human-robot interaction, etc.) in challenging real-world settings; hence, further boosting their widespread use, including the following main/common application domains:

- **Agentic mobility**: FMs significantly extend the capabilities of the conventional autonomous driving stack (i.e., perception, prediction, planning, and control), by transforming it into a single, cohesive, end-to-end decision-making framework (Wu et al., 2024; Xu et al., 2024b; Wang et al., 2024c).
- **Industrial manipulation**: FMs revolutionize industrial automation pipelines, by converting rigid, task-specific solutions into flexible, general-purpose agents that are capable of handling dynamic, unstructured manufacturing tasks (Sohn et al., 2024; Inc., 2025; Kim et al., 2025; Zitkovich et al., 2023).
- **Supply operations**: FMs dramatically increase flexibility, intelligence, and generalization, enabling robots to move beyond repetitive, structured tasks to handle unstructured, dynamic, and complex operations (Nicoletti & Appolloni, 2024; Xu et al., 2024a; Nicoletti, 2025).
- **Service robots**: Robots are more efficient in operating safely and intelligently in challenging household environments, while greatly capitalizing on their ability of receiving instructions in natural language (Wu et al., 2023a; Mon-Williams et al., 2025).
- **Medical robots**: Robotic platforms incorporate comprehensive, fine-grained medical knowledge, enabling them to robustly undertake high-stake, high-variability tasks, to provide context-aware intelligence, and to support consistent, precise assistance (Cui et al., 2024; Zeinoddin et al., 2024; He et al., 2024).
- **Cognitive agrisystems**: Robotic platform capabilities evolve from conventional, field-level task execution to efficient, resilient, and sustainable precision farming, i.e., moving beyond simple automation to genuine, autonomous intelligence (Yin et al., 2025b).
- **Crisis agents**: FMs enable robot operations to elaborate from conventional, remote-controlled settings to the handling of autonomous reasoning and adaptation circumstances in unpredictable, highly dangerous environments (Driess et al., 2023; Kim et al., 2025; Zitkovich et al., 2023).
- **Maritime robotics**: Robots are reinforced with advanced capabilities so as to efficiently overcome typical, extreme environmental challenges in under- and open-water settings, in principle relying on robust, generalized, and multi-sensorial intelligence/reasoning pipelines (Zheng et al., 2024b).
- **Space robotics**: FMs enable the functioning of robots under extreme operating conditions, involving limited resource availability and highly variable, unknown environments, largely relying on their enhanced autonomous decision-making capabilities (Giannakis et al., 2024; Zhao & Ye, 2024).

5 Foundation model types

FMs used in robotics can be organized into groups with respect to the number and the nature of the input-output modalities involved. The latter also largely affects their exhibited capabilities. In particular, the main types of FMs are: a) Large Language Models (LLMs), b) Vision Foundation Models (VFMs), c) Vision-Language Models (VLMs), and d) Vision-Language-Action Models (VLAs), as discussed in Section 4 and further detailed below.

5.1 Large Language Models (LLMs)

In the context of robotics, LLMs are primarily used as high-level, cognitive task planners and reasoning engines that generate the sequence of operations that are necessary for accomplishing a stated goal (Brohan et al., 2023b; Liang et al., 2023; Chen et al., 2024c). So far, they have been shown to be particularly effective for tasks that require sophisticated, logical reasoning and complex decision-making. In practice, LLMs translate high-level natural language inputs to low-level robot behaviors, by turning free-form instructions and short state summaries into typed goals, multi-step plans, executable code, constraints, and run-time feedback. Their main advantages are (Li et al., 2025a): a) Robustness in instruction translation, where free-form, flexible, human-like natural language instructions are transformed into formal, actionable representations, b) Efficiency in task decomposition and sequencing, where complex, long-horizon natural language instructions are mapped to concrete sequences of robot sub-tasks, and c) Increased generalization ability, where the large

amount of general-purpose knowledge incorporated in an LLM can boost the generation of robust action plans in diverse, complex, real-world environments. On the contrary, LLMs exhibit the following critical limitations (Tayyab Khan & Waheed, 2025): a) Lack of embodiment and grounding, where LLMs operate in a semantic space that does not incorporate connections to the physical environment; hence, leading to action plans that may not be feasible, b) Presence of hallucinations, where the LLM inference procedure may result into semantically/syntactically correct outcomes that, however, correspond to irrational actions, c) Increased latency, where the inherently high computational complexity of LLMs may not be suitable for real-time operational settings, and d) Input bias and sensitivity, where obscure or ambiguous input instructions may lead to unpredictable or biased action plans.

In terms of supported functionality/operation, LLM-based systems can be classified into the following main categories:

- Goal/constraint grounding and context: LLMs serve as a reasoning and translation interface that maps high-level, abstract human instructions (i.e., goal and context) to low-level, physical robot actions (i.e., grounding). In particular, SayCan (Brohan et al., 2023b) employs a base LLM mechanism (PaLM (Kim et al., 2024)) for filtering using value-based affordances, so that the selected skills to be maintained within capabilities and context. Additionally, LM-Nav (Shah et al., 2023a) converts instructions into visually grounded way-points that a planner can subsequently follow in a navigation setting. For larger spaces, SayPlan (Rana et al., 2023) associates language with 3D scene graphs, so that plans can remain consistent across different rooms and floors.

- Command interpretation and code synthesis: LLMs in principle act as natural language to code translators, enabling robots to receive high-level, human-like flexible instructions, instead of requiring the writing of conventional code. Specifically, Code-as-Policies (Liang et al., 2023) compiles instructions into robot Application Programming Interface (API) code using code-generating LLMs, which renders robot behavior easy to inspect and to reuse. Similarly, AutoTAMP (Chen et al., 2024c) translates requests into task-and-motion planning (TAMP) algorithmic specs that a symbolic planner checks for geometric and kinematic feasibility. Additional works focus on generating behavioral trees or decomposing tasks into formalized subproblems (Ao et al., 2025; Kwon et al., 2025; Liu et al., 2025c), synthesizing policy code from video-plus-text prompts (Xie et al., 2025), and constructing language-conditioned 3D value maps to guide placement and grasping (Huang et al., 2023c).

- Task planning and long-horizon reasoning: LLMs implement a high-level, cognitive mechanism that decomposes abstract human goals into sequential, grounded actions and maintains contextual awareness over multiple subsequent steps. In particular, SELP (Wu et al., 2025b) performs the mapping of language instructions to temporal logic representations, while combining constrained decoding with domain tuning, so that the generated plans to eventually satisfy safety and efficiency constraints. Similarly, LLM-GROP (Zhang et al., 2025g) cross-checks language-driven task instructions with motion feasibility and perceptual information for improving execution in cluttered settings. Moreover, hierarchical goal decomposition can speed up planning and robustness of long-horizon task execution (Kwon et al., 2025; Liu et al., 2025c; 2024b).

- Perception-aware and multimodal integration: The LLM core language processing functionality is enhanced by taking into account the robot's surrounding visual and physical world. In this context, PaLM-E (Driess et al., 2023) incorporates visual and proprioceptive information streams into the inference process, so that the model decisions to better correspond to real-world observations and actions. In a similar way, Chain-of-Modality (Wang et al., 2025a) employs prompts originating from both human-performing videos and auxiliary, associated signals (i.e., muscle or audio), in order to define both a task plan and corresponding control parameters from a single demonstration. Moreover, associating language inputs with visual grounding and 3D value maps can also facilitate more accurate robot manipulation (Huang et al., 2023c).

- Navigation and spatial understanding: LLMs translate abstract navigation commands into physically grounded, actionable paths within a map, primarily through the estimation of structured textual representations of the surrounding environment. For instance, LM-Nav (Shah et al., 2023a) links

human instructions with spatial landmarks and routes, via a CLIP-based grounding mechanism, and passes way-points to a low-level navigation module. Additionally, SayPlan (Rana et al., 2023) partitions a plan over a 3D scene graph into layers and, subsequently, performs re-planning when an action step is infeasible, keeping in this way long-horizon missions on track in large-scale environments.

- Conversational interfaces and teleoperation: LLMs create intuitive, conversational interfaces and enable shared-control teleoperation, making robots accessible to non-expert users. In particular, TidyBot (Wu et al., 2023a) learns user-specific tidying conventions through conversation and transfers them to novel domestic scenarios. Additionally, LAMS (Tao et al., 2025) predicts human intent and automatically switches teleoperation modes in assistive settings, aiming at lowering cognitive load during extended task executions.

- Execution-time validation, error handling, and recovery: LLMs translate technical run-time failures into semantic problems, enabling the definition of dynamic, common-sense solutions, instead of relying solely on pre-defined, baseline, brittle recovery routines. Specifically, STATLER (Yoneda et al., 2024) collects and interprets robot state and tool feedback information, allowing the suggestion of targeted repairs without restarting the overall execution pipeline. Similarly, CAPE (Raman et al., 2024) performs re-prompting and proposes concrete fixes in case of preconditioned failures, while CoPAL (Joublin et al., 2024) realizes re-planning when divergence incidents are detected. Moreover, uncertainty estimation over planning proposals can provide additional mitigation measures, prior to acting on hardware (Yin et al., 2025a).

- Adaptation, efficiency, and safety: LLMs facilitate robotic systems to be adaptive, efficient, and safe during task execution, by robustly handling novel situations and operational disruptions. In particular, Eureka (Ma et al., 2024a) writes and refines reward code using GPT-4 (Achiam et al., 2023), accelerating in this way skill acquisition across diverse robot platforms. Additionally, DrEureka (Ma et al., 2024b) co-designs rewards and domain randomization for efficient sim-to-real transfer. AutoRT (Ahn et al., 2024) implements simultaneously instruction handling, task assignment, and safety checking across multiple robots. Moreover, planners can encode safety rules directly (e.g., SELP's use of temporal logic) (Wu et al., 2025b), while recent works study jailbreak-style exploits and propose corresponding defense measures (Ravichandran et al., 2025).

- Knowledge retrieval and memory: LLMs leverage their ability to access, synthesize, and store massive amounts of information from external knowledge and past execution episodes, allowing systems to overcome limitations of their immediate sensorial data or within a single planning session. For instance, ELLMER (Mon-Williams et al., 2025) combines GPT-4 with a retrieval-augmented memory mechanism, so that a mobile manipulator can incorporate context, adapt plans on the fly, and complete multi-step household tasks as conditions change. Similar architectures are capable of performing inference over planning route and tool knowledge for on-demand lookup (Temiraliev et al., 2026; Anwar et al., 2025).

5.2 Vision Foundation Models (VFMs)

The ultimate goal of VFMs is to satisfy the perceptual requirements of embodied AI systems, by providing generalized, high-quality visual representations necessary for interaction with the physical world (Kirillov et al., 2023; Oquab et al., 2024). In the robotics setting, VFMs aim at distilling raw pixel information into rich, transferable visual features or embeddings. The latter enables a robust and generalized visual understanding that serves as the input information stream for modulating downstream policies, enabling the direct mapping of visual observations to specific control actions (Shang et al., 2024). Their main advantages are (Tayyab Khan & Waheed, 2025): a) Increased transfer learning capabilities, where visuomotor robot policies can efficiently generalize and acquire new skills, without requiring task-specific perception module learning from scratch, b) Open-world perception, where robots are enabled to recognize and to process previously unseen object and scene categories, c) Improved spatial awareness, where especially the incorporation of depth perception is crucial for enabling the execution of intricate actions with precision, and d) Increased robustness, where VFMs are shown to be less susceptible to visual distortions (e.g., presence of noise, variations in lighting conditions, etc.), compared to previous visual information processing modules. On the other hand, the

main limitations of VFMs are (Awais et al., 2025): a) Domain specificity, where VFMs can often perform sufficiently well in a relatively narrow range of domains, b) Incomplete physical world dynamics modeling, where VFMs often exhibit limitations to generalize robustly to subtle physical dynamics, long-range temporal correlations, causal coherence, and geometric properties, and c) Increased computational cost, which may result into critical constraints regarding real-time, uninterrupted operation.

In terms of supported functionality/operation, VFM-based systems can be classified into the following main categories:

- Object recognition: VFMs enable generalized visual recognition, by providing transferable representations that reduce the need for task-specific perception training. In particular, SAM (Kirillov et al., 2023) provides class-agnostic, promptable segmentation masks that facilitate isolating objects and parts, while DINOv2 (Oquab et al., 2024) provides robust dense visual features that transfer well across scenes and support recognition and retrieval under domain shift (Kirillov et al., 2023; Oquab et al., 2024). Additionally, 3D-MVP (Qian et al., 2025a) uses a multi-view encoder to learn object and part-level representations that are useful for recognition and downstream manipulation. Moreover, ZeroGrasp (Iwase et al., 2025) couples recognition with reconstruction, estimating object geometry and predicting grasp poses from a single RGB-D observation in near real-time.

- Localization: VFMs enhance robot localization by estimating robust, semantic, and globally consistent visual representations, outperforming conventional geometric methods. Specifically, DINO-VO (Azhari & Shim, 2025) utilizes DINOv2-generated features and a ViT-based keypoint estimation scheme for improving robustness and generalization in monocular visual odometry at high throughput rates. Additionally, LiteVLoc (Jiao et al., 2025) and ZeroVO (Lai et al., 2025) support long-range re-localization for image-goal navigation and zero-shot cross-camera visual odometry, respectively.

- Object tracking: VFMs provide trackers with rich, semantic understanding and robust long-term memory, increasing robustness with respect to occlusion, viewpoint change, and category drift. In particular, Zhong et al. (2024) use a pre-trained VFM to extract semantic segmentation masks with text prompts, while a recurrent policy network with offline reinforcement learning is subsequently trained from the collected demonstrations. Additional approaches make use of open-vocabulary or frozen-backbone cues for instance tracking and segmentation, which is particularly useful for handling novel objects over time (Guo et al., 2025a; Fang et al., 2025a).

- Depth perception: VFMs enable robust, generalized, and high-fidelity depth estimation, especially in scenarios where specialized sensors or traditional methods tend to under-perform. Specifically, Metric3D v2 (Hu et al., 2024) employs geometric priors (namely, canonical camera transformations and joint depth–normal optimization) and supports zero-shot metric depth estimation across diverse camera settings. Additionally, Prompt-Depth-Anything (Lin et al., 2025a) demonstrates that a small LiDAR 'metric prompt' can steer a FM to accurate, high-resolution metric depth estimation. Similarly, DepthCrafter (Hu et al., 2025b) generates temporally consistent long depth sequences with intricate details for open-world videos.

- Semantic map creation: VFMs provide robust, semantic-aware features that enhance accuracy, robustness, and informativeness of the generated maps. In particular, Busch et al. (2025) create reusable open-vocabulary feature maps, capable of supporting probabilistic-semantic updating for informed multi-object exploration. In a similar way, VFM feature representations can be combined with Gaussian-splatting or factor-graph mechanisms for robust long-horizon missions in dynamic environments (Zheng et al., 2025a; Yugay et al., 2025).

- Visual-inertial fusion: VFMs enhance the visual part of visual-inertial odometry (VIO) systems, which is crucial for drift correction and estimation of metric scale. Specifically, features that improve VO (Azhari & Shim, 2025) or metric priors from depth FMs (Hu et al., 2024) are combined with Inertial Measurement Unit (IMU) data in standard VIO estimators. Additionally, depth-foundation priors are injected to stereo/VIO pipelines for stabilizing scale and short-horizon pose, while being complemented by IMU FMs for cross-platform generalization (Jiang et al., 2025a; Zhao et al., 2025b).

- Environment mapping: VFMs enable map building by producing dense visual embeddings that can be fused into persistent, semantically enriched scene representations. FMGS (Zuo et al., 2025) combines FM features with 3D Gaussian splatting for semantic 3D scene reconstruction and open-vocabulary scene understanding. Similarly, OpenGS-SLAM (Yang et al., 2025a) and FeatureSLAM (Thirgood et al., 2026) extend Gaussian-splatting-based mapping with FM-derived semantic features, enabling open-set scene understanding and improving the robustness of real-time tracking and mapping.

5.3 Vision-Language Models (VLMs)

VLMs combine computer vision with natural language processing capabilities for establishing a coherent, concrete semantic understanding of the world, enabling interpretation and generation of language descriptions of the observed visual entities (Radford et al., 2021). In the context of robotics, VLMs enable robots to simultaneously interpret visual data and natural language commands, allowing for intuitive human-robot interaction and robust task execution in unstructured environments (Zhou et al., 2025c). Their main advantages are (Tayyab Khan & Waheed, 2025): a) Richer semantic information and environment understanding, where VLMs generate detailed, interpretable semantic outputs that boost robots to handle complex and rare/novel scenarios more efficiently, b) Open-vocabulary object recognition, where the use of the VLMs' shared visual-language embedding space significantly facilitates open-world perception, c) Flexible perception interface, where embodied agents are capable of supporting nuanced queries and generalized reasoning, and d) Enhanced generalization capability, where VLMs allow for efficient handling of novel visual and linguistic combinations. On the contrary, the main limitations of VLMs are (Tayyab Khan & Waheed, 2025): a) Inability to define precise actions, where VLMs are capable of interpreting instructions and identifying visual entities, but they lack the intrinsic capability to translate this semantic understanding into precise, executable motor control commands, b) Incomplete semantic grounding, where VLMs exhibit difficulties in connecting abstract, human-like commands to concrete, actionable physical locations and poses required for robot manipulation tasks in real-world environments, and c) Dependency on external policy generators, where VLMs inherently require a dedicated external policy or low-level planner to provide their semantic output for implementing robot actions.

With respect to supported functionalities/operations, VLM-based systems can be classified into the following main categories:

- Manipulation grounding and control signals: VLMs map high-level, semantic intents into concrete, physically-actionable constraints in the proximity of the robot's operating environment. In particular, OmniManip (Pan et al., 2025c) translates VLM reasoning outcomes to object-centric interaction primitives and, subsequently, implements dual closed-loop procedures (namely, planning and execution) to produce precise 3D spatial constraints. Additionally, RoboGround (Huang et al., 2025a) provides grounded masks for both targets and placement regions into a low-level policy for improving generalization across different types of scenes. KUDA (Liu et al., 2025d) poses queries to a VLM for task keypoints and, subsequently, converts them into optimization costs for model-based planning, enabling open-vocabulary manipulation over rigid, deformable, and granular objects. Moreover, chain-of-modality-style schemes prompt a VLM on human videos and auxiliary signals for extracting step-wise plans and control parameters (Wang et al., 2025a). In a similar way, IKER (Patel et al., 2024) and ReWiND (Zhang et al., 2025d) refine visually-grounded rewards or costs over long-horizon tasks for stabilizing execution, respectively.

- Semantic mapping, referring expressions, and navigation: VLMs are capable of creating human-readable maps, understanding complex spatial language, and grounding navigation goals in the physical world, materializing the transition from purely geometric to semantic navigation. In particular, One-Map-to-Find-Them-All (Busch et al., 2025) creates a reusable open-vocabulary feature map for real-time, zero-shot, multi-object-oriented navigation, while supporting probabilistic semantic updates. Additionally, functional and relational reasoning at scene level is enabled by the use of open-vocabulary functional 3D scene graphs (Zhang et al., 2025b) and open-scene graphs (Loo et al., 2025) for open-world object-based navigation. OpenVIS (Guo et al., 2025a) incorporates open-vocabulary video instance segmentation and tracking for pursuing tools or novel objects. Moreover,

Vision-Language Fly (VLFly) (Zhang et al., 2025j) supports grounded vision-language navigation for Unmanned Aerial Vehicles (UAVs) with open-vocabulary goal understanding, without requiring localization or active ranging sensors.

- Execution-time check and progress verification: VLMs enable robot, semantic, self-monitoring within a closed-loop control framework, materializing the translation of low-level, sensor feedback to high-level, human-understandable checks. Specifically, ExploreVLM (Lou et al., 2025) deploys a closed-loop task planning framework for real-time integration of perception, planning, and execution validation. Additionally, Ahmad et al. (2025) introduce a unified framework for real-time failure recovery, where a VLM acts as a monitoring tool for verifying pre- and post-conditions of individual skills, inferring missing pre-conditions, and suggesting new skills for recovery. Moreover, Guardian (Pacaud et al., 2025) enhances the robot abilities to detect manipulation planning and execution errors, by identifying fine-grained failure modes (e.g., object slippage, incorrect action sequencing, etc.).

- Closed-loop mobile manipulation: VLMs provide the necessary cognitive capabilities regarding continuous feedback and adaptation, so as to directly address the inherent long-horizon execution challenge in unstructured, large-scale, and dynamic environments. For instance, COME-robot (Zhi et al., 2025) uses GPT-4 for situated reasoning and iterative feedback, in order to recover from failures. Additionally, HomeRobot (Yenamandra et al., 2023) relies on an agent that is capable of navigating through household environments for grasping novel objects and placing them on target receptacles.

5.4 Vision-Language-Action models (VLAs)

VLAs aim at integrating multi-modal understanding with direct physical execution, targeting to serve as the basis for autonomous embodied task execution (Ma et al., 2024c). In particular, a VLA model receives multi-modal inputs (typically, vision, language, and robot state) and generates real-world physical actions or control policies in real-time, often designed in an end-to-end way. Their main advantages are (Sapkota et al., 2025): a) End-to-end implementation and operational simplicity, where VLAs perform the direct mapping from perception signals to control actions, eliminating the need for complex interconnection of multiple, distinct planning modules, b) Robust generalization ability, where VLAs are shown to demonstrate increased generalization performance across different operating environments and robotic platforms, and c) Integration of action and reasoning, where VLAs by design address the multi-modal, semantic grounding problem internally. On the other hand, the main limitations of VLAs are (Sapkota et al., 2025): a) High demand for large-scale training data, where VLAs require outstanding amounts of high-quality, heterogeneous action data, b) Decreased efficiency across multiple robotic platforms, where unavailability of not always sufficiently-broad robotic training datasets leads to suboptimal task performance across different robotic setups, c) Increased operational latency, where the large-scale nature of the underlying VLA model architectures leads to critical challenges in real-time control settings, and d) Increased complexity in failure management, where errors in end-to-end VLA systems may affect significantly the model's overall behavior.

In terms of supported functionality/operation, VLA-based systems can be classified into the following main categories:

- Scaling and web-to-robot transfer: VLAs target the transferring of vast semantic, visual, and common sense knowledge acquired from the internet to the robots' physical-world control policies, while reinforcing generalization across different tasks and embodiments. In particular, RT-1 (Brohan et al., 2023a) scales-up imitation learning capabilities using tokenized action sequences corresponding to language formalized goals, improving robustness in long-tail household task execution. Its successor RT-2 (Zitkovich et al., 2023) incorporates web-scale vision–language pretraining in the control pipeline, so that open-vocabulary knowledge transfer to be performed to real-world robots, through the use of a language-aligned visual encoder. Additionally, OpenVLA (Kim et al., 2025) integrates a vision–language encoder and a relatively small action head, while supporting different robotic platforms and maintaining increased zero/few-shot generalization capabilities. Similar approaches incorporate stronger inductive biases and specific 3D spatial priors (Li et al., 2025c), object-centric

adapters for few-shot tuning (Li et al., 2025b), standardized cross-embodiment corpora (O'Neill et al., 2024), multiple low-rank adaptation modules (Zhao et al., 2025a), and policy distillation schemes (Xu et al., 2025).

- Fusion and action parameterization: VLAs aim at unifying perception, reasoning, and control by combining a VLM with an action decoder. In particular, GR00T N1 (Bjorck et al., 2025) tightly couples a VLM for interpreting the environment, through vision and language instructions, with a subsequent diffusion transformer for generating motor actions in real-time. Additionally, $\pi_{0.5}$ (Black et al., 2025) formalizes action generation as a flow matching process on top of a VLM, supporting stable, continuous control and increased generalization at moderate computing requirements. Moreover, similar systems further improve the interaction between the VLM and the action decoder through several complementary design choices, including diffusion models for enhanced dexterity (Wen et al., 2025b), unified autoregressive–diffusion heads (Liu et al., 2026a), rectified-flow policies (Reuss et al., 2025), state-space formulations (Liu et al., 2024c), short-horizon video prediction (Hu et al., 2025c), and embodied reasoning mechanisms (Team et al., 2025).

- Specialization, adapters, and mixture-of-experts: VLAs accomplish to leverage massive pre-trained backbones without the need/cost of parsing/running the entire network for every predicted action. Specifically, MoRE (Zhao et al., 2025a) makes use of sparse Low-Rank Adaptation (LoRA) experts, selecting only a few of them per step for expanding the model's capacity without increased inference cost. Additionally, OpenVLA (Kim et al., 2025) combines a Llama 2 language model with a visual encoder and implements efficient fine-tuning for new tasks, boosting robust, generalizable policies for visuo-motor control. Moreover, RLDG (Xu et al., 2025) combines task-specific Reinforcement Learning (RL) with generalist policy distillation for more efficient robotic manipulation.

- Navigation and locomotion: VLAs serve as semantic, navigation planners that translate abstract, natural-language instructions into physically executable movements by mobile and legged robots, often adopting a hierarchical design. In particular, NaVILA (Cheng et al., 2025) initially generates mid-level actions with spatial information in the form of language and, subsequently, utilizes this as input to a visual locomotion RL policy generator for execution. Additionally, VAMOS (Castro et al., 2025) comprises a hierarchical VLA that decouples semantic planning from embodiment grounding, where a generalist planner learns from diverse, open-world data and a specialist affordance model encodes the robot's physical constraints and capabilities. Moreover, Humanoid-VLA (Ding et al., 2025) integrates language-motion pre-alignment (using non-egocentric human motion data paired with textual descriptions), and egocentric visual context (through parameter efficient video-conditioned fine-tuning).

- Operations, deployment, and safety: VLAs aim at combining their high-level, flexible, and generalized reasoning capabilities with conventional safety guarantees for real-world deployment. Specifically, SafeVLA (Zhang et al., 2025a) formalizes safety alignment as a constraint learning problem, targeting the VLA operation to respect task rules and safety measures, and not relying only on post-hoc filtering. VLATest (Wang et al., 2025h) comprises a framework designed to generate robotic manipulation scenes for testing VLAs. Moreover, fleet orchestration frameworks combine language interfaces with guardrails and human-in-the-loop supervision, while mechanistic steering implements zero-shot behavior shaping (Ahn et al., 2024; Häon et al., 2025).

5.5 Comparative analysis and key insights

Having discussed in detail the various types of FMs (Sections 5.1-5.4), this section systematically examines the literature methods, providing a comparative analysis and critical insights for each methodological category. In this respect, Table 3 summarizes for each type of FM its: a) Primary functions, b) Input types, c) Output types, d) Key strengths, e) Critical limitations, and f) Indicative models. Among the various observations and insights, it can be seen that LLMs excel at high-level, long-horizon planning (e.g., task decomposition), whereas VFMs are optimized for real-time, low-level perception (e.g., feature extraction). Additionally, VLMs serve as the connection between human language and visual perception (e.g., object grounding), while VLAs

Table 3: Foundation model types: Comparative analysis and key insights.

Aspect	LLMs	VFMs	VLMs	VLAs
Primary functions	• High-level cognitive task planning • Symbolic reasoning • Translation of high-level natural language inputs to low-level robot actions • Complex task decomposition	• Generalized, high-quality visual representations • Dense, rich visual features or embeddings • Object/instance differentiation • Geometric scene reconstruction	• Natural language grounding in vision • Open-vocabulary recognition • Visual reasoning • Multi-modal semantic alignment	• End-to-end policy execution • Direct hardware actuation • Embodied task execution • Alignment of multi-modal understanding with physical interaction
Input	• Text tokens • Natural language instructions and task goals • Reasoning traces • Code snippets • Environment descriptions • Previous conversations and memory	• RGB images • RGBD video streams • 3D point clouds • LiDAR data • Camera specs	• Aligned image-text pairs • Visual prompts • RGBD video streams • Task-conditioned scene descriptions	• RGB-D video • Text instructions • Proprioceptive states • Haptic/tactile feedback • Action execution trajectories • Success/failure signals
Output	• Logic-based sub-goals • Code snippets • Symbolic language plans • Safety constraints • Feedback messages	• High-level visual features and embeddings • Pixel-level segmentation masks • Object detection and tracking • Depth estimates • Surface keypoint detection	• Image/video captions • Semantic textual descriptions • VQA answers • Semantically grounded maps • Multi-modal vector alignment	• Low-level motor commands • End-effector poses • Discrete/continuous action policies • Failure management routines
Strengths	• Robust natural language instruction translation • Efficient complex task decomposition and sequencing • Increased generalization ability • Easy interaction with humans	• Increased transfer learning capabilities • Open-world perception • Improved spatial awareness • Increased robustness to distortions	• Rich semantic information and environment understanding • Open-vocabulary object recognition • Flexible perception interface • Enhanced generalization to novel entities	• End-to-end implementation and operational simplicity • Robust generalization across different platforms • Integration of action and reasoning pipelines
Limitations	• Lack of embodiment and grounding • Presence of hallucinations • Increased latency for real-time control • Input bias and sensitivity	• Domain specificity • Incomplete physical world dynamics modeling • Increased computational cost	• Inability to define precise actions • Incomplete semantic grounding • Dependency on external policy generators	• High demand for large-scale training data • Decreased efficiency across multiple robotic platforms • Increased operational latency • Increased complexity in failure management
Indicative models	• BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), Llama 3 (Grattafiori et al., 2024), DeepSeek-V3 (Liu et al., 2024a)	• ViT (Yuan et al., 2021), DINOv2 (Oquab et al., 2024), SAM (Kirillov et al., 2023), R3M (Nair et al., 2023), VC-1 (Majumdar et al., 2023)	• CLIP (Radford et al., 2021), OWL-ViT (Minderer et al., 2022), BLIP (Li et al., 2022a), SigLIP (Zhai et al., 2023)	• RT-1 (Brohan et al., 2023a), PaLM-E (Driess et al., 2023), RT-2 (Zitkovich et al., 2023), OpenVLA (Kim et al., 2025), Octo (Mees et al., 2024)

represent the complete, end-to-end control policy that directly maps multi-modal inputs to robot actions. Moreover, representative literature methods per FM type are illustrated in Fig. 4.

6 Neural network architectures

The type of the underlying neural network architecture that is employed in a FM solution largely dictates its capabilities. The main categories of NNs used in robotic FM methods are: a) Transformers, b) State-Space Models (SSMs), c) Diffusion Models (DMs), d) Convolutional and hybrid encoders, and e) Graphical models, as discussed in Section 4 and further detailed below.

6.1 Transformers

In the context of robotics, transformers are widely used for different/diverse tasks (e.g., high-level task planning, low-level policy learning, perception, human-robot interaction, etc.), relying on the fundamental principle of formalizing them as a sequence modeling problem (Firoozi et al., 2025). The latter is grounded on converting input data (e.g., states, actions, images, etc.) to numerical tokens, which are then processed as a sequence for estimating a prediction. Their main advantages are (Sanghai & Brown, 2024): a) Increased long-range dependency modeling, where transformers enable policies to capture complex, long-range dependencies between states, actions, and rewards over long time horizons, b) Architectural homogenization, where a single, well-understood transformer architecture can be used as the backbone for diverse robotic tasks and input modalities, and c) Increased efficiency in high-level planning and reasoning, where the primary origin of the transformer architecture in language modeling also makes it suitable for high-level, language-conditioned task planning. On the other hand, the main limitations of transformers are (Firoozi et al., 2025): a) Increased

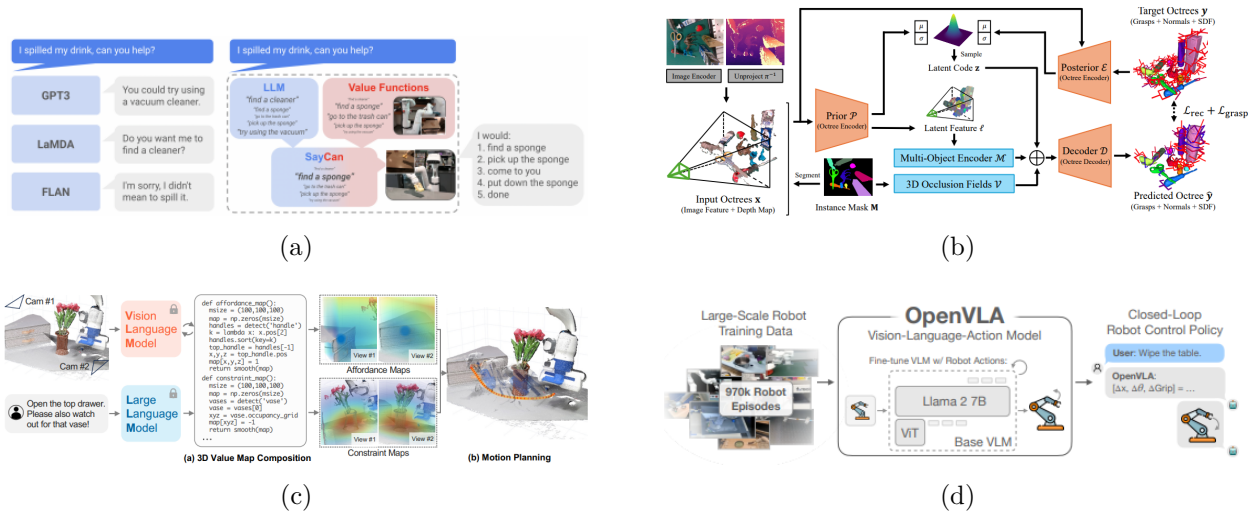


Figure 4: Representative literature methods per FM type: (a) LLMs (SayCan (Brohan et al., 2023b)), which receive natural-language instructions as input and output grounded skill selections or high-level plans; (b) VFMs (ZeroGrasp (Iwase et al., 2025)), which receive visual observations as input and output perceptual representations, including object reconstruction and grasp pose predictions; (c) VLMs (VoxPoser (Huang et al., 2023c)), which receive language and scene observations as input and output grounded affordance and constraint representations; and (d) VLAs (OpenVLA (Kim et al., 2025)), which receive multimodal inputs and directly output robot-executable actions.

computational complexity, where the transformer self-attention mechanism compute and memory requirements scale quadratically with the input sequence length; hence, often making pure transformers unsuitable for low-latency, real-time control loops, b) Requirement for discrete tokenization, where transformers inherently operate on discrete tokens, which in turn introduces quantization error and is a fundamentally unnatural representation of physical dynamics, and c) Infinite context contradiction, where a pure-form transformer exhibits outstanding reasoning performance over a finite context, but it is architecturally unsuited for infinite-horizon, continuous processing required by embodied agents.

With respect to the input data modality, transformer-based systems can be classified into the following main categories:

- Vision Transformers (ViTs): ViTs comprise the fundamental architecture of multiple VFMs in robotics. They use self-attention to capture global image relationships, unlike traditional local convolutional methods (Yuan et al., 2021). Models like DINOv2 use self-supervised learning on web-scale datasets to estimate general-purpose visual features that transfer across tasks without extensive fine-tuning (Oquab et al., 2024). As a consequence, ViTs have a central role in several perception pipelines in robotics. One such key area is scene understanding, which focuses on the immediate, local environment (including 3D scene understanding) for realizing manipulation, localization, and visual odometry (Azhari & Shim, 2025; Martins et al., 2025a). Additionally, in the context of semantic mapping, ViT features facilitate the construction/maintenance of long-term, global, semantic maps, supporting open-vocabulary representations for multi-object navigation, SLAM, and targeted exploration (Busch et al., 2025; Laina et al., 2025; Martins et al., 2025b; Jiang et al., 2025b; Deng et al., 2025b). Moreover, ViTs also provide essential geometric cues for low-level control, by generating dense spatial signals (such as zero-shot metric depth and surface normals) from a single image (Hu et al., 2024; Yang et al., 2024b; Guo et al., 2025b); these signals, along with grounded spatial constraints, support closed-loop execution and grasp planning in manipulation systems (Huang et al., 2025a).
- Text transformers: Text transformers act as the language interface, planner, programmer, memory, and supervisor across robotic pipelines, providing a common semantic layer that connects high-level

intent to grounded perception and control. In particular, bidirectional encoders (e.g., BERT) learn transferable semantics for downstream modules (Devlin et al., 2019) and decoder-only models (e.g., PaLM (Kim et al., 2024) and GPT-4 (Achiam et al., 2023)) benefit from scale, improving robotic reasoning and planning performance. Also open-weight LLaMA backbones enable practical on-device deployment (Touvron et al., 2023). In terms of exhibited functionality, text transformers can translate free-form language into structured plans and formal artefacts, by decomposing long-term goals, filling behavioural trees, and generating PDDL templates that pass feasibility checks (Ao et al., 2025; Liu et al., 2025c; Zhang et al., 2025g). Similarly, the same models can generate or adapt policy code from language or video inputs, enabling low-level controllers to remain auditable and closing the gap between high-level intent and executable actions (Liang et al., 2023; Xie et al., 2025; Ji et al., 2026). Additionally, retrieval-augmented setups can ground the planner in external memory and past context, in order to improve long-horizon behaviours, to sustain states across sub-goals, and to reduce drift during environmental changes (Mon-Williams et al., 2025; Gu et al., 2024; Anwar et al., 2025). In case of action execution divergence from the planned one, text transformers are able to propose corrections, to reason about failed preconditions, and to maintain explicit state estimations to prevent cascading errors/effects (Yoneda et al., 2024; Raman et al., 2024; Joublin et al., 2024). Moreover, even in generalist vision–language–action frameworks, the textual component remains on top of the formed reasoning layer, efficiently modulating the respective perception and control ones (Zitkovich et al., 2023; Mees et al., 2024; Bjorck et al., 2025; Kim et al., 2025; Team et al., 2025).

- Multi-modal transformers: These models transform each individual sensor stream into a sequence of tokens, project them to a common/shared latent space, and subsequently align them using a cross-attention or a gated fusion mechanism. The training process combines alignment and generative objectives, so that a single formed backbone network to be capable of understanding scenes, following instructions, and selecting actions (Kim et al., 2025; Zitkovich et al., 2023; Driess et al., 2023; Team et al., 2025; Mees et al., 2024; Bjorck et al., 2025). Multi-modal transformers are also capable of combining vision with proprioception information, where the fusion of visual and robot state tokens allows the generation of embodiment-aware policies that can scale across multiple/diverse robots and tasks (Wang et al., 2024b; Mees et al., 2024; Zitkovich et al., 2023). An additional capability comprises the enrichment of vision with geometric cues, where the integration of metric depth and surface normals can strengthen mapping and grasp planning with dense geometry that can generalize to novel scenes without requiring task-specific labels (Hu et al., 2024; Yang et al., 2024b; Huang et al., 2025a). Moreover, multi-modal transformers enable the generation of unified tactile–vision embeddings that can boost reasoning capabilities regarding contact, texture, and stability; hence, improving manipulation under uncertainty and occlusion (Yang et al., 2024a; Feng et al., 2025). On another direction, joint audio–visual tokens can support navigation in noisy, multi-source settings and improve robustness in the presence of distractions (Shi et al., 2025; Park et al., 2026). Furthermore, transformer-based fusion of thermal and RGB information streams can improve perception and localization, especially under low light and adverse weather conditions (Puttagunta et al., 2024; Skorokhodov et al., 2026). More recently, lightweight adapter modules are widely used for enabling the incorporation of additional sensors without retraining the full model, which comprises a common pattern shared across VLA and heterogeneous pre-training schemes (Kim et al., 2025; Wang et al., 2024b; Team et al., 2025).

6.2 State–space models

SSMs are increasingly adopted in robotics for realizing real-time control and long-horizon reasoning, by treating the sensorial input streams as latent states and subsequently producing output predictions in a step-by-step way (Liu et al., 2024c). The latter is in practice performed by learning end-to-end system matrices with diagonal-plus-low-rank parameterizations and hardware-aware modeling (Gu & Dao, 2024; Dao & Gu, 2024). Their main advantages are (Gu et al., 2022; Smith et al., 2023): a) Linear scaling of complexity, where computation and memory requirements grow linearly with sequence length, b) Stable long-horizon memory, where the latent state maintains temporal context without large activation caches, and c) Deployment efficiency, where low latency and steady throughput are suitable for embedded robotic solutions. On the contrary, the main limitations of SSMs are (Lenz et al., 2025): a) Reduced cross-token

check, since explicit/direct correlation between tokens is not inherently supported, b) Reduced global context modeling, compared to full attention-based counterparts, and c) Frequent need for hybrid designs, which typically integrate attention or retrieval blocks for implementing tool use and long-range planning.

With respect to the input modalities, SSM-based systems can be classified into the following main categories:

- Visual SSMs: Visual SSMs can replace attention mechanisms with corresponding selection-based ones, resulting into linear-complexity encoders that can be plugged into recognition, dense prediction, and tracking heads (Liu et al., 2024f; Xiao et al., 2025a). Additionally, scene understanding and object tracking can benefit from long temporal context modeling at constant cost, which is particularly suitable for long video streams and multi-camera setups (Park et al., 2024). Moreover, event-driven perception can be boosted to handle irregular sampling and rapid environmental dynamics, which is beneficial for agile navigation and manipulation in low light or high motion settings (Zubic et al., 2024).

- Policy/control SSMs: SSMs can be used for realizing data-driven nonlinear reduction in complex systems, such as modeling hysteresis and memory effects. In particular, table-top manipulation can adopt region-aware selective-state policies with flow-matching objectives to learn precise, real-world skills from limited demonstrations (Wang et al., 2025c). Additionally, hybrid selective-state diffusion policies may reduce parameters while maintaining performance, in order to improve sample efficiency under multi-view inputs and long horizons (Cao et al., 2025a). Moreover, vision-driven locomotion can incorporate depth and proprioception information, through stacked selective-state formalisms and end-to-end reinforcement learning (Wang & Tao, 2026).

- Multimodal SSMs: A single SSM trunk can be simultaneously pretrained on long videos, robot logs, and demonstrations, in order to support perception, planning interfaces, and action heads. In particular, SSM-based VLA models can fuse vision, language, and proprioception inputs as token streams and, subsequently, generate actions with lower latency and memory than attention-only implementations (Tsuji, 2025), which makes the application of long-horizon policies more efficient and practical (Liu et al., 2024c). When explicit/direct lookup, tool use, or long-range queries are required, attention or retrieval layers can be integrated on top of the main SSM model (Lenz et al., 2025; Wang et al., 2025f).

6.3 Diffusion models

DMs can generate robot behaviours by reversing a gradual noising process, where a forward pass increasing adds Gaussian noise to the input data and, subsequently, a learned reverse model denoises back to the original input space (e.g., actions, trajectories, sub-goals, etc.) (Ho et al., 2020; Song et al., 2021). In particular, DMs implement control policies as a conditional denoising process and have been shown robust across different manipulation tasks (Chi et al., 2025), while they can also serve as generative heads on top of pretrained vision and multimodal backbones (Kapelyukh et al., 2024; Zeng et al., 2024). Their main advantages are (Liang, 2025): a) Multi-modal feature modelling, where sampling can produce diverse, uncertainty-aware candidates that can facilitate planning under partial observability and contact variability (Janner et al., 2022; Chi et al., 2025), b) Composite conditioning, where a single denoising process can employ frozen vision, language, depth, and proprioceptive backbones to modulate goals and action chunks without task-specific labels (Kapelyukh et al., 2024; Zeng et al., 2024; Ze et al., 2024), and c) Trajectory-level decision making, where value or constraint guidance can steer full roll-outs towards safe and feasible plans, while improving long-horizon behavior (Janner et al., 2022). On the contrary, the main limitations of DMs are (Wolf et al., 2025): a) Sampling latency and energy cost, stemming from the inherent iterative denoising process (Dong et al., 2024), b) Lack of built-in safety guarantees, where constraint checks or guided objectives are needed to avoid collisions and dynamics violations (Janner et al., 2022; Wolf et al., 2025), and c) Sensitivity to conditioning drift, where errors/noise in visual or language features can mislead the sampling process (Chi et al., 2025; Ze et al., 2024).

With respect to the conditioning type, DM-based systems can be classified into the following main categories:

- Vision-conditioned DMs: DMs can translate visual goals into usable, structured information for control purposes. In particular, image-goal generation and rearrangement of priors can estimate object- and scene-level targets that downstream controllers can subsequently follow (Kapelyukh et al., 2023; Zeng et al., 2024). Additionally, pretrained image-editing DMs can generate sub-goal images from language instructions and current camera views, guiding goal-conditioned policies in real-world settings (Black et al., 2024). Moreover, compact 3D visual tokens can be employed for improving spatial grounding and robustness across different viewpoints for manipulation planning (Chi et al., 2025; Ze et al., 2024; Kapelyukh et al., 2024).
- Proprioception-, force-, and haptic-conditioned DMs: Visuomotor diffusion policies treat action sequences as denoised samples conditioned on images and robot states, which facilitates in handling multi-modal actions and improving stability for manipulation tasks (Chi et al., 2025). When force or contact requirements are present, conditioning on haptics and force signals can be incorporated, for example, in visual–tactile slow–fast policies for contact-rich skills (Shukla et al., 2025; Xue et al., 2025). In order to maintain control loops short, progressive refinement can increase prediction rates up to real-time performance (Dong et al., 2024).
- Language-conditioned DMs: Textual inputs can serve as a guiding signal, where the denoising mechanism can modulate goals, trajectories, or sub-goals, simplifying task setup and execution (Bjorck et al., 2025). Additional works demonstrate the growing use of language prompts for manipulation and planning tasks based on diffusion backbones (Wolf et al., 2025; Liang, 2025).
- Human behaviour-conditioned DMs: Diffusion objectives can be defined so that early-stage human motion detection can facilitate the accurate prediction of intent, improving intuitiveness and comfort in human–robot interaction without changing the controller structure. In this context, the Legibility Diffuser (Bronars et al., 2024) demonstrates that a policy trained on offline demonstrations can generate intent-expressive collaborative motions that humans find easier to understand, while still completing a given task more efficiently (Ng et al., 2023).

6.4 Convolutional and hybrid encoders

Visual encoders typically comprise the main perception module of any robotic solution, translating raw pixel data into latent representations that a robot can use for subsequent planning procedures (Nair et al., 2023). The selection between CNNs and hybrid CNN-transformer implementations essentially comprises a decision on the trade-off between local spatial precision and global context modeling, respectively. Their main advantages are (Brohan et al., 2023a; Tan & Le, 2019): a) Zero-shot generalization, where due to the fact that the models are trained on internet-scale datasets, previously unseen objects can often be recognized, b) Robustness to noise, exhibiting increased resilience to changes in lighting, shadows, or cluttered backgrounds, and c) Reduced need for training data, where 'frozen' pre-trained encoders often exhibit increased performance, without re-training. On the contrary, their main limitations are (Ma et al., 2023): a) High computational latency, due to the typical high-scale of FMs, b) Loss of fine-grained details, where encoders often divide processing of images in patches for reducing memory requirements, c) Prone to distribution shifts, where there might be a significant discrepancy between the internet-scale training data and the specific application images, and d) Lack of temporal consistency, due to many visual encoders processing video streams in an independent frame-by-frame fashion.

With respect to the encoder and integration type, the following main categories can be identified:

- CNN encoders: CNNs excel at capturing low-level spatial details like edges, textures, and object boundaries, due to their local receptive fields. In particular, ResNet-series encoders are used for high-fidelity state representations in diffusion-based exploration tasks in (Cao et al., 2025b). Additionally, EfficientNet-B3 is employed as a visual encoder to facilitate real-time, goal-conditioned navigation and exploration in (Sridhar et al., 2024). Additionally, R3M (Nair et al., 2023) freezes a ResNet–50 trained on Ego4D for improving manipulation for both simulation and real-world scenarios. Moreover, language-reasoning segmentation masks generated by internet-scale trained encoders are leveraged to condition robot manipulation tasks in (Yang et al., 2025b).

- CNN–transformer hybrids: Hybrid architectures often combine a CNN for handling pixel-level information and a transformer one for addressing context and action aspects. In particular, RT–1 (Brohan et al., 2023a) encodes frames with a FiLM–conditioned EfficientNet (Tan & Le, 2019), compresses them with TokenLearner, and then predicts discrete actions using a transformer, enabling real–world task control. Additionally, BC–Z/PaLM–SayCan (Jang et al., 2022; Brohan et al., 2023b) employ a lightweight ResNet coupled with a shallow attention network for supporting instruction–conditioned policies. In a similar way, diffusion–transformer policies may also adopt convolutional components as image tokenizers prior to respective transformer layers (Dasari et al., 2025).
- CNN tokenizers inside generalist agents: Generalist agents typically convert high–resolution images into compact tokens, prior to sequence modeling. In particular, Gato (Reed et al., 2022) employs a small ResNet image tokenizer and feeds visual, text, and proprioception information to a single transformer for controlling multiple skills. RoboCat (Bousmalis et al., 2024) makes use of a pretrained VQ–GAN image tokenizer and a transformer network, in order to adapt across robots and tasks according to a sequence of self–improvement cycles. Moreover, CNN tokenizers achieve to maintain low information bandwidth and to preserve an efficient interface to large sequence models for various navigation and manipulation tasks (Shah et al., 2023c).
- CNN–conditioned diffusion policies: Diffusion policies often condition a temporal U–Net on CNN features for generating action chunks that are diverse, yet feasible. In particular, Diffusion Policy (Chi et al., 2025) employs ResNet–18 features for robot manipulation, while preserving low added latency. DiffuserLite (Dong et al., 2024) demonstrates that progressive refinement with a frozen MobileNet–V3 encoder can increase prediction rates towards real-time performance on embedded platforms. Additional works concentrate on similar CNN–based diffusion frameworks for imitation and reinforcement learning purposes (Liang et al., 2025; Dasari et al., 2025).

6.5 Graphical models

In the context of robotic FM methods, the use of graphs introduces additional capabilities towards the goal of connecting low-level, raw sensorial data and high-level, structured reasoning (Maggio et al., 2024; Booker et al., 2024). Unlike other NN architectures that process input data as matrices (e.g., images) or sequences (e.g., text tokens), graphs enable the modeling and interpretation of the surrounding environment as a set of interconnected entities. For example, scene graphs often decompose the visual world into nodes (e.g., objects, parts, or agents) and edges (e.g., spatial, semantic, or functional relationships), essentially aiming at modeling their inter-dependencies. Their main advantages are (Gu et al., 2024): a) Combinatorial generalization, where the learning of relationships among entities (instead of specific instances) can boost the generalization to previously unseen cases, b) Permutation invariance, due to the inherent ability of graphs to learn the structure (and not the order) of the data, c) Sample efficiency, which derives from the strong inductive bias natively incorporated in a graph model, and d) Increased explainability, due to the improved efficiency in interpreting the reasoning process of a graphical model. On the contrary, their main limitations are (Rana et al., 2023): a) Computational overhead, where the complexity of the message-passing algorithm in large graphs can introduce critical latency for real-time control applications, b) Dynamic topology, which relates to the mathematical difficulty in modeling real-world, dynamic environments in a stable way, and c) Need for integration with latent spaces, where graphs typically need to connect and operate on precomputed, high-dimensional vector spaces.

With respect to the graph and function type, the following main categories can be identified:

- Scene graphs: Open–vocabulary 3D scene graphs associate vision–language features with real-world entities (often in a hierarchical way), while remaining compact compared to dense map representations. This structure enables robots to query targets, to reason about relationships, and to define sub-goals to planners in large-scale environments (Gu et al., 2024; Rana et al., 2023; Yan et al., 2025b). More recently, graphs are constructed online directly from RGB–D streams, using hierarchical structures for language–grounded navigation and adding functional links to the incorporated entities (Werby et al., 2024; Yin et al., 2024; Zhang et al., 2025b).

Table 4: Neural network architectures: Comparative analysis and key insights.

Architecture	Transformers	SSMs	DMs	CNNs/hybrid	Graphical models
Primary functions	• Multi-modal alignment • High-level reasoning • Hierarchical sub-task decomposition • Cross-embodiment skill transfer	• Temporal sequence modeling • Real-time edge control • High-frequency state estimation • Long-range contextual memory	• Multi-modal action generation • High-precision manipulation • Receding horizon control • Trajectory score estimation	• Local feature detection • Pixel-level spatial grounding • Multi-objective perception • Object classification	• Causal reasoning • Structured planning • Relational grounding • State transitions
Main mechanisms	• Global self-attention • Positional encoding • Autoregressive prediction • Chain-of-thought	• Dynamics discretization • Selective scan operators • Hardware-aware kernels • Input-dependent gating	• Score-based denoising • Langevin dynamics • Action chunking • Latent space diffusion	• Learnable convolutional layers • Local connectivity • Parameter sharing • Pooling operators	• Graph neural networks • Symbolic reasoning • Entity masking • Scene graph serialization
Strengths	• Long-range dependency • Architectural homogenization • Planning efficiency	• Linear scaling complexity • Stable long-horizon memory • Deployment efficiency	• Multi-modal modelling • Composite conditioning • Trajectory decision making	• Zero-shot generalization • Robustness to noise • Reduced training data	• Combinatorial generalization • Permutation invariance • Sample efficiency
Limitations	• Computational complexity • Discrete tokenization • Context contradiction	• Reduced cross-token check • Global context modeling • Need for hybrid designs	• Sampling latency/cost • No safety guarantees • Conditioning drift	• High computational latency • Loss of fine-grained details • Distribution shifts	• Computational overhead • Dynamic topology • Latent space integration
Indicative models	• RT-2 (Zitkovich et al., 2023), PaLM-E (Driess et al., 2023), Gato (Reed et al., 2022), OpenVLA (Kim et al., 2025), Octo (Mees et al., 2024)	• RoboMamba (Liu et al., 2024c), Mamba (Gu & Dao, 2024), AnoleVLA (Takagi et al., 2026), Decision Mamba (Huang et al., 2024b)	• Diffusion Policy (Chi et al., 2025), Diffuser (Janner et al., 2022), Motion Planning Diffusion (Carvalho et al., 2023), M2Diffuser (Yan et al., 2025a)	• RT-1 (Brohan et al., 2023a), R3M (Nair et al., 2023), VC-1 (Majumdar et al., 2023), MVP (Wei et al., 2022)	• GRID (Ni et al., 2024), ConceptGraphs (Gu et al., 2024), HOV-SG (Werby et al., 2024), Open3DSG (Koch et al., 2024)

- Shared graphs: Compressed-form scene graphs allow bandwidth–limited sharing and map merging, while maintaining open–vocabulary query capabilities. In particular, decentralized visual FMs can estimate peer poses and can produce local Bird's-Eye View (BEV) maps on embedded hardware, while reducing communication requirements without losing key semantic information (Blumenkamp et al., 2025; Gu et al., 2025). Such designs make robot-team perception and planning feasible in large-scale environments.

- Graph neural networks: Graph Neural Networks (GNNs) enable message passing over task, object, and agent graphs for performing allocation, scheduling, and policy conditioning in a data-driven way. Recent hybrid, cognitive pipelines couple GNN–based scene graphs with LLM or symbolic planners, achieving to maintain plans physically feasible, while at the same time still following language goals (Tong et al., 2026; Strader et al., 2025).

- Embodiment graphs: Embodiment graphs encode robot joint information, as well as links between them, allowing a single learned policy to adapt across different platforms. In this respect, attention or message passing algorithms follow the learned graph connectivity, which in turn boosts zero–shot transfer to new morphologies and supports reusable controllers across different robots (Patel & Song, 2025).

6.6 Comparative analysis and key insights

Having discussed in detail the various types of neural network architectures (Sections 6.1-6.5), this section systematically examines the literature methods, providing a comparative analysis and critical insights for each category. In this respect, Table 4 summarizes for each type of architecture its: a) Primary functions, b) Main mechanisms, c) Key strengths, d) Critical limitations, and e) Indicative models. Among the various observations and insights, it can be seen that transformers are widely adopted for multi-modal integration and high-level reasoning; however, their quadratic computational complexity motivates the exploration of more efficient alternatives, such as SSMs. In parallel, DMs accomplish to model complex, multi-modal

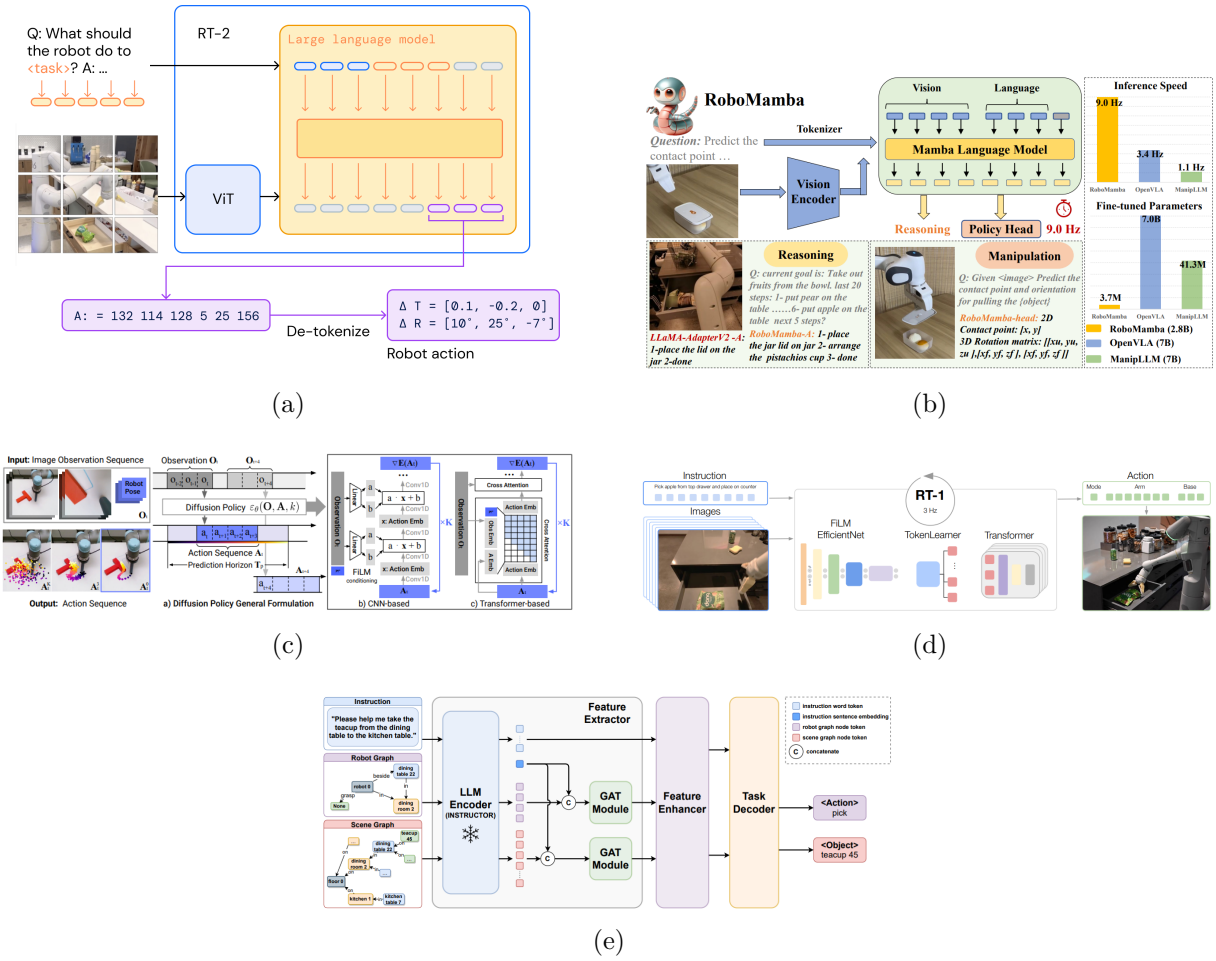


Figure 5: Representative literature methods incorporating different NN architecture types: (a) Transformers (RT-2 (Zitkovich et al., 2023)), which model robot control as token-based sequence prediction from visual and language inputs, (b) State-space models (RoboMamba (Liu et al., 2024c)), which enable efficient multimodal reasoning and action generation through Mamba-based sequence modeling, (c) Diffusion models (Diffusion Policy (Chi et al., 2025)), which generate action sequences through iterative denoising conditioned on observations, (d) Convolutional and hybrid encoders (RT-1 (Brohan et al., 2023a)), which combine a CNN encoder with token compression and a transformer policy for robot control, and (e) Graphical models (GRID (Ni et al., 2024)), which encode scene and robot relations as graphs and use graph-based reasoning for instruction-driven task decoding.

probability distributions, while CNN/hybrid encoders serve as robust perceptual backbones required for physical grounding. Moreover, graphical models support structured relational reasoning and causal grounding, combining the merits of data-driven and symbolic logic approaches. Moreover, representative literature methods incorporating different NN architecture types are illustrated in Fig. 5.

7 Learning paradigms

In order to develop robust, real-world robotic FM solutions, different/diverse learning techniques, principles, and approaches can be adopted. The most commonly met learning paradigms in the literature, which are typically combined in a comprehensive learning methodology, are: a) Pre-training, b) Self-Supervised Learning (SSL), c) Fine-tuning, d) Domain Adaptation (DA), e) Imitation Learning (IL), f) Reinforcement Learning

(RL), g) In-Context/prompt Learning (ICL), h) World Model (WM) learning, and i) Generative Learning (GL), as discussed in Section 4 and further detailed below.

7.1 Pre-training

Pre-training in robotic FM development aims at realizing a fundamental shift from constructing specialized, task-specific robots to creating generalist agents. In particular, instead of training a robot from scratch for a single/specific task, pre-training pipelines allow a model to learn a broad understanding of physics, semantics, and movement from massive datasets, prior to be applied to a specific robotic hardware platform. Its main advantages are (Xiao et al., 2025b): a) Data efficiency, since it significantly reduces the total number of real-world demonstrations needed for a downstream application, b) Zero-shot generalization, where the robust learned patterns can often facilitate the execution of tasks in previously unseen environments, c) Knowledge transfer, where knowledge gained from a given robotic platform can be transferred to improve performance of a completely different one, and d) Emergent reasoning capabilities, where large-scale pre-training allows models to perform sophisticated, semantic reasoning. On the contrary, its main limitations are (Zitkovich et al., 2023): a) Embodiment gap, where models are capable of understanding what to do, but not how to implement it in the physical world, b) Sim-to-real gap, where the physics learned in simulation environments may deviate from real-world ones, c) High computational cost, where typically massive GPU resources are required, and d) Safety concerns, where the presence of hallucinations in the model behavior leads to lacking of formal safety guarantees. In this context, OpenVLA (Kim et al., 2025) trains a 7B model on $\sim$ 970K real robot episodes for learning vision–language–action alignment across different embodiments. Additionally, Octo (Mees et al., 2024) develops a transformer policy, using $\sim$ 800K trajectories, for estimating general motor routines and language-conditioned control. Moreover, Gemini robotics (Team et al., 2025) adopts web-scale multi-modal pre-training, adapted for embodied execution with safety filters.

7.2 Self-supervised learning

SSL techniques (that are widely adopted under the pre-training setting) employ a set of 'pretext tasks' (e.g., predicting the next video frame, reconstructing a masked image patch, etc.) and large quantities of unlabeled data streams (e.g., visual, depth, force, audio, robot logs, etc.) for enabling FMs to acquire common sense knowledge regarding physics, object permanence, and spatial relationships (He et al., 2022; Tong et al., 2022; Oquab et al., 2024). Their main advantages are (Nair et al., 2023; Wu et al., 2023b; Assran et al., 2025): a) Massive scalability, where robots can learn from internet-scale datasets, without requiring human supervision, b) Sample efficiency, where only a handful of training samples is needed for adapting to new tasks, c) Zero-shot generalization, where SSL-trained models often exhibit increased performance in novel settings, and d) Autonomous improvement, where SSL methods can infer reward signals for training directly from the data itself. On the contrary, their main limitations are (Nair et al., 2023; Wu et al., 2023b): a) Embodiment gap, where most large-scale SSL-employed sources lack proprioceptive data (e.g., joint torques and forces), b) High computational cost, which typically requires massive GPU resources, c) Presence of hallucinations, where the learned models may infer physically impossible actions, and d) Evaluation difficulty, where it is inherently challenging to measure the success of SSL-learned representations, until actual model deployment.

The most common SSL techniques used for developing robotic FM solutions are:

- Masked Autoencoder (MAE): The model learns by reconstructing missing or corrupted parts of the input data, aiming at modeling robust spatial and temporal features. Image MAE (He et al., 2022) and its video counterpart VideoMAE (Tong et al., 2022) are widely used for producing robust visual backbone networks. Additionally, robotics-specific masked-pretraining approaches, such as 3D-MVP (Qian et al., 2025a), adapt the MAE paradigm to robot learning for manipulation.

- Contrastive learning: The fundamental principle relies on aligning matched views (and separating mismatched ones) for creating discriminative features and supporting open-vocabulary grounding, so that robots can link language to perception and retrieval skills. CLIP (Radford et al., 2021)

establishes vision–language alignment at scale, while egocentric robot features, like R3M (Nair et al., 2023), can improve manipulation sample-efficiency using recorded human-performing videos.

- Autoregressive sequence modeling: The ultimate goal is grounded on predicting the next token in vision, language, or action streams, in order to capture long-horizon structure and to enable unified perception-to-policy modeling. Generalist agents, such as Gato (Reed et al., 2022), demonstrate how one sequence model can condition on images, text, and proprioception to produce actions across many tasks. Additionally, LLMs, like GPT-3 (Brown et al., 2020) and GPT-4 (Achiam et al., 2023), showcase the scalability and cross-domain reasoning capabilities of autoregressive transformers in embodied pipelines (Turcato et al., 2025; Mon-Williams et al., 2025).

- World model learning: Also often reported as an individual learning paradigm (Section 7.8), WMs aim at learning to predict how the world will change in response to specific actions, essentially encoding causal relations. Various individual WM-based methods are introduced for varying environmental settings, including Dreamer-style agents on physical robots (Wu et al., 2023b), JEPA-based designs (such as AdaWorld (Gao et al., 2025b) and ACT-JEPA (Vujinovic & Kovacevic, 2025)), and compositional video world models (like RoboDreamer (Zhou et al., 2024a)).

7.3 Fine-tuning

Fine-tuning aims at adapting the internet-scale acquired knowledge to specific physical-world execution settings. In particular, it targets to adjust the rich, common-sense knowledge structures of a pretrained FM to a specific robot, sensor suite, or task, making use of a smaller, in-domain, annotated dataset. Its main advantages are (Yu et al., 2025b): a) Sample efficiency, where only a relatively reduced set of training data is needed for the new/specific robot task, b) Domain adaptation, allowing the pretrained FM to handle different/specific application settings that are not present in the original training set, and c) Improved precision, where the learned policies are enabled to become more accurate and robust. On the contrary, its main limitations are (Mees et al., 2024): a) Catastrophic forgetting, related to the risk of the FM losing part of its general-purpose skills, b) Overfitting, which may occur when the fine-tuning dataset is relatively small and the FM may extensively adapt to specific operational settings, and c) Need for annotated data, where the requirement for high-quality robotic data is still present.

The most common fine-tuning techniques in robotics are:

- Full fine-tuning: The goal is to update all FM weights using a relatively small, in-domain robot dataset, in order to re-target a pretrained policy to a new robot, sensor setup, or task. Recent generalist policies report fast adaptation to new observation and action spaces on standard GPUs, using full fine-tuning as the baseline learning method (Mees et al., 2024; Kim et al., 2025).

- Low-Rank Adaptation (LoRA): The fundamental aim relies on maintaining the original FM weights unchanged and adding two low-rank matrices that learn/incorporate the required model modifications. In particular, OpenVLA (Kim et al., 2025) demonstrates LoRA-based tuning on the large-scale Open-X-Embodiment dataset. Additionally, more recent approaches (including quantized LoRA versions) target model adaptation in resource-constrained robotic platforms (Williams et al., 2026; Kim et al., 2026).

- Quantized parameter-efficient fine-tuning: This extends the original Parameter-Efficient Fine-Tuning (PEFT) approach, by combining low-precision weights with LoRA-style adapters for maintaining latency and memory requirements low on embedded or field hardware, while at the same time retaining most of the full-precision performance. In particular, LiteVLA (Williams et al., 2026) and similar approaches (Williams et al., 2025) report that an 4–8 bit quantization plus adapters can preserve high recognition rates, while enabling real-time control on smaller platforms.

- Action space remapping: This relies on integrating lightweight encoders/decoders or tokenizers, so that the core policy can be fine-tuned to new sensors or actuators without retraining the FM from scratch. Generalist policies, such as Octo (Mees et al., 2024) and RT-2 (Zitkovich et al., 2023), rely

on this technique to re-target a given FM to multiple robots and grippers with modest additional data.

7.4 Domain adaptation

Domain Adaptation (DA) aims at bridging the critical gap between FMs being pre-trained on massive internet-scale data and their subsequent deployment in real-world environments. In practice, DA targets to equip FMs with the required physical-world grounding or specific environmental awareness capabilities for given application scenarios, which is typically termed as the 'sim-to-real' transfer challenge. Its main advantages are (Da et al., 2025): a) Reduced need for real-world samples, where the use of simulation data accommodates the need for extensive amounts of real-world ones, b) Robustness to sensor shifts, where models are boosted to remain stable in case of robot hardware changes, and c) Safety guarantee, where, since training is performed in simulation environments, the risk of robots damaging their surroundings during the learning phase is reduced. On the contrary, its main limitations are (Tayyab Khan & Waheed, 2025): a) Risk of negative transfer, which may occur when the difference between the simulation and the real-world environments is large, b) Training instability, where the process of defining multiple hyper-parameters in the simulation environment renders the whole training process sensitive to their selection, and c) Lack of predictability, where it is difficult to assess the real-world situations where the trained robot might succeed or fail.

The most common DA techniques in robotics are:

- Sim-to-real transfer: This aims at training a FM model using large corpora of simulated/synthetic data, evaluating their performance in such simulation suites, and eventually calibrating them on real-world hardware (typically employing some real-world data). In particular, humanoid and manipulation solutions demonstrate the efficiency of this approach (Luo et al., 2026; Deng et al., 2025a).
- Real-to-sim-to-real transfer: This scenario involves the replay of real-world robot trajectories in high-fidelity simulations, the diversification of the depicted scenes and objects, the synthesis of new training data, the evaluation of the developed models in simulation (often employing domain randomization techniques), and the eventual calibration of the FM model to the real-world specifications. Various works demonstrate the validity of this approach in diverse operational settings (Zhu et al., 2025; Fang et al., 2025b).

7.5 Imitation learning

The fundamental consideration of Imitation Learning (IL) relies on enabling a model to learn directly from a (human) expert via teleoperation or video demonstrations of the desired skills. The main usefulness of IL lies on its efficient multi-modal alignment, which allows the mapping of high-level language instructions and/or visual inputs directly to low-level motor commands. Its main advantages are (Zare et al., 2024): a) No reward signal definition, where the robot is only required to mimic the expert behavior, b) Reduced training data, where IL is shown to require a reduced number of expert demonstrations to achieve robust performance, and c) Easy training supervision, which is performed directly through teleoperation and/or expert demonstrations. On the contrary, its main limitations are (Kawaharazuka et al., 2025): a) High dependency on data quality, where the quality of the observed demonstrations has a direct impact on the robot learning process, b) Covariate shift, where the robot is highly likely to fail if its actions deviate from the observed demonstrations, and c) Causal confusion, where the model might incorrectly learn unintended factors being present in the demonstrations.

The most common IL techniques in robotics are:

- Behavioral cloning: This serves as a particular type of Supervised Fine-Tuning (SFT), by imitating human expert demonstrations. In particular, RT-1 (Brohan et al., 2023a) learns a direct mapping of observations and language goals to actions, using large-scale transformer networks trained on diverse demonstration data. RT-2 (Zitkovich et al., 2023) extends this with web-scale vision–language

pretraining to open-vocabulary, instruction-following control. Moreover, vision FMs for embodiment- and environment-agnostic scene representation further decouple perception from control to facilitate cross-robot transfer (Riou et al., 2024).

- Diffusion-based IL: The fundamental consideration relies on representing a robot's behavior as a conditional denoising process. In this way, diffusion policies treat actions as a data distribution that can be iteratively refined from random noise. In particular, Diffusion Policy (Chi et al., 2025) and Diff-Dagger (Lee et al., 2025b) generate action sequences that match expert behavior and handle multi-modal inputs, while improving stability for long-horizon manipulation.
- In-context IL: The main goal is for the robot to learn novel tasks on the fly, i.e., to enable zero- or few-shot task adaptation, but without updating the model weights. ICRT (Fu et al., 2025) instantiates this idea by using next-token prediction over sensorimotor streams for real-robot in-context imitation, extending a similar previous approach that is based on a sequence modeling formalism. Similarly, prompt demonstrations are augmented using explicit visual reasoning traces in (Nguyen et al., 2026), allowing the model to infer task intent more reliably in complex and ambiguous environments, while jointly predicting reasoning and low-level actions in an autoregressive way.
- Continual IL: This focuses on addressing the long-term memory and evolution challenges of robotic FMs. The core functionality is to enable a robot to sequentially acquire new skills over time, without forgetting previously learned ones. In this context, LOTUS (Wan et al., 2024) introduces a continual imitation learning framework for skill acquisition by a real robot.

7.6 Reinforcement learning

Reinforcement Learning (RL) serves as the optimization formalism that aims at bridging the gap between high-level, semantic reasoning (supported by FMs) and low-level, physical robot actions. In practice, the goal of RL is to involve the FM in a continuous, self-improvement cycle of subsequent perception, action, and evaluation steps. Its main advantages are (Tang et al., 2025): a) Self-improvement capability, where RL allows a robot to explore and to surpass the quality of its training data, b) Increased generalization ability, where RL encourages the model to investigate alternative strategies, making it more robust to novel situations, and c) Efficient sim-to-real implementation, where the RL component can be extensively trained in simulation, prior to be deployed in real operational settings. On the contrary, its main limitations are (Ter et al., 2025): a) Sample inefficiency, where RL often requires a very large number of training trajectories to converge, b) Careful reward engineering, which may require detailed definition of the reward function for avoiding misleads during training, and c) Difficulty in credit assignment, which corresponds to the inherent challenge of identifying incorrect robot behaviors over long-horizon tasks.

The most common RL techniques in robotics are:

- SFT-to-RL: This involves a two-stage process, where BC is initially applied for learning a policy and subsequently RL is employed for further improving it. In this context, RT-1 (Brohan et al., 2023a) and RT-2 (Zitkovich et al., 2023) demonstrate how large-scale BC may produce powerful priors that can be refined further through interaction. Additionally, ExploRLLM (Ma et al., 2025) combines an LLM-guided exploration policy with a residual RL head to improve sample efficiency and robustness.
- LLM-guided reward design: This leverages the capabilities of LLMs for writing/refining the RL reward code and tuning domain randomization procedures, in order to improve the robustness of learning and knowledge transfer. In particular, Eureka (Ma et al., 2024a) automates reward design and outperforms expert rewards on multiple robotic tasks, while DrEureka (Ma et al., 2024b) extends this approach to the sim-to-real setting, by jointly optimizing rewards and randomization for locomotion and dexterous manipulation. Moreover, Gen2Sim (Katara et al., 2024) increases the application scope, by using generative models and LLMs to synthesize tasks, scenes, and reward functions for large-scale RL in simulation.
- Preference-based RL: This relies on replacing detailed/numeric RL rewards with preferences produced by VLMs (or adapted ones, by involving small-scale human intervention). RL-VLM-F (Wang et al.,

2024d) models rewards from VLM comparisons over image observations and task-related text, in order to improve manipulation without human guidance. Additionally, VARP (Singh et al., 2025) regularizes VLM-derived preferences with the agent's own rollouts for reducing misalignment and hallucinations in vision–language feedback.

- Offline-to-online RL: This involves a 3-step process, where a) The model is initially trained offline on massive, heterogeneous datasets, b) It subsequently undergoes an offline RL refinement step using an appropriate reward function, and c) Eventually, it follows online RL for real-world deployment. Indicatively, embodied visual tracking is combined with a text-promptable encoder and offline RL for improved perception in (Zhong et al., 2024). Additionally, FLaRe (Hu et al., 2025a) applies large-scale RL fine-tuning on a pre-trained VLA for adaptive manipulation across diverse tasks.

- World-model RL: World-Model Reinforcement Learning (WM-RL) aims at learning a generative world model with language-aware structure and, subsequently, using it for RL-based policy improvement. In this respect, RoboDreamer (Zhou et al., 2024a) factors video generation into compositional parts, conditioned by language and visual goals, and exhibits robust performance on long-horizon tasks.

7.7 In-context/prompt learning

ICL and prompt learning enable the adaptation of FMs at inference time by conditioning their behavior on demonstrations, examples, or task-specific instructions, without requiring updates to the underlying model weights. While closely related, the two are not identical: ICL relies on task-relevant examples provided within the context window, whereas prompt learning more broadly focuses on steering a pre-trained model through textual or multimodal instructions. Their main advantages are (Fu et al., 2025; Yin et al., 2025c): a) Generalization to novel settings, where a small number of demonstrations or prompts can enable adaptation to previously unseen tasks and environments, b) High-level task planning, where natural-language guidance can be translated into structured reasoning steps and, subsequently, into low-level physical actions, and c) Multimodal task specification, where context can be expressed through language, visual observations, or sensorimotor demonstrations. On the contrary, their main limitations are (Yao et al., 2023): a) Prompt sensitivity, where even minor changes in the input context may lead to unstable or suboptimal behavior, b) Limited grounding, where inference-time reasoning does not inherently guarantee consistency with real-world constraints, and c) Inference overhead, where long context windows and multi-step prompting strategies may increase latency and computational cost during deployment.

The most common in-context/prompt learning techniques in robotics are:

- Language prompting: This relies on the use of natural language instructions for guiding a robot's behavior, decision-making, and physical actions. In particular, few-shot language prompts can encode demonstrations or templates, so that an LLM can output low-level actions or to acquire new skills (Yin et al., 2025c; Liang et al., 2023; Huang et al., 2023b).

- Reason–act prompting (ReAct): This interleaves natural language reasoning with physical actions, allowing a robot to decompose complex goals, to validate its progress, and to dynamically adjust its plan. In this respect, planning, execution, and re-planning can be performed in a single loop, also involving an LLM-based verification checking step (Yao et al., 2023; Grigorev et al., 2025).

- In-context imitation: This enables a FM to perform a novel task by observing a few videos or sensorimotor demonstrations, without applying any permanent changes to its internal weights. In this context, a causal policy can parse short teleoperation trajectories as a prompt and, subsequently, to predict the next action for new tasks without the need for fine-tuning (Fu et al., 2025).

7.8 World model learning

The fundamental functionality of World Models (WMs) in robotic FM application is that they allow robots to predict environmental changes in response to their actions. In particular, their primary role is to decouple perception from action, where, instead of directly operating only on pixel values, the robot can learn the

underlying physics of the world. Their main advantages are (Li et al., 2025d): a) Incorporation of 'imagined' experiences, which reduces the need for physical-world trials, b) Modeling rules of physics, where the model learns the impact of factors like gravity, friction, and collisions in the real environment, and c) Reduced delays in operation, where the robot is able to predict future states and to maintain smooth actions. On the contrary, their main limitations are (Zhang et al., 2025f): a) Presence of hallucinations, where the robot may converge to misleading actions in case of slight inaccuracies in the WM, b) Cumulative errors, where small prediction errors can be accumulated in long-horizon tasks, and c) Increased computational cost, where the training of a high-fidelity WM may require excessive GPU resources.

The most common WM learning techniques in robotics are:

- Feature-space WMs: Instead of performing a prediction for each pixel, which is computationally expensive and often noisy, feature-space WMs map visual inputs to an abstract feature space and perform future predictions there. In particular, future DINOv2 patch embeddings are predicted from offline trajectories and, subsequently, action sequences are optimized in the embedding space for zero-shot planning in (Zhou et al., 2025b).
- Latent-action WMs: Latent-action WMs aim at learning robots to model the underlying physics and intent of actions, instead of focusing on representing specific skills. In this respect, continuous latent actions are discovered from videos, while an auto-regressive WM is trained that conditions on those actions to transfer skills across scenes and embodiments with small-scale finetuning in (Gao et al., 2025b).
- Compositional video WMs: These adopt a modular approach, where the model aims at breaking down (factorizing) the surrounding environment into its constituent parts (namely, objects, relationships, and action primitives) and, subsequently, reconnecting them for generating future scenarios. In this context, videos are factorized into objects and relations so that the model can synthesize plans for unseen combinations of goals-scenes and guide long-horizon decisions in (Zhou et al., 2024a).
- JEPA-style WMs: The main focus lies on predicting an abstract meaning of what will happen next, which enables robots to plan complex tasks, without getting distracted by irrelevant noise. In particular, joint predictions of short-horizon actions and abstract observations are realized in (Vujinovic & Kovacevic, 2025), in order to couple imitation with predictive learning and to reduce control error accumulation.

7.9 Generative learning

Generative Learning (GL) enables robots to imagine future states, to synthesize training data, and to propose complex action sequences. Its fundamental use lies on the capability of generative FMs of producing large quantities of data samples, alleviating from the need for extensive high-quality robotic interaction samples. Its main advantages are (Zhang et al., 2025e): a) Increased zero-shot generalization, where generative models are suitable for handling objects or environments previously unseen, b) Handling multi-modality, due to the ability of generative models of predicting missing information between different sensorial data, and c) Long-horizon planning, where generative models enable robots to perform accurate predictions for multiple future steps. On the contrary, its main limitations are (Liu et al., 2025a): a) Sim-to-real gap, where the generated data can miss subtle physical nuances, b) Presence of hallucinations, where generative robots are prone to hallucinating a physical capability, and c) Increased inference latency, where generative models typically require extensive computations.

The most common GL techniques in robotics are:

- Autoregressive sequence modeling: The main goal lies on predicting the next action or state based on previous observations. In this respect, PACT (Bonatti et al., 2023) trains a causal transformer to predict the next observation–action token, so that a single model can capture long-horizon structure across different tasks. Additionally, long-horizon manipulation is modeled through sequential generation of action tokens in (Zhang et al., 2025i).

Table 5: Learning paradigms: Comparative analysis and key insights.

Learning paradigm	Primary function	Main mechanisms	Primary data source	Strengths	Limitations	Indicative models
Pre-training	• Acquisition of general semantic and sensorial representations	• Large-scale (self-) supervised learning	• Internet-scale data	• Data efficiency • Zero-shot generalization • Knowledge transfer • Emergent reasoning	• Embodiment gap • Sim-to-real gap • High computational cost • Safety concerns	• PaLM-E (Driess et al., 2023), RT-2 (Zitkovich et al., 2023), Octo (Mees et al., 2024), OpenVLA (Kim et al., 2025)
Self-supervised learning	• Representational learning without manual labels	• Pretext tasks	• Unlabeled robot trajectories • Egocentric videos • Raw sensor data	• Massive scalability • Sample efficiency • Zero-shot generalization • Autonomous improvement	• Embodiment gap • High computational cost • Presence of hallucinations	• R3M (Nair et al., 2023), MVP (Wei et al., 2022), DINOv2 (Oquab et al., 2024), Masked Autoencoders (He et al., 2022)
Fine-tuning	• Adaptation of pre-trained models to specific tasks and domains	• Supervised update on specialized labeled data	• Action-labeled robot demonstrations • Domain-specific instructions	• Sample efficiency • Domain adaptation • Improved precision	• Catastrophic forgetting • Overfitting • Need for annotated data	• OpenVLA (Kim et al., 2025), Octo (Mees et al., 2024), RoboCat (Bousmalis et al., 2024), RT-2 (Zitkovich et al., 2023)
Domain adaptation	• Bridging the sim-to-real gap	• Feature alignment • Adversarial training • Distribution reweighting	• Paired synthetic data with sparse real-world observations	• Reduced need for real-world samples • Robustness to sensor shifts • Improved transferability	• Risk of negative transfer • Training instability • Lack of predictability	• DrEureka (Ma et al., 2024b), Gen2Sim (Katara et al., 2024), ReBot (Fang et al., 2025b), VR-Robo (Zhu et al., 2025)
Imitation learning	• Replication of expert behaviors from demonstrations	• Behavioral cloning	• Expert teleoperation • Kinesthetic teaching • Human videos	• No reward signal definition • Reduced training data • Easy training supervision	• High dependency on data quality • Covariate shift • Causal confusion	• RT-1 (Brohan et al., 2023a), Gato (Reed et al., 2022), Octo (Mees et al., 2024), BC-Z (Jang et al., 2022)
Reinforcement learning	• Policy learning through environment interaction	• Reward-driven exploration and policy optimization	• Interaction data from simulation or real-world deployment	• Self-improvement capability • Increased generalization ability • Efficient sim-to-real implementation	• Sample inefficiency • Careful reward engineering • Difficulty in credit assignment	• Eureka (Ma et al., 2024a), DrEureka (Ma et al., 2024b), RL-VLM-F (Wang et al., 2024d), ExploRLLM (Ma et al., 2025)
In-context/ prompt learning	• Task adaptation via demonstrations in the input prompt	• Inference on a frozen model using contextual examples	• Multi-modal instructions • Observation-action pairs	• Generalization to novel settings • Human-like task planning • Flexible use of multi-modal instructions	• Prompt sensitivity • Lack of grounding • Inference latency	• SayCan (Brohan et al., 2023b), VIMA (Jiang et al., 2023), ICRT (Fu et al., 2025), Instruct2Act (Huang et al., 2023b)
World model learning	• Prediction of environment dynamics and future states	• Transition functions in latent space	• Interaction data	• Incorporation of imagined experiences • Modeling rules of physics • Reduced delays in operation	• Presence of hallucinations • Cumulative errors • Increased computational cost	• Dreamer (Hafner et al., 2020), RoboDreamer (Zhou et al., 2024a), ACT-JEPA (Vujinovic & Kovacevic, 2025), AdaWorld (Gao et al., 2025b)
Generative learning	• Synthesis of data, plans, or control trajectories	• Data distribution modeling	• Massive multi-modal datasets	• Increased zero-shot generalization • Handling multi-modality • Long-horizon planning	• Sim-to-real gap • Presence of hallucinations • Increased inference latency	• Diffuser (Janner et al., 2022), Diffusion Policy (Chi et al., 2025), DALL-E-Bot (Kapelyukh et al., 2023), Gen2Sim (Katara et al., 2024)

- Diffusion-based action policies: This employs a diffusion model for generating a chunk of actions at once, by gradual application of a denoising process. In particular, Diffusion Policy (Chi et al., 2025) learns a conditional denoising process that samples action sequences for handling multi-modal behaviors and supporting stable visuomotor control. Similarly, Legibility Diffuser (Bronars et al., 2024) comprises an intent-expressive variant of the latter.

- Generative video and scene synthesis: The fundamental aim comprises the creation of a model of the physical reality, which in turn enables robots to imagine, to simulate, and to plan actions prior to their execution in the real world. In this context, RoboDreamer (Zhou et al., 2024a) employs compositional video WMs, while ReBot (Fang et al., 2025b) makes use of a real-to-sim-to-real synthesis approach.

7.10 Comparative analysis and key insights

Having discussed in detail the various types of learning paradigms (Sections 7.1-7.9), this section systematically examines the literature methods, providing a comparative analysis and critical insights for each type. In this respect, Table 5 summarizes for each learning paradigm type its: a) Primary function, b) Main mechanisms, c) Primary data source, d) Key strengths, e) Critical limitations, and f) Indicative models. Among the various observations and insights, it can be seen that each paradigm addresses a specific aspect of the robot skill acquisition gap, balancing the need for massive data ingestion with the requirement for precise, real-time physical execution. In particular, pre-training and self-supervised learning provide the necessary general-purpose representational learning capabilities, while imitation and reinforcement learning emphasize on behavioral precision. Additionally, domain adaptation and fine-tuning aim for learned, generalist models to be deployed on specific hardware in real-world environments. More recent paradigms, like in-context, world model, and generative learning target the development of robotic agents that can learn continuously and reason about the future environmental states.

8 Learning stages

During the overall learning process of a robotic FM, the particular phase at which knowledge is incorporated largely defines the type/nature of the acquired skills, algorithmic/development details, and key assumptions about the model behavior. In this context, the main learning stages identified in the literature are: a) Pre-training, b) Offline fine-tuning, c) Online adaptation, and d) Continuous learning, as discussed in Section 4 and further detailed below.

8.1 Pre-training

The ultimate goal of the pre-training stage is to estimate robust, general-purpose representations of robotic data. In particular, instead of training a robotic agent for a specific task or application, pre-training aims at processing massive (often internet-scale) amounts of diverse data (e.g., human demonstration videos, simulation data, sensorial data streams, etc.) from multiple robotic platforms, in order to acquire knowledge and to model the cross-correlations among vision, language, and action. Its main advantages are (Li et al., 2024a): a) Increased generalization, where pre-trained models are more likely to adapt to new environments, than undergoing training from scratch, b) Robust zero-shot capability, where pre-trained models can often perform robustly tasks that they weren't explicitly trained for, without the need for extra training samples, and c) Accurate multi-modal mapping, where due to the usual large-scale datasets used, pre-trained models are proven to robustly map across textual words, visual concepts, and physical actions. On the contrary, its main limitations are (Kawaharazuka et al., 2025): a) Excessive computational cost, where pre-training requires massive GPU resources to be implemented, b) Reduced performance in real-world, where if a model is pre-trained largely on internet or simulation data, it may not always perform robustly in real-world, physical circumstances, and c) Safety concerns, where training using internet resources does not always take into consideration strict safety constraints.

The most common learning paradigms adopted during the pre-training stage are:

- Pre-training: Supervised pre-training aims at equipping a model with a foundational understanding of the world from high-quality, diverse, labeled data. In particular, PaLM-E (Driess et al., 2023) is jointly trained using both web-scale multi-modal data and embodied experiences. Similarly, RT-2 (Zitkovich et al., 2023) and Gemini Robotics (Team et al., 2025) are constructed using web and robot manipulation data.
- Self-supervised learning: SSL targets to enable robots to operate beyond narrow, task-specific programming, aiming at accomplishing generalized intelligence capabilities. In particular, DINOv2 (Oquab et al., 2024) and VideoMAE (Tong et al., 2022) can learn perception priors from unlabeled data. Moreover, R3M (Nair et al., 2023) extends this approach, by incorporating egocentric features.
- Imitation learning: The goal of IL is to learn a prior distribution of successful behaviors directly from expert demonstrations. In particular, RT-1 (Brohan et al., 2023a) treats robot control as a next-token

prediction problem over multi-modal streams, so that the learned policies to inherit both semantic and sensorimotor skills. Additionally, RT-2 (Zitkovich et al., 2023) is trained using vision–language and robot action tokens, while Octo (Mees et al., 2024) makes use of the Open-X-Embodiment trajectories for deriving a generalist policy. Moreover, Seer (Tian et al., 2025) investigates the scaling laws of multi-task IL, aiming at drastically reducing the required target-domain data.

8.2 Offline fine-tuning

Offline fine-tuning aims at bridging the knowledge gap between general-purpose representations (learned during the pre-training step) and the specificities of a given physical-world application. In short, the primary purpose of this stage is task and embodiment specialization. Its main advantages are (Hu et al., 2023): a) Reduced need for training data, where the need for training samples is significantly lower than the pre-training step, b) Increased training stability, where models are likely to converge to robust performance states, provided that sufficient training examples are available, and c) Knowledge distillation, where only the necessary parts of the general-purpose representations learned during the pre-training stage can be used for the given application at hand. On the contrary, its main limitations are (Firoozi et al., 2025): a) Distribution shift, where if a robot encounters novel state challenges, it is difficult for it to handle, b) Dependency on data quality, where the presence of suboptimal or erroneous demonstrations can mislead the training process, and c) Lack of online exploration, where the model can only employ skills present in the training dataset, while not being able to adapt to online challenges.

The most common learning paradigms adopted during the offline fine-tuning stage are:

- Imitation learning: IL aims at equipping pre-trained models with the necessary low-level precision skills for a specific application. In particular, by mimicking expert demonstration, it targets to learn specific motor commands with respect to a given task or robotic platform. In particular, Octo (Mees et al., 2024) and OpenVLA (Kim et al., 2025) employ large-scale Open X Embodiment pretraining and, subsequently, focus on new robotic platforms, making use of LoRA-style adapters. Similarly, LiteVLA (Williams et al., 2025) demonstrates that NF4 quantized LoRA can be tuned on CPU-only hardware.
- Reinforcement learning: RL enables robots to learn from a reward signal, by focusing on performed actions that lead to successful task executions. In particular, a recurrent tracker is trained, using conservative offline RL, on VFM annotated trajectories, prior to deployment in (Zhong et al., 2024). Additionally, FLaRe (Hu et al., 2025a) applies large-scale RL fine-tuning to transformer-based policies, in order to improve long-horizon mobile manipulation.
- Generative learning: GL facilitates the specialization of pre-trained models to specific environments, especially in the presence of sparse constraints on the target task. In particular, ReBot (Fang et al., 2025b) replays real trajectories in simulation and, subsequently, composes them into inpainted real backgrounds to adapt to new domains. Additionally, RoboDreamer (Zhou et al., 2024a) makes use of compositional WMs to generate imagined video plans, which serve as additional training data.

8.3 Online adaptation

Online adaptation targets to equip robots with the appropriate routines for learning in real-time from their own experiences. In practice, at this stage robots aim to handle the distribution shift between the offline training data and the ones encountered during online deployment. Its main advantages are (Firoozi et al., 2025): a) High precision performance, where robots typically accomplish superior task execution accuracy for the specific environments to which they are adapted, b) Continuous improvement, where the robot continuously increases its performance as it constantly learns from its experiences, and c) Robustness to distribution shifts, where the models achieve to maintain performance in the presence of environmental changes. On the contrary, its main limitations are (Yuan et al., 2025): a) Catastrophic forgetting, which denotes the risk that might occur as the robot learns new skills to loose some of its general-purpose ones, b) Computational latency, where the algorithmic operations involved need to be performed in (near) real-time, and c) Prone to noise, where real-time, noisy sensorial data may mislead the adaptation process.

The most common learning paradigms adopted during the online adaptation stage are:

- Domain adaptation: This aims at handling the particular physical-world constraints and sensorial noise for a specific robot deployment scenario. In practice, this enables the model to re-calibrate its knowledge structures to the perceived environment. In this context, Test-Time Adaptation (TTA)-Nav (Piriyajitakonkij et al., 2024) incorporates a reconstruction decoder on top of a pre-trained policy, so that the agent can denoise corrupted frames without gradient updates, while restoring point-goal navigation under severe corruptions. Additionally, Phys2Real (Wang et al., 2025b) targets to bridge sim-to-real gaps, by combining FM priors with interaction-based estimations.

- Reinforcement learning: The aim of RL lies on allowing the robot to perform micro-adjustments to its general-purpose knowledge, based on sensory feedback and exploration. In this context, self-improving embodied FMs refine pre-trained policies, based on reward and success estimation from model predictions, across a robot fleet in (Ghasemipour et al., 2025). Additionally, RL-VLM-F (Wang et al., 2024d) estimates rewards, using a VLM model that compares trajectory snippets with language goals. Similarly, VARP (Singh et al., 2025) regularizes VLM-derived preferences with the agent's own rollouts for reducing misalignment. Moreover, Eureka (Ma et al., 2024a) and DrEureka (Ma et al., 2024b) are LLM-guided RL frameworks that automate reward design and, in the case of DrEureka, domain randomization, in order to improve policies for locomotion and dexterous manipulation.

- In-context/prompt learning: This paradigm is particularly suitable for online adaptation, since it enables zero- or few-shot specialization at deployment time, without requiring updates to the underlying model weights. In particular, ICRT (Fu et al., 2025) employs in-context imitation policies that condition on a small number of recent demonstration trajectories for adapting to novel manipulation tasks. Additionally, LLM-based control stacks can interleave reasoning and acting through ReAct-style prompting, allowing robots to monitor execution progress and to revise plans when needed (Yao et al., 2023). Similarly, verifier-based frameworks can check high-level task plans prior to execution, improving reliability in dynamic deployment settings (Grigorev et al., 2025).

8.4 Continuous learning

Continuous Learning (CL) (often termed lifelong learning) aims at enabling robots to acquire new skills or to adapt to new environments incrementally, without requiring to be fully retrained from scratch. In practice, it employs conventional learning paradigms (e.g., IL, RL, etc.) in slower outer loops for improving the robot's performance. Its main advantages are (Xiao et al., 2025b): a) Increased adaptability, where robot capabilities can continuously evolve in response to changes in the surrounding environment, making them suitable for long-term deployment, b) Improved scalability, where data from a single robot can be used to update the knowledge structures of an entire fleet, and c) Reduced downtime, where robots that continuously update their underlying models are less likely to become non-sufficiently robust/operational for long periods of time. On the contrary, its main limitations are (Firoozi et al., 2025): a) Catastrophic forgetting, where the continuous update in the robot's knowledge structures may result into overwriting previously learned skills, b) Stability-plasticity dilemma, which corresponds to a critical trade-off between the ability to integrate new skills and the capability to maintain the old ones, and c) Memory overhead, where the model needs to maintain increased past data for robustly updating its behavior in subsequent steps.

The most common learning paradigms adopted during the continuous learning stage are:

- Domain adaptation: DA enables a model to continuously adjust its internal knowledge to dynamic, real-world settings. In particular, Action Flow Matching for Lifelong Learning (Murillo-González & Liu, 2025) develops a lifelong robot-learning framework that incrementally aligns robot dynamics across sequential tasks, supporting efficient and safe continual adaptation. On another direction, VR-Robo (Zhu et al., 2025) utilizes a real-to-sim-to-real framework for constructing photorealistic digital twins from logged data, retraining navigation or locomotion policies in them, and transferring the acquired skills to the real-world.

Table 6: Learning stages: Comparative analysis and key insights.

Learning stage	Pre-training	Offline fine-tuning	Online adaptation	Continuous learning
Primary function	• Learning of general-purpose world knowledge	• Specialized grounding for specific application settings	• Real-time adjustment to dynamic environments	• Lifelong knowledge acquisition and retention
Data requirements	• Internet-scale, multi-modal, multi-embodiment robot data	• High-quality, in-domain expert demonstrations	• Robot interaction history and recent deployment context	• Sequential, non-stationary streams of demonstrations and interactions
Learning paradigms	• Pre-training • Self-supervised learning • Imitation learning	• Imitation learning • Reinforcement learning • Generative learning	• Domain adaptation • Reinforcement learning • In-context/prompt learning	• Domain adaptation • Imitation learning • Reinforcement learning
Computational requirements	• Extreme GPU resources	• Moderate to high GPU resources	• Low (edge/onboard, real-time inference)	• Moderate (periodic, incremental updates)
Generalization capability	• Zero-shot and cross-domain reasoning	• Task-specific precision and robust motor control	• Sample-efficient adjustment to local perturbations	• Knowledge retention and transfer across sequential tasks
Strengths	• Increased generalization • Robust zero-shot capability • Robust multi-modal mapping	• Reduced need for training data • Increased training stability • Knowledge distillation	• High precision performance • Continuous improvement • Robustness to distribution shifts	• Increased adaptability • Improved scalability • Reduced downtime
Limitations	• Excessive computational cost • Reduced performance in real-world settings • Safety concerns	• Distribution shift • Dependency on data quality • Lack of online exploration	• Catastrophic forgetting • Computational latency • Sensitivity to noise	• Catastrophic forgetting • Stability-plasticity dilemma • Memory overhead
Indicative models	• Open X-Embodiment/RT-X (O'Neill et al., 2024), GR00T N1 (Bjorck et al., 2025), PaLM-E (Driess et al., 2023), CLIP (Radford et al., 2021)	• Octo (Mees et al., 2024), OpenVLA (Kim et al., 2025), BC-Z (Jang et al., 2022), RT-2 (Zitkovich et al., 2023)	• TTA-Nav (Piriyajitakonkij et al., 2024), Phys2Real (Wang et al., 2025b), RL-VLM-F (Wang et al., 2024d), ICRT (Fu et al., 2025)	• DrEureka (Ma et al., 2024b), VR-Robo (Zhu et al., 2025), LOTUS (Wan et al., 2024), Self-Improving Embodied FMs (Ghasemipour et al., 2025)

- Imitation learning: IL aims at constantly bridging the gap between a model's general knowledge and the specific, high-precision requirements of a real-world application case, on the condition of the availability of few expert demonstrations. In particular, SkillsCrafter (Wang et al., 2026a) realizes lifelong language-conditioned robot learning across multiple sequential manipulation skills, while reducing catastrophic forgetting through symbolic skill distillation. Additionally, LOTUS (Wan et al., 2024) models and refines manipulation skills from demonstration streams, supporting long-term expansion of its skill repertoire.

- Reinforcement learning: RL supports continuous learning by enabling FM robotic policies to improve through repeated interaction, autonomous practice, and reward-driven post-training over extended deployment horizons. In particular, self-improving embodied FMs refine pretrained policies through autonomous practice based on self-predicted rewards and success signals across a robot fleet, thereby enabling downstream skill acquisition with minimal human supervision in (Ghasemipour et al., 2025). Similarly, LiReN (Stachowicz et al., 2024) demonstrates that navigation FMs can improve lifelong learning, through the employment of online RL pipelines, in open-world settings. More recently, VLA models have been adapted through reinforcement fine-tuning, showing improved retention and adaptation to new tasks, while mitigating catastrophic forgetting in sequential manipulation tasks (Liu et al., 2026b).

8.5 Comparative analysis and key insights

Having discussed in detail the different learning stages (Sections 8.1-8.4), this section systematically examines the literature methods, providing a comparative analysis and critical insights for each stage. In this respect, Table 6 summarizes for each learning stage its: a) Primary function, b) Data requirements, c) Learning paradigms, d) Computational requirements, e) Generalization capability, f) Key strengths, g) Critical limitations, and h) Indicative models. Among the various observations and insights, it can be seen that the different learning stages are essentially involved in a complex/sophisticated interplay/trade-off between general-purpose knowledge acquisition and specialized precision refinement. In particular, pre-training on internet-scale data constructs multi-modal data representations for supporting general-purpose reasoning, while offline fine-tuning grounds the acquired skills to the specificities of physical-world settings. On the

other hand, online adaptation and continuous learning aim to enhance robot resilience and evolution aspects, targeting to enable agents to operate robustly in dynamic, real-world environments.

9 Robotic tasks

The introduction of FMs has led to transformative effects in the materialization and execution of all core robotic tasks, mainly by shifting the field from task-specific programming to general-purpose, multi-task agents. In this context, the main robotic tasks identified in the literature, where FM-based solutions have been applied, are: a) Perception, b) Planning, c) Navigation, d) Manipulation, and e) Human-robot interaction, as discussed in Section 4 and further detailed below.

9.1 Perception

Perception aims at creating rich, semantic maps of the surrounding environment, which subsequently enable robots to execute individual actions. In particular, robot perception enables the realization of semantic grounding (i.e., the connection of visual stimuli with real-world entities), discovery of object affordances (i.e., the tasks that can be performed with different objects), and contextual awareness (i.e., the identification of the different types of semantic entities and their location). The main advantages of the incorporation of FMs in robot perception are (Kawaharazuka et al., 2024): a) Open-vocabulary recognition, where models can identify entities for which they are not specifically trained for, b) Zero-shot generalization, where robots can handle novel environments or object types, c) Multi-modal fusion, where robotic agents can efficiently combine multiple information streams (e.g., visual, language, proprioception, etc.), and d) Robustness to noise, where FMs are shown to be reliable in the presence of noisy data. On the contrary, the main limitations of the integration of FMs in robot perception are (Hu et al., 2023): a) High latency, due to the typical extreme scale of the underlying models employed, b) Decreased explainability, where the main factors leading to a particular robot decision is difficult to be precisely defined, c) Presence of hallucinations, where model predictions can be misled and to result in failures in the physical world, and d) Spatial imprecision, which relates to inaccurate localization of (even correct) entity predictions.

The main categories of perception methods are:

- Language-grounded detection and segmentation: This aims at identifying (detection) and precisely outlining (segmentation) the objects present in the robot's surrounding environment based on natural language prompts. In particular, GLIP (Li et al., 2022b) and Grounding DINO (Liu et al., 2024d) provide phrase-based detections that remain robust in the presence of clutter as well as in the zero-shot setting; such detections can subsequently feed language-conditioned manipulation and navigation pipelines, such as CLIPort (Shridhar et al., 2022) and similar solutions (Unlu et al., 2024; Hao et al., 2025). Additionally, SAM-style promptable segmenters (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026) allow the incorporation of box, click, or text prompt information into control pipelines in an interactive way, like SAM-6D (Lin et al., 2024) for 6D pose estimation and RoG-SAM (Mei et al., 2025) for instance-level robotic grasping detection.

- Open-vocabulary 3D semantic mapping: This allows robots to perceive and to localize objects of previously unseen categories in the 3D space, making use of natural language inputs. In this respect, ConceptFusion (Jatavallabhula et al., 2023) builds open-set, language-searchable maps that support uncommon and previously unseen concepts, and multi-modal queries. Additionally, ConceptGraphs (Gu et al., 2024) estimates object nodes and their relations, so that task planners can subsequently operate on top of a semantic scene graph (instead of raw pixels). Moreover, radiance-field (e.g., LERF (Kerr et al., 2023)) and 3D-Gaussian (e.g., FMGS (Zuo et al., 2025)) grounded approaches embed CLIP/DINO features into neural fields for estimating consistent, view-invariant labels. Furthermore, OpenFusion++ (Jin et al., 2025), OpenVox (Deng et al., 2025b), and MR-COGraphs (Gu et al., 2025) focus on real-time, open-vocabulary voxel mapping and multi-robot scene graphs for efficient robot exploration.

- Pose estimation and affordance prediction: Pose estimation aims at aligning an object's local coordinate system to the global, world one, while affordance prediction aims at the detection of the ways that an object can be manipulated; in both cases, FMs are particularly useful due to their inherent ability of linking semantics (language) with spatial geometry (pixels). In particular, OV9D (Cai et al., 2024a) estimates category-agnostic 9-DoF pose and object size without relying on the use of CAD models. Similarly, Oryon (Corsetti et al., 2024) aligns CLIP-guided segments across different views, in order to recover relative 6D pose for unseen objects. On the other hand, OpenAD (Nguyen et al., 2023) models zero-shot 3D affordances in a shared vision–language embedding space, while OVA-Fields (Su et al., 2025) extends this direction to weakly supervised open-vocabulary affordance fields for robot operational part detection in 3D scenes. Moreover, one-shot open affordance learning transforms a single example into dense, class-agnostic affordance masks in (Li et al., 2024b).

- Contact-centric and visuotactile perception: This category of methods focuses on analyzing and modeling the physics of robot interactions for enhancing perception. In particular, contact-centric approaches consider junction points as the primary state representation, while visuotactile ones integrate exocentric vision with local tactile sensing. In this respect, TLA (Hao et al., 2026) and Tactile-VLA (Huang et al., 2025b) combine tactile information streams with vision and language, in order to improve insertion, assembly, and material reasoning tasks. Large tactile–vision–language models (e.g., TALON (Jiang et al., 2024b)) further extend this idea to richer contact semantics. Moreover, visuotactile systems, such as NeuralFeels (Suresh et al., 2024), can enhance in-hand pose and shape estimation, when visual cues are uncertain.

- Long-term object tracking: This aims at locating a given object or point across a sequence of video frames, while maintaining prediction stability over time and capability to recall objects when they become occluded or exit the field of view. The latter is a particular requirement for long-horizon tasks. In this respect, OVTrack (Li et al., 2023b) employs language and diffusion priors to generalize multi-object tracking to unseen categories without explicit video pre-training. Similarly, DINO-MOT (Lee et al., 2024a) combines DINOv2 features with a memory mechanism for robust pedestrian tracking, while COVTrack (Qian et al., 2025b) further strengthens open-vocabulary tracking through improved temporal association across continuous trajectories.

9.2 Planning

Planning serves as the fundamental bridge between high-level (semantic) reasoning and low-level motor control procedures. In particular, the primary usefulness of robot planning lies in long-horizon task decomposition, where a high-level task goal needs to be broken down into multiple, sequential, individual sub-goals over an extended period, prior to their actual execution. The main advantages of the incorporation of FMs in robot planning are (Hu et al., 2023): a) Increased interpretability, where FMs enable a planned sequence of robot actions to be represented in human-like form, b) Increased generalization ability, where robots can often adapt to novel settings by leveraging general-purpose, world knowledge stored in a FM, c) Reduced need for training data, where pre-trained FMs exhibit a decreased need for large-scale, expensive, visuomotor robot data, and d) Safety contraints integration, where FMs enable the efficient incorporation of safety constraints directly into the planning loop. On the contrary, the main limitations of the integration of FMs in robot planning are (Firoozi et al., 2025): a) Logical gaps, where FM-based planners may estimate action steps that are not physically possible, b) Lack of grounding, where FM inference may result into deviations between a high-level plan and the low-level capacities of the robot at hand, c) Increased latency, where the high-computing nature of FMs may be proven restrictive for high-pace tasks, and d) Increased closed-loop complexity, where FM-based solutions face challenges in adjusting their behavior in real-time settings, as a response to constant environmental changes.

The main categories of planning methods are:

- Language-driven task decomposition: This aims at breaking down high-level, long-horizon goals into logical sequences of robot primitive actions (atomic skills), often in the form of structured, executable code (program synthesis). In particular, SayCan (Brohan et al., 2023b) and SayPlan (Rana et al.,

2023) ground each action step on an affordance map or a 3D scene graph, respectively, so that abstract sub-goals to correspond to concrete objects and locations. On the other hand, Code-as-Policies (Liang et al., 2023) and similar approaches generate directly short, Python-like programs for integrating existing software libraries, rendering planning easier to inspect, test, and modify (Singh et al., 2023; Huang et al., 2023c).

- Neuro-symbolic closed-loop reasoning: This category of methods combines the general-purpose knowledge of a FM with formal, logical checking procedures, so as to guarantee the successful operation of robot agents in the real world. In this respect, ISR-LLM (Zhou et al., 2024b) converts natural language instructions to PDDL ones and iteratively refines the estimated plans using a symbolic validation scheme, until a feasible sequence of actions is determined. Additionally, AutoTAMP (Chen et al., 2024c) employs an LLM for generating or translating high-level language instructions into intermediate representations suitable for a task-and-motion planning (TAMP) solving mechanism, while making use of autoregressive re-prompting to correct syntactic and semantic errors. Moreover, safety- and feasibility-oriented planners, such as SELP (Wu et al., 2025b) and LLM-GROP (Zhang et al., 2025g), translate LLM proposals into explicit constraint checking and task-and-motion reasoning procedures, in order to avoid unsafe or dead-end policies.

- Multi-modal policy generation: This category focuses on generating high-level plans or robot actions by integrating multiple input modalities, such as language, vision, and embodied state information. In particular, PaLM-E (Driess et al., 2023) combines language, vision, and proprioception for embodied reasoning and action generation, while RT-2 (Zitkovich et al., 2023) and OpenVLA (Kim et al., 2025) demonstrate that language-aligned visual representations can transfer web-scale semantic knowledge to real-world robot control.

- Execution-time validation and failure recovery: This category focuses on monitoring generated plans during deployment, verifying action preconditions, detecting inconsistencies or failures, and triggering corrective re-planning when needed. In this respect, Code-as-Monitor (Zhou et al., 2025a) introduces constraint-aware visual programming for reactive and proactive robotic failure detection, SELP (Wu et al., 2025b) incorporates explicit safety and feasibility constraints into the planning loop, while VLM-based monitoring frameworks such as Guardian (Pacaud et al., 2025) and unified real-time failure-handling approaches (Ahmad et al., 2025) support execution-time failure detection and recovery in robotic manipulation.

- Semantic multi-robot coordination: This category of methods capitalizes on the broad knowledge base of FMs for coordinating multi-robot setups, where accurate reasoning about task dependencies, resources, and scheduling is needed. In this direction, LiP-LLM (Obata et al., 2024) employs an LLM to create a skill list and a corresponding dependency graph from language instructions, while subsequently relying on linear programming techniques to allocate tasks across robotic platforms. Additionally, SMART-LLM (Kannan et al., 2024) assigns agents task-specific roles based on structured representations of their skills and capabilities, reducing in this way instruction drift and enabling coalition formation and task allocation, based on a single high-level instruction. Moreover, large-scale orchestration systems, such as AutoRT (Ahn et al., 2024), combine language-based task assignment with human oversight and monitoring, demonstrating that FM-based planners can coordinate dozens of physical robots in real-world environments.

9.3 Navigation

The fundamental usefulness of FMs in robot navigation lies on providing the necessary spatial common-sense knowledge in embodied AI settings. The latter is mainly accomplished due to the capacity of FMs to process the robot's surrounding environment as a high-level semantic space, instead of considering rigid, inflexible spatial maps. The main advantages of the incorporation of FMs in robot navigation are (Pan et al., 2025b): a) Open-world navigation, where robots can operate in environments including previously unseen entities, b) Cross-embodiment transfer, where a single pretrained model can be deployed in different/diverse hardware platforms, and c) Semantic reasoning capability, which relies on the inherent ability of FMs to combine vision with language understanding for instruction interpretation. On the contrary, the main limitations of the

integration of FMs in robot navigation are (Firoozi et al., 2025): a) Sim-to-real gap, where FMs trained in simulation may exhibit difficulties in operating in real-world environments, b) Lack of training data, where large-scale, high-quality, diverse 3D navigation data, required for FM training, is difficult and expensive to collect, c) Increased latency, where the high computational needs of FMs can lead to challenging situations in real-world applications, and d) Safety and ethical concerns, where FMs need to be equipped with appropriate social norms for operating in human-shared spaces.

The main categories of navigation methods are:

- Semantic spatial grounding: This category of methods aims at registering the objects present in the environment, but also their relative position (with respect to the robots) and a concrete action plan for reaching them (in natural language form). In particular, VLMaps (Huang et al., 2023a) builds CLIP-indexed spatial memories that enable the system to query objects and rooms without task-specific retraining, directly ingesting visual–language features into a 3D map. Zero-shot localization schemes, such as PixNav (Cai et al., 2024c) and VLTNet (Wen et al., 2025a), guide navigation through pixel-level target cues or construct semantic navigation maps and rank exploration frontiers from language prompts. Moreover, open-vocabulary mapping methods (like One Map to Find Them All (Busch et al., 2025), DualMap (Jiang et al., 2025b), and scene graph-based approaches (Loo et al., 2025)) support multi-object navigation, dynamic environments, and functional queries, so that a single map can accommodate for many language goals and robots.

- Instruction-following policies: This is based on the increased capability of FMs in interpreting natural language instructions, without requiring a pre-defined map or hard-coded scripts for every possible object present in the environment. In this respect, generalist navigation models, such as GNM (Shah et al., 2023b) and ViNT (Shah et al., 2023c), formalize navigation as a sequence prediction problem over images and poses, relying on a single model across different robots and environments. Additionally, NaviLLM (Zheng et al., 2024a) unifies instruction following and embodied QA with schema-tuned prompts, while NavFormer (Wang et al., 2024a) learns target-driven policies in unknown, dynamic environments. Moreover, FASTNav (Chen et al., 2024d) demonstrates that compact, LoRA-adapted language models can operate in real-time on embedded hardware, offering a practical path from large offline training to on-board controllers.

- End-to-end policies: This aims at developing a unified architecture that can map raw sensorial data directly to motor commands, instead of adopting the conventional modular design of breaking down navigation into individual steps (like mapping, localization, and path planning). In particular, ViNT (Shah et al., 2023c) learns a generalizable visuomotor navigation policy across multiple robots and environments, while NavFoM (Zhang et al., 2026) extends this idea towards cross-embodiment and cross-task navigation. In driving-oriented settings, end-to-end frameworks, such as DiffusionDrive (Liao et al., 2025) and DriveGPT-4 (Xu et al., 2024b), further demonstrate that multi-modal models can predict low-level control signals directly from visual- and language-conditioned inputs.

9.4 Manipulation

The primary utility of FMs in robot manipulation lies on providing the required knowledge for implementing the translation from high-level, semantic, human-like instructions to precise physical pressures and movements needed to manipulate an object. The main contribution of FMs for achieving the latter comprises their cross-embodiment learning capability, where the same model pretrained on data from multiple different platforms can be deployed to diverse setups. The main advantages of the incorporation of FMs in robot manipulation are (Li et al., 2024a): a) Increased generalization, where FMs are shown to be robust in handling previously unseen object types, b) Improved robustness, where FMs can make use of real-time (visual) feedback to adjust their manipulation strategy on the fly, and c) Enhanced embodiment capability, where FM solutions enable the understanding of the physical properties of the objects, prior to their manipulation. On the contrary, the main limitations of the integration of FMs in robot manipulation are (Sapkota et al., 2025): a) Limited training data, where sufficient quantities of high-quality, labeled robotic manipulation data is difficult to collect, b) Safety concerns, where the robot policies are not always guaranteed to result into

feasible and safe manipulations, and c) Low action precision, where the increased generalization ability of FMs is accompanied with corresponding decrease in physical task execution for specialized domains.

The main categories of manipulation methods are:

- Language-to-action models: This category aims to interpret the semantic meaning of a natural language command and to map it to the physical world setting, without requiring task-specific programming. In particular, RT-1 (Brohan et al., 2023a) and RT-2 (Zitkovich et al., 2023) approach manipulation as a sequence modeling problem over multi-modal tokens, while PaLI-X (Chen et al., 2024b) aims at incorporating broad visual knowledge. Additionally, OpenVLA (Kim et al., 2025) comprises a large vision–language–action policy that adapts to new robotic platforms, based on small-scale fine-tuning. GR00T N1 (Bjorck et al., 2025) combines a deliberative VLM with a diffusion motor policy for bimanual humanoid skills acquisition. Moreover, state-space variants, such as RoboMamba (Liu et al., 2024c), replace the core transformer component with a Mamba-based selective state-space model for lowering latency, while maintaining increased visuomotor skill performance.

- Retrieval-augmented imitation learning: This relies on receiving guidance from a large database of relevant previous demonstrations for predicting future actions. In this context, DINOBot (Di Palo & Johns, 2024) detects similar demonstrations based on DINO feature correspondence and subsequently estimates dense manipulation trajectories for realizing one- and few-shot generalization to novel objects. Additionally, STRAP (Memmel et al., 2025) retrieves relevant manipulation sub-trajectories from prior demonstrations and uses them to augment few-shot imitation learning, improving generalization to novel objects and tasks.

- Constraint-aware policy synthesis: This category employs a FM to generate high-level control code or mathematical objectives, which are explicitly bounded by physical, safety, and environmental constraints. In this context, CoPa (Huang et al., 2024a) detects task-relevant parts using a multi-modal LLM and estimates spatial constraints that a conventional planner translates into 6-DoF actions. ReKep (Huang et al., 2025c) represents tasks as sequences of relational keypoint constraints and optimizes them hierarchically for single- and dual-arm assembly tasks. Moreover, part-centric perception methods, like PartSLIP++ (Zhou et al., 2025d), estimate part-affordance correlations that are required for robust task execution.

- Semantic spatial maps: This category aims at generating representations of the environment that combine 3D spatial geometry with semantic information about the involved objects. In this direction, VoxPoser (Huang et al., 2023c) estimates constraints and object affordances from natural language inputs, creates 3D value maps from vision-language information cues, and applies common motion planning routines in a zero-shot fashion. Additionally, AdaRPG (Zhang et al., 2025h) leverages VLMs to infer part affordances and operational constraints that guide primitive skills for articulated-object manipulation.

9.5 Human-robot interaction

The fundamental usefulness of FMs in HRI is grounded on their increased ability for realizing semantic, human-like reasoning and interpreting human intent. In particular, robots are capable of reacting to conversational instructions, asking clarification questions, understanding contextual settings, and providing explanations for their actions, which greatly simplifies communication (especially) with non-expert users. The main advantages of the incorporation of FMs in HRI are (Zhao et al., 2025c): a) Rapid generalization, where a human user can demonstrate a new task to a robot and it can adapt instantly, b) Intuitive control, where due to the increased human behavior interpretation capabilities of FMs, robot control becomes more efficient and intuitive, c) Increased safety alignment, where user-provided feedback can boost robots to learn/incorporate social norms, and d) Efficient error recovery, where human provided feedback can be rapidly exploited by a robot for recovering from a fault state. On the contrary, the main limitations of the integration of FMs in HRI are (Xiao et al., 2025b): a) Lack of training data, where data collection involving human feedback/interactions is typically costly, b) Incorporation of human bias, where robots learn in an unconstrained way from their

human-provided feedback, and c) Semantic drift occurrences, where human intent is not always easy to interpret as contextual/environmental conditions may change.

The main categories of HRI methods are:

- Conversational policy alignment: This category of methods focuses on using natural language dialogue to dynamically adjust a robot's behavior in real-time, so that it matches a human's specific intent, preferences, and safety boundaries. In particular, DRAGON (Liu et al., 2024e) comprises a dialogue-based navigation framework that grounds free-form commands in visual landmarks, describes the environment, and asks clarification questions when the reference is unclear. Additionally, PlanCollabNL (Izquierdo-Badiola et al., 2024) translates spoken instructions into editable plans so that users can insert, remove, or reorder steps in collaborative manipulation and assembly settings.

- Reciprocal social tuning: This aims at developing a high-level, semantic communication framework, where both humans and robots continuously adjust their behaviors, social cues, and expectations to harmonize each other. In this context, incremental system updates combine natural-language feedback with on-robot sensing so that the controller can refine prompts, code snippets, or skill graphs after each mistake and also to reuse the acquired knowledge at a later stage (Bärmann et al., 2024). Additionally, design studies on conversational companion robots provide concrete guidelines for everyday HRI settings, such as clear grounding, turn-taking, and repair under uncertainty (Irfan et al., 2024). TidyBot (Wu et al., 2023a) demonstrates that LLMs can learn user-specific preferences for household clean-up from a few examples and then generalize these preferences to new objects and scenes. Similarly, LAMS (Tao et al., 2025) extends this idea to assistive teleoperation, by using an LLM to switch control modes based on task context and improving performance over time based on user feedback.

- Active alignment and mitigation: This targets to ensure that a robot remains synchronized with human intent, while proactively handling errors or deviations. In this respect, RoboVQA (Sermanet et al., 2024) queries egocentric video to check preconditions, to verify whether a sub-task has succeeded, and to proactively request human intervention in case of identified failures. Additionally, embodied LLM controllers also incorporate human feedback for adjusting their plans when they detect that the current state drifts from the expected goal (Mon-Williams et al., 2025). Overall, the common practice relies on combining a latency-aware VLM with a dialogue-based manager, so that the robot can provide concise explanations, ask specific clarifications, and implement targeted re-planning (Bärmann et al., 2024; Sermanet et al., 2024).

9.6 Comparative analysis and key insights

Having discussed in detail the different robotic tasks (Sections 9.1-9.5), this section systematically examines the literature methods, providing a comparative analysis and critical insights for each task. In this respect, Table 7 summarizes for each robotic task its: a) Primary function, b) FM type, c) Input types, d) Output types, e) Key strengths, f) Critical limitations, and h) Indicative models. Among the various observations and insights, it can be seen that for all tasks a fundamental transition is being implemented from isolated, task-specific modules to integrated, generalist architectures that leverage internet-scale pretraining. Additionally, though perception achieves robust open-vocabulary recognition and semantic fluency, task planning largely depends on sophisticated grounding mechanisms to ensure high-level reasoning remains physically feasible. On the other hand, navigation and manipulation put particular focus on the embodiment aspects; however, they face a critical trade-off between large-scale, semantic reasoning latency and high-speed requirements for dexterous, real-time motor control. Moreover, HRI facilitates intuitive communication, yet its practical deployment is being hindered by the performance-latency trade-off of existing FMs and the emerging challenge of adhering to safety and social norms. Overall, a unifying observation across all tasks comprises the lack of large-scale, physical-world, training data, which emerges as the primary bottleneck for further extending universal robotic intelligence capabilities. Moreover, representative literature methods per robotic task are illustrated in Fig. 6.

Table 7: Robotic tasks: Comparative analysis and key insights.

Robotic task	Perception	Planning	Navigation	Manipulation	Human-robot interaction
Primary function	• Semantic scene recognition • Object detection • 3D spatial grounding	• High-level goal decomposition • Long-horizon reasoning • Constraint-aware task sequencing	• Goal-conditioned pathfinding • Obstacle avoidance • Topological exploration	• Fine-motor control • Grasping actions • Sequential object relocation	• Context-aware communication • Socially compliant collaboration • Natural-language command interpretation
FM type	• VFMs • VLMs	• LLMs • VLMs	• LLMs • VFMs • VLMs • VLAs	• VLMs • VLAs	• LLMs • VLMs • VLAs
Input	• RGB-D images • 3D point clouds • Textual prompts	• Natural-language goals • Scene graphs • Symbolic state • Execution feedback	• RGB images • Spatial maps • Spatial memory • Odometry measurements • Language instructions	• Natural-language goals • Proprioception and gripper state • Video demonstrations • Scene descriptions • Tactile/force data	• Human speech • Human gestures • Interaction history • Social context
Output	• Segmentation masks • Scene graphs • Semantic labels • 3D maps • Poses and affordances	• Plans, sub-goals, and code • Constraints and preconditions • Verifiable task decompositions	• Velocity commands • Spatial trajectories • Target locations • Route descriptions	• Executable action sequences • Grasps and placements • 6-DOF end-effector poses • Gripper states • Adapted policies	• Natural-language speech • Socially-aware motion • Refined goals and corrections • Explanations and progress reports • User models and preferences
Strengths	• Open-vocabulary recognition • Zero-shot generalization • Multi-modal fusion • Robustness to noise	• Increased interpretability • Increased generalization ability • Reduced need for training data • Safety-constraint integration	• Open-world navigation • Cross-embodiment transfer • Semantic reasoning capability	• Increased generalization • Improved robustness • Enhanced embodiment capability	• Rapid generalization • Intuitive control • Increased safety alignment • Efficient error recovery
Limitations	• High latency • Decreased explainability • Presence of hallucinations • Spatial imprecision	• Logical gaps • Lack of grounding • Increased latency • Increased closed-loop complexity	• Sim-to-real gap • Lack of training data • Increased latency • Safety and ethical concerns	• Limited training data • Safety concerns • Low action precision	• Lack of training data • Incorporation of human bias • Semantic drift occurrences
Indicative models	• CLIP (Radford et al., 2021), SAM (Kirillov et al., 2023), Grounding DINO (Liu et al., 2024d), DINOv2 (Oquab et al., 2024), ConceptGraphs (Gu et al., 2024)	• SayCan (Brohan et al., 2023b), Code-as-Policies (Liang et al., 2023), SayPlan (Rana et al., 2023), AutoTAMP (Chen et al., 2024c)	• LM-Nav (Shah et al., 2023a), GNM (Shah et al., 2023b), ViNT (Shah et al., 2023c), DRAGON (Liu et al., 2024e)	• RT-1 (Brohan et al., 2023a), OpenVLA (Kim et al., 2025), GR00T N1 (Bjorck et al., 2025), Diffusion Policy (Chi et al., 2025)	• PaLM-E (Driess et al., 2023), Gemini Robotics (Team et al., 2025), RoboVQA (Sermanet et al., 2024), TidyBot (Wu et al., 2023a), LAMS (Tao et al., 2025)

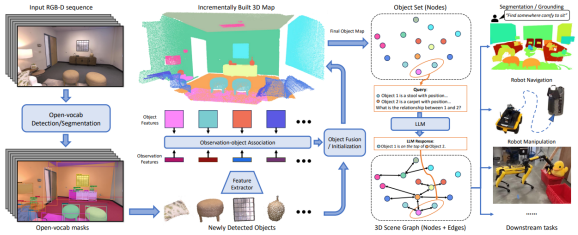
10 Application domains

The incorporation of FMs in robotic solutions has greatly boosted multiple aspects of their core technologies (e.g., autonomy, complex decision-making, semantic reasoning, etc.). This has in turn facilitated their widespread use in a wide range of challenging real-world application domains, including: a) Agentic mobility, b) Industrial manipulation, c) Supply operations, d) Service robots, e) Medical robots, f) Cognitive agrisystems, g) Crisis agents, h) Maritime robotics, and i) Space robotics, as discussed in Section 4 and further detailed below. Additionally, representative literature methods per application domain are illustrated in Fig. 7.

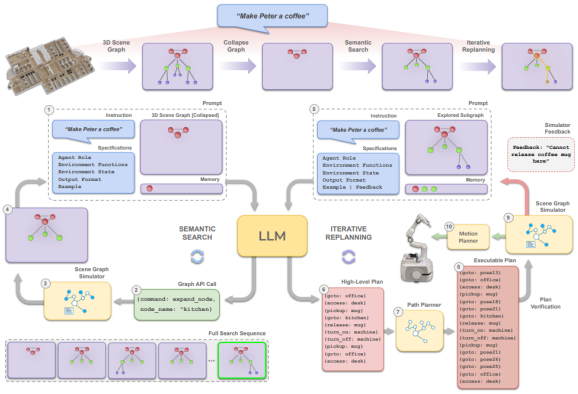
10.1 Agentic mobility

The integration of FMs has induced a paradigm shift in autonomous movement, transitioning from rigid path-following routines to autonomous agentic solutions, where robots leverage high-level reasoning to understand mission objectives and to adapt to environmental changes.

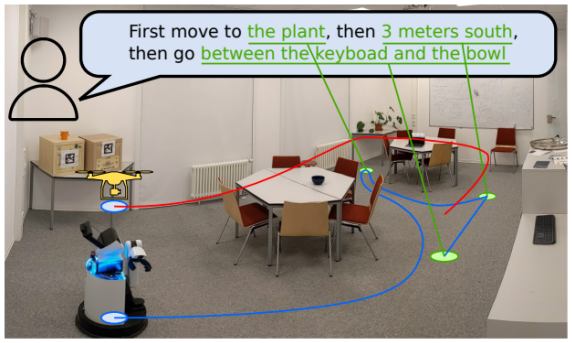
A core application of these models is in semantic spatial grounding, which replaces traditional, inflexible spatial maps with queryable memories. For instance, VLMaps (Huang et al., 2023a) constructs CLIP-indexed spatial memories by directly ingesting visual-language features into 3D maps; this allows robotic systems to navigate to specific rooms or objects via natural language queries, without requiring task-specific retraining. Similarly, methods like EffoNAV (Shen et al., 2025) pair CLIP-based goal detection with exploration routines and low-level controllers to handle visual targets in challenging settings.



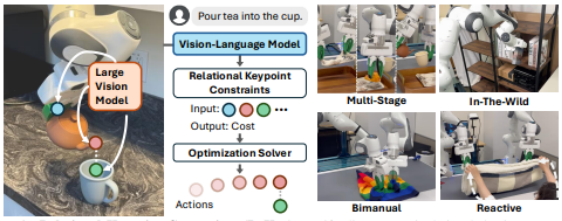
(a)



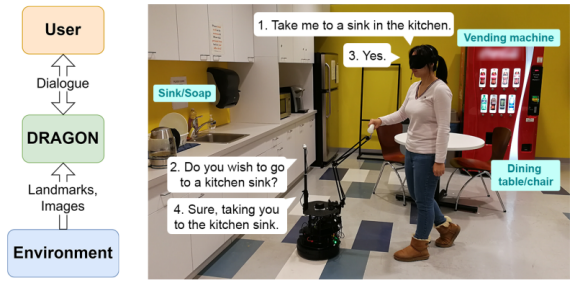
(b)



(c)



(d)



(e)

Figure 6: Representative literature methods per robotic task: (a) Perception: Open-vocabulary 3D scene understanding (ConceptGraphs (Gu et al., 2024)), (b) Planning: Grounded long-horizon task decomposition (SayPlan (Rana et al., 2023)), (c) Navigation: Semantic spatial grounding in language-indexed maps (VLMaps (Huang et al., 2023a)), (d) Manipulation: Visually grounded relational keypoint constraint estimation (ReKep (Huang et al., 2025c)), and (e) Human-robot interaction: Dialogue-based assistive navigation (DRAGON (Liu et al., 2024e)).

FMs also address the per-robot engineering bottleneck through cross-embodiment and cross-site generalization. Generalist models, such as GNM (Shah et al., 2023b) and ViNT (Shah et al., 2023c), formalize navigation as a sequence prediction problem over images and poses. This enables a single visual backbone to drive diverse mobile platforms across varied environments, significantly facilitating the rollout of new robots in novel sites. This scalability is further demonstrated by City Walker (Liu et al., 2025b), which learns policies that transfer across road networks in different cities.

For ensuring real-time feasibility on embedded hardware, particular focus has been given on model efficiency. In particular, FASTNav (Chen et al., 2024d) utilizes compact, LoRA-adapted language models to provide instruction-following capabilities that run on-board in real-time. Additionally, frameworks like NaviLLM (Zheng et al., 2024a) unify instruction following with embodied question-answering to provide a flexible interface for operators.

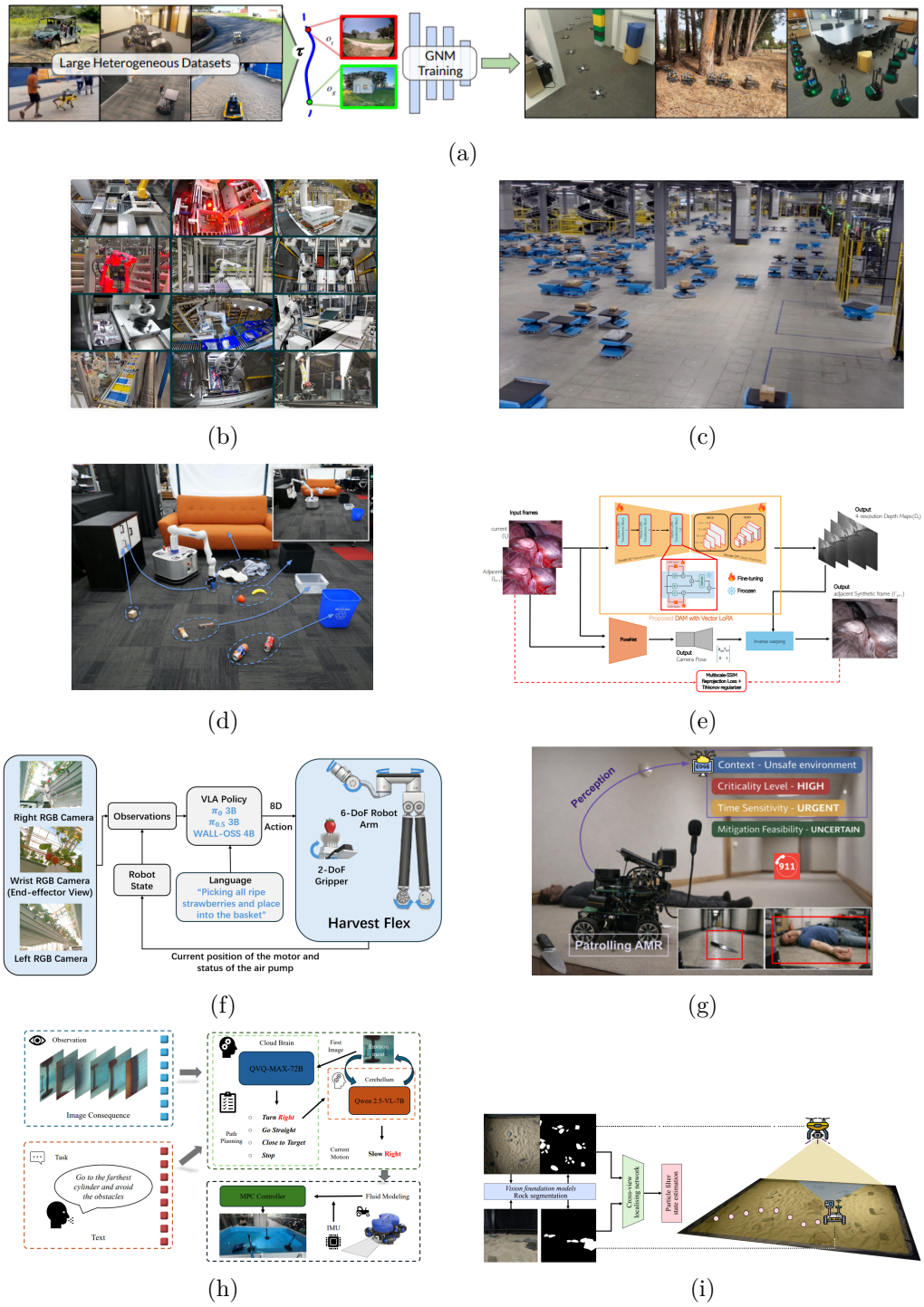


Figure 7: Representative literature methods per application domain: (a) Agentic mobility (GNM (Shah et al., 2023b)), (b) Industrial manipulation (RFM-1 (Sohn et al., 2024)), (c) Supply operations (DeepFleet (Agaskar et al., 2025)), (d) Service robots (TidyBot (Wu et al., 2023a)), (e) Medical robots (DARES (Zeinoddin et al., 2024)), (f) Cognitive agrisystems (HarvestFlex (Zhao et al., 2026b)), (g) Crisis agents (See Something, Say Something (Cancelli et al., 2026)), (h) Maritime robotics (UnderwaterVLA (Wang et al., 2025g)), and (i) Space robotics (Planetary cross-view localisation (Holden et al., 2026)).

In road environments, FMs enhance safety and transparency, by conditioning planning on linguistic reasoning or graph-based visual question answering. Frameworks like DriveGPT-4 (Xu et al., 2024b), Drive Anywhere (Wang et al., 2024c), and DriveLM (Sima et al., 2024) go beyond black-box end-to-end policies, by producing human-readable rationales and intermediate decisions that are easier to inspect, to debug, and to align with safety standards. For assistive scenarios, DRAGON (Liu et al., 2024e) uses dialogue-based navigation to allow users to describe targets and constraints in natural language, where the robot provides verbal explanations of its planned route.

10.2 Industrial manipulation

FMs are fundamentally redefining industrial automation by replacing rigid, task-specific routines with flexible, general-purpose agents that excel in unstructured manufacturing environments. In particular, models like RFM-1 (Sohn et al., 2024) are deployed within high-throughput production cells, leveraging visuo-motor backbones trained on millions of real-world pick actions to achieve zero-shot generalization, when encountering novel objects. For large-scale logistics, PRIME-1 (Inc., 2025) provides a foundation tailored to parcel workflows, facilitating site-specific adaptation and the rapid rollout of autonomous systems across diverse distribution centers. In order to address hardware variability, OpenVLA (Kim et al., 2025) utilizes parameter-efficient fine-tuning to simplify transfer across different robotic platforms, effectively reducing the need for extensive per-line retraining, while enabling operators to issue high-level, language-conditioned tasks. Moreover, RT-2 (Zitkovich et al., 2023) capitalizes on web-to-robot knowledge transfer to significantly boost robustness against long-tail objects and complex instructions on assembly and kitting lines, ensuring that robots can adapt to items outside their original training data.

10.3 Supply operations

FMs are drastically evolving autonomous supply operations, by providing robots with the flexibility and intelligence needed to handle unstructured, dynamic, and complex logistics tasks. This paradigm shift allows robotic agents to adapt to new items, facility layouts, and operational disruptions without requiring manual reprogramming. At a broader scale, logistics operations utilize FMs for sustainable planning, multi-agent coordination, and end-to-end decision support, specifically targeting 'Logistics 5.0' goals, such as greener routing (Nicoletti & Appolloni, 2024). By integrating multi-agent systems with foundation backbones, autonomous supply chains can effectively link demand forecasting, resource allocation, and transport routing through shared representations and agentic reasoning (Xu et al., 2024a). Consequently, FM-driven solutions can support a wide array of critical industrial tasks, including demand sensing, inventory positioning, warehouse tasking, and transport orchestration (Nicoletti, 2025; Agaskar et al., 2025).

10.4 Service robots

The integration of FMs represents a major paradigm shift for service robots, transforming them to adaptable agents, capable of navigating through unstructured domestic environments and interpreting natural language instructions. In this context, TidyBot (Wu et al., 2023a) exemplifies the use of few-shot LLM reasoning to learn personalized tidying preferences directly from user conversations, allowing a mobile manipulator to ground abstract 'put-away' commands into specific physical actions across novel scenes. Extending these capabilities to more complex, multi-step procedures, ELLMER (Mon-Williams et al., 2025) performs tasks such as coffee brewing, by integrating visual, force, and linguistic feedback; specifically, ELLMER utilizes a retrieval-augmented memory mechanism to adapt its plans on the fly if environmental conditions change or failures are detected, ensuring robust execution under uncertainty.

10.5 Medical robots

FMs in the medical domain serve as a critical bridge between high-level clinical knowledge and low-level robotic execution. These models are increasingly integrated into surgical perception and decision-making pipelines to handle high-stake, high-variability tasks. A key application is monocular depth estimation; for instance, Surgical-DINO (Cui et al., 2024) adapts DINOv2 using a LoRA adapter specifically for endoscopic imagery.

Table 8: Public datasets concerning in principle perception tasks. Abbreviations used: Environment type (Simulation (S), Real (R)), Temporality (Static (ST), Video (V), Sequential (SE)), and Embodiment (Yes (Y), No (N)).

Dataset	Year	Scale	Classes	Modalities	Annotation type	Domain	Env. type	Temp.	Emb.	Description
Matterport3D (Chang et al., 2017)	2017	• 90 building scenes • 10,800 panoramas • 194,400 RGB-D images • 50,811 object instances • 38,328 3D bounding boxes	• 20 structural • 20 objects	• RGB-D panoramas • 3D meshes • Camera poses	• 3D reconstruction • 2D masks • 3D semantic & instance labels	Indoors	R	ST	N	RGB-D dataset of entire buildings for global scene understanding
2D-3D-S (Armeni et al., 2017)	2017	• 6K m^2 coverage • 270 rooms / 6 areas • 70,496 regular RGB images • 1,413 equirectangular RGB images • 695M 3D points	• 7 structural • 6 objects	• RGB-D images • Surface normals • 3D meshes & point clouds • Camera poses & metadata	• 2D & 3D semantic & instance labels • 3D bounding boxes • Room categorizations	Indoors	R	ST	N	Jointly registered 2D and 3D data for office-style indoor scene understanding
Semantic–KITTI (Behley et al., 2019)	2019	• 22 sequences • 43,552 LiDAR scans • 360° FOV coverage	• 11 structural • 9 objects • 8 actors	• LiDAR point clouds • 3D point trajectories • Sensor poses	• Point-wise semantic labels • Sequence-level ID tracking • Scene completion targets	Driving	R	SE	Y	Large-scale LiDAR driving dataset with point-level semantic labels for sequences
BDD100K (Yu et al., 2020)	2020	• 100K HD video clips • 40M video frames • 100M total distance (km) • 10 vision tasks	• 11 structural • 8 objects	• RGB video • GPS/IMU metadata • Time/ Weather/ Lighting tags	• 2D bounding boxes • Instance & driveable masks • Lane marking • Object tracking	Driving	R	V	Y	Diverse driving dataset covering varying weather, times, and city environments
nuScenes (Caesar et al., 2020)	2020	• 1K scenes • 1.4M images • 390K LiDAR • 1.4M radar sweeps • 1.4M 3D object boxes	• 9 structural • 23 objects	• 6 RGB cams (360°) • 32-beam LiDAR • 5 long-range radars • GPS/IMU metadata	• 3D bounding boxes • LiDAR semantic labels	Driving	R	SE	N	Driving dataset providing full AV sensor suite (LiDAR, radar, 6 cameras)
Waymo Open (Sun et al., 2020)	2020	• 1K sequences • 200K LiDAR frames • 1M RGB images • 12M 3D boxes • 12.6M 2D boxes	• 15 structural • 13 objects • 8 actors	• 5 LiDAR • 5 RGB cameras • Sensor poses	• 2D/3D tracking boxes • Global 3D point IDs • Cross-camera 2D labels	Driving	R	SE	N	High-resolution, multi-sensor driving data focused on perception and motion prediction
Ego4D (Grauman et al., 2022)	2022	• 3,670 hours of video • 931 participants • 74 locations • 9 countries • 5M+ episodic annotations	• N/A (open)	• Egocentric RGB video • Multi-channel audio • 3D eye gaze • IMU • Stereo	• Episodic memory tags • Hand-object interactions • Audio-visual diarization • Social interaction labels	Daily life	R	V	N	Large-scale egocentric video captured by hundreds of people worldwide
HM3D-SEM (Yadav et al., 2023)	2022	• 216 3D spaces • 3,100 rooms • 142,646 object instances • 14,200+ human hours	• 12 structural • 28 objects	• Textured 3D meshes • Semantic textures • Room-level metadata	• Pixel-level semantics • 40 Matterport categories • Instance-level labels	Indoors	R	ST	N	Dataset of semantically annotated building-scale 3D indoor reconstructions
ScanNet++ (Yeshwanth et al., 2023)	2023	• 460 scenes • 1,858 laser scans • 280K 33MP DSLR images • 3.7M RGB-D frames	• 50 structural • 950+ objects	• RGB-D video • 3D meshes • Camera poses	• 3D semantic & instance labels • Multi-label ambiguity tags	Indoors	R	ST	N	Sub-millimeter fidelity indoor scans paired with high-resolution 33MP DSLR imagery

Similarly, DARES (Zeinoddin et al., 2024) tailors a Depth Anything Model, using self-supervised Vector-LoRA, to better align with surgical scene statistics. Additionally, EndoDDC (Lin et al., 2026) addresses sparse-to-dense depth reconstruction for endoscopic robotic navigation through diffusion-based depth completion, supporting more reliable 3D reconstruction and safe instrument guidance. These advancements allow for the robust implementation of surgical assistance systems, including real-time instrument tracking, tool detection, and intraoperative guidance within hospital settings (He et al., 2024). Ultimately, this integration provides

Table 9: Public datasets concerning in principle planning tasks. Abbreviations used: Environment type (Simulation (S), Real (R)), Temporality (Static (ST), Video (V), Sequential (SE)), and Embodiment (Yes (Y), No (N)).

Dataset	Year	Scale	Classes	Modalities	Annotation type	Domain	Env. type	Temp.	Emb.	Description
AI2-THOR (Kolve et al., 2017)	2017	• 120 room scenes • 4 room categories • 3,578 interactive objects	• 100+ object types	• Egocentric RGB-D videos • Object metadata	• Object state changes • Navigation actions • Arm manipulation	Household	S	SE	Y	Near photo-realistic 3D environments for visual AI agents navigation and interaction
VirtualHome (Puig et al., 2018)	2018	• 2,821 action programs • 6 furnished houses • 357 objects per house	• ∼300 objects • 70 actions	• Natural language • Synthetic video • 3D poses	• Action programs • Timestamps • Atomic interactions	Household	S	SE	Y	Simulation of complex daily activities via executable programs and scripts
BabyAI (Chevalier-Boisvert et al., 2019)	2018	• 19 difficulty levels • 2.48×10^{19} instructions	• 6 object types	• Synthetic language • 2D grid	• Expert demonstrations • Sub-goal decompositions	Indoors	S	SE	Y	Dataset focused on sample efficiency and grounded language learning in grid worlds
ALFRED (Shridhar et al., 2020)	2020	• 120 indoor scenes • 25,743 language directives • 8,055 expert demos	• 7 task types • 80 objects	• Egocentric RGB videos • Natural language	• High/low-level instructions • Action sequences • Pixel-wise interaction masks	Household	S	V,SE	Y	Mapping of natural language to sequences of actions for visual AI agents
CALVIN (Mees et al., 2022)	2022	• 4 environments • 389 instructions	• 34 tasks	• RGB-D videos • Vision-based tactile • Proprioception • Natural language	• Language goals • Pre-task locomotion behavior • Precomputed MiniLM embeddings	Tabletop	S	SE	Y	Long-horizon, language-conditioned policy learning for continuous control
BEHAVIOR-1K (Li et al., 2023a)	2024	• 50 scenes	• 1K activities • 1.2K objects	• Egocentric RGB-D videos • Proprioception	• Logic (BDDL) • Transition rules • Semantic properties	Household	S	SE	Y	Human-centered activities with realistic physics and state changes
LAMBDA (Jaafar et al., 2025)	2024	• 31 rooms • 8 environments	• 571 tasks	• Natural language • Egocentric RGB-D videos • Robot poses & actions	• Human-collected trajectories • Semantic maps	Household	S,R	SE	Y	Focus on data-efficiency for multi-room, multi-floor mobile manipulation

context-aware intelligence and precise assistance, enabling robotic platforms to robustly support complex clinical procedures.

10.6 Cognitive agrisystems

The integration of FMs into cognitive agrisystems marks a transition towards adaptive, embodied intelligence in field operations. By utilizing VLA architectures, these systems can ingest high-level semantic instructions and translate them into precise motor outputs in real-time. Unlike traditional agricultural robots that rely on hard-coded rules, FMs bridge the semantic gap, by using massive pre-trained datasets to reason through the unpredictable variability of biological environments, such as shifting light, overlapping foliage, and irregular crop shapes (Yin et al., 2025b). For instance, HarvestFlex (Zhao et al., 2026b) employs VLA policies for real greenhouse tabletop strawberry harvesting, a long-horizon, unstructured task, challenged by occlusion and specular reflections. Additionally, FM-based reasoning supports task planning and action selection in crop monitoring and field-management scenarios in (Cuaran et al., 2026).

10.7 Crisis agents

FMs are fundamentally reforming disaster response and public safety, by enabling crisis agents to transition from rigid, remote-controlled setups to autonomous systems capable of high-level reasoning in unpredictable and hazardous environments. In particular, SafeGuard ASF (Canh et al., 2026) combines multi-modal hazard perception with agentic reasoning for real-time fire-risk detection and disaster recovery. Additionally, a robotic fire-risk detection system based on dynamic knowledge graphs and LLM-enhanced multi-modal reasoning is

Table 10: Public datasets concerning in principle navigation tasks. Abbreviations used: Environment type (Simulation (S), Real (R)), Temporality (Static (ST), Video (V), Sequential (SE)), and Embodiment (Yes (Y), No (N)).

Dataset	Year	Scale	Classes	Modalities	Annotation type	Domain	Env. type	Temp.	Emb.	Description
REVERIE (Qi et al., 2020)	2020	• 10,567 panoramas • 86 scenes • 23,536 instructions • 4,140 objects	• 489 objects	• Natural language • RGB panoramas	• Instructions • Target boxes • Nav-graph sequences	Indoors	S	SE	Y	Remote embodied visual referring expressions in unseen environments
VLN-CE (Krantz et al., 2020)	2020	• 90 scenes • 7,189 trajectories • 21,567 instructions	• N/A (path-based)	• Natural language • Egocentric RGB-D videos	• Low-level continuous actions • Navigation paths	Indoors	S	SE	Y	Navigation in continuous environments, emphasizing fine-grained control
RxR (Ku et al., 2020)	2020	• 126,069 instructions • 16.5M total words • 3 languages	• N/A (path-based)	• Multilingual text • Virtual pose traces • RGB panoramas	• Dense spatiotemporal grounding • Time-aligned text-to-pose	Indoors	S	SE	Y	Large-scale multilingual vision-and-language navigation
Habitat-Web (Ramrakhya et al., 2022)	2022	• 80K navigation trajectories • 12K Pick&Place trajectories • 29.3M actions • 22,600 hours	• 21 object trajectories	• Egocentric RGB-D videos • Teleoperation traces	• Human task trajectories • Implicit search heuristics	Indoors	S	SE	Y	Imitation learning from large-scale human demonstrations collected on the web
ScaleVLN (Wang et al., 2023)	2023	• 1.2K+ scenes • 4.9M trajectories	• N/A (path-based)	• Synthesized language • Egocentric RGB-D videos	• Synthesized trajectory-instruction pairs	Indoors	S	SE	Y	Navigation using automatically generated large-scale synthetic data
GOAT-Bench (Khanna et al., 2024)	2024	• 90 scenes • 9 tasks • 3K+ goal entities	• N/A (open-vocabulary)	• Multi-modal goals • Egocentric RGB-D videos	• Target object locations • Sequential goal sequences	Indoors	S	SE	Y	Lifelong navigation to open-vocabulary goals
HM3D-OVON (Yokoyama et al., 2024)	2024	• 216 scenes • 15K+ object instances	• 379 objects	• Free-form language • Egocentric RGB-D videos	• Open-vocabulary object goals • 3D bounding boxes	Indoors	S	SE	Y	Open-vocabulary object goal navigation, based on free-form natural language

presented in (Pan et al., 2025a), demonstrating how FM-based reasoning can support emergency response in safety-critical settings.

10.8 Maritime robotics

The incorporation of FM intelligence has significantly bolstered the capabilities of autonomous maritime systems, allowing them to perceive, reason, and act more effectively within complex aquatic environments. These models are specifically engineered to navigate typical underwater challenges, such as high turbidity, limited visibility, and severe communication constraints that often degrade traditional robotic sensors. In particular, UnderwaterVLA (Wang et al., 2025g) introduces a dual-brain VLA architecture for autonomous underwater navigation, combining multi-modal reasoning with embodied control for improving robustness under degraded visual and communication conditions. Additionally, MarineInst (Zheng et al., 2024b) and MarineGPT (Zheng et al., 2023) demonstrate FM capabilities in bridging raw marine visual data, semantic understanding, and domain-specific natural-language knowledge, thereby supporting richer perception and reasoning modules for maritime robotic platforms.

10.9 Space robotics

The integration of FMs is critically transforming the field of astro-embodied intelligence, enabling robotic agents to reason, to adapt, and to perceive within unstructured, off-world environments, where human intervention is physically impossible. These models provide strong priors and zero-shot generalization capabilities that are crucial for operating under the extreme conditions and data scarcity typical of planetary missions. A primary application involves the usage of SAM for universal crater detection (Giannakis et al.,

Table 11: Public datasets concerning in principle manipulation tasks. Abbreviations used: Environment type (Simulation (S), Real (R)), Temporality (Static (ST), Video (V), Sequential (SE)), and Embodiment (Yes (Y), No (N)).

Dataset	Year	Scale	Classes	Modalities	Annotation type	Domain	Env. type	Temp.	Emb.	Description
RoboNet (Dasari et al., 2020)	2019	• 15M frames • 162K trajectories • 4 locations	• N/A (pushing)	• RGB videos • Robot actions • Gripper states	• Action trajectories • Video targets	Tabletop	R	SE	Y	Multi-robot dataset on learning visual foresight and video prediction for non-prehensile manipulation
RLBench (James et al., 2020)	2020	• Infinite demonstrations	• 100+ tasks	• RGB-D videos • Segmentation • Proprioception	• Motion-planned trajectories • Target way-points	Tabletop	S	SE	Y	Tasks algorithmically generated using ground-truth state information
CALVIN (Mees et al., 2022)	2022	• 4 environments • 2.4M interaction steps	• 34 tasks	• RGB-D videos • Tactile • Proprioception	• Language goals • Pre-task locomotion	Tabletop	S	SE	Y	Long-horizon language-conditioned continuous control
LIBERO (Liu et al., 2023a)	2023	• 6.5K trajectories	• 130 tasks	• RGB video • Proprioception • Language instructions	• Expert demonstrations • Task completion tags	Tabletop	S	SE	Y	Evaluation of knowledge transfer across sequentially learned task suites
RH20T (Fang et al., 2024)	2023	• 110K trajectories	• 150+ skills	• RGB video • Force • Tactile • Audio	• Action trajectories • Language descriptions	Tabletop	R	SE	Y	Multi-modal dataset including force and audio, targeting contact-rich skills
BridgeData V2 (Walke et al., 2023)	2023	• 60,096 trajectories • 24 environments	• 13 skills	• RGB videos • Proprioceptio	• Goal images • Language instructions	Tabletop	R	SE	Y	Use of low-cost robots across 24 environments to boost generalization
Open X-Embodimer (O'Neill et al., 2024)	2024	• 1M+ episodes • 60 datasets	• 500+ skills	• RGB video • End-effector poses • Language instructions	• Action trajectories	Multi-domain	R,S	SE	Y	Aggregation of 60+ datasets and 22 robotic platforms into a unified format for cross-embodiment scenarios
DROID (Khazatsky et al., 2024)	2024	• 76K trajectories • 564 scenes	• 86 tasks	• RGB-D videos • Proprioception	• Teleoperated actions • Language instructions	In-the-wild	R	SE	Y	In-the-wild dataset collected by 50 people across 52 buildings to maximize scene and lighting diversity
RoboMIND (Wu et al., 2025a)	2024	• 107K trajectories • 96 objects	• 279 tasks	• RGB-D videos • Proprioception • Natural language	• Expert teleoperation • Failure demonstrations • Fine-grained instructions	Indoors	S,R	SE	Y	Unified-standard dataset covering humanoids and dual-arm robots for multi-embodiment intelligence
AgiBot World (Bu et al., 2025)	2025	• 1,001,552 trajectories • 3K+ objects	• 217 tasks	• RGB-D video • Visuo-tactile • Proprioception	• Human-in-the-loop teleoperation actions	Multi-domain	S,R	SE	Y	Large-scale facility-based platform using 100 humanoid robots to collect high-fidelity, bi-manual, long-horizon task data

2024), which utilizes promptable segmentation to identify features across diverse planetary imagery without requiring domain-specific retraining. Beyond basic detection, such FMs are being extended to facilitate autonomous terrain understanding and complex geological analysis (Giannakis et al., 2024; Zhao & Ye, 2024; Holden et al., 2026), allowing robots to make high-stake decisions independently in remote and hazardous space settings.

11 Public datasets

Having systematically investigated the robotic FM literature using different criteria (Sections 4-10), this section outlines the main public datasets that have been introduced so far for developing and evaluating robotic FM methods. In particular, Tables 8-12 group the various benchmarks with respect to the main robotic task concerned, while they also include information about the following aspects for each entry: a) Year, b) Scale, c) Semantic classes, d) Modalities involved, e) Annotation type, f) Domain, g) Environment

Table 12: Public datasets concerning in principle human-robot interaction tasks. Abbreviations used: Environment type (Simulation (S), Real (R)), Temporality (Static (ST), Video (V), Sequential (SE)), and Embodiment (Yes (Y), No (N)).

Dataset	Year	Scale	Classes	Modalities	Annotation type	Domain	Env. type	Temp.	Emb.	Description
CVDN (Thomason et al., 2020)	2020	• 2,050 human dialogues • 7K+ trajectories • 83 scenes	• N/A (goal-driven)	• Natural language dialogues • RGB panoramas	• Dialogue history • Shortest-path actions • Navigation traces	Household	S	SE	Y	Multi-turn human-human dialogues for navigation, where agents ask for help
TEACh (Padmakumar et al., 2022)	2022	• 3,047 dialogues • 39.5K utterances	• 12 tasks	• Natural language dialogues • Egocentric RGB videos • Discrete actions	• Dialogue history • Human demonstrations • Object state changes	Household	S	V,SE	Y	Task-driven agents that communicate to complete complex household tasks
DialFRED (Gao et al., 2022)	2022	• 53K Q&A pairs	• 25 subgoals	• Natural language dialogues • Egocentric RGB videos	• Q&A pairs • Action sequences • Oracle responses	Household	S	V,SE	Y	Active questioning framework where agents ask humans for clarifications to solve household tasks
RoboVQA (Sermanet et al., 2024)	2023	• 829,502 video-text pairs • 29,520 instructions	• N/A (open-ended QA)	• RGB videos • Natural language • Robot actions	• VQA pairs • Multi-embodiment demonstrations	Daily life	R	V	Y	Large-scale reasoning dataset using interleaved vision-text-action for long-horizon robot planning
NatSGD (Sneshesh Shrestha 2025)	2024	• 1,143 commands • 18 participants	• 11 actions • 20 objects	• Speech • Audio • Gestures • Robot actions	• Intent labels • Time-aligned behavior	Cooking	S,R	V,SE	Y	Synchronized speech, gestures, and robot demonstrations for natural human-robot interaction
HA-R2R (Li et al., 2024c)	2024	• 21,567 instructions • 486 motion sequences	• 145 activities	• Natural language • RGB-D & fisheye videos • Human activity data	• Human activity descriptions • Navigation trajectories	Indoors	S,R	V,SE	Y	Human-aware navigation focusing on social constraints and dynamic human interactions

type (Simulation (S), Real (R)), h) Temporality (Static (ST), Video (V), Sequential (SE)), i) Embodiment (Yes (Y), No (N)), and j) Short description. Apart from per task remarks, the following global observations can be made:

- Focus on massive cross-embodiment datasets: In order to address the need for constructing generalist policies that can perform robustly under different kinematic structures and environmental conditions, dataset creation activities have shifted from specialized, single-robot setups to large-scale, multi-robot, heterogeneous ones. For example, Open X-Embodiment (O'Neill et al., 2024) and AgiBot World (Bu et al., 2025) contain over a million trajectories across dozens of robotic platforms.
- Emphasis on vision-language-action capturings: Following the research trend of developing embodied VLA solutions, data collection procedures increasingly support sensorial types that aim at integrating perception, reasoning, and control actions. For example, BridgeData V2 (Walke et al., 2023) and CALVIN (Mees et al., 2022) include synchronized information streams of RGB-D video, natural language instructions, and low-level action tokens.
- Translation to high-fidelity, real-world capturing settings: In order to robustly address the sim-to-real gap in the FM era, intense efforts have been devoted on creating massive, real-world benchmarks, instead of simulation ones. In this respect, datasets like Ego4D (Grauman et al., 2022) and DROID (Khazatsky et al., 2024) support geographic and environmental diversity at unprecedented scale.
- Incorporation of human reasoning aspects: In order to enable FMs to develop common-sense, human-like reasoning and interaction capabilities, such aspects are increasingly incorporated in the dataset formation procedures. For instance, RoboVQA (Sermanet et al., 2024) and TEACh (Padmakumar et al., 2022) include hundreds of thousands of video-text pairs and clarification-oriented Q&A dialogues.

- Lack of tactile sensing data: Despite the high availability of visual and textual data, tactile (touch) and force-sensing information streams remain critically underrepresented in the current benchmarks. Incorporation of the aforementioned modalities is essential for developing efficient, high-precision dexterous manipulation.

- Lack of failure and recovery data: A common observation across all datasets comprises the typical lack of recordings corresponding to rare failure modes and successful recovery actions. The latter is also essential during the training phase for developing models that are likely to be robust in real-world execution settings.

12 Current challenges

Despite the large body of works that have recently been introduced in the field of robotic FM-based methods and the tremendous advancements accomplished, significant challenges and open research problems still remain, which if robustly addressed will further increase the efficiency, reliability, acceptance, and adoption of such solutions in real-world deployment scenarios. In the remaining of this section, the main challenges identified in the literature are systematically examined and outlined.

12.1 Data aspects

Unlike benchmark requirements and availability in fields like NLP and CV, robotic data is physically-grounded, high-dimensional, multi-modal, and typically expensive to collect. In this respect, the following specific challenges related to data aspects in robotic FM research are present:

- Scarcity of physical-world data: Despite the availability of internet-scale visual/text benchmarks, robotic solution development lacks comparable datasets of high-quality, diverse, real-world robot trajectories (Brohan et al., 2023a). The main factor for the latter comprises the data collection cost, which requires physical hardware, human teleoperation/demonstration/instruction, and significant time. Additionally, the added difficulty of collecting 'long-tail scenarios' (i.e., rare and unexpected events with huge impact on the system performance) makes the learning of safe recovery behaviors even more challenging (Park et al., 2025).

- Embodiment heterogeneity: Robotic data is generally not uniform, since it is produced by different robots with diverse hardware and physical configurations (e.g., degrees of freedom, sensor suites, kinematics, etc.) (O'Neill et al., 2024). Additionally, there is no single format for continuous action data across diverse embodiments, without losing physical meaning/properties (Reed et al., 2022). The above result into an inherent difficulty to transfer knowledge between different robotic platforms (Liu et al., 2025a).

- Maintaining high data quality: Increasing the data scale usually leads to improved performance, conditioned on the fact of maintaining a high-quality in the captured data (Zitkovich et al., 2023). However, human-captured teleoperation data often includes suboptimal movements, hesitation, or outright failures (Hu et al., 2023). Additionally, manual inspection of immense amounts of expert demonstration videos is tremendously expensive.

- Domain transfer gap: Typical approaches to overcome the data scarcity problem is to make use of simulation or data from another (similar) domain. However, simulation engines often fail to accurately reproduce complex physics (e.g., friction, object deformations, etc.), leading to low performance in real-world settings (Xiao et al., 2024). On the other hand, incorporation of data from multiple, diverse environments is shown to have a more crucial impact on training, than simply scaling the available datasets (O'Neill et al., 2024; Zitkovich et al., 2023).

- Multi-modal and temporal alignment: Robots need to efficiently integrate vision, language, and proprioception streams to accomplish robust performance in real-world settings. However, the high and different frequency of the various information sequences, along with the inherent difficulty in

mapping continuous physical-world actions to discrete information tokens, results into precision degradation (e.g., in dexterous tasks) (Zhou et al., 2025e).

12.2 Computation aspects

In any real-world robotic application case, time performance and resource requirements constitute a cornerstone regarding fundamental safety and stability specifications. In this respect, the following specific challenges related to computation aspects in robotic FM research are present:

- Need for real-time inference: Robot control loops typically operate at high frequencies (e.g., often more than 50Hz), in order to achieve stability and safe execution (Chi et al., 2025). However, the inherently extreme scale of FMs naturally introduces significant inference latencies, ranging from some hundreds of milliseconds up to multiple seconds (Firoozi et al., 2025; Zitkovich et al., 2023). This is highly likely to result into FM-based solutions not to be applicable in multiple cases or to lead to inaccurate policy executions (Ameperosa et al., 2025).

- Need for safety bound interval: Apart from the inference time itself, safe robot operation requires the application of certain safety routines or the execution of corrective actions (Sinha et al., 2024). The latter impose additional time constraints during execution, i.e. a strict upper bound for overall end-to-end latency (Firoozi et al., 2025; Sinha et al., 2024).

- Constrained onboard resources: Robotic platforms typically pose specific and strict size, weight, and energy specifications, setting particular limitations to the onboard embedded GPUs and their operation. Contrary to FM execution on cloud, resource-abundant environments, edge devices exhibit limitations in terms of available memory, processing time, and energy/thermal tolerance, prohibiting the deployment of full-scale and best-performing versions of robotic FMs (O'Neill et al., 2024; Yue et al., 2024).

12.3 Safety and security aspects

The fundamental advantage of FMs in robotic applications relies on their ability to combine high-level semantic reasoning with low-level, physical planning/execution procedures. The latter though raise particular safety and security concerns, which extend beyond conventional/traditional robotic settings. In this respect, the following specific challenges related to safety and security aspects in robotic FM research are present:

- Semantic-physical space mismatch: It is very common for FMs to generate policies that are linguistically, logically, and semantically sound, but can likely lead to dangerous or even physically impossible scenarios (with the extreme case being that of the presence of hallucinations) (Lin et al., 2025b; Yin et al., 2025a). This is mainly due to the typical limitation of most FMs to account for physical-world parameters (e.g., friction, torque limits, material properties, etc.) in their reasoning process (Firoozi et al., 2025). To make matters worse, FMs typically lack the ability to estimate the consequences of their (erroneous) actions in the physical environment (Zitkovich et al., 2023; Lin et al., 2025b).

- Adversarial vulnerabilities: As the number of modality streams and data scale that a single FM can process rises, the respective (cyber) attack surface of the (often cloud-hosted) model itself increases proportionally (Radanliev et al., 2026). In particular, even relatively minor perturbations in the data can result into significant behavior deviations or incorrect/hazardous actions (Wang et al., 2025d).

- Inaccurate uncertainty quantification: Especially in human-robot collaboration settings, the robotic agent needs to constantly maintain a precise estimate of its own state/policy uncertainty (Wang et al., 2025e). Current FMs though do not reliably balance their reactions with respect to aleatoric (environmental noise) and epistemic (model ignorance) uncertainty (Marques & Berenson, 2024). This bottleneck becomes even more evident in onboard deployment settings, where increased inference latency or compressed models are used (due to resource constraints) (Zitkovich et al., 2023).

12.4 Embodiment aspects

FMs have equipped robots with unprecedented perception, reasoning, and execution capabilities under real-world, dynamic environments. However, the critical issue that arises comprises that of grounding robot tasks on a physical platform, which poses specific sensor, actuator, and hardware constraints. In this respect, the following specific challenges related to embodiment aspects in robotic FM research are present:

- Heterogeneity of robot action spaces: Unlike the case of other AI deployment scenarios, robots exhibit a vast variety in terms of physical forms, morphologies, and capabilities. The latter renders difficult to deploy a model trained on a specific hardware setup (e.g., dual-arm, mobile platform) to another (e.g., quadruped, drone, etc.), due to difference in key robotic specifications (e.g., varying degrees-of-freedom) (O'Neill et al., 2024; Mees et al., 2024). Additionally, the absence of a universally applicable, standardized control interface across different robots, makes the development of robust, generalized, platform-independent FMs particularly difficult (Zheng et al., 2025b).

- Sim-to-real gap: In an attempt to alleviate from the need for large-scale, high-quality robotic interaction data, simulation engines are often used for data generation. However, employing simulation settings does not always result into the assembly of benchmarks with sufficient diversity in both task execution and environment settings (Jonnarth et al., 2025). Moreover, simulation engines are typically prone to not model accurately complex physical-world dynamics (e.g., friction, deformation, fluid interactions, etc.), which in turn results into reduced robot performance (Makoviychuk et al., 2021; Ai et al., 2025).

- Limited physical space grounding: Despite the unprecedented semantic capabilities of FMs, the mapping of such high-level, semantic representations to low-level, real-world physics is not guaranteed. In particular, current FMs exhibit limitations in precise spatial, geometric, and physical interactions (Qi et al., 2025). Additionally, the lack of sufficient contextual, common-sense knowledge in FMs, may result into the generation of failure modes in the real-world (Chen et al., 2024a; Huang et al., 2023c).

- Constrained haptic capabilities: The so-called 'final frontier' problem of physical intelligence comprises the difficulty of robots to replicate nuanced sensorial feedback, like humans do in the real world (Yang et al., 2024a). The latter constitutes inherently a multi-faceted problem, including perspectives related to the scarcity of large-scale haptic/tactile data, robot hardware heterogeneity, sim-to-real variance, and high-frequency temporal reaction demands (Zhang et al., 2025c).

12.5 Reasoning aspects

While inference capabilities of FMs in other AI domains (e.g., CV, NLP, etc.) has successfully achieved the incorporation of rich, abstract logic in most cases, their application to physically-grounded, robot agents exhibits significant obstacles. In this respect, the following specific challenges related to reasoning aspects in robotic FM research are present:

- Lack of physical common sense knowledge: The usual embodiment-agnostic understanding of the world by FMs (e.g., gravity, object permanence, material properties, etc.) makes their real-world deployment difficult (Firoozi et al., 2025; Kawaharazuka et al., 2024). Additionally, the lack of sufficient causal reasoning capabilities hinders robots to predict the physical consequences of their actions (Töberg et al., 2024).

- Constrained long-horizon planning: The reasoning performance of robotic FMs tends to degrade exponentially, as the number of required steps increases (Driess et al., 2023). The latter has a great impact, especially in cases of large-scale operational environments (Lisondra et al., 2026). Moreover, the training process itself poses difficulties to FMs to connect long-term goals to specific sequences of decision steps (Brohan et al., 2023b).

- Imprecise explanations of robot behaviors: FMs in robotics comprise by nature massive architectures that integrate multi-modal information streams, while connecting high-level reasoning with low-level, physical control (Brohan et al., 2023a). In contrast to traditional modular robotic solutions, where failure modes can be traced to specific components, FMs implement end-to-end generalist policies that make it particularly difficult to specify the factors that have led to the execution of a specific physical action (Kawaharazuka et al., 2024).

12.6 Evaluation aspects

Provided that FMs bridge the entire gap between high-level, semantic reasoning and low-level, physical execution in dynamic environments, the assessment of their performance and robustness exhibits particular characteristics, compared to other conventional AI solutions. In this respect, the following specific challenges related to evaluation aspects in robotic FM research are present:

- Lack of a unified evaluation framework: Despite the availability of a series of standardized, task-oriented, intuitive performance metrics, no holistic and integrated evaluation protocol is currently present to assess the multi-factored robotic failure cases (Firoozi et al., 2025). In particular, the available (and typically binary) metrics are often proven to be coarse, failing to clearly highlight the underlying factors leading to low performance (e.g., inefficient bi-manual coordination, asymmetric arm usage, etc.) (Jiang et al., 2023; O'Neill et al., 2024). Additionally, the development of independent, domain-specific metrics across different fields makes even more difficult to assess robot performance with respect to different architectures or motion parameters (Brohan et al., 2023a).

- Imprecise generalization ability assessment: Despite the key driving force of FMs to exhibit increased (especially zero-shot) generalization ability, its comprehensive, accurate, and robust assessment is especially difficult (Zitkovich et al., 2023; Liang et al., 2023). In particular, the introduction of even minimal distribution shifts in task descriptions or observation domains can have a great impact on the resulting robot behavior (or even failure) (Kube et al., 2026).

13 Future research directions

The introduction of FMs in the field of robotics has led to unprecedented accomplishments and advances in all core robotic technologies, as discussed in Sections 3-10. Despite these tremendous developments, several open challenges are still present, which pose restrictions in the wider deployment of robotic solutions in real-world scenarios, as outlined in Section 12. In this respect, this section discusses the main and most promising future research directions towards achieving the goal of developing efficient, robust, and general-purpose robotic agents, in correspondence to the challenges described in Section 12.

13.1 Architectural evolution

VLAs comprise one of the most promising types of NN architectures, which implement a unified neural function for mapping visual observations, linguistic instructions, and proprioceptive states directly to low-level, robot control policies. Their core characteristic is their end-to-end integration nature, which enables robotic agents to perceive/reason about their surrounding environment and, subsequently, to react upon the input stimuli in a single forward/inference pass. In this respect, the following promising research directions emerge:

- Heterogeneous action spaces: An ambitious goal of current research efforts focuses on the development of universal robotic FMs, capable of operating across different robotic platforms. This requires the robust addressing of the action heterogeneity (or cross-embodiment generalization, stated alternatively) problem (O'Neill et al., 2024). In this context, future research could emphasize on the developed of embodiment-agnostic action spaces, which would enable FMs to predict desired end-effector trajectories/forces, while their eventual mapping to particular, platform-specific, low-level operations could be controlled by a hardware-specific modulation component (Mees et al., 2024).

- Sophisticated action sequence tokenization: One of the main challenges in the design of VLA transformer architectures comprises the discrete, tokenized representation of continuous space robotic actions. The conventional approach of binning each dimension to a fixed number of values inherently leads to loss of precision, especially for dexterous tasks (Zitkovich et al., 2023; Reed et al., 2022). In this context, future research could emphasize on more sophisticated tokenization methods capable of capturing the detailed dynamics of continuous actions, while maintaining the efficiency of autoregressive decoding.

- Diffusion and flow-based action modeling: When a robot attempts to perform an action, there is a set of practically infinite trajectories to select from. Common architectures (e.g. transformers) tend to learn the average of such possible motions, leading to performance degradation or failure (Florence et al., 2022). On the contrary, flow- and diffusion-based objectives facilitate the modeling of temporal dynamics in a continuous latent space. In this context, future research could emphasize on the latter aspects, targeting to equip the robots with the ability to predict the consequences of their actions in a continuous space representation (Chi et al., 2025).

13.2 Multi-modal embodied intelligence

The main driving-force of current FMs lies on the incorporation of vision and language information, while true embodied intelligence requires a holistic understanding of the physical world through the integration of touch, sound, and force sensing. The latter is essential especially for tasks requiring dexterity and delicate interaction. In this respect, the following promising research directions emerge:

- Tactile information: Vision is often insufficient for performing contact-rich tasks and reaching human-level dexterity capability, which relies on incorporating feedback regarding the surface texture, slip, and force (Yu et al., 2024). For achieving the latter, the robust integration of tactile information is essential, capitalizing on the current/early-works on developing tactile FMs. In this context, future research could emphasize on the latter aspects, aiming at enhancing robots to efficiently manipulate deformable objects (Wu et al., 2020).

- Proprioception and force control: Beyond tactile sensing, the incorporation of proprioception (i.e., the robot's sense of its own body state) and force control is crucial for realizing contact-rich manipulaton tasks. Current solutions usually focus only on position control, while exploratory works, like the integration of joint torque sensors, emerges as a promising approach (Kumar et al., 2021). In this context, future research could emphasize on further exploiting dense proprioception and force technologies, in order to boost more fine-grained manipulation and efficient human-robot collaboration (Liu et al., 2023b).

- Auditory feedback: The sound signal can often bear significant information regarding the occurrence of critical events during robot execution, complementing the corresponding visual and force feedback streams (Dimiccoli et al., 2022). Additionally, audio processing requires significantly less resources than the respective visual stream. In this context, future research could emphasize on further investigating the exploitation of auditory feedback, enabling robots to operate more accurately and reliably in dynamic and noisy environments (Mejia et al., 2024).

13.3 Reasoning and long-horizon autonomy

Current VLA-based solutions have demonstrated reliable performance on immediate reactive tasks. However, the implementation of multiple real-world activities requires the robust handling of complex, long-horizon, multi-stage missions. In this respect, the following promising research directions emerge:

- Long-horizon memory frameworks: Due to constraints in the memory capacity of current FMs, robots often tend to repeat unsuccessful policies, due to the relatively limited context window that they maintain and which might not include information about the previous policy failure (Firoozi et al., 2025). To this end, extending the memory capabilities of robotic solutions would have a

great impact on long-horizon goal achievement. In this context, future research could emphasize on enhancing long-horizon autonomy, by incorporating structured summaries of past interactions (instead of full-sequence replays) (Wu et al., 2026), latent-graph memory formalisms to maintain experiences over longer contexts (Chen et al., 2026b), and stage-aware reward signals for more efficient policy learning in multi-step setups (Chen et al., 2026a).

- Hierarchical semantic representations: In case of robot operation in cluttered environments, the usage of hierarchical abstractions (e.g., scene graphs) of the processed semantic information is proven to be efficient in simplifying and strengthening decision-making (Gu et al., 2024). In this context, future research could emphasize on realizing task planning and reasoning on such hierarchical abstractions, so as to both reduce computational cost (i.e., avoid intense computations at the pixel or joint angle level) and to achieve more robust decision making (Gao et al., 2025a; Hughes et al., 2022).

13.4 World foundation models

The deployment of FMs in robotic applications requires high-quality, diverse, and large-scale embodied interaction data. Given the fact that the collection of the latter datasets is typically expensive and time-consuming, world foundation models are being investigated as an alternative solution for trajectory generation. In this respect, the following promising research directions emerge:

- Physics-informed generative models: WMs have already been proven to generate photorealistic data of high quality. However, there still exists a gap between simulated physics and real-world dynamics, which would enable robots to learn and to act more reliably in actual operational settings (Ai et al., 2025; Lee et al., 2025a). In this context, future research could emphasize on explicitly incorporating constraints like gravity, friction, and fluid dynamics into the video generation process (Xie et al., 2024). Another closely related aspect concerns the integration of sophisticated physics-grounded predictive pipelines, which would serve as value functions for model-based planning, especially in long-horizon tasks (Zhou et al., 2024a; Yang et al., 2023).

- Action-conditioned scenario generation: Research in world models regarding action-conditioned scenario generation has advanced from plain video prediction towards the creation of unified models, which function as physically grounded, interactive cognitive engines (Bruce et al., 2024). In particular, significant efforts are devoted on integrating differentiable physics and unified geometric representations for ensuring the spatio-temporal consistency and accuracy of the generated predictions over long horizons (Lee et al., 2025a; Tu et al., 2025). In this context, future research could emphasize on further extending current capabilities, especially focusing on cross-embodiment generalization and 'long-tail' scenarios, so as to boost the translation of human-centric video data into physically plausible robotic trajectories across multiple hardware platforms (O'Neill et al., 2024; Mees et al., 2024).

13.5 Safety and verification

Given the increasing deployment of robots in dynamic, unstructured environments, their interaction and collaboration with humans requires the robust addressing of safety risks. For achieving this, FM-based solutions need to incorporate a combined approach, consisting of both high-level semantic understanding and low-level, deterministic safety aspects. In this respect, the following promising research directions emerge:

- Adaptive safety: The incorporation of adaptive safety measures in robotic FMs is gradually shifting from a reactive post-processing approach towards an integrated one, where constraints are directly projected to the model's training and reasoning loops (Ren et al., 2023). This enables robots to predict and to mitigate physical risks through grounded reasoning, prior to action execution (Zha et al., 2024). In this context, future research could emphasize on further enhancing safety alignment approaches through constrained learning and process reward models for evaluating individual reasoning steps and environmental affordances in real-time (Yu et al., 2025a; Anand et al., 2026).

- Formal verification: Research in formal verification of robotic FMs is currently translating from conventional offline proofs towards modular, runtime-based assurance frameworks. In particular, current efforts largely focus on control barrier functions and reachability analysis, in order to intercept model outputs in real-time (Miyaoka & Inoue, 2025; Wang & Wen, 2025). In this context, future research could emphasize on investigating neuro-symbolic integration approaches for mapping neural outputs to formal logics (Cunnington et al., 2024) and automated specification mining pipelines that employ LLMs for translating natural language instructions into precise mathematical formulas (Rabiei et al., 2025).

14 Conclusion

The introduction of Foundation Models (FMs) has resulted into transformative effects in the field of robotics, transitioning the current practice from rigid, single-task solutions towards adaptive, multi-sensory, and generalist agents, capable of operating in complex, dynamic, open-world environments. The current review has provided a holistic, thorough, systematic, and in-depth analysis of the research landscape by delineating five distinct evolution phases, starting from early Natural Language Processing (NLP) and Computer Vision (CV) model integration to the current frontier of multi-sensory generalization and real-world deployment. Through a highly-granular, multi-criteria, taxonomic literature investigation, this work has analyzed the interplay between different foundation model types (LLMs, VFMs, VLMs, and VLAs), underlying neural network architectures, adopted learning paradigms, learning stages for skill acquisition, robotic tasks (perception, planning, navigation, manipulation, and human-robot interaction), and real-world application domains. The above discussion, has been accompanied by a methodical comparative analysis of the various categories of approaches and critical insights per defined criterion. Moreover, a comprehensive, task-oriented report on the publicly available datasets required for model training and evaluation was provided, while a detailed and hierarchical discussion on the current open challenges and promising future research directions in the field was incorporated.

Overall, the current analysis reveals that while internet-scale pre-training provides unprecedented reasoning priors, significant bottlenecks still remain, including, among others, the scarcity of high-quality physical-world data, the inherent 'embodiment gap', and the computational latency that hinders real-time control. For addressing these, future research needs to prioritize the development of universal, hardware-agnostic architectures and the integration of nuanced sensorial modalities, such as tactile and auditory feedback. Additionally, the advancement of physics-informed world models and modular, runtime-based formal verification techniques are essential for ensuring that these massive architectures remain safe and predictable in high-stake human-robot interaction environments. Robust addressing of the latter would facilitate the bridging of the gap between high-level, semantic seasoning and low-level, physical grounding, boosting in this way the development of more autonomous and reliable robotic systems.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ameya Agaskar, Sriram Siva, William Pickering, Kyle O'Brien, Charles Kekeh, Ang Li, Brianna Gallo Sarker, Alicia Chua, Mayur Nemade, Charun Thattai, et al. Deepfleet: Multi-agent foundation models for mobile robots. *arXiv preprint arXiv:2508.08574*, 2025.

Faseeh Ahmad, Hashim Ismail, Jonathan Styrud, Maj Stenmark, and Volker Krueger. A unified framework for real-time failure handling in robotics using vision-language models, reactive planner and behavior trees. In *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, pp. 887–894. IEEE, 2025.

Michael Ahn, Debidatta Dwibedi, Chelsea Finn, Montserrat Gonzalez Arenas, Keerthana Gopalakrishnan, Karol Hausman, brian ichter, Alex Irpan, Nikhil J Joshi, Ryan Julian, Sean Kirmani, Isabel Leal, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, sharath maddineni, Kanishka Rao, Dorsa Sadigh, Pannag R

Sanketi, Pierre Sermanet, Quan Vuong, Stefan Welker, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, and Zhuo Xu. AutoRT: Embodied foundation models for large scale orchestration of robotic agents. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

Bo Ai, Stephen Tian, Haochen Shi, Yixuan Wang, Tobias Pfaff, Cheston Tan, Henrik I Christensen, Hao Su, Jiajun Wu, and Yunzhu Li. A review of learning-based dynamics models for robotic manipulation. *Science Robotics*, 10(106):eadt1497, 2025.

Anurag Ajay, Seungwook Han, Yilun Du, Shuang Li, Abhi Gupta, Tommi Jaakkola, Josh Tenenbaum, Leslie Kaelbling, Akash Srivastava, and Pulkit Agrawal. Compositional foundation models for hierarchical planning. *Advances in Neural Information Processing Systems*, 36:22304–22325, 2023.

Ezra Ameperosa, Jeremy A Collins, Mrinal Jain, and Animesh Garg. Rocoda: Counterfactual data augmentation for data-efficient robot learning from demonstrations. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13250–13256. IEEE, 2025.

Yashwanthi Anand, Nnamdi Nwagwu, Kevin Sabbe, Naomi T Fitter, and Sandhya Saisubramanian. Adaptive querying for reward learning from human feedback. *Frontiers in Robotics and AI*, 12:1734564, 2026.

Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2838–2845. IEEE, 2025.

Jicong Ao, Fan Wu, Yansong Wu, Abdalla Swiki, and Sami Haddadin. Llm-as-bt-planner: Leveraging llms for behavior tree generation in robot task planning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1233–1239. IEEE, 2025.

Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Maulana Bisyir Azhari and David Hyunchul Shim. Dino-vo: A feature-based visual odometry leveraging a visual foundation model. *IEEE Robotics and Automation Letters*, 2025.

Leonard Bärmann, Rainer Kartmann, Fabian Peller-Konrad, Jan Niehues, Alex Waibel, and Tamim Asfour. Incremental learning of humanoid robot behavior from natural interaction and large language models. *Frontiers in Robotics and AI*, 11:1455375, 2024.

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9297–9307, 2019.

Aude Billard, Alin Albu-Schaeffer, Michael Beetz, Wolfram Burgard, Peter Corke, Matei Ciocarlie, Ravinder Dahiya, Danica Kragic, Ken Goldberg, Yukie Nagai, et al. A roadmap for ai in robotics. *Nature Machine Intelligence*, 7(6):818–824, 2025.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Rich Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, brian ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *9th Annual Conference on Robot Learning*, 2025.

Jan Blumenkamp, Steven Morad, Jennifer Gielis, and Amanda Prorok. Covis-net: A cooperative visual spatial foundation model for multi-robot applications. In *Conference on Robot Learning*, pp. 3780–3808. PMLR, 2025.

Rogerio Bonatti, Sai Vemprala, Shuang Ma, Felipe Frujeri, Shuhang Chen, and Ashish Kapoor. Pact: Perception-action causal transformer for autoregressive robotics pre-training. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3621–3627. IEEE, 2023.

Meghan Booker, Grayson Byrd, Bethany Kemp, Aurora Schmidt, and Corban Rivera. Embodiedrag: Dynamic 3d scene graph retrieval for efficient and scalable robot task planning. *arXiv preprint arXiv:2410.23968*, 2024.

Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Manon Devin, Alex X. Lee, Maria Bauza Villalonga, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, Antoine Laurens, Claudio Fantacci, Valentin Dalibard, Martina Zambelli, Murilo Fernandes Martins, Rugile Pevceviciute, Michiel Blokzijl, Misha Denil, Nathan Batchelor, Thomas Lampe, Emilio Parisotto, Konrad Zolna, Scott Reed, Sergio Gómez Colmenarejo, Jonathan Scholz, Abbas Abdolmaleki, Oliver Groth, Jean-Baptiste Regli, Oleg Sushkov, Thomas Rothörl, Jose Enrique Chen, Yusuf Aytar, David Barker, Joy Ortiz, Martin Riedmiller, Jost Tobias Springenberg, Raia Hadsell, Francesco Nori, and Nicolas Heess. Robocat: A self-improving generalist agent for robotic manipulation. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *Robotics: Science and Systems XIX*, 2023a.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pp. 287–318. PMLR, 2023b.

Matthew Bronars, Shuo Cheng, and Danfei Xu. Legibility diffuser: Offline imitation for intent expressive motion. *IEEE Robotics and Automation Letters*, 2024.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 1877–1901, 2020.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Xindong He, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025.

Finn Lukas Busch, Timon Homberger, Jesús Ortega-Peimbert, Quantao Yang, and Olov Andersson. One map to find them all: Real-time open-vocabulary mapping for zero-shot multi-object navigation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14835–14842. IEEE, 2025.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Junhao Cai, Yisheng He, Weihao Yuan, Siyu Zhu, Zilong Dong, Liefeng Bo, and Qifeng Chen. Ov9d: Open-vocabulary category-level 9d object pose and size estimation. *arXiv preprint arXiv:2403.12396*, 2024a.

Shaofei Cai, Bowei Zhang, Zihao Wang, Xiaojian Ma, Anji Liu, and Yitao Liang. GROOT: Learning to follow instructions by watching gameplay videos. In *The Twelfth International Conference on Learning Representations*, 2024b.

Wenzhe Cai, Siyuan Huang, Guangran Cheng, Yuxing Long, Peng Gao, Changyin Sun, and Hao Dong. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5228–5234. IEEE, 2024c.

Marco Cancelli, Federico Stefani, Brittany Tinker, Tsogbadrakh Erdenebayar, Minae Kwon, Maya Cakmak, Anca D. Dragan, and Dorsa Sadigh. See something, say something: Context-criticality-aware mobile robot communication for hazard mitigations. *arXiv preprint arXiv:2603.28901*, 2026.

Thanh Nguyen Canh, Thang Tran Viet, Thanh Tuan Tran, and Ben Wei Lim. Safeguard asf: Sr agentic humanoid robot system for autonomous industrial safety. *arXiv preprint arXiv:2603.25353*, 2026.

Jiahang Cao, Qiang Zhang, Jingkai Sun, Jiaxu Wang, Hao Cheng, Yulin Li, Jun Ma, Kun Wu, Zhiyuan Xu, Yecheng Shao, et al. Mamba policy: Towards efficient 3d diffusion policy with hybrid selective state models. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11359–11366. IEEE, 2025a.

Yuhong Cao, Jeric Lew, Jingsong Liang, Jin Cheng, and Guillaume Sartoretti. Dare: Diffusion policy for autonomous robot exploration. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11987–11993. IEEE, 2025b.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026.

Joao Carvalho, An T Le, Mark Baierl, Dorothea Koert, and Jan Peters. Motion planning diffusion: Learning and planning of robot motions with diffusion models. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1916–1923. IEEE, 2023.

Mateo Guaman Castro, Sidharth Rajagopal, Daniel Gorbatov, Matt Schmittle, Rohan Baijal, Octi Zhang, Rosario Scalise, Sidharth Talia, Emma Romig, Celso de Melo, et al. Vamos: A hierarchical vision-language-action model for capability-modulated and steerable navigation. *arXiv preprint arXiv:2510.20818*, 2025.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pp. 667–676. IEEE, 2017.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a.

Qianzhong Chen, Justin Yu, Mac Schwager, Pieter Abbeel, Fred Shentu, and Philipp Wu. SARM: Stage-aware reward modeling for long horizon robot manipulation. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. On scaling up a multilingual vision and language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14432–14444, 2024b.

Yongchao Chen, Jacob Arkin, Charles Dawson, Yang Zhang, Nicholas Roy, and Chuchu Fan. Autotamp: Autoregressive task and motion planning with llms as translators and checkers. In *2024 IEEE International conference on robotics and automation (ICRA)*, pp. 6695–6702. IEEE, 2024c.

Yuxuan Chen, Yixin Han, and Xiao Li. Fastnav: Fine-tuned adaptive small-language-models trained for multi-point robot navigation. *IEEE Robotics and Automation Letters*, 2024d.

Zhisheng Chen, Tingyu Wu, Zijie Zhou, Zhengwei Xie, Ziyan Weng, and Yingwei Zhang. Polarmem: A training-free polarized latent graph memory for verifiable multimodal agents. *arXiv preprint arXiv:2602.00415*, 2026b.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. NaVILA: Legged Robot Vision-Language-Action Model for Navigation. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.018.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. BabyAI: First steps towards grounded language learning with a human in the loop. In *International Conference on Learning Representations*, 2019.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Jaime Corsetti, Davide Boscaini, Changjae Oh, Andrea Cavallaro, and Fabio Poiesi. Open-vocabulary object 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18071–18080, 2024.

Jose Cuaran, Kendall Koe, Aditya Potnis, Naveen Kumar Uppalapati, and Girish Chowdhary. Visual-language-guided task planning for horticultural robots. *arXiv preprint arXiv:2601.11906*, 2026.

Beilei Cui, Mobarakol Islam, Long Bai, and Hongliang Ren. Surgical-dino: adapter learning of foundation models for depth estimation in endoscopic surgery. *International Journal of Computer Assisted Radiology and Surgery*, 19(6):1013–1020, 2024.

Daniel Cunnington, Mark Law, Jorge Lobo, and Alessandra Russo. The role of foundation models in neuro-symbolic learning and reasoning. In *International Conference on Neural-Symbolic Learning and Reasoning*, pp. 84–100. Springer, 2024.

Longchao Da, Justin Turnau, Thirulogasankar Pranav Kutralingam, Alvaro Velasquez, Paulo Shakarian, and Hua Wei. A survey of sim-to-real methods in rl: Progress, prospects and challenges with foundation models. *arXiv preprint arXiv:2502.13187*, 2025.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, pp. 10041–10071. PMLR, 2024.

Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. In *Conference on Robot Learning*, pp. 885–897. PMLR, 2020.

Sudeep Dasari, Oier Mees, Sebastian Zhao, Mohan Kumar Srirama, and Sergey Levine. The ingredients for robotic diffusion transformers. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15617–15625. IEEE, 2025.

Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Heming Cui, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. In *Conference on Robot Learning*, pp. 1004–1029. PMLR, 2025a.

Yinan Deng, Bicheng Yao, Yihang Tang, Tianxing Zhou, Yi Yang, and Yufeng Yue. Openvox: Real-time instance-level open-vocabulary probabilistic voxel representation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1305–1311. IEEE, 2025b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Norman Di Palo and Edward Johns. Dinobot: Robot manipulation via retrieval and alignment with vision foundation models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2798–2805. IEEE, 2024.

Mariella Dimiccoli, Shubhan Patni, Matej Hoffmann, and Francesc Moreno-Noguer. Recognizing object surface material from impact sounds for robot manipulation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9280–9287. IEEE, 2022.

Pengxiang Ding, Jianfei Ma, Xinyang Tong, Binghong Zou, Xinxin Luo, Yiguo Fan, Ting Wang, Hongchao Lu, Panzhong Mo, Jinxin Liu, et al. Humanoid-vla: Towards universal humanoid control with visual integration. *arXiv preprint arXiv:2502.14795*, 2025.

Zibin Dong, Jianye Hao, Yifu Yuan, Fei Ni, Yitian Wang, Pengyi Li, and Yan Zheng. Diffuserlite: Towards real-time diffusion planning. *Advances in Neural Information Processing Systems*, 37:122556–122583, 2024.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 8469–8488, 2023.

Hao Fang, Runmin Cong, Xiankai Lu, Xiaofei Zhou, Sam Kwong, and Wei Zhang. Decoupled motion expression video segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13821–13831, 2025a.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 653–660. IEEE, 2024.

Yu Fang, Yue Yang, Xinghao Zhu, Kaiyuan Zheng, Gedas Bertasius, Daniel Szafir, and Mingyu Ding. Rebot: Scaling robot learning with real-to-sim-to-real robotic video synthesis. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11351–11358. IEEE, 2025b.

Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. In *ICLR*, 2025.

Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, 44(5):701–739, 2025.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on robot learning*, pp. 158–168. PMLR, 2022.

Max Fu, Huang Huang, Gaurav Datta, Lawrence Yunliang Chen, Will Panitch, Fangchen Liu, Hui Li, and Ken Goldberg. Icrt: In-context imitation learning via next-token prediction. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5937–5944. IEEE, 2025.

Chongkai Gao, Zixuan Liu, Zhenghao Chi, Junshan Huang, Xin Fei, Yiwen Hou, Yuxuan Zhang, Yudi Lin, Zhirui Fang, and Lin Shao. VLA-OS: Structuring and dissecting planning representations and paradigms in vision-language-action models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions. In *Forty-second International Conference on Machine Learning*, 2025b.

Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4): 10049–10056, 2022.

Seyed Kamyar Seyed Ghasemipour, Ayzaan Wahid, Jonathan Tompson, Pannag R Sanketi, and Igor Mordatch. Self-improving embodied foundation models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Iraklis Giannakis, Anshuman Bhardwaj, Lydia Sam, and Georgios Leontidis. A flexible deep learning crater detection scheme using segment anything model (sam). *Icarus*, 408:115797, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Danil S Grigorev, Alexey K Kovalev, and Aleksandr I Panov. Verifyllm: Llm-based pre-execution task plan verification for robots. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 18489–18496. IEEE, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028. IEEE, 2024.

Qiuyi Gu, Zhaocheng Ye, Jincheng Yu, Jiahao Tang, Tinghao Yi, Yuhan Dong, Jian Wang, Jinqiang Cui, Xinlei Chen, and Yu Wang. Mr-cographs: Communication-efficient multi-robot open-vocabulary mapping system via 3d scene graphs. *IEEE Robotics and Automation Letters*, 2025.

Pinxue Guo, Hao Huang, Peiyang He, Xuefeng Liu, Tianjun Xiao, and Wenqiang Zhang. Openvis: Open-vocabulary video instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3275–3283, 2025a.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. In *International Conference on Learning Representations (ICLR)*, 2026.

Yuliang Guo, Sparsh Garg, S Mahdi H Miangoleh, Xinyu Huang, and Liu Ren. Depth any camera: Zero-shot metric depth estimation from any camera. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26996–27006, 2025b.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Peng Hao, Chaofan Zhang, Dingzhe Li, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. Tla: tactile-language-action model for contact-rich manipulation. *Robot Learning*, 3(1):17–18, 2026.

Yu Hao, Fan Yang, and Nicholas Fang. Cora: A chain of robotic actions reasoning model for autonomous robotic arm manipulation. In *2025 11th International Conference on Automation, Robotics, and Applications (ICARA)*, pp. 165–169, 2025. doi: 10.1109/ICARA64554.2025.10977668.

Bear Häon, Kaylene Caswell Stocking, Ian Chuang, and Claire Tomlin. Mechanistic interpretability for steering vision-language-action models. In *Conference on Robot Learning*, pp. 2743–2762. PMLR, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Yuting He, Fuxiang Huang, Xinrui Jiang, Yuxiang Nie, Minghao Wang, Jiguang Wang, and Hao Chen. Foundation model for advancing healthcare: Challenges, opportunities and future directions. *IEEE Reviews in Biomedical Engineering*, 2024.

Nick Heppert, Minh Quang Nguyen, and Abhinav Valada. Scaling single human demonstrations for imitation learning using generative foundational models. In *IEEE International Conference on Robotics Automation (ICRA)*, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Lachlan Holden, Feras Dayoub, Alberto Candela, David Harvey, and Tat-Jun Chin. Vision foundation models for domain generalisable cross-view localisation in planetary ground-aerial robotic teams. *arXiv preprint arXiv:2601.09107*, 2026.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 1643–1653, 2021.

Jiaheng Hu, Rose Hendrix, Ali Farhadi, Aniruddha Kembhavi, Roberto Martín-Martín, Peter Stone, Kuo-Hao Zeng, and Kiana Ehsani. Flare: Achieving masterful and adaptive robot policies with large-scale reinforcement learning fine-tuning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3617–3624. IEEE, 2025a.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2005–2015, 2025b.

Yafei Hu, Quanting Xie, Vidhi Jain, Jonathan Francis, Jay Patrikar, Nikhil Keetha, Seungchan Kim, Yaqi Xie, Tianyi Zhang, Hao-Shu Fang, et al. Toward general-purpose robots via foundation models: A survey and meta-analysis. *arXiv preprint arXiv:2312.08782*, 2023.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In *Forty-second International Conference on Machine Learning*, 2025c.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10608–10615. IEEE, 2023a.

Haifeng Huang, Xinyi Chen, Yilun Chen, Hao Li, Xiaoshen Han, Zehan Wang, Tai Wang, Jiangmiao Pang, and Zhou Zhao. Roboground: Robotic manipulation with grounded vision-language priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22540–22550, 2025a.

Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9488–9495. IEEE, 2024a.

Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-vla: Unlocking vision-language-action model's physical knowledge for tactile generalization. *arXiv preprint arXiv:2507.09160*, 2025b.

Sili Huang, Jifeng Hu, Zhejian Yang, Liwei Yang, Tao Luo, Hechang Chen, Lichao Sun, and Bo Yang. Decision mamba: Reinforcement learning via hybrid selective sequence modeling. *Advances in Neural Information Processing Systems*, 37:72688–72709, 2024b.

Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv preprint arXiv:2305.11176*, 2023b.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *Conference on Robot Learning*, pp. 540–562. PMLR, 2023c.

Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In *Conference on Robot Learning*, pp. 4573–4602. PMLR, 2025c.

N. Hughes, Y. Chang, and L. Carlone. Hydra: A real-time spatial perception system for 3D scene graph construction and optimization. 2022.

Ambi Robotics Inc. Ambi robotics launches prime-1 foundation model for warehouse robots, 2025.

Bahar Irfan, Sanna Kuoppamäki, and Gabriel Skantze. Recommendations for designing conversational companion robots with older adults through foundation models. *Frontiers in Robotics and AI*, 11:1363713, 2024.

Shun Iwase, Muhammad Zubair Irshad, Katherine Liu, Vitor Guizilini, Robert Lee, Takuya Ikeda, Ayako Amma, Koichi Nishiwaki, Kris Kitani, Rares Ambrus, et al. Zerograsp: Zero-shot shape reconstruction enabled robotic grasping. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17405–17415, 2025.

Silvia Izquierdo-Badiola, Gerard Canal, Carlos Rizzo, and Guillem Alenyà. Plancollabnl: Leveraging large language models for adaptive plan generation in human-robot collaboration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 17344–17350. IEEE, 2024.

Ahmed Jaafar, Shreyas Sundara Raman, Sudarshan Harithas, Yichen Wei, Sofia Juliani, Anneke Wernerfelt, Benedict Quartey, Ifrah Idrees, Jason Xinyu Liu, and Stefanie Tellex. λ: A benchmark for data-efficiency in long-horizon indoor mobile manipulation robotics. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8942–8949. IEEE, 2025.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Dae-Sung Jang, Doo-Hyun Cho, Woo-Cheol Lee, Seung-Keol Ryu, Byeongmin Jeong, Minji Hong, Minjo Jung, Minchae Kim, Minjoon Lee, SeungJae Lee, et al. Unlocking robotic autonomy: A survey on the applications of foundation models. *International Journal of Control, Automation and Systems*, 22(8): 2341–2384, 2024.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pp. 991–1002. PMLR, 2022.

Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, pp. 9902–9915. PMLR, 2022.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. *Robotics: Science and Systems (RSS)*, 2023.

Wenkang Ji, Huaben Chen, Mingyang Chen, Guobin Zhu, Lufeng Xu, Roderich Groß, Rui Zhou, Ming Cao, and Shiyu Zhao. Genswarm: Scalable multi-robot code-policy generation and deployment via language models. *npj Robotics*, 4(1):5, 2026.

Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. Defom-stereo: Depth foundation model based stereo matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21857–21867, 2025a.

Jiajun Jiang, Yiming Zhu, Zirui Wu, and Jie Song. Dualmap: Online open-vocabulary semantic mapping for natural language navigation in dynamic changing scenes. *IEEE Robotics and Automation Letters*, 2025b.

Li Jiang, Shaoshuai Shi, and Bernt Schiele. Open-vocabulary 3d semantic segmentation with foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21284–21294, 2024a. doi: 10.1109/CVPR52733.2024.02011.

Xinyi Jiang, Guoming Wang, Huanhuan Li, Qinghua Xia, Rongxing Lu, and Siliang Tang. Talon: Improving large language model cognition with tactility-vision fusion. In *2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1–6. IEEE, 2024b.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: Robot manipulation with multimodal prompts. In *International Conference on Machine Learning*, pp. 14975–15022. PMLR, 2023.

Jianhao Jiao, Jinhao He, Changkun Liu, Sebastian Aegidius, Xiangcheng Hu, Tristan Braud, and Dimitrios Kanoulas. Litevloc: Map-lite visual localization for image goal navigation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5244–5251. IEEE, 2025.

Xiaofeng Jin, Matteo Frosi, and Matteo Matteucci. Openfusion++: An open-vocabulary real-time scene understanding system. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 634–641. IEEE, 2025.

Arvi Jonnarth, Ola Johansson, Jie Zhao, and Michael Felsberg. Sim-to-real transfer of deep reinforcement learning agents for online coverage path planning. *IEEE Access*, 2025.

Frank Joublin, Antonello Ceravola, Pavel Smirnov, Felix Ocker, Joerg Deigmoeller, Anna Belardinelli, Chao Wang, Stephan Hasler, Daniel Tanneberg, and Michael Gienger. Copal: corrective planning of robot actions with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8664–8670. IEEE, 2024.

Shyam Sundar Kannan, Vishnunandan LN Venkatesh, and Byung-Cheol Min. Smart-llm: Smart multi-agent robot task planning using large language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12140–12147. IEEE, 2024.

Ivan Kapelyukh, Vitalis Vosylius, and Edward Johns. Dall-e-bot: Introducing web-scale diffusion models to robotics. *IEEE Robotics and Automation Letters*, 8(7):3956–3963, 2023.

Ivan Kapelyukh, Yifei Ren, Ignacio Alzugaray, and Edward Johns. Dream2real: Zero-shot 3d object rearrangement with vision-language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4796–4803. IEEE, 2024.

Pushkal Katara, Zhou Xian, and Katerina Fragkiadaki. Gen2sim: Scaling up robot learning in simulation with generative models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6672–6679. IEEE, 2024.

Kento Kawaharazuka, Tatsuya Matsushima, Andrew Gambardella, Jiaxian Guo, Chris Paxton, and Andy Zeng. Real-world robot applications of foundation models: A review. *Advanced Robotics*, 38(18):1232–1254, 2024.

Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*, 2025.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19729–19739, 2023.

Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. Goat-bench: A benchmark for multi-modal lifelong navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16373–16383, 2024.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems*, 2024.

Donghoon Kim, Minji Bae, Unghui Nam, Gyeonghun Kim, Suyun Lee, Kyuhong Shim, and Byonghyo Shim. Adaptive capacity allocation for vision language action fine-tuning. *arXiv preprint arXiv:2603.07404*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, pp. 2679–2713. PMLR, 2025.

Sanghwan Kim, Daoji Huang, Yongqin Xian, Otmar Hilliges, Luc Van Gool, and Xi Wang. Palm: Predicting actions through language models. In *European Conference on Computer Vision*, pp. 140–158. Springer, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14183–14193, 2024.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pp. 104–120. Springer, 2020.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4392–4412, 2020.

David Kube, Simon Hadwiger, and Tobias Meisen. Robotic foundation models for industrial control: A comprehensive survey and readiness assessment framework. *arXiv preprint arXiv:2603.06749*, 2026.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots. *Robotics: Science and Systems XVII*, 2021.

Minseo Kwon, Yaesol Kim, and Young J Kim. Fast and accurate task planning using neuro-symbolic language models and multi-level goal decomposition. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16195–16201. IEEE, 2025.

Lei Lai, Zekai Yin, and Eshed Ohn-Bar. Zerovo: Visual odometry with minimal assumptions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17092–17102, 2025.

Sebastián Barbas Laina, Simon Boche, Sotiris Papatheodorou, Simon Schaefer, Jaehyung Jung, and Stefan Leutenegger. Findanything: Open-vocabulary and object-centric mapping for robot exploration in any environment. *arXiv preprint arXiv:2504.08603*, 2025.

Min Young Lee, Christina Dao Wen Lee, Jianghao Li, and Marcelo H Ang Jr. Dino-mot: 3d multi-object tracking with visual foundation model for pedestrian re-identification using visual memory mechanism. *IEEE Robotics and Automation Letters*, 2024a.

Sangmin Lee, Sungyong Park, and Heewon Kim. Dynscene: Scalable generation of dynamic robotic manipulation scenes for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12166–12175, 2025a.

Sung-Wook Lee, Xuhui Kang, and Yen-Ling Kuo. Diff-dagger: Uncertainty estimation with diffusion policy for robotic manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4845–4852. IEEE, 2025b.

Taeyoon Lee, Jaewoon Kwon, Patrick M Wensing, and Frank C Park. Robot model identification and learning: A modern perspective. *Annual Review of Control, Robotics, and Autonomous Systems*, 7, 2024b.

Barak Lenz, Opher Lieber, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, Daniel Gissin, Daniel Jannai, Dor Muhlgay, Dor Zimberg, Edden M. Gerber, Elad Dolev, Eran Krakovsky, Erez Safahi, Erez Schwartz, Gal Cohen, Gal Shachaf, Haim Rozenblum, Hofit Bata, Ido Blass, Inbal Magar, Itay Dalmedigos, Jhonathan Osin, Julie Fadlon, Maria Rozman, Matan Danos, Michael Gokhman, Mor Zusman, Naama Gidron, Nir Ratner, Noam Gat, Noam Rozen, Oded Fried, Ohad Leshno, Omer Antverg, Omri Abend, Or Dagan, Orit Cohavi, Raz Alon, Ro'i Belson, Roi Cohen, Rom Gilad, Roman Glozman, Shahar Lev, Shai Shalev-Shwartz, Shaked Haim Meirom, Tal Delbari, Tal Ness, Tomer Asida, Tom Ben Gal, Tom Braude, Uriya Pumerantz, Josh Cohen, Yonatan Belinkov, Yuval Globerson, Yuval Peleg Levy, and Yoav Shoham. Jamba: Hybrid transformer-mamba language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chen Li, Zhantao Yang, Han Zhang, Fangyi Chen, Chenchen Zhu, Anudeepsekhar Bolimera, and Marios Savvides. Metavla: Unified meta co-training for efficient embodied adaptation. In *International Conference on Learning Representations (ICLR)*, 2026a.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pp. 80–93. PMLR, 2023a.

Dingzhe Li, Yixiang Jin, Yuhao Sun, Yong A, Hongze Yu, Jun Shi, Xiaoshuai Hao, Peng Hao, Huaping Liu, Xiang Li, et al. What foundation models can bring for robot learning in manipulation: A survey. *The International Journal of Robotics Research*, pp. 02783649251390579, 2024a.

Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3086–3096, 2024b.

Heng Li, Minghan Li, Zhi-Qi Cheng, Yifei Dong, Yuxuan Zhou, Jun-Yan He, Qi Dai, Teruko Mitamura, and Alexander G Hauptmann. Human-aware vision-and-language navigation: Bridging simulation to reality with dynamic human interactions. *Advances in Neural Information Processing Systems*, 37:119411–119442, 2024c.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10965–10975, 2022b.

Peihan Li, Zijian An, Shams Abrar, and Lifeng Zhou. Large language models for multi-robot systems: A survey. *arXiv preprint arXiv:2502.03814*, 2025a.

Puhao Li, Yingying Wu, Ziheng Xi, Wanlin Li, Yuzhe Huang, Zhiyuan Zhang, Yinghan Chen, Jianan Wang, Song-Chun Zhu, Tengyu Liu, and Siyuan Huang. ControlVLA: Few-shot object-centric adaptation for pre-trained vision-language-action models. In *9th Annual Conference on Robot Learning*, 2025b.

Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5567–5577, 2023b.

Xiaoqi Li, Liang Heng, Jiaming Liu, Yan Shen, Chenyang Gu, Zhuoyang Liu, Hao Chen, Nuowei Han, Renrui Zhang, Hao Tang, et al. 3ds-vla: A 3d spatial-aware vision language action model for robust multi-task manipulation. In *9th Annual Conference on Robot Learning*, 2025c.

Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A comprehensive survey on world models for embodied ai. *arXiv preprint arXiv:2510.16732*, 2025d.

Yuyang Li, Yinghan Chen, Zihang Zhao, Puhao Li, Tengyu Liu, Siyuan Huang, and Yixin Zhu. Simultaneous tactile-visual perception for learning multimodal robot manipulation. *IEEE Robotics and Automation Letters*, 11(4):5254–5261, 2026b.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500. IEEE, 2023.

Jessica E Liang. Diffusion models for robotics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 29587–29589, 2025.

Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. In *Conference on Robot Learning*, pp. 3943–3960. PMLR, 2025.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12037–12047, 2025.

Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17070–17080, 2025a.

Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27906–27916, 2024.

Xixun Lin, Yucheng Ning, Jingwen Zhang, Yan Dong, Yilong Liu, Yongxuan Wu, Xiaohua Qi, Nan Sun, Yanmin Shang, Kun Wang, et al. Llm-based agents suffer from hallucinations: A survey of taxonomy, methods, and directions. *arXiv preprint arXiv:2509.18970*, 2025b.

Yinheng Lin, Yiming Huang, Beilei Cui, Long Bai, Huxin Gao, Hongliang Ren, and Jiewen Lai. Endoddc: Learning sparse to dense reconstruction for endoscopic robotic navigation via diffusion depth completion. *arXiv preprint arXiv:2602.21893*, 2026.

Matthew Lisondra, Beno Benhabib, and Goldie Nejat. Embodied ai with foundation models for mobile service robots: A systematic review. *Robotics*, 15(3):55, 2026.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36: 44776–44791, 2023a.

Haokun Liu, Yaonan Zhu, Kenji Kato, Atsushi Tsukahara, Izumi Kondo, Tadayoshi Aoyama, and Yasuhisa Hasegawa. Enhancing the llm-based robot manipulation through human-robot collaboration. *IEEE Robotics and Automation Letters*, 9(8):6904–6911, 2024b.

Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Pengju An, Xiaoqi Li, Kaichen Zhou, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. *Advances in Neural Information Processing Systems*, 37:40085–40110, 2024c.

Jiaming Liu, Hao Chen, Zhuoyang Liu, Pengju An, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, Chengkai Hou, Mengdi Zhao, KC alex Zhou, Pheng-Ann Heng, and Shanghang Zhang. Unifying diffusion and autoregression for generalizable vision-language-action model. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024d.

Shuijing Liu, Aamir Hasan, Kaiwen Hong, Runxuan Wang, Peixin Chang, Zachary Mizrachi, Justin Lin, D Livingston McPherson, Wendy A Rogers, and Katherine Driggs-Campbell. Dragon: A dialogue-based robot for assistive navigation with visual language grounding. *IEEE Robotics and Automation Letters*, 9 (4):3712–3719, 2024e.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1b: a diffusion foundation model for bimanual manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Xiaofeng Liu, Ziyang Wang, Jie Li, Angelo Cangelosi, and Chenguang Yang. Demonstration learning and generalization of robotic motor skills based on wearable motion tracking sensors. *IEEE Transactions on Instrumentation and Measurement*, 72:1–15, 2023b.

Xinhao Liu, Jintong Li, Yicheng Jiang, Niranjan Sujay, Zhicheng Yang, Juexiao Zhang, John Abanes, Jing Zhang, and Chen Feng. Citywalker: Learning embodied urban navigation from web-scale videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6875–6885, 2025b.

Yuan Liu, Haoran Li, Shuai Tian, Yuxing Qin, Yuhui Chen, Yupeng Zheng, Yongzhen Huang, and Dongbin Zhao. Towards long-lived robots: Continual learning vla models via reinforcement fine-tuning. *arXiv preprint arXiv:2602.10503*, 2026b.

Yuchen Liu, Luigi Palmieri, Sebastian Koch, Ilche Georgievski, and Marco Aiello. Delta: Decomposed efficient long-term robot task planning using large language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10995–11001. IEEE, 2025c.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37: 103031–103063, 2024f.

Zixian Liu, Mingtong Zhang, and Yunzhu Li. Kuda: Keypoints to unify dynamics learning and visual prompting for open-vocabulary robotic manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10561–10569. IEEE, 2025d.

Joel Loo, Zhanxin Wu, and David Hsu. Open scene graphs for open-world object-goal navigation. *The International Journal of Robotics Research*, pp. 02783649251369549, 2025.

Zhichen Lou, Kechun Xu, Zhongxiang Zhou, and Rong Xiong. Explorevlm: Closed-loop robot exploration task planning with vision-language models. *arXiv preprint arXiv:2508.11918*, 2025.

Yuankai Luo, Woping Chen, Tong Liang, Baiqiao Wang, and Zhenguo Li. Simvla: A simple vla baseline for robotic manipulation. *arXiv preprint arXiv:2602.18224*, 2026.

Runyu Ma, Jelle Luijkx, Zlatan Ajanović, and Jens Kober. Explorllm: Guiding exploration in reinforcement learning with large language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9011–9017. IEEE, 2025.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *The Eleventh International Conference on Learning Representations*, 2023.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *The Twelfth International Conference on Learning Representations*, 2024a.

Yecheng Jason Ma, William Liang, Hungju Wang, Sam Wang, Yuke Zhu, Linxi Fan, Osbert Bastani, and Dinesh Jayaraman. Dreureka: Language model guided sim-to-real transfer. In *Robotics: Science and Systems (RSS)*, 2024b.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024c.

Dominic Maggio, Yun Chang, Nathan Hughes, Matthew Trang, Dan Griffith, Carlyn Dougherty, Eric Cristofalo, Lukas Schmid, and Luca Carlone. Clio: Real-time task-driven open-set 3d scene graphs. *IEEE Robotics and Automation Letters*, 2024.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, 36:655–677, 2023.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance GPU based physics simulation for robot learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Luís Marques and Dmitry Berenson. Quantifying aleatoric and epistemic dynamics uncertainty via local conformal calibration. In *16th International Workshop on the Algorithmic Foundations of Robotics (WAFR)*, 2024. doi: 10.48550/arXiv.2409.08249. arXiv preprint arXiv:2409.08249.

Tomas Berriel Martins, Martin R Oswald, and Javier Civera. Open-vocabulary online semantic mapping for slam. *IEEE Robotics and Automation Letters*, 2025a.

Tomas Berriel Martins, Martin R Oswald, and Javier Civera. Open-vocabulary online semantic mapping for slam. *IEEE Robotics and Automation Letters*, 2025b.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

Oier Mees, Dibya Ghosh, Karl Pertsch, Kevin Black, Homer Rich Walke, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

Yunpeng Mei, Jian Sun, Zhihong Peng, Fang Deng, Gang Wang, and Jie Chen. Rog-sam: A language-driven framework for instance-level robotic grasping detection. *IEEE Transactions on Multimedia*, 2025.

Jared Mejia, Victoria Dean, Tess Hellebrekers, and Abhinav Gupta. Hearing touch: Audio-visual pretraining for contact-rich manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6912–6919. IEEE, 2024.

Marius Memmel, Jacob Berg, Bingqing Chen, Abhishek Gupta, and Jonathan Francis. STRAP: Robot sub-trajectory retrieval for augmented policy learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pp. 728–755. Springer, 2022.

Yuya Miyaoka and Masaki Inoue. Control barrier function for aligning large language models. *arXiv preprint arXiv:2511.03121*, 2025.

Saad Mokssit, Daniel Bonilla Licea, Bassma Guermah, and Mounir Ghogho. Deep learning techniques for visual slam: A survey. *IEEE Access*, 11:20026–20050, 2023.

Ruaridh Mon-Williams, Gen Li, Ran Long, Wenqian Du, and Christopher G Lucas. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence*, pp. 1–10, 2025.

Alejandro Murillo-González and Lantao Liu. Action flow matching for continual robot learning. In *Robotics: Science and Systems (RSS)*, 2025.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning*, pp. 892–909. PMLR, 2023.

Taewook Nam, Juyong Lee, Jesse Zhang, Sung Ju Hwang, Joseph J Lim, and Karl Pertsch. Lift: Unsupervised reinforcement learning with foundation models as teachers. *CoRR*, 2023.

Mohammad Nazeri, Aniket Datar, Anuj Pokhrel, Chenhui Pan, Garrett Warnell, and Xuesu Xiao. Verticoder: Self-supervised kinodynamic representation learning on vertically challenging terrain. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6536–6543. IEEE, 2025.

Rhys Newbury, Morris Gu, Lachlan Chumbley, Arsalan Mousavian, Clemens Eppner, Jürgen Leitner, Jeannette Bohg, Antonio Morales, Tamim Asfour, Danica Kragic, et al. Deep learning approaches to grasp synthesis: A review. *IEEE Transactions on Robotics*, 39(5):3994–4015, 2023.

Eley Ng, Ziang Liu, and Monroe Kennedy. Diffusion co-policy for synergistic human-robot collaborative tasks. *IEEE Robotics and Automation Letters*, 9(1):215–222, 2023.

Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4018–4028, 2024.

Toan Nguyen, Minh Nhat Vu, An Vuong, Dzung Nguyen, Thieu Vo, Ngan Le, and Anh Nguyen. Open-vocabulary affordance detection in 3d point clouds. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5692–5698. IEEE, 2023.

Toan Nguyen, Weiduo Yuan, Songlin Wei, Hui Li, Daniel Seita, and Yue Wang. Iclr: In-context imitation learning with visual reasoning. *arXiv preprint arXiv:2603.07530*, 2026.

Zhe Ni, Xiaoxin Deng, Cong Tai, Xinyue Zhu, Qinghongbing Xie, Weihang Huang, Xiang Wu, and Long Zeng. Grid: Scene-graph-based instruction-driven robotic task planning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13765–13772. IEEE, 2024.

Bernardo Nicoletti. Foundation models-driven support to logistics. In *Artificial Intelligence for Logistics 5.0: From Foundation Models to Agentic AI*, pp. 133–162. Springer, 2025.

Bernardo Nicoletti and Andrea Appolloni. Green logistics 5.0: a review of sustainability-oriented innovation with foundation models in logistics. *European Journal of Innovation Management*, 27(9):542–561, 2024.

Kazuma Obata, Tatsuya Aoki, Takato Horii, Tadahiro Taniguchi, and Takayuki Nagai. Lip-llm: Integrating linear programming and dependency graph with large language models for multi-robot task planning. *IEEE Robotics and Automation Letters*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Paul Pacaud, Ricardo Garcia Pinel, Shizhe Chen, and Cordelia Schmid. Guardian: Detecting robotic planning and execution errors with vision-language models. In *Workshop on Making Sense of Data in Robotics: Composition, Curation, and Interpretability at Scale at CoRL 2025*, 2025.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2017–2025, 2022.

Haimei Pan, Jiyun Zhang, Qinxi Wei, Xiongnan Jin, Chen Xinkai, and Jie Cheng. Robotic fire risk detection based on dynamic knowledge graph reasoning: An llm-driven approach with graph chain-of-thought. *arXiv preprint arXiv:2509.00054*, 2025a.

Jia Pan, Weizi Li, Wenxi Liu, Iftekharul Islam, Ke Guo, Yajue Yang, Shuai Zhang, Xuebo Ji, and Dawei Wang. Mixed crowd navigation: Perception, interaction, planning, and control. *Annual Review of Control, Robotics, and Autonomous Systems*, 9, 2025b.

Mingjie Pan, Jiyao Zhang, Tianshu Wu, Yinghao Zhao, Wenlong Gao, and Hao Dong. Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17359–17369, 2025c.

Meenal Parakh, Alisha Fong, Anthony Simeonov, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Lifelong robot learning with human assisted language planners. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 523–529. IEEE, 2024.

Daehee Park, Monu Surana, Pranav Desai, Ashish Mehta, Reuben John, and Kuk-Jin Yoon. Generative active learning for long-tail trajectory prediction via controllable diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27839–27850, 2025.

Hyoungseob Park, Lipeng Ke, Pritish Mohapatra, Huajun Ying, sankar venkataraman, and Alex Wong. Entropy-monitored kernelized token distillation for audio-visual compression. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jinyoung Park, Hee-Seon Kim, Kangwook Ko, Minbeom Kim, and Changick Kim. Videomamba: Spatio-temporal selective state space model. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Austin Patel and Shuran Song. GET-zero: Graph embodiment transformer for zero-shot embodiment generalization. In *ICRA 2025 Workshop: Beyond Pick and Place*, 2025.

Shivansh Patel, Xinchen Yin, Wenlong Huang, Shubham Garg, Hooshang Nayyeri, Li Fei-Fei, Svetlana Lazebnik, and Yunzhu Li. A real-to-sim-to-real approach to robotic manipulation with VLM-generated iterative keypoint rewards. In *2nd CoRL Workshop on Learning Effective Abstractions for Planning*, 2024.

Maytus Piriyajitakonkij, Mingfei Sun, Mengmi Zhang, and Wei Pan. Tta-nav: Test-time adaptive reconstruction for point-goal navigation under visual corruptions. *CoRR*, 2024.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8494–8502, 2018.

Raghunath Sai Puttagunta, Birendra Kathariya, Zhu Li, and George York. Multi-scale feature fusion using channel transformers for guided thermal image super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3086–3095, 2024.

Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. *arXiv preprint arXiv:2503.17349*, 2025.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9982–9991, 2020.

Shengyi Qian, Kaichun Mo, Valts Blukis, David F Fouhey, Dieter Fox, and Ankit Goyal. 3d-mvp: 3d multiview pretraining for manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22530–22539, 2025a.

Zekun Qian, Ruize Han, Zhixiang Wang, Junhui Hou, and Wei Feng. Covtrack: Continuous open-vocabulary tracking via adaptive multi-cue fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10054–10063, 2025b.

Behrad Rabiei, Mahesh Kumar AR, Zhirui Dai, Surya LSR Pilla, Qiyue Dong, and Nikolay Atanasov. Ltlcodegen: Code generation of syntactically correct temporal logic for robot task planning. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 19240–19247. IEEE, 2025.

Petar Radanliev, Omar Santos, and Carsten Maple. Threats and vulnerabilities in artificial intelligence and agentic ai models. *Frontiers in Artificial Intelligence*, 9:1731566, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Javlonbek Rakhmatillaev, Vytautas Bucinskas, and Nozimjon Kabulov. An integrative review of control strategies in robotics. *Robotic Systems and Applications*, July 2025. ISSN 2669-2473. doi: 10.21595/rsa.2025.25014.

Shreyas Sundara Raman, Vanya Cohen, Ifrah Idrees, Eric Rosen, Raymond Mooney, Stefanie Tellex, and David Paulius. Cape: Corrective actions from precondition errors using large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14070–14077. IEEE, 2024.

Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek Das. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5173–5183, 2022.

Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Sünderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In *Proceedings of the 7th Conference on Robot Learning (CoRL)*, pp. 23–72. Proceedings of Machine Learning Research, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zachary Ravichandran, Alexander Robey, Vijay Kumar, George J. Pappas, and Hamed Hassani. Safety guardrails for LLM-enabled robots. In *RSS 2025 Workshop on Reliable Robotics: Safety and Security in the Face of Generative AI*, 2025.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. Featured Certification, Outstanding Certification.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *7th Annual Conference on Robot Learning*, 2023.

Moritz Reuss, Hongyi Zhou, Marcel Rühle, Ömer Erdinç Yağmurlu, Fabian Otto, and Rudolf Lioutikov. FLOWER: Democratizing generalist robot policies with efficient vision-language-action flow policies. In *7th Robot Learning Workshop: Towards Robots with Human-Level Abilities*, 2025.

Kevin Riou, Kevin Subrin, and Patrick Le Callet. Vision foundation models for an embodiment and environment agnostic scene representation for robotic manipulation. In *International Conference on Intelligent Robots and Systems (IROS), on Brain over Brawn Workshop (BoB)(https://bob-workshop. github. io/)*, 2024.

Sandeep Routray, Hengkai Pan, Unnat Jain, Shikhar Bahl, and Deepak Pathak. ViPRA: Video prediction for robot actions. In *The Fourteenth International Conference on Learning Representations*, 2026.

Nikunj Sanghai and Nik Bear Brown. Advances in transformers for robotic applications: A review. *arXiv preprint arXiv:2412.10599*, 2024.

Ranjan Sapkota, Yang Cao, Konstantinos I Roumeliotis, and Manoj Karkee. Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769*, 2025.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 645–652. IEEE, 2024.

Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pp. 492–504. PMLR, 2023a.

Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. Gnm: A general navigation model to drive any robot. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7226–7233. IEEE, 2023b.

Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. Vint: A foundation model for visual navigation. In *Conference on Robot Learning*, pp. 711–733. PMLR, 2023c.

Jinghuan Shang, Karl Schmeckpeper, Brandon B. May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. In *8th Annual Conference on Robot Learning*, 2024.

Wangtian Shen, Pengfei Gu, Haijian Qin, and Ziyang Meng. Effonav: An effective foundation-model-based visual navigation approach in challenging environment. *IEEE Robotics and Automation Letters*, 2025.

Zhanbo Shi, Lin Zhang, Linfei Li, and Ying Shen. Towards audio-visual navigation in noisy environments: A large-scale benchmark dataset and an architecture considering multiple sound-sources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14673–14680, 2025.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10740–10749, 2020.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, pp. 894–906. PMLR, 2022.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pp. 785–799. PMLR, 2023.

R Shukla, S Moode, R Talan, and SK Gupta. Learning force-conditioned visuomotor diffusion policy from human demonstrations for complex robotic assembly tasks. In *North American Manufacturing Research Conference (NAMRC)*, volume 2025, 2025.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*, pp. 256–274. Springer, 2024.

Anukriti Singh, Amisha Bhaskar, Peihong Yu, Souradip Chakraborty, Ruthwik Dasyam, Amrit Bedi, and Pratap Tokekar. Varp: Reinforcement learning from vision-language model feedback with agent regularized preferences. *arXiv preprint arXiv:2503.13817*, 2025.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11523–11530. IEEE, 2023.

Rohan Sinha, Amine Elhafsi, Christopher Agia, Matthew Foutter, Edward Schmerling, and Marco Pavone. Real-time anomaly detection and reactive planning with large language models. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024. doi: 10.15607/RSS.2024.XX.114.

Vsevolod Skorokhodov, Chenghao Xu, Shuo Sun, Olga Fink, and Malcolm Mielle. Sear: Simple and efficient adaptation of visual geometric transformers for rgb+ thermal 3d reconstruction. *arXiv preprint arXiv:2603.18774*, 2026.

Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. In *ICLR*, 2023.

Saketh Banagiri Ge Gao Yiannis Aloimonos Cornelia Fermüller Snehesh Shrestha, Yantian Zha. NatSGLD: A dataset with Speech, Gestures, Logic, and demonstrations for robot learning in Natural human-robot interaction. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2025.

Andrew Sohn, Anusha Nagabandi, Carlos Florensa, Daniel Adelberg, Di Wu, Hassan Farooq, Ignasi Clavera, Jeremy Welborn, Juyue Chen, Nikhil Mishra, Peter Chen, Peter Qian, Pieter Abbeel, Rocky Duan, Varun Vijay, and Yang Liu. Introducing rfm-1: Giving robots human-like reasoning capabilities. Covariant blog, Mar 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Mohsen Soori, Behrooz Arezoo, and Roza Dastres. Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics*, 3:54–70, 2023.

Ajay Sridhar, Dhruv Shah, Catherine Glossop, and Sergey Levine. Nomad: Goal masked diffusion policies for navigation and exploration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 63–70. IEEE, 2024.

Kyle Stachowicz, Lydia Ignatova, and Sergey Levine. Lifelong autonomous improvement of navigation foundation models in the wild. In *8th Annual Conference on Robot Learning*, 2024.

Jared Strader, Aaron Ray, Jacob Arkin, Mason B Peterson, Yun Chang, Nathan Hughes, Christopher Bradley, Yi Xuan Jia, Carlos Nieto-Granda, Rajat Talak, et al. Language-grounded hierarchical planning and execution with multi-robot 3d scene graphs. *arXiv e-prints*, pp. arXiv–2506, 2025.

Heng Su, Mengying Xie, Nieqing Cao, Yan Ding, Beichen Shao, Xianlei Long, Fuqiang Gu, and Chao Chen. Ova-fields: Weakly supervised open-vocabulary affordance fields for robot operational part detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6385–6395, 2025.

Fan-Yun Sun, Shengguang Wu, Christian Jacobsen, Thomas Yim, Haoming Zou, Alex Zook, Shangru Li, Yu-Hsin Chou, Ethem F. Can, Xunlei Wu, Clemens Eppner, Valts Blukis, Jonathan Tremblay, Jiajun Wu, Stan Birchfield, and Nick Haber. 3d-GENERALIST: Vision-language-action models for crafting 3d worlds. In *Thirteenth International Conference on 3D Vision*, 2026.

Jiankai Sun, De-An Huang, Bo Lu, Yun-Hui Liu, Bolei Zhou, and Animesh Garg. Plate: Visually-grounded planning with transformers in procedural tasks. *IEEE Robotics and Automation Letters*, 7(2):4924–4930, 2022.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, et al. Neuralfeels with neural fields: Visuotactile perception for in-hand manipulation. *Science Robotics*, 9(96):eadl0628, 2024.

Yusuke Takagi, Motonari Kambara, Daichi Yashima, Koki Seno, Kento Tokura, and Komei Sugiura. Anolevla: Lightweight vision-language-action model with deep state space models for mobile manipulation. *arXiv preprint arXiv:2603.15046*, 2026.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martín-Martín, and Peter Stone. Deep reinforcement learning for robotics: A survey of real-world successes. *Annual Review of Control, Robotics, and Autonomous Systems*, 8(1):153–188, 2025.

Yiran Tao, Jehan Yang, Dan Ding, and Zackory Erickson. Lams: Llm-driven automatic mode switching for assistive teleoperation. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 242–251. IEEE, 2025.

Muhammad Tayyab Khan and Ammar Waheed. Foundation model driven robotics: A comprehensive review. *arXiv e-prints*, pp. arXiv–2507, 2025.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Izat Temiraliev, Diji Yang, and Yi Zhang. Retrieval-augmented robots via retrieve-reason-act. *arXiv preprint arXiv:2603.02688*, 2026.

Kumater Ter, Ore-Ofe Ajayi, and Daniel Udekwe. Taxonomy and trends in reinforcement learning for robotics and control systems: A structured review. *arXiv preprint arXiv:2510.21758*, 2025.

Christopher Thirgood, Oscar Mendez, Erin Ling, Jon Storey, and Simon Hadfield. Featureslam: Feature-enriched 3d gaussian splatting slam in real time. *arXiv preprint arXiv:2601.05738*, 2026.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pp. 394–406. PMLR, 2020.

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jan-Philipp Töberg, Axel-Cyrille Ngonga Ngomo, Michael Beetz, and Philipp Cimiano. Commonsense knowledge in cognitive robotics: a systematic literature review. *Frontiers in Robotics and AI*, 11:1328934, 2024.

Xiaodong Tong, Ke Li, and Jinsong Bao. Gnn-llm hybrid cognitive architectures for generative task adaptation in multi-human multi-robot collaborative disassembly. *Robotics and Computer-Integrated Manufacturing*, 98:103169, 2026.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Toshiaki Tsuji. Mamba as a motion encoder for robotic imitation learning. *IEEE Access*, 2025.

Sifan Tu, Xin Zhou, Dingkang Liang, Xingyu Jiang, Yumeng Zhang, Xiaofan Li, and Xiang Bai. The role of world models in shaping autonomous driving: A comprehensive survey. *arXiv preprint arXiv:2502.10498*, 2025.

Niccolò Turcato, Matteo Iovino, Aris Synodinos, Alberto Dalla Libera, Ruggero Carli, and Pietro Falco. Towards autonomous reinforcement learning for real-world robotic manipulation with large language models. *IEEE Robotics and Automation Letters*, 2025.

Georgios Tziafas and Hamidreza Kasaei. Lifelong robot library learning: Bootstrapping composable and generalizable skills for embodied control with language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 515–522. IEEE, 2024.

Halil Utku Unlu, Shuaihang Yuan, Congcong Wen, Hao Huang, Anthony Tzes, and Yi Fang. Reliable semantic understanding for real world zero-shot object goal navigation. In *International Conference on Pattern Recognition*, pp. 135–150. Springer, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Aleksandar Vujinovic and Aleksandar Kovacevic. Act-jepa: Joint-embedding predictive architecture improves policy representation learning. *arXiv preprint arXiv:2501.14622*, 2025.

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023.

Weikang Wan, Yifeng Zhu, Rutav Shah, and Yuke Zhu. Lotus: Continual imitation learning for robot manipulation through unsupervised skill discovery. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 537–544. IEEE, 2024.

Chen Wang, Fei Xia, Wenhao Yu, Tingnan Zhang, Ruohan Zhang, C Karen Liu, Li Fei-Fei, Jie Tan, and Jacky Liang. Chain-of-modality: Learning manipulation programs from multimodal human videos with vision-language-models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6527–6535. IEEE, 2025a.

Haitong Wang, Aaron Hao Tan, and Goldie Nejat. Navformer: A transformer architecture for robot target-driven navigation in unknown and dynamic environments. *IEEE Robotics and Automation Letters*, 2024a.

Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *Advances in neural information processing systems*, 37:124420–124450, 2024b.

Maggie Wang, Stephen Tian, Aiden Swann, Ola Shorinwa, Jiajun Wu, and Mac Schwager. Phys2real: Fusing vlm priors with interactive online adaptation for uncertainty-aware sim-to-real manipulation. *arXiv preprint arXiv:2510.11689*, 2025b.

Sen Wang, Le Wang, Sanping Zhou, Jingyi Tian, Jiayi Li, Haowen Sun, and Wei Tang. Flowram: Grounding flow matching policy with region-aware mamba framework for robotic manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12176–12186, 2025c.

Shengbo Wang and Shiping Wen. Safe control against uncertainty: A comprehensive review of control barrier function strategies. *IEEE Systems, Man, and Cybernetics Magazine*, 11(1):34–47, 2025.

Taowen Wang, Cheng Han, James Liang, Wenhao Yang, Dongfang Liu, Luna Xinyu Zhang, Qifan Wang, Jiebo Luo, and Ruixiang Tang. Exploring the adversarial vulnerabilities of vision-language-action models in robotics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6948–6958, 2025d.

Tianyang Wang, Yunze Wang, Jun Zhou, Benji Peng, Xinyuan Song, Charles Zhang, Xintian Sun, Qian Niu, Junyu Liu, Silin Chen, et al. From aleatoric to epistemic: Exploring uncertainty quantification techniques in artificial intelligence. *arXiv preprint arXiv:2501.03282*, 2025e.

Tsun-Hsuan Wang, Alaa Maalouf, Wei Xiao, Yutong Ban, Alexander Amini, Guy Rosman, Sertac Karaman, and Daniela Rus. Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6687–6694. IEEE, 2024c.

Xudong Wang, Zebin Han, Zhiyu Liu, Gan Li, Jiahua Dong, Baichen Liu, Lianqing Liu, and Zhi Han. Lifelong language-conditioned robotic manipulation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18629–18637, 2026a.

Yinuo Wang and Xiaowen Tao. Locomamba: Vision-driven locomotion via end-to-end deep reinforcement learning with mamba. *Advanced Engineering Informatics*, 70:104230, 2026.

Yuan Wang, Yuxin Chen, Zhongang Qi, Lijun Liu, Jile Jiao, Xuetao Feng, Yujia Liang, Ying Shan, and Zhipeng Zhang. Mamba-3vl: Taming state space model for 3d vision language learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6273–6283, 2025f.

Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. Rl-vlm-f: reinforcement learning from vision language foundation model feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 51484–51501, 2024d.

Yuqi Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. Unified vision-language-action model. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Zhangyuan Wang, Yunpeng Zhu, Yuqi Yan, Xiaoyuan Tian, Xinhao Shao, Meixuan Li, Weikun Li, Guangsheng Su, Weicheng Cui, and Dixia Fan. Underwatervla: Dual-brain vision-language-action architecture for autonomous underwater navigation. *arXiv preprint arXiv:2509.22441*, 2025g.

Zhijie Wang, Zhehua Zhou, Jiayang Song, Yuheng Huang, Zhan Shu, and Lei Ma. Vlatest: Testing and evaluating vision-language-action models for robotic manipulation. *Proceedings of the ACM on Software Engineering*, 2(FSE):1615–1638, 2025h.

Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling data generation in vision-and-language navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12009–12020, 2023.

Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. In *European conference on computer vision*, pp. 337–353. Springer, 2022.

Congcong Wen, Yisiyuan Huang, Hao Huang, Yanjia Huang, Shuaihang Yuan, Yu Hao, Hui Lin, Yu-Shen Liu, and Yi Fang. Zero-shot object navigation with vision-language models reasoning. In *International Conference on Pattern Recognition*, pp. 389–404. Springer, 2025a.

Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. DexVLA: Vision-language model with plug-in diffusion expert for general robot control. In *9th Annual Conference on Robot Learning*, 2025b.

Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

Justin Williams, Kishor Datta Gupta, Roy George, and Mrinmoy Sarkar. Lite vla: Efficient vision-language-action control on cpu-bound edge robots. *arXiv preprint arXiv:2511.05642*, 2025.

Justin Williams, Kishor Datta Gupta, Roy George, and Mrinmoy Sarkar. Litevla-edge: Quantized on-device multimodal control for embedded robotics. *arXiv preprint arXiv:2603.03380*, 2026.

Rosa Wolf, Yitian Shi, Sheng Liu, and Rania Rayyes. Diffusion models for robotic manipulation: A survey. *arXiv preprint arXiv:2504.08438*, 2025.

Ian Wu, Yuxiao Qu, Amrith Setlur, and Aviral Kumar. Reasoning cache: Continual improvement over long horizons via short-horizon rl. *arXiv preprint arXiv:2602.03773*, 2026.

Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102, 2023a.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. In *Robotics: Science and Systems (RSS) 2025*. Robotics: Science and Systems Foundation, 2025a.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pp. 2226–2240. PMLR, 2023b.

Wansen Wu, Tao Chang, Xinmeng Li, Quanjun Yin, and Yue Hu. Vision-language navigation: a survey and taxonomy. *Neural Computing and Applications*, 36(7):3291–3316, 2024.

Yi Wu, Zikang Xiong, Yiran Hu, Shreyash S Iyengar, Nan Jiang, Aniket Bera, Lin Tan, and Suresh Jagannathan. Selp: Generating safe and efficient task plans for robot agents with large language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2599–2605. IEEE, 2025b.

Yilin Wu, Wilson Yan, Thanard Kurutach, Lerrel Pinto, and Pieter Abbeel. Learning to manipulate deformable objects without demonstrations. *Robotics: Science and Systems XVI*, 2020.

Chaodong Xiao, Minghan Li, Zhengqiang Zhang, Deyu Meng, and Lei Zhang. Spatial-mamba: Effective visual state space models via structure-aware state fusion. In *13th International Conference on Learning Representations, ICLR 2025*, pp. 44892–44910. International Conference on Learning Representations, ICLR, 2025a.

Ruihong Xiao, Chenguang Yang, Yiming Jiang, and Hui Zhang. One-shot sim-to-real transfer policy for robotic assembly via reinforcement learning with visual demonstration. *Robotica*, 42(4):1074–1093, 2024.

Xuan Xiao, Jiahang Liu, Zhipeng Wang, Yanmin Zhou, Yong Qi, Shuo Jiang, Bin He, and Qian Cheng. Robot learning in the era of foundation models: A survey. *Neurocomputing*, pp. 129963, 2025b.

Senwei Xie, Hongyu Wang, Zhanqi Xiao, Ruiping Wang, and Xilin Chen. Robotic programmer: Video instructed policy code generation for robotic manipulation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 14923–14930. IEEE, 2025.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4389–4398, 2024.

Charles Xu, Qiyang Li, Jianlan Luo, and Sergey Levine. RLDG: Robotic Generalist Policy Distillation via Reinforcement Learning. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.028.

Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2228–2238, 2023.

Liming Xu, Sara Almahri, Stephen Mak, and Alexandra Brintrup. Multi-agent systems and foundation models enable autonomous supply chains: Opportunities and challenges. *IFAC-PapersOnLine*, 58(19): 795–800, 2024a.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024b.

Zhiyuan Xu, Kun Wu, Junjie Wen, Jinming Li, Ning Liu, Zhengping Che, and Jian Tang. A survey on robotics with foundation models: toward embodied ai. *arXiv preprint arXiv:2402.02385*, 2024c.

Han Xue, Jieji Ren, Wendi Chen, Gu Zhang, Fang Yuan, Guoying Gu, Huazhe Xu, and Cewu Lu. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. In *ICRA 2025 Workshop: Beyond Pick and Place*, 2025.

Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4927–4936, 2023.

Yajat Yadav, Zhiyuan Zhou, Andrew Wagenmaker, Karl Pertsch, and Sergey Levine. Robust fine-tuning of vision-language-action robot policies via parameter merging. In *International Conference on Learning Representations (ICLR)*, 2026.

Kashu Yamazaki, Taisei Hanyu, Khoa Vo, Thang Pham, Minh Tran, Gianfranco Doretto, Anh Nguyen, and Ngan Le. Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9411–9417. IEEE, 2024.

Sixu Yan, Zeyu Zhang, Muzhi Han, Zaijin Wang, Qi Xie, Zhitian Li, Zhehan Li, Hangxin Liu, Xinggang Wang, and Song-Chun Zhu. M 2 diffuser: Diffusion-based trajectory optimization for mobile manipulation in 3d scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Zhijie Yan, Shufei Li, Zuoxu Wang, Lixiu Wu, Han Wang, Jun Zhu, Lijiang Chen, and Jihong Liu. Dynamic open-vocabulary 3d scene graphs for long-term language-guided mobile manipulation. *IEEE Robotics and Automation Letters*, 2025b.

Dianyi Yang, Yu Gao, Xihan Wang, Yufeng Yue, Yi Yang, and Mengyin Fu. Opengs-slam: Open-set dense semantic slam with 3d gaussian splatting for object-level scene understanding. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8486–8492. IEEE, 2025a.

Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26340–26353, 2024a.

Jiange Yang, Wenhui Tan, Chuhao Jin, Keling Yao, Bei Liu, Jianlong Fu, Ruihua Song, Gangshan Wu, and Limin Wang. Transferring foundation models for generalizable robotic manipulation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1999–2010. IEEE, 2025b.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024b.

Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1389–1399, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin S Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander Clegg, John M Turner, et al. Homerobot: Open-vocabulary mobile manipulation. In *Conference on Robot Learning*, pp. 1975–2011. PMLR, 2023.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in neural information processing systems*, 37:5285–5307, 2024.

Shiyuan Yin, Chenjia Bai, Zhang Zihao, Junwei Jin, Xinxin Zhang, Chi Zhang, and Xuelong Li. Towards reliable LLM-based robots planning via combined uncertainty estimation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Shuolei Yin, Yejing Xi, Xun Zhang, Chengnuo Sun, and Qirong Mao. Foundation models in agriculture: A comprehensive review. *Agriculture*, 15(8):847, 2025b.

Yida Yin, Zekai Wang, Yuvan Sharma, Dantong Niu, Trevor Darrell, and Roei Herzig. In-context learning enables robot action prediction in llms. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8972–8979. IEEE, 2025c.

Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, Dhruv Batra, and Sehoon Ha. Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5543–5550. IEEE, 2024.

Takuma Yoneda, Jiading Fang, Peng Li, Huanyu Zhang, Tianchong Jiang, Shengjie Lin, Ben Picker, David Yunis, Hongyuan Mei, and Matthew R Walter. Statler: State-maintaining language models for embodied reasoning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15083–15091. IEEE, 2024.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

Kelin Yu, Yunhai Han, Qixian Wang, Vaibhav Saxena, Danfei Xu, and Ye Zhao. Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation. In *8th Annual Conference on Robot Learning*, 2024.

Rui Yu, Shenghua Wan, Yucen Wang, Chen-Xiao Gao, Le Gan, Zongzhang Zhang, and De-Chuan Zhan. Reward models in deep reinforcement learning: A survey. *arXiv preprint arXiv:2506.15421*, 2025a.

Zhaoshu Yu, Bo Wang, Pengpeng Zeng, Haonan Zhang, Ji Zhang, Lianli Gao, Jingkuan Song, Nicu Sebe, and Heng Tao Shen. A survey on efficient vision-language-action models. *arXiv preprint arXiv:2510.24795*, 2025b.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 558–567, 2021.

Mingqi Yuan, Tao Yu, Wenqi Ge, Xiuyong Yao, Dapeng Li, Huijiang Wang, Jiayu Chen, Bo Li, Wei Zhang, Wenjun Zeng, et al. A survey of behavior foundation model: Next-generation whole-body control system of humanoid robots. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yang Yue, Yulin Wang, Bingyi Kang, Yizeng Han, Shenzhi Wang, Shiji Song, Jiashi Feng, and Gao Huang. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. *Advances in Neural Information Processing Systems*, 37:56619–56643, 2024.

Vladimir Yugay, Theo Gevers, and Martin R Oswald. Magic-slam: Multi-agent gaussian globally consistent slam. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6741–6750, 2025.

Maryam Zare, Parham M Kebria, Abbas Khosravi, and Saeid Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, 2024.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation*, 2024.

Mona Sheikh Zeinoddin, Chiara Lena, Jiongqi Qu, Luca Carlini, Mattia Magro, Seunghoi Kim, Elena De Momi, Sophia Bano, Matthew Grech-Sollars, Evangelos B Mazomenos, et al. Dares: Depth anything in robotic endoscopic surgery with self-supervised vector-lora of the foundation model. *CoRR*, 2024.

Yiming Zeng, Mingdong Wu, Long Yang, Jiyao Zhang, Hao Ding, Hui Cheng, and Hao Dong. Lvdiffusor: Distilling functional rearrangement priors from large models into diffusor. *IEEE Robotics and Automation Letters*, 2024.

Lihan Zha, Yuchen Cui, Li-Heng Lin, Minae Kwon, Montserrat Gonzalez Arenas, Andy Zeng, Fei Xia, and Dorsa Sadigh. Distilling and retrieving generalizable knowledge for robot manipulation via language corrections. In *2024 IEEE international conference on robotics and automation (ICRA)*, pp. 15172–15179. IEEE, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Borong Zhang, Yuhao Zhang, Jiaming Ji, Yingshan Lei, Josef Dai, Yuanpei Chen, and Yaodong Yang. SafeVLA: Towards safety alignment of vision-language-action model via constrained learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Chenyangguang Zhang, Alexandros Delitzas, Fangjinhua Wang, Ruida Zhang, Xiangyang Ji, Marc Pollefeys, and Francis Engelmann. Open-vocabulary functional 3d scene graphs for real-world indoor spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19401–19413, 2025b.

Chi Zhang, Penglin Cai, Haoqi Yuan, Chaoyi Xu, and Zongqing Lu. Unitachand: Unified spatio-tactile representation for human to robotic hand skill transfer. *arXiv preprint arXiv:2512.21233*, 2025c.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023.

Jiahui Zhang, Yusen Luo, Abrar Anwar, Sumedh Anand Sontakke, Joseph J Lim, Jesse Thomason, Erdem Biyik, and Jesse Zhang. RewiND: Language-guided rewards teach robot policies without new demonstrations. In *Second Workshop on Out-of-Distribution Generalization in Robotics at RSS 2025*, 2025d.

Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing LI, Yuxin Fan, Wenjun Li, Zhibo Chen, Fei Gao, Qi Wu, Zhizheng Zhang, and He Wang. Embodied navigation foundation model. In *The Fourteenth International Conference on Learning Representations*, 2026.

Kun Zhang, Peng Yun, Jun Cen, Junhao Cai, Didi Zhu, Hangjie Yuan, Chao Zhao, Tao Feng, Michael Yu Wang, Qifeng Chen, et al. Generative artificial intelligence in robotic manipulation: A survey. *CoRR*, 2025e.

Peng-Fei Zhang, Ying Cheng, Xiaofan Sun, Shijie Wang, Fengling Li, Lei Zhu, and Heng Tao Shen. A step toward world models: A survey on robotic manipulation. *arXiv preprint arXiv:2511.02097*, 2025f.

Xiaohan Zhang, Yan Ding, Yohei Hayamizu, Zainab Altaweel, Yifeng Zhu, Yuke Zhu, Peter Stone, Chris Paxton, and Shiqi Zhang. Llm-grop: Visually grounded robot task and motion planning with large language models. *The International Journal of Robotics Research*, pp. 02783649251378196, 2025g.

Xiaojie Zhang, Yuanfei Wang, Ruihai Wu, Kunqi Xu, Yu Li, Liuyu Xiang, Hao Dong, and Zhaofeng He. Adaptive articulated object manipulation on the fly with foundation model reasoning and part grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13032–13042, 2025h.

Xinyu Zhang, Yuhan Liu, Haonan Chang, Liam Schramm, and Abdeslam Boularias. Autoregressive action sequence learning for robotic manipulation. *IEEE Robotics and Automation Letters*, 2025i.

Yuhang Zhang, Haosheng Yu, Jiaping Xiao, and Mir Feroskhan. Grounded vision-language navigation for uavs with open-vocabulary goal understanding. *arXiv preprint arXiv:2506.10756*, 2025j.

Han Zhao, Wenxuan Song, Donglin Wang, Xinyang Tong, Pengxiang Ding, Xuelian Cheng, and Zongyuan Ge. More: Unlocking scalability in reinforcement learning for quadruped vision-language-action models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11212–11218. IEEE, 2025a.

Runyi Zhao, Sheng Xu, Ruixing Jin, Yueci Deng, Yunxin Tai, Kui Jia, and Guiliang Liu. Sim2real-vla: Zero-shot generalization of synthesized skills to realistic manipulation. In *International Conference on Learning Representations (ICLR)*, 2026a.

Shibo Zhao, Sifan Zhou, Raphael Blanchard, Yuheng Qiu, Wenshan Wang, and Sebastian Scherer. Tartan imu: A light foundation model for inertial positioning in robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22520–22529, 2025b.

Wenzheng Zhao, Kruthika Gangaraju, and Fengpei Yuan. Multimodal perception-driven decision-making for human-robot interaction: a survey. *Frontiers in Robotics and AI*, 12:1604472, 2025c.

Yaqi Zhao and Hongxia Ye. Crater detection and population statistics in tianwen-1 landing area based on segment anything model (sam). *Remote Sensing*, 16(10):1743, 2024.

Ziyang Zhao, Shuheng Wang, Zhonghua Miao, and Ya Xiong. Harvestflex: Strawberry harvesting via vision-language-action policy adaptation in the wild. *arXiv preprint arXiv:2603.05982*, 2026b.

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13624–13634, 2024a.

Jianhao Zheng, Zihan Zhu, Valentin Bieri, Marc Pollefeys, Songyou Peng, and Iro Armeni. Wildgs-slam: Monocular gaussian splatting slam in dynamic environments. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11461–11471, 2025a.

Jinliang Zheng, Jianxiong Li, Dongxiu Liu, Yinan Zheng, Zhihao Wang, Zhonghong Ou, Yu Liu, Jingjing Liu, Ya-Qin Zhang, and Xianyuan Zhan. Universal actions for enhanced embodied foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22508–22519, 2025b.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, Ya-Qin Zhang, Jiangmiao Pang, Jingjing Liu, Tai Wang, and Xianyuan Zhan. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *International Conference on Learning Representations (ICLR)*, 2026.

Ziqiang Zheng, Jipeng Zhang, Tuan-Anh Vu, Shizhe Diao, Yue Him Wong Tim, and Sai-Kit Yeung. Marinegpt: Unlocking secrets of ocean to the public. *arXiv preprint arXiv:2310.13596*, 2023.

Ziqiang Zheng, Yiwei Chen, Huimin Zeng, Tuan-Anh Vu, Binh-Son Hua, and Sai-Kit Yeung. Marineinst: A foundation model for marine image analysis with instance visual description. In *European Conference on Computer Vision*, pp. 239–257. Springer, 2024b.

Peiyuan Zhi, Zhiyuan Zhang, Yu Zhao, Muzhi Han, Zeyu Zhang, Zhitian Li, Ziyuan Jiao, Baoxiong Jia, and Siyuan Huang. Closed-loop open-vocabulary mobile manipulation with gpt-4v. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4761–4767. IEEE, 2025.

Fangwei Zhong, Kui Wu, Hai Ci, Churan Wang, and Hao Chen. Empowering embodied visual tracking with visual foundation models and offline rl. In *European Conference on Computer Vision*, pp. 139–155. Springer, 2024.

Enshen Zhou, Qi Su, Cheng Chi, Zhizheng Zhang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and He Wang. Code-as-monitor: Constraint-aware visual programming for reactive and proactive robotic failure detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6919–6929, 2025a.

Gaoyue Zhou, Hengkai Pan, Yann Lecun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning*, pp. 79115–79135. PMLR, 2025b.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. In *International Conference on Machine Learning*, pp. 61885–61896. PMLR, 2024a.

Weijie Zhou, Manli Tao, Chaoyang Zhao, Haiyun Guo, Honghui Dong, Ming Tang, and Jinqiao Wang. Physvlm: Enabling visual language models to understand robotic physical reachability. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6940–6949, 2025c.

Yuchen Zhou, Jiayuan Gu, Xuanlin Li, Minghua Liu, and Hao Su. PartSLIP++: Enhancing low-shot 3d part segmentation via multi-view instance segmentation and maximum likelihood estimation. In *ICCV 2025 Workshop on Wild 3D: 3D Modeling, Reconstruction, and Generation in the Wild*, 2025d.

Zhehua Zhou, Jiayang Song, Kunpeng Yao, Zhan Shu, and Lei Ma. Isr-llm: Iterative self-refined large language model for long-horizon sequential task planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2081–2088. IEEE, 2024b.

Zhongyi Zhou, Yichen Zhu, Minjie Zhu, Junjie Wen, Ning Liu, Zhiyuan Xu, Weibin Meng, Yaxin Peng, Chaomin Shen, Feifei Feng, et al. Chatvla: Unified multimodal understanding and robot control with vision-language-action model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5377–5395, 2025e.

Shaoting Zhu, Linzhan Mou, Derun Li, Baijun Ye, Runhan Huang, and Hang Zhao. Vr-robo: A real-to-sim-to-real framework for visual robot navigation and locomotion. *IEEE Robotics and Automation Letters*, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

Nikola Zubic, Mathias Gehrig, and Davide Scaramuzza. State space models for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5819–5828, 2024.

Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. *International Journal of Computer Vision*, 133(2):611–627, 2025.