10 Open Challenges Steering the Future of Vision-Language-Action Models

Soujanya Poria[1], Navonil Majumder[2], Chia-Yu Hung[1], Amir Ali Bagherzadeh[3], Chuan Li[3], Kenneth Kwok[5], Ziwei Wang[1], Cheston Tan[5], Jiajun Wu[6], David Hsu[4]

[1]Nanyang Technological University, [2]SUTD, [3]Lambda Labs, [4]National University of Singapore
[5]A*STAR, [6]Stanford University

Abstract

Due to their ability of follow natural language instructions, vision-language-action (VLA) models are increasingly prevalent in the embodied AI arena, following the widespread success of their precursors—LLMs and VLMs. In this paper, we discuss 10 principal milestones in the ongoing development of VLA models—multimodality, reasoning, data, evaluation, cross-robot action generalization, efficiency, whole-body coordination, safety, agents, and coordination with humans. Furthermore, we discuss the emerging trends of using spatial understanding, modeling world dynamics, post training, and data synthesis—all aiming to reach these milestones. Through these discussions, we hope to bring attention to the research avenues that may accelerate the development of VLA models into wider acceptability.

1 Introduction to VLA Models

Enabling robots to perform a wide range of manipulation tasks in unstructured real-world environments is a central goal of modern robotics. Robotic policy models, which aims to generate sequences of low-level actions (e.g., end-effector poses or joint torques) for complex tasks, have been at the forefront of this effort. Traditional reinforcement learning-based approaches have achieved impressive results on narrowly-defined tasks within fixed environments (Ma et al. 2024), but they often struggle to generalize beyond their training scenarios (Brohan et al. 2023c; Chi et al. 2023b). The immense diversity and complexity of real-world scenarios pose a significant challenge for robotic policies, as they must possess strong generalization capabilities to perform effectively across multiple tasks and environments in the real world.

Recent progress in the foundation models for vision and language has demonstrated remarkable capabilities in scene understanding and task planning (Radford et al. 2021; Zhai et al. 2023; Touvron et al. 2023b). Visual-Language Models (VLMs) have emerged as powerful tools for interpreting and reasoning about the visual world, demonstrating proficiency in tasks from object detection to video understanding. Their capacity to use chain-of-thought to decompose complex visual or language tasks into a sequence of logical steps makes them a promising platform for high-level robotic planning. Yet, a fundamental gap remains: VLMs are not designed and trained to produce the robot-executable policies needed to navigate the specific physical constraints and control dynamics of a given robot.

To bridge this gap, **Visual-Language-Action (VLA) models** have emerged that combine visual observations and language instructions to produce generalized robotic actions across diverse environments. The tasks are learned through Imitation Learning (IL), a paradigm where a robot learns to execute tasks by observing and replicating expert demonstrations. This approach allows the policy to generalize across a wide range of environments and tasks. These models can be broadly categorized into two groups, based on their action modeling strategy:

- **Continuous-action models** (Octo Model Team et al. 2024), which typically leverage diffusion processes to generate smooth trajectories in continuous action spaces.
- **Discrete-action models** (Brohan et al. 2023b,c; Kim et al. 2024; Sun et al. 2024), where robot actions are represented as sequences of discrete tokens. Due to compression through quantization, this approach risks lossy tokenization and detokenization.

In the standard VLA setting for imitation learning, the robot's state at time t is described by a multimodal observation (O_t, L_t, S_t), including current observation O_t, textual instructions L_t, and prior context S_t. The objective is to maximize the log probability of a sequence of actions A_t generated by the equation $\max_\theta \mathbb{E}_{(A_t,O_t,L_t,S_t)\sim\mathcal{D}} \log \pi_\theta(A_t|O_t, L_t, S_t)$, where D is the demonstration dataset.$\pi_\theta(A_t \mid O_t, L_t, S_t)$. For cases with discrete action representations, this can be framed as a standard next-token prediction problem over a sequence of action tokens, a straightforward approach analogous to autoregressive language modeling. Another family of approaches utilizes diffusion (Ho, Jain, and Abbeel 2020) or flow matching (Lipman et al. 2023) to model the continuous action distribution (Octo Model Team et al. 2024; Chi et al. 2023a). This dichotomy has led to two primary schools of thought for modeling the action space: discrete action representation and continuous action representation, each with distinct trade-offs. Discrete action models can be easily implemented with transformer-based language models, allowing them to leverage next-token prediction. However, this approach introduces quantization errors and its auto-regression incurs slow inference speeds, making it unsuitable for high-frequency control (Kim, Finn, and Liang 2025). In contrast, diffusion-based continuous action models jointly predict an action chunk, better preserving the fidelity of robotic movements and better suiting high-frequency control, but requires a significantly higher compute budget for convergence.

1.1 Discrete Action Models

Robot actions are inherently continuous, spanning multiple degrees of freedom such as 3D translation (x, y, z) and rotation (roll, pitch, yaw). To integrate these actions with transformer-based language models, it is common to discretize them into finite bins (Brohan et al. 2023c,b). A typical approach maps each action dimension to one of 256 discrete bins using a quantile-based strategy, which balances granularity with robustness to outliers.

OpenVLA (Kim et al. 2024) incorporates these action tokens directly into the language model vocabulary by replacing the 256 least-used tokens of the LLaMA tokenizer. This enables the model to perform next-token prediction over action sequences in the same way it predicts text. To further improve efficiency, the FAST+ tokenization method (Pertsch et al. 2025) applies a discrete cosine transform (DCT) across action dimensions at each timestep, decorrelating joint action components. The resulting sequences can then be compressed using byte-pair encoding (BPE), reducing vocabulary size and sequence length while preserving essential structure. This allows robot action to be represented with a few tokens, which accelerates training and inference.

1.2 Continuous Action Models

In contrast to discrete action tokenization, continuous-action models directly parameterize and predict a more expressive continuous action distribution. This approach aims to preserve the inherent fidelity of robotic movements, avoiding the quantization errors that can arise from binning. Discrete-Action Models are also unsuitable for high-frequency controls due to slow inference speed (3-5Hz) (Kim, Finn, and Liang 2025) and unreliable task execution on bimanual manipulators. Prominent approaches (Black et al. 2024; Intelligence et al. 2025) add an additional small action expert that models the continuous actions conditioned on the internal hidden states of the larger VLM, effectively decoupling the action generation task from the VLM's primary function of multimodal representation. However, a key limitation is that training diffusion policies requires a significantly higher computational budget for convergence (Pertsch et al. 2025). This disparity has motivated the development of hybrid approaches that combine the strengths of both paradigms. Specifically, VLA models are first pretrained autoregressively on discrete action tokens, after which a specialized action expert is integrated and further trained to produce continuous action outputs. This methodology has demonstrated superior performance and significantly faster convergence (Intelligence et al. 2025). Furthermore, to improve the generalization capabilities of the model, knowledge insulation (Driess et al. 2025) was introduced to preserve the core semantic knowledge of the VLM while training the action expert.

2 The 10 Open Challenges

Despite this great progress in this research field, there remain certain challenges that we discuss below.

2.1 Multimodal Sensing and Perception

Depth Perception. Many VLA models, except MolmoAct (Lee et al. 2025) and SpatialVLA (Qu et al. 2025), ignore explicit depth information to robustly operate in a 3D environment. While depth perception could be approximated through comparing visual frames across time, such an approximation loses sensitivity with the object size, distance from the camera, and fineness of movements, which may precipitate into erroneous reasoning by the VLA.

Even though MolmoAct and SpatialVLA uses depth information, it is limited to the training phase where the VLA explicitly learns to impute depth from the RGB frames. This naturally eliminates the need for a specialized depth-gauging camera, but the limitation still remains due to the above reasons.

Environmental Noise and Artifacts. Current evaluations of VLA models are evaluated under very controlled environments, be it in real world or simulation. To this end, SimplerEval (Li et al. 2024) applies distribution shift across background, lighting, distractors, texture, and camera pose. However, a broad range of noise and artifacts, such as, reflections, lens flares, environment-induced noisy camera feed—due to water, dust, or debris—, are still unaccounted for in the current evaluation frameworks.

Beyond Vision Modality. At this nascent stage, VLA models, as named, are limited to visual and language modalities. However, there is strong incentive to expand the domain of these models into audio, speech, touch among other

sensory information. For instance, embodied AI systems equipped with such multimodal action models could be deployed into hazardous disaster zones for rescue operations—audio signal could be instrumental here to listen to victims' call for help, respond to explosive sounds or falling debris, etc. On the other hand, touch modality would allow VLA to perform delicate tasks that require careful application of force through the joints and gripper—handling of glass and ceramic items, cooking, assemble and disassembly of electronic and electrical devices, etc. Recently, Bi et al. (2025) proposed a touch capable VLA that has shown to enhance overall performance on various pick, place, wipe tasks over vision-only VLA models.

2.2 Robust Reasoning

Owing to being pre-trained on a large amount of text data and subsequent large-scale post-training, LLMs (Bai et al. 2023; OpenAI 2023; Touvron et al. 2023a) and, by extension, VLMs (Bai et al. 2025; Chen et al. 2023; OpenAI 2023) are reasonably adept at high-level reasoning and planning to solve complex problems. However, such reasoning performance do not translate that well into the derived VLA models on seemingly much simpler robot tasks. The recent VLA models, such as, Emma-X (Sun et al. 2024), CoT-VLA (Zhao et al. 2025b), and MolmoAct (Lee et al. 2025), that are trained on both high-level linguistic and low-level action-level reasoning traces still produce imperfect results in both real-life and simulations, such as, LIBERO (Liu et al. 2024a) and SimplerEnv (Li et al. 2024). These evaluations usually involve simple tasks like picking and placing items, operating a drawer, pointing to an item, etc. To be deployable in more complex and sensitive environments, the error rate on such simpler tasks must approach near perfection. Robust reasoning becomes even more important for long-horizon tasks where the performance generally drops with increasing horizon. Furthermore, robust chain-of-thought reasoning has been shown to improve LLM performance on out-of-distribution tasks, which is also shown to be the trend for VLA models like Emma-X, CoT-VLA, and MolmoAct.

Another key challenge lies in the effective use of tools. In everyday life, even simple goals—such as making coffee—require identifying and using the right tools (e.g., a cup, a spoon, and so on). Equipping VLAs with the ability to understand, select, and utilize tools in a similar manner remains an important open problem.

2.3 Quality Training Data

Open-X-Embodiment (Collaboration et al. 2023) is a unification of around 70 and growing smaller datasets that cover more than 1M episodes of diverse robots performing various tasks involving many items and scenes. Despite being trained on such a vast dataset, VLA models are often brittle to out-of-distribution environments and robot setups, requiring the collection of additional episodes for fine-tuning.

There have also been attempts at using simulations to collect trajectories at a cheaper rate, like the Sim2Real (Kadian et al. 2020) challenge, where models are trained on simulation data and evaluated on real robots. However, constructing real-enough simulations has been a challenge for both data collection and evaluation.

Finally, a major challenge arises from the variance and noise in the collected data (Zheng et al. 2025b). Such variability can stem from differences in embodiments, camera positions and angles, as well as inconsistencies in the cognitive behaviors of human data collectors.

2.4 Evaluation of VLA Models

Evaluation of VLA models is a great challenge due to general limited availability of robots, environments, and objects. Thus, most VLA models are evaluated on a few real benchmarks that involve WidowX (Sun et al. 2024; Kim et al. 2024) or Franka (Kim et al. 2024) robots in a predefined environment, such as a kitchen with related objects therein. However, such evaluations obviously do not capture the overall capability of the VLA models spanning a broad range of robots and settings. Simulation-based evaluations (Li et al. 2024; Liu et al. 2023) somewhat mitigate this issue by evaluating the models on a finite but much larger set of environments with variable texture, lighting, camera pose, backdrop, etc.

However, the environments simulated in such tools often fail to capture enough details of their real-life counterparts, leading to poor correlation between in-simulation and real-life performance. The gap could be in the environmental details, such as, lighting, reflection, texture and consistency of the object surfaces, as well as the PD parameters (stiffness and damping parameters) of the various robot joints and gripper. SimplerEnv (Li et al. 2024) made great strides in minimizing this distribution shift in the robot motion through simulated annealing and in environments with image in-painting. However, there still remains much room for improvement, not only in terms of environmental and embodiment diversity and fidelity, but also incorporation of modalities, such as, audio and touch that have lately been gaining much research interest.

2.5 Cross-Robot Action Generalization

Cross-robot action generalization remains a fundamental challenge for VLA models, primarily due to action heterogeneity. Zheng et al. (2025b) shows that training on action data from a fixed set of embodiments often fails to generalize to others with distinct action spaces—for example, robots with higher degree of freedom or structural differences across robotic arms, quadrupeds, and autonomous vehicles. The diversity of control interfaces further exacerbates this issue. Addressing action heterogeneity is therefore a critical first step toward building VLA models that can adapt seamlessly and achieve zero-shot generalization, analogous to the generality demonstrated by VLMs and LLMs in multimodal tasks.

2.6 Resource Efficiency

Embodiments/robots are generally much more compute-limited than the training infrastructure due to the constraints on space and energy requirements. To circumvent this limitation, the embodiments often act as thin clients that col-

lect input and observations from the environment and relay them to a server with ample compute to infer the next actions from a VLA model. However, this setting is limited by the network communication latency and disruption—for instance, disasters zones are often cutoff from the internet and telecommunication services. Thus, embodiments are equipped with smaller and more resource efficient VLA models capable of running on the hardware on-board. Unfortunately, these smaller models are generally less performant as compared to their larger counterparts (Octo Model Team et al. 2024; Kim et al. 2024; Brohan et al. 2023a; Lee et al. 2025). Thus, striking the right balance between VLA model capacity and resource efficiency remains a key hurdle in the adoption of VLA models.

2.7 Whole-Body Coordination

Real-world VLA tasks often require an agent's entire body to act in concert. A mobile manipulator may reposition its base while moving its arm to reach, grasp, or carry objects in a safe manner. Enabling such whole-body coordination by coupling locomotion and manipulation under uncertainty is a core challenge for embodied VLA agents. Broadly, two approaches exist: (1) model-based control and (2) learning-based control. Model-based controllers, such as Model Predictive Control (MPC), optimize short-horizon trajectories for base and arm subject to dynamic balance, contact feasibility, and joint/torque limits (Le Cleac'h et al. 2024; Kim et al. 2025). By explicitly modeling robot dynamics, MPC yields precise, interpretable coordination and safety guarantees in structured settings. However, it relies on accurate models, faces real-time bottlenecks as dimensionality and non-smooth contact events grow, and often requires embodiment-specific tuning, complicating generalization and integration with high-level vision–language plans (Wang, Chen, and Zhao 2025; Heins et al. 2023). Data-driven policies learn unified locomotion–manipulation behaviors from multimodal observations (e.g., vision, proprioception, force/tactile) and demonstration (Fu, Zhao, and Finn 2024; Ha et al. 2024; Chen et al. 2025). Reinforcement and imitation learning can decide when to move the base versus the arm, and recent systems expand the action space to include base motion alongside end-effector control. While adaptable to unstructured settings, these methods struggle with exploration and credit assignment, require careful reward shaping, and may generalize poorly across embodiments (Liu et al. 2024b; Ha et al. 2024). Future progress will likely hinge on hybrid control frameworks that combine analytical structure with learning-based flexibility—for example, using constraint-aware planners to ensure feasibility and safety while learned policies provide goal-directed coordination, or learning to tune costs and propose feasible whole-body poses (Zheng et al. 2025a; Xie et al. 2025; Ji et al. 2024). A central need is objective and reward design that explicitly couples locomotion with manipulation so base motion contributes to stable, precise end-effector control (Sundaresan et al. 2025; Dugar et al. 2024; Zhao et al. 2025a; Du et al. 2024; Xie et al. 2025). Ultimately, the dominant challenge is the high-dimensional search space of coupled actions, motivating compact action representations and hierarchical planning.

2.8 Safety Assurances

The recent proliferation and wide accessibility of powerful LLMs like GPT, Claude, and Gemini have raised alarm bells across many institutions, from education and law to healthcare and security. One of the major concern is the LLMs producing harmful responses, truthful or otherwise, having negative influence on the users or the society as a whole. Such concern translates to the embodied AI systems as well, except the harm could be inflicted very directly through the actions, without any direct human supervision. For instance, in a rescue operation in a disaster zone, a robot equipped with embodied AI may harm the victims while saving them due to imperfect actions. To avoid such scenarios that may harm people and the overall trust in embodied AI systems, the community must devise appropriate guardrails and failsafes. To this end, Zhang et al. (2025) recently proposed an reinforcement learning-based approach for safety alignment of VLA models by constraining the actions while maintaining general performance.

2.9 VLA Models in Agentic Frameworks

Human achievements are often made possible through collaboration in groups. Inspired by this, Agentic AI is a powerful framework increasingly relevant due to their general ability to solve problems through inter-agent communication, without being constrained to the local tools available to an individual agent. Such multi-agent framework is yet to be thoroughly explored through the lens of VLA models. Adoption of such framework could address the agent-level resource constraints discussed in Section 2.6. For instance, certain compute load can be delegated to a nearby idle robot. On the other hand, another robot equipped with certain special sensor or at a different vantage point could collect observations for the benefit of another robot for improved decision making. VLA models capable of such inter-agent communication could be instrumental to the applicability and growth of embodied AI systems. At the same time, an important challenge lies in determining the appropriate level of agency and autonomy to assign to each agent. More broadly, realizing such multi-agent VLA systems requires the development of trustworthy, safe, and verifiable workflow generation to reliably execute complex tasks.

Recently (Yang et al. 2025b) proposed a heterogeneous agentic framework of VLM/LLM/VLA models that consists of an LLM as a high-level planner and another VLM as a verifier to assist the VLA model carry out the given tasks. This early work shows a promising path toward truly-embodied VLA agents.

2.10 Human-Robot Coordination

The explicit communication between users and current VLA models is unidirectional, i.e., human to robot through natural language instructions. Going forward, natural language and visual output—containing reasoning traces and questions for the user seeking missing information—from the VLA models could be fruitful for interpretation of the actions and user interaction. Recently, CoT-VLA (Zhao et al.

2025b) has been trained to produce a visual output of the intended state prior to action decoding; this has been shown to be improving the overall model performance across the standard benchmarks. Emma-X (Sun et al. 2024) trained to generate high-level rationale in natural language, has shown to outperform direct action generation.

3 Emerging Trends to Face the Challenges

We cover the emerging trends, pivotal to the progress of VLA models.

3.1 Hierarchical Planning

We cast the class of Embodied-AI tasks addressed by VLA models as hierarchical planning problems that could be addressed by the following high-level framework (see Fig. 1):

Algorithm 1: Multi-Agent VLA Planning with Orchestrator, Workflow Generation, and Safety Guardrails

Input: Initial observation/state s_0 (e.g., image, joint readings, etc.);
Goal specification G in natural language;
Library of expert skill policies $\{\pi_1, \ldots, \pi_K\}$;
High-level planner (also serving as orchestrator and workflow generator).
Output: Successful execution achieving goal G if $s^{(N)} \models G$.

Step 1: High-level orchestration and workflow generation
The high-level planner/orchestrator decomposes G into a workflow of subtasks:

$$P = [\,\tau^{(1)}, \tau^{(2)}, \ldots, \tau^{(N)}\,],$$

where each $\tau^{(i)}$ is assigned to an expert VLA specialized in that skill.

Step 2: Subgoal execution with expert VLAs
for $i = 1$ **to** N **do**
Each subtask $\tau^{(i)}$ is executed as a sequence of low-level actions:

$$A^{(i)} = [a_1^{(i)}, a_2^{(i)}, \ldots, a_{T_i}^{(i)}],$$

where for each $j = 1, \ldots, T_i$:

$$a_j^{(i)} \sim \pi_{k_i}\big(s_{t_j}^{(i-1)}, \tau^{(i)}\big), \quad s_{t_j+1}^{(i-1)} = \mathcal{W}(s_{t_j}^{(i-1)}, a_j^{(i)}).$$

Execution yields state $s^{(i)}$ after satisfying $\tau^{(i)}$.

Step 2.1: Safety guardrails and evaluators
After completion of $A^{(i)}$, invoke safety checks through world model simulation $\mathcal{W}$ and evaluators. If feedback indicates plan inconsistency or safety risks, trigger **re-planning** by returning control to Step 1.

Step 3: Coordination
The resulting $s^{(i)}$ becomes the input to subtask $\tau^{(i+1)}$. Continue until $i = N$.

Step 4: Goal verification
If $s^{(N)} \models G$, the multi-agent VLA workflow succeeds.

High-Level Planner. LLMs and VLMs excel at high-level planning due to their large-scale training, which equips them with extensive world knowledge and commonsense reasoning skills. One may leverage this capability to generate abstract plans; as $\pi_{0.5}$ (Black et al. 2024) used a VLM to encode text and image and guide the low-level action generation expert. In multi-agent scenarios, a high-level planner may also serve as an orchestrator, assigning tasks to the low-level action experts.

Low-Level Action Expert. A low-level action expert generates a sequence of actions to achieve each subgoal, that may have been generated by the high-level planner. It can be implemented either as a diffusion-based or an autoregressive model trained on real robot action datasets, such as, Open-X-Embodiment (Collaboration et al. 2023) or DROID (Khazatsky et al. 2024). The low-level action expert may also take input from multiple modalities—e.g., vision, text, and touch among others.

Reasoning before Actions. Rather than directly sampling low-level actions, one can introduce an intermediate reasoning layer that grounds each subtask $\tau^{(i)}$ in structured, interpretable guidance. This paradigm—termed *reasoning before actions*—encourages the policy to articulate how a subgoal should be achieved prior to generating the corresponding motor commands (Sun et al. 2024; Zawalski et al. 2024).

Consider the robot as a *delivery driver* tasked with delivering a package to a destination (the goal). The driver first decomposes the journey into subgoals: "Reach Main Street", "Cross the river bridge", "Arrive at the central library". For each subgoal, the driver reasons about the appropriate strategy: to reach the Main Street, "follow Elm Road until the first intersection"; to cross the river bridge, "merge onto the highway ramp". These reasoning steps ground the low-level driving actions—such as turning the wheel left, pressing the accelerator, or braking—which are executed in sequence to realize the subgoal.

Formally, a subtask $\tau^{(i)}$ is realized through a sequence of low-level actions

$$A^{(i)} = \big[a_1^{(i)},\, a_2^{(i)}, \ldots, a_{T_i}^{(i)}\big], \quad a_j^{(i)} \sim \pi_{k_i}\Big(s_{t_j}^{(i-1)},\, \tau^{(i)}\Big),$$

with state transitions

$$s_{t_j+1}^{(i-1)} \;=\; \mathcal{W}\Big(s_{t_j}^{(i-1)},\, a_j^{(i)}\Big).$$

Before producing $A^{(i)}$, the system generates a reasoning trace

$$r^{(i)} = f_{\text{reason}}\Big(\tau^{(i)},\, s_{t_0}^{(i-1)}\Big),$$

which provides both linguistic guidance (e.g., "move up by 22 steps, then left by 23 steps") and semantic rationale (e.g., "close the gripper around the pot-lid's handle to secure the grasp"). This reasoning trace constrains and grounds the subsequent action generation.

Example. Consider a high-level plan for cooking:
- *Plan*: Poise above pot-lid → Grasp pot-lid → Lift lid.
- *Subtask* $\tau^{(i)}$: Grasp pot-lid.
- *Reasoning* $r^{(i)}$: "Align the gripper with the pot-lid's handle and close it firmly to ensure a stable grasp."
- *Low-level actions* $A^{(i)}$: move up 22 steps, move left 23 steps, move down 23 steps, close gripper.

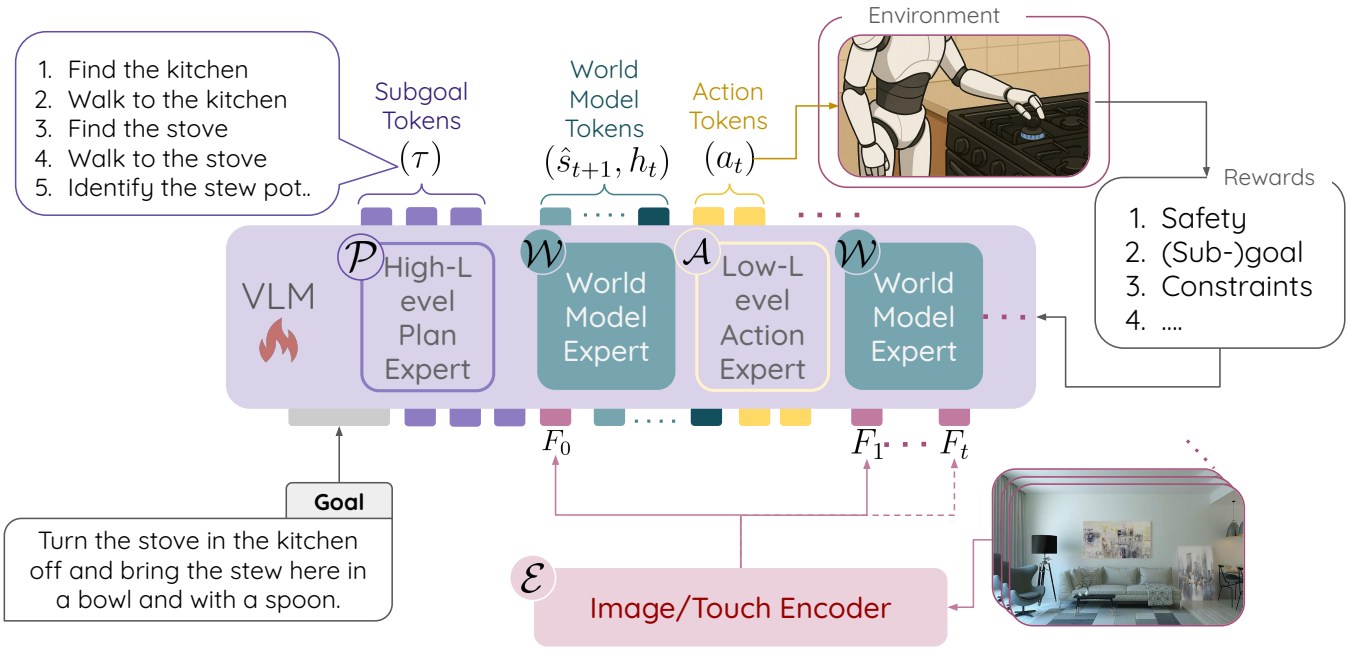

Figure 1: A high-level emerging VLA framework.

By explicitly generating $r^{(i)}$ before $A^{(i)}$, the policy links linguistic intent with motor execution, yielding more robust and interpretable action generation.

3.2 Improving Perception by Spatial Understanding

As discussed in Section 2.1, most VLA models are limited in their spatial understanding due to limited depth perception, being reliant on two-dimensional RGB-only images as input. To remedy this, the VLAs could be trained to understand and reason with the precise depth information obtained from depth-gauging cameras, LIDAR sensors, or multi-camera setup, should they be available. Approaches using depth estimation or depth-aware perception tokens (Lee et al. 2025; Qu et al. 2025), could only be used as a fallback option.

Unfortunately, most VLM backbones of VLAs lack inherent depth-aware spatial understanding. Thus, VLMs could be fine-tuned with both real and synthesized RGB-D frame data to inculcate depth-aware spatial understanding. Locate 3D (Arnaud et al. 2025) framework could be used to synthesize such RGB-D frames from RGB frames. Such data could be sourced from many Visual QA datasets, but the constituent queries may not necessarily require depth understanding. Thus, one may leverage Locate 3D to insert new objects from a predefined set into the frames and construct new QA pairs requiring depth awareness. Alternatively, powerful video generation platforms and world models like Veo3 could be leveraged to construct diverse scenes and form QA pairs from them. The RGB-D frames could be constructed from the moving camera positions—farther objects move slower than the closer objects in the frame.

Following a pre-training regimen on synthetic depth-containing visual data, the depth-aware VLM could be trained on synthetically constructed RGB-D frames from the Open-X-Embodiment (Collaboration et al. 2023) dataset. For standard depth-devoid RGB frames, one can keep a separate expert that generates estimated-depth-aware encodings of the frames like MolmoAct and SpatialVLA.

3.3 Universal Action Representation

As highlighted in Section 2.5, cross-robot action generalization remains a central challenge in VLA research. One promising direction is to learn unified atomic representations of actions. For instance, Zheng et al. (2025b) proposed learning universal atomic actions via a codebook and then decoding them into robot-specific actions through a decoder. Their results suggest that this approach significantly reduces the effort and data required to adapt to new robots. However, it still falls short of achieving true zero-shot generalization. We envision that this gap could be addressed by teaching VLA models about a new robot's action space through prompting—analogous to few-shot prompting that teaches LLMs new tasks, generally referred to as in-context learning. Realizing this capability, however, would require a fundamental shift in the pre-training paradigm of VLA models.

3.4 Modeling World Dynamics

World models are integral to robotics frameworks, as they model the cause-and-effect relationships essential for predicting the outcomes of actions. Recent studies have highlighted that current visual-language models (VLMs) often face challenges in executing plans. These difficulties may primarily arise from interpreting the current state, anticipating desired outcomes, and grounding both in the physical world to determine the required actions. These challenges underscore the need for explicit world models.

The two major approaches to world models are
1. **Generative modeling:** In this approach, a world model $\mathcal{W}$ takes a current state s and an action a, and predicts the resulting next state: $s' \leftarrow \mathcal{W}_\theta(s, a)$. Such a model can be initialized from a pretrained VLM or LLM. To enforce self-consistency, additional objectives—such as predicting the action that causally connects two states—can be included. These objectives are often formulated contrastively, promoting correct state predictions while demoting incorrect ones, and can be implemented via specialized prediction heads to avoid interfering with next-state prediction. Flow-based architectures are one possible instantiation of this paradigm.
2. **Embedding prediction:** This approach, proposed by Yann LeCun and realized in models such as JEPA (Dawid and LeCun 2023) and V-JEPA (Bardes et al. 2024), predicts the masked patches of visual frames as pre-training objective to learn the embeddings. V-JEPA-2 (Assran et al. 2025) extends this idea by post-training a hierarchical predictor: given the current observation/visual context and an action a_t, predict the latent embedding of the future frame. Such approach circumvents explicit pixel-level reconstruction by abstracting away irrelevant visual details, allowing efficient long-horizon prediction. This could be particularly suitable for robotics where reasoning about high-level state changes is more important than exact pixel reproduction.

3.5 Data Synthesis with Visual Generative Models

Large Language Models are trained on trillions of tokens that are readily available from internet-scale corpora. In contrast, collecting data at scale for Embodied AI models, including Vision-Language-Action (VLA) systems, is far more challenging, as it typically requires costly robot teleoperation. A common workaround is to use simulators and then mix simulated trajectories with a limited amount of real robot data for sim-to-real transfer. However, achieving robust sim-to-real generalization remains a long-standing challenge in robotics.

Recent advances in video generative models provide a promising alternative. Such models can synthesize videos of robots accomplishing tasks in diverse environments, where the environments themselves can be generated from textual prompts at scale. This opens new opportunities for VLA research, but introduces a key limitation: generated videos do not contain explicit robot action sequences. To address this, one can extract *latent actions* (Yang et al. 2025a) from video via world modeling, as detailed below.

Joint Learning from Video and Robot Data. We introduce a latent action variable z for world model $\mathcal{W}$. Concretely, we train an encoder E to infer $z_t = E(s_t, s_{t+1})$ from consecutive states, and optimize the world model to predict dynamics $s_{t+1} = \mathcal{W}_\theta(s_t, z_t)$. Since ground-truth future states are available in both simulated and video data, we can compute a reconstruction loss and backpropagate to jointly train E and $\mathcal{W}$. This formulation sidesteps the need for manual action annotations by grounding policy learning in a self-supervised latent space.

However, the latent representation z does not directly correspond to the real robot action space. To bridge this gap, we leverage datasets with ground-truth robot actions a_t that can be used to optimize E and $\mathcal{W}$, while jointly learning z_t. This joint learning can then be realized through
1. Minimizing the distance between the latent representation z_t and the true action a_t.
2. Encoding a_t into the latent space and ensuring $\mathcal{W}$ predicts s_{t+1} to either z_t or a_t as action input.
This hybrid training scheme allows video-derived latent actions to be aligned with real robot control signals, enabling scalable pretraining on synthetic video while remaining grounded in executable action spaces.

Video Generative Models and World Models as Simulators. Video generative models and world models such as V-JEPA-2 offer a powerful means of producing simulated data for training embodied agents. These models can be fine-tuned to predict future states conditioned on the current state and a sequence of actions. To generate large-scale simulated datasets, one can initialize from random frames and sample actions from a predefined distribution to produce corresponding future states. An evaluator can then filter the generated states to retain only those that satisfy specific subgoals, ensuring data quality. The resulting cleaned dataset can subsequently be used to train low-level action experts.

3.6 Post-Training

In LLM research, significant gains in System-II level intelligence and task performance have been achieved through extensive post-training using reinforcement learning. The key idea is that the model explore the solution space by generating multiple rollouts. A reward model then evaluates these rollouts based on criteria such as task completion, efficiency, and optimality. The reward signals are used to update the policy, enabling the model to gradually improve its action selection and favor strategies that achieve higher rewards. This process effectively combines exploration of possible actions with guided learning from feedback, allowing the model to discover increasingly effective behaviors. Extending this paradigm to Visual-Language-Action (VLA) models faces a fundamental challenge: how can we define and provide reward signals for these models?

A straightforward approach might involve running the VLA model N times to generate action sequences $A^{(i)}$, executing them in a simulator, and then using evaluators to assess the outcomes and assign rewards. However, this approach depends on the availability of highly accurate and efficient robot-specific simulators—a requirement that is often difficult and costly to meet.

Recent advances in world models and video generative models offer a promising alternative. These models can serve as implicit reward estimators by predicting the consequences of actions and evaluating whether desired subgoals are achieved. Leveraging such learned models as reward functions could enable scalable post-training of VLA policies without the need for fully engineered simulators, providing a practical path forward for reinforcement learning in embodied settings.

Improving the Action Expert through Rewards. Given N rollouts from the action expert, action-conditioned world

dynamics models can be used to predict the resulting future states. These states can then be compared to the ground-truth subgoal states using a suitable metric to obtain reward signals for the action expert. Alternatively, LLMs can assess if the predicted outcomes satisfy the intended subgoals, providing feedback to guide policy improvement. Note that one can also use different other rewards, both verifiable or unverifiable. Preference optimization techniques, such as, Direct Preference Optimization (DPO) (Rafailov et al. 2023), RL, and GRPO (Group Reward Preference Optimization) (Shao et al. 2024), can be adopted for improving the action expert.

Safety Check through Evaluators. Ensuring safety and task compliance is critical to deploying VLA models. One approach is to leverage action-conditioned world models or specialized evaluators—or reward models for post-training—as virtual guardrails to simulate the proposed actions and detect potential safety violations or deviations from task objectives or trigger human interventions. If a safety or compliance issue is identified, the VLA model can replan to reach the goal, while avoiding unsafe behaviors. Such evaluators are especially important in multi-agent settings, where interactions between autonomous VLA agents can introduce emergent risks. By incorporating these safety checks, the system can maintain reliable and controlled behavior even in complex, collaborative environments.

4 Conclusion

Vision-Language-Action (VLA) models are central to the development of Embodied AI. In this work, we revisit several key challenges in this area—such as multimodal sensing and perception, and cross-robot action generalization—that we believe will shape future research directions. In addition, we propose an exploratory framework aimed at addressing these challenges.

Acknowledgement

This work was supported in part by A*STAR SERC CRF funding to C.T. This work is also supported by the National Research Foundation, Singapore, under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-005), NTU SUG project #025628-00001:Post-training to Improve Embodied AI Agents.

References

Arnaud, S.; McVay, P.; Martin, A.; Majumdar, A.; Jatavallabhula, K. M.; Thomas, P.; Partsey, R.; Dugas, D.; Gejji, A.; Sax, A.; Berges, V.-P.; Henaff, M.; Jain, A.; Cao, A.; Prasad, I.; Kalakrishnan, M.; Rabbat, M.; Ballas, N.; Assran, M.; Maksymets, O.; Rajeswaran, A.; and Meier, F. 2025. Locate 3D: Real-World Object Localization via Self-Supervised Learning in 3D. arXiv:2504.14151.

Assran, M.; Bardes, A.; Fan, D.; Garrido, Q.; Howes, R.; Mojtaba; Komeili; Muckley, M.; Rizvi, A.; Roberts, C.; Sinha, K.; Zholus, A.; Arnaud, S.; Gejji, A.; Martin, A.; Hogan, F. R.; Dugas, D.; Bojanowski, P.; Khalidov, V.; Labatut, P.; Massa, F.; Szafraniec, M.; Krishnakumar, K.; Li, Y.; Ma, X.; Chandar, S.; Meier, F.; LeCun, Y.; Rabbat, M.; and Ballas, N. 2025. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. arXiv:2506.09985.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Bardes, A.; Garrido, Q.; Ponce, J.; Chen, X.; Rabbat, M.; LeCun, Y.; Assran, M.; and Ballas, N. 2024. Revisiting Feature Prediction for Learning Visual Representations from Video. arXiv:2404.08471.

Bi, J.; Ma, K. Y.; Hao, C.; Shou, M. Z.; and Soh, H. 2025. VLA-Touch: Enhancing Vision-Language-Action Models with Dual-Level Tactile Feedback. arXiv:2507.17294.

Black, K.; Brown, N.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Groom, L.; Hausman, K.; Ichter, B.; et al. 2024. π0: A vision-language-action flow model for general robot control, 2024. *URL https://arxiv. org/abs/2410.24164*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; Florence, P.; Fu, C.; Arenas, M. G.; Gopalakrishnan, K.; Han, K.; Hausman, K.; Herzog, A.; Hsu, J.; Ichter, B.; Irpan, A.; Joshi, N.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Leal, I.; Lee, L.; Lee, T.-W. E.; Levine, S.; Lu, Y.; Michalewski, H.; Mordatch, I.; Pertsch, K.; Rao, K.; Reymann, K.; Ryoo, M.; Salazar, G.; Sanketi, P.; Sermanet, P.; Singh, J.; Singh, A.; Soricut, R.; Tran, H.; Vanhoucke, V.; Vuong, Q.; Wahid, A.; Welker, S.; Wohlhart, P.; Wu, J.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yu, T.; and Zitkovich, B. 2023a. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *arXiv preprint arXiv:2307.15818*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; et al. 2023b. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Hsu, J.; Ibarz, J.; Ichter, B.; Irpan, A.; Jackson, T.; Jesmonth, S.; Joshi, N. J.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Leal, I.; Lee, K.-H.; Levine, S.; Lu, Y.; Malla, U.; Manjunath, D.; Mordatch, I.; Nachum, O.; Parada, C.; Peralta, J.; Perez, E.; Pertsch, K.; Quiambao, J.; Rao, K.; Ryoo, M.; Salazar, G.; Sanketi, P.; Sayed, K.; Singh, J.; Sontakke, S.; Stone, A.; Tan, C.; Tran, H.; Vanhoucke, V.; Vega, S.; Vuong, Q.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yu, T.; and Zitkovich, B. 2023c. RT-1: Robotics Transformer for Real-World Control at Scale. arXiv:2212.06817.

Chen, S.; Liu, J.; Qian, S.; Jiang, H.; Li, L.; Zhang, R.; Liu, Z.; Gu, C.; Hou, C.; Wang, P.; Wang, Z.; and Zhang, S. 2025.

AC-DiT: Adaptive Coordination Diffusion Transformer for Mobile Manipulation. *arXiv preprint arXiv:2507.01961*.

Chen, X.; Wang, X.; Beyer, L.; Kolesnikov, A.; Wu, J.; Voigtlaender, P.; Mustafa, B.; Goodman, S.; Alabdulmohsin, I. M.; Padlewski, P.; Salz, D. M.; Xiong, X.; Vlasic, D.; Pavetic, F.; Rong, K.; Yu, T.; Keysers, D.; Zhai, X.-Q.; and Soricut, R. 2023. PaLI-3 Vision Language Models: Smaller, Faster, Stronger. *arXiv preprint arXiv:2310.09199*.

Chi, C.; Feng, S.; Du, Y.; Xu, Z.; Cousineau, E.; Burchfiel, B.; and Song, S. 2023a. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*.

Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2023b. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 02783649241273668.

Collaboration, O. X.-E.; O'Neill, A.; Rehman, A.; Gupta, A.; Maddukuri, A.; Gupta, A.; Padalkar, A.; Lee, A.; Pooley, A.; Gupta, A.; Mandlekar, A.; Jain, A.; Tung, A.; Bewley, A.; Herzog, A.; Irpan, A.; Khazatsky, A.; Rai, A.; Gupta, A.; Wang, A.; Kolobov, A.; Singh, A.; Garg, A.; Kembhavi, A.; Xie, A.; Brohan, A.; Raffin, A.; Sharma, A.; Yavary, A.; Jain, A.; Balakrishna, A.; Wahid, A.; Burgess-Limerick, B.; Kim, B.; Schölkopf, B.; Wulfe, B.; Ichter, B.; Lu, C.; Xu, C.; Le, C.; Finn, C.; Wang, C.; Xu, C.; Chi, C.; Huang, C.; Chan, C.; Agia, C.; Pan, C.; Fu, C.; Devin, C.; Xu, D.; Morton, D.; Driess, D.; Chen, D.; Pathak, D.; Shah, D.; Büchler, D.; Jayaraman, D.; Kalashnikov, D.; Sadigh, D.; Johns, E.; Foster, E.; Liu, F.; Ceola, F.; Xia, F.; Zhao, F.; Frujeri, F. V.; Stulp, F.; Zhou, G.; Sukhatme, G. S.; Salhotra, G.; Yan, G.; Feng, G.; Schiavi, G.; Berseth, G.; Kahn, G.; Yang, G.; Wang, G.; Su, H.; Fang, H.-S.; Shi, H.; Bao, H.; Amor, H. B.; Christensen, H. I.; Furuta, H.; Bharadhwaj, H.; Walke, H.; Fang, H.; Ha, H.; Mordatch, I.; Radosavovic, I.; Leal, I.; Liang, J.; Abou-Chakra, J.; Kim, J.; Drake, J.; Peters, J.; Schneider, J.; Hsu, J.; Vakil, J.; Bohg, J.; Bingham, J.; Wu, J.; Gao, J.; Hu, J.; Wu, J.; Wu, J.; Sun, J.; Luo, J.; Gu, J.; Tan, J.; Oh, J.; Wu, J.; Lu, J.; Yang, J.; Malik, J.; Silvério, J.; Hejna, J.; Booher, J.; Tompson, J.; Yang, J.; Salvador, J.; Lim, J. J.; Han, J.; Wang, K.; Rao, K.; Pertsch, K.; Hausman, K.; Go, K.; Gopalakrishnan, K.; Goldberg, K.; Byrne, K.; Oslund, K.; Kawaharazuka, K.; Black, K.; Lin, K.; Zhang, K.; Ehsani, K.; Lekkala, K.; Ellis, K.; Rana, K.; Srinivasan, K.; Fang, K.; Singh, K. P.; Zeng, K.-H.; Hatch, K.; Hsu, K.; Itti, L.; Chen, L. Y.; Pinto, L.; Fei-Fei, L.; Tan, L.; Fan, L. J.; Ott, L.; Lee, L.; Weihs, L.; Chen, M.; Lepert, M.; Memmel, M.; Tomizuka, M.; Itkina, M.; Castro, M. G.; Spero, M.; Du, M.; Ahn, M.; Yip, M. C.; Zhang, M.; Ding, M.; Heo, M.; Srirama, M. K.; Sharma, M.; Kim, M. J.; Kanazawa, N.; Hansen, N.; Heess, N.; Joshi, N. J.; Suenderhauf, N.; Liu, N.; Palo, N. D.; Shafiullah, N. M. M.; Mees, O.; Kroemer, O.; Bastani, O.; Sanketi, P. R.; Miller, P. T.; Yin, P.; Wohlhart, P.; Xu, P.; Fagan, P. D.; Mitrano, P.; Sermanet, P.; Abbeel, P.; Sundaresan, P.; Chen, Q.; Vuong, Q.; Rafailov, R.; Tian, R.; Doshi, R.; Mart'in-Mart'in, R.; Baijal, R.; Scalise, R.; Hendrix, R.; Lin, R.; Qian, R.; Zhang, R.; Mendonca, R.; Shah, R.; Hoque, R.; Julian, R.; Bustamante, S.; Kirmani, S.; Levine, S.; Lin, S.; Moore, S.; Bahl, S.; Dass, S.; Sonawani, S.; Tulsiani, S.; Song, S.; Xu, S.; Haldar, S.; Karamcheti, S.; Adebola, S.; Guist, S.; Nasiriany, S.; Schaal, S.; Welker, S.; Tian, S.; Ramamoorthy, S.; Dasari, S.; Belkhale, S.; Park, S.; Nair, S.; Mirchandani, S.; Osa, T.; Gupta, T.; Harada, T.; Matsushima, T.; Xiao, T.; Kollar, T.; Yu, T.; Ding, T.; Davchev, T.; Zhao, T. Z.; Armstrong, T.; Darrell, T.; Chung, T.; Jain, V.; Kumar, V.; Vanhoucke, V.; Zhan, W.; Zhou, W.; Burgard, W.; Chen, X.; Chen, X.; Wang, X.; Zhu, X.; Geng, X.; Liu, X.; Liangwei, X.; Li, X.; Pang, Y.; Lu, Y.; Ma, Y. J.; Kim, Y.; Chebotar, Y.; Zhou, Y.; Zhu, Y.; Wu, Y.; Xu, Y.; Wang, Y.; Bisk, Y.; Dou, Y.; Cho, Y.; Lee, Y.; Cui, Y.; Cao, Y.; Wu, Y.-H.; Tang, Y.; Zhu, Y.; Zhang, Y.; Jiang, Y.; Li, Y.; Li, Y.; Iwasawa, Y.; Matsuo, Y.; Ma, Z.; Xu, Z.; Cui, Z. J.; Zhang, Z.; Fu, Z.; and Lin, Z. 2023. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. https://arxiv.org/abs/2310.08864.

Dawid, A.; and LeCun, Y. 2023. Introduction to Latent Variable Energy-Based Models: A Path Towards Autonomous Machine Intelligence. arXiv:2306.02572.

Driess, D.; Springenberg, J. T.; Ichter, B.; Yu, L.; Li-Bell, A.; Pertsch, K.; Ren, A. Z.; Walke, H.; Vuong, Q.; Shi, L. X.; and Levine, S. 2025. Knowledge Insulating Vision-Language-Action Models: Train Fast, Run Fast, Generalize Better. arXiv:2505.23705.

Du, W.; et al. 2024. An Efficient Representation of Whole-body Model Predictive Control for Dual-Arm Mobile Manipulators. *arXiv preprint arXiv:2410.22910*.

Dugar, P.; Shrestha, A.; Yu, F.; van Marum, B.; and Fern, A. 2024. Learning Multi-Modal Whole-Body Control for Real-World Humanoid Robots. *arXiv preprint arXiv:2408.07295*.

Fu, Z.; Zhao, T. Z.; and Finn, C. 2024. Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation. *arXiv preprint arXiv:2401.02117*.

Ha, H.; Gao, Y.; Fu, Z.; Tan, J.; and Song, S. 2024. UMI on Legs: Making Manipulation Policies Mobile with Manipulation-Centric Whole-body Controllers. *arXiv preprint arXiv:2407.10353*.

Heins, A.; Kim, Y.-J.; Bjelonic, M.; and Hutter, M. 2023. Keep it Upright: Model Predictive Control for Object Balancing with a Mobile Manipulator. *arXiv preprint arXiv:2305.17484*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.

Intelligence, P.; Black, K.; Brown, N.; Darpinian, J.; Dhabalia, K.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Galliker, M. Y.; Ghosh, D.; Groom, L.; Hausman, K.; Ichter, B.; Jakubczak, S.; Jones, T.; Ke, L.; LeBlanc, D.; Levine, S.; Li-Bell, A.; Mothukuri, M.; Nair, S.; Pertsch, K.; Ren, A. Z.; Shi, L. X.; Smith, L.; Springenberg, J. T.; Stachowicz, K.; Tanner, J.; Vuong, Q.; Walke, H.; Walling, A.; Wang, H.; Yu, L.; and Zhilinsky, U. 2025. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. arXiv:2504.16054.

Ji, M.; et al. 2024. Advanced Expressive Humanoid Whole-Body Control. *arXiv preprint arXiv:2412.13196*.

Kadian, A.; Truong, J.; Gokaslan, A.; Clegg, A.; Wijmans, E.; Lee, S.; Savva, M.; Chernova, S.; and Batra, D. 2020. Sim2Real Predictivity: Does Evaluation in Simulation Predict Real-World Performance? *IEEE Robotics and Automation Letters*, 5(4): 6670–6677.

Khazatsky, A.; Pertsch, K.; Nair, S.; Balakrishna, A.; Dasari, S.; Karamcheti, S.; Nasiriany, S.; Srirama, M. K.; Chen, L. Y.; Ellis, K.; et al. 2024. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*.

Kim, G.; Kang, D.; Kim, J.-H.; Hong, S.; and Park, H.-W. 2025. Contact-Implicit Model Predictive Control: Controlling Diverse Quadruped Motions Without Pre-Planned Contact Modes or Trajectories. *The International Journal of Robotics Research*. ArXiv:2312.08961.

Kim, M. J.; Finn, C.; and Liang, P. 2025. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success. arXiv:2502.19645.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E. P.; Sanketi, P. R.; Vuong, Q.; Kollar, T.; Burchfiel, B.; Tedrake, R.; Sadigh, D.; Levine, S.; Liang, P.; and Finn, C. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. In *8th Annual Conference on Robot Learning*.

Le Cleac'h, S.; Howell, T. A.; Yang, S.; Lee, C.-Y.; Zhang, J. Z.; Bishop, A. L.; Schwager, M.; and Manchester, Z. 2024. Fast Contact-Implicit Model Predictive Control. *IEEE Transactions on Robotics*, 40: 1617–1629.

Lee, J.; Duan, J.; Fang, H.; Deng, Y.; Liu, S.; Li, B.; Fang, B.; Zhang, J.; Wang, Y. R.; Lee, S.; Han, W.; Pumacay, W.; Wu, A.; Hendrix, R.; Farley, K.; VanderBilt, E.; Farhadi, A.; Fox, D.; and Krishna, R. 2025. MolmoAct: Action Reasoning Models that can Reason in Space. arXiv:2508.07917.

Li, X.; Hsu, K.; Gu, J.; Pertsch, K.; Mees, O.; Walke, H. R.; Fu, C.; Lunawat, I.; Sieh, I.; Kirmani, S.; Levine, S.; Wu, J.; Finn, C.; Su, H.; Vuong, Q.; and Xiao, T. 2024. Evaluating Real-World Robot Manipulation Policies in Simulation. *arXiv preprint arXiv:2405.05941*.

Lipman, Y.; Chen, R. T. Q.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2023. Flow Matching for Generative Modeling. arXiv:2210.02747.

Liu, B.; Zhu, Y.; Gao, C.; Feng, Y.; Liu, Q.; Zhu, Y.; and Stone, P. 2023. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. *arXiv preprint arXiv:2306.03310*.

Liu, B.; Zhu, Y.; Gao, C.; Feng, Y.; Liu, Q.; Zhu, Y.; and Stone, P. 2024a. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Ma, Y.; Song, Z.; Zhuang, Y.; Hao, J.; and King, I. 2024. A Survey on Vision-Language-Action Models for Embodied AI. *arXiv preprint arXiv:2405.14093*.

Octo Model Team; Ghosh, D.; Walke, H.; Pertsch, K.; Black, K.; Mees, O.; Dasari, S.; Hejna, J.; Xu, C.; Luo, J.; Kreiman, T.; Tan, Y.; Sanketi, P.; Vuong, Q.; Xiao, T.; Sadigh, D.; Finn, C.; and Levine, S. 2024. Octo: An Open-Source Generalist Robot Policy. In *Proceedings of Robotics: Science and Systems*. Delft, Netherlands.

OpenAI. 2023. GPT-4 Technical Report. arxiv:2303.08774.

Pertsch, K.; Stachowicz, K.; Ichter, B.; Driess, D.; Nair, S.; Vuong, Q.; Mees, O.; Finn, C.; and Levine, S. 2025. FAST: Efficient Action Tokenization for Vision-Language-Action Models. arXiv:2501.09747.

Qu, D.; Song, H.; Chen, Q.; Yao, Y.; Ye, X.; Ding, Y.; Wang, Z.; Gu, J.; Zhao, B.; Wang, D.; et al. 2025. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Model. *arXiv preprint arXiv:2501.15830*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Sun, Q.; Hong, P.; Pala, T. D.; Toh, V.; Tan, U.-X.; Ghosal, D.; and Poria, S. 2024. Emma-X: An Embodied Multimodal Action Model with Grounded Chain of Thought and Look-ahead Spatial Reasoning. arXiv:2412.11974.

Sundaresan, P.; Malhotra, R.; Miao, P.; Yang, J.; Wu, J.; Hu, H.; Antonova, R.; Engelmann, F.; Sadigh, D.; and Bohg, J. 2025. HoMeR: Learning In-the-Wild Mobile Manipulation via Hybrid Imitation and Whole-Body Control. *arXiv preprint arXiv:2506.01185*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. LLaMA: Open and Efficient Foundation Language Models. arxiv:2302.13971.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, Y.; Chen, R.; and Zhao, M. 2025. Whole-Body Model Predictive Control for Mobile Manipulation with Task Priority Transition. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Xie, W.; et al. 2025. KungfuBot: Physics-Based Humanoid Whole-Body Control for Highly-Dynamic Skills. *arXiv preprint arXiv:2506.12851*.

Yang, J.; Shi, Y.; Zhu, H.; Liu, M.; Ma, K.; Wang, Y.; Wu, G.; He, T.; and Wang, L. 2025a. CoMo: Learning Continuous Latent Motion from Internet Videos for Scalable Robot Learning. arXiv:2505.17006.

Yang, Z.; Chen, Y.; Zhou, X.; Yan, J.; Song, D.; Liu, Y.; Li, Y.; Zhang, Y.; Zhou, P.; Chen, H.; and Sun, L. 2025b. Agentic Robot: A Brain-Inspired Framework for Vision-Language-Action Models in Embodied Agents. arXiv:2505.23450.

Zawalski, M.; Chen, W.; Pertsch, K.; Mees, O.; Finn, C.; and Levine, S. 2024. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid Loss for Language Image Pre-Training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11941–11952. IEEE.

Zhang, B.; Zhang, Y.; Ji, J.; Lei, Y.; Dai, J.; Chen, Y.; and Yang, Y. 2025. SafeVLA: Towards Safety Alignment of Vision-Language-Action Model via Safe Reinforcement Learning. *arXiv preprint arXiv:2503.03480*.

Zhao, J.; Teng, T.; De Momi, E.; and Ajoudani, A. 2025a. Reactive Whole-body Locomotion-Integrated Manipulation Based on Combined Learning and Optimization. *Machine Intelligence Research*, 22: 627–640.

Zhao, Q.; Lu, Y.; Kim, M. J.; Fu, Z.; Zhang, Z.; Wu, Y.; Li, Z.; Ma, Q.; Han, S.; Finn, C.; et al. 2025b. CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models. *arXiv preprint arXiv:2503.22020*.

Zheng, C.; Li, Y.; Song, Z.; Bi, Z.; Zhou, J.; Zhou, B.; and Ma, J. 2025a. Local Reactive Control for Mobile Manipulators with Whole-Body Safety in Complex Environments. *arXiv preprint arXiv:2501.02815*.

Zheng, J.; Li, J.; Liu, D.; Zheng, Y.; Wang, Z.; Ou, Z.; Liu, Y.; Liu, J.; Zhang, Y.-Q.; and Zhan, X. 2025b. Universal actions for enhanced embodied foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22508–22519.