

An Anatomy of Vision-Language-Action Models: From Modules to Milestones and Challenges

Chao Xu, Suyu Zhang, Yang Liu, Baigui Sun, Weihong Chen, Bo Xu, Qi Liu, Juncheng Wang, Shujun Wang, Shan Luo, Jan Peters, Athanasios V. Vasilakos, Stefanos Zafeiriou, Jiankang Deng

Abstract—Vision-Language-Action (VLA) models are driving a revolution in robotics, enabling machines to understand instructions and interact with the physical world. This field is exploding with new models and datasets, making it both exciting and challenging to keep pace with. This survey offers a clear and structured guide to the VLA landscape. We design it to follow the natural learning path of a researcher: we start with the basic Modules of any VLA model, trace the history through key Milestones, and then dive deep into the core Challenges that define recent research frontier. Our main contribution is a detailed breakdown of the five biggest challenges in: (1) Representation, (2) Execution, (3) Generalization, (4) Safety, and (5) Dataset and Evaluation. This structure mirrors the developmental roadmap of a generalist agent: establishing the fundamental perception-action loop, scaling capabilities across diverse embodiments and environments, and finally ensuring trustworthy deployment—all supported by the essential data infrastructure. For each of them, we review existing approaches and highlight future opportunities. We position this paper as both a foundational guide for newcomers and a strategic roadmap for experienced researchers, with the dual aim of accelerating learning and inspiring new ideas in embodied intelligence. A live version of this survey, with continuous updates, is maintained on our [project page](#).

Index Terms—Vision-Language-Action Model, Artificial Intelligence, Embodied Intelligence, Robotics, Foundation models

arXiv:2512.11362v3 [cs.RO] 19 Dec 2025

1 INTRODUCTION

The quest for general-purpose robots that can operate in real-world human environments is a central goal of artificial intelligence. In recent years, a new approach has emerged as one of the most promising paths toward this goal: Vision-Language-Action (VLA) models. By connecting vision, language, and physical action, these models have catalyzed rapid progress, making the field of embodied intelligence both exciting and increasingly complex.

To help navigate this rapidly growing landscape, numerous survey papers have recently emerged, covering the field from various perspectives. On the one hand, several works provide focused, in-depth reviews on specific technical subareas, such as action tokenization [1], efficient training paradigms [2], and post-training methodologies [3], offering granular insights into individual system components. On the other hand, broader surveys [4]–[9] offer comprehensive system overviews. These works typically serve as structured taxonomies, organizing the VLA landscape by model architectures, input modalities, or training objectives, providing readers with a systematic list of the core components.

C. Xu, S. Zhang, Y. Liu, B. Sun, W. Chen, B. Xu, Q. Liu are with IROOTECH TECHNOLOGY (e-mail: chaoxuc@gmail.com, sunbaigui85@gmail.com).

C. Xu, S. Zhang, Y. Liu, B. Sun are with Wolf 1069 b Lab, Sany Group.

Y. Liu and S. Luo are with the Department of Engineering, King's College London (e-mail: yang.15.liu@kcl.ac.uk, shan.luo@kcl.ac.uk).

J. Wang and S. Wang are with the Hong Kong Polytechnic University (e-mail: wjc2830@gmail.com, shu-jun.wang@polyu.edu.hk).

J. Peters is with the Computer Science Department of the Technische Universität Darmstadt (e-mail: peters@ias.tu-darmstadt.de).

A. Vasilakos is with Department of ICT and Center for AI Research, University of Agder (UiA) (e-mail: th.vasilakos@gmail.com).

S. Zafeiriou and J. Deng are with the Department of Computing, Imperial College London (e-mail: j.deng16@imperial.ac.uk, s.zafeiriou@imperial.ac.uk).

C. Xu, S. Zhang and Y. Liu contributed equally to this work.

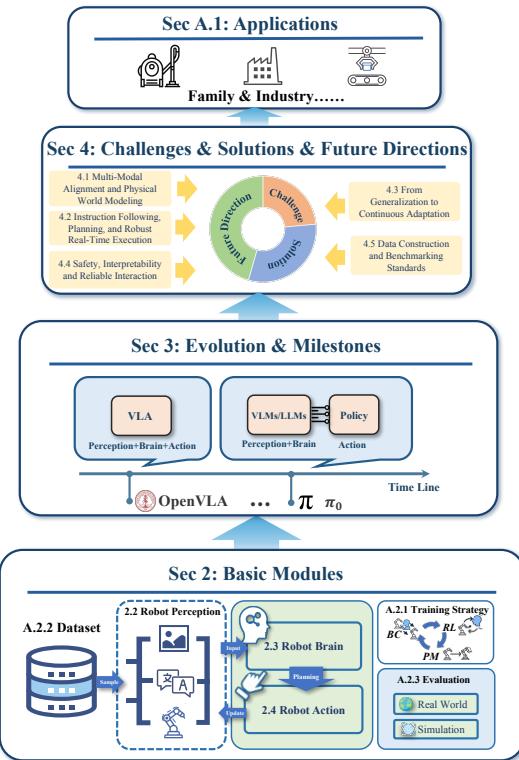


Fig. 1: The structure of this survey in a pyramid format. Section 2 lays the foundational knowledge by deconstructing the core components of any VLA model. Building upon this, the second stage, Section 3, traces the historical evolution of the field through its most representative works, providing context and intuition. The deepest stage, Section 4, serves as the intellectual core, offering an in-depth analysis of the grand open problems and outlining actionable future research directions. The final section depicts the various applications, which are included in Appendix A.1.

However, we identify two key gaps that this survey aims to address. First, existing surveys often relegate research challenges to a concluding section—a high-level overview appended at the end of the paper. The field still lacks a unified resource that places these challenges at its core, systematically breaking them down, comparing alternative solution paths, and charting clear directions for future work. For researchers aiming to make novel contributions, a mere list of problems is insufficient; what is needed is a deep, structured analysis of the problem space. Second, the structure of most surveys does not align how researchers learn a new field. Most existing works simply list and group methods by category—like grouping visual-based approaches in one chapter and control strategies in another. While this facilitates quick reference, it presents a fragmented view of the field. It provides extensive information but fails to illustrate how these pieces integrate into a coherent, evolving research timeline. Consequently, such surveys do not guide newcomers from foundational concepts to recent breakthroughs along a clear, progressive learning trajectory.

This survey makes two core contributions to address these gaps. Our primary contribution is a deep and systematic analysis of the core challenges in VLA research. Rather than appearing as a brief concluding section, our challenge analysis forms the central pillar of this survey. We identify five key challenges following the developmental roadmap of VLA: (1) Multi-Modal Alignment and Physical World Modeling, (2) Instruction Following, Planning, and Robust Real-Time Execution, (3) From Generalization to Continuous Adaptation, (4) Safety, Interpretability, and Reliable Interaction, (5) Data Construction and Benchmarking Standards. For each, we provide an in-depth review of competing solutions and outline concrete avenues for future research. Our goal is twofold: to help researchers efficiently navigate the vast landscape of existing work and to position this section as a direct catalyst for novel research ideas.

Our second contribution is the unique structure of this survey, designed to mirror the natural learning journey of a researcher. We intentionally structure this survey as a step-by-step roadmap. We begin with a detailed breakdown of the foundational *Modules* that constitute any VLA model, establishing a shared vocabulary. We then trace the historical evolution through key *Milestones*, providing context for how the field has arrived at its current state. This journey culminates in our deep dive into the core *Challenges*, demonstrating recent trends and pointing out future directions. This structure allows newcomers to build expertise from the ground up, while allowing experienced researchers to access the sections most relevant to their interests. The structure of this survey is illustrated in Fig. 1. This work is designed as a living resource, and [project page](#) will be *continuously* updated to reflect advances at the research frontier.

2 BASIC MODULES

2.1 Overall and Architectural Trend

Vision-Language-Action (VLA) systems integrate perception, reasoning, and control to translate abstract instructions into physical actions. Typically, a VLA system comprises three core modules: the perception module extracts grounded observations, the brain module fuses multimodal

inputs for planning, and the action module executes motor commands. Recently, these components are undergoing a fundamental shift: Perception (Sec. 2.2) is evolving from standard visual backbones to Language-Aligned Transformers (e.g., SigLIP) to bridge the semantic gap, increasingly augmented by geometric representations (e.g., DINOv2) to ensure manipulation precision. The Brain (Sec. 2.3) is converging toward pre-trained VLMs, leveraging internet-scale knowledge to enable zero-shot generalization and unified token processing. Finally, Action (Sec. 2.4) is pivoting from discrete tokenization towards continuous generative modeling (e.g., Diffusion), achieving smooth, multi-modal distribution modeling. Notably, to prioritize the in-depth analysis of challenges (Sec. 4), we provide a streamlined overview here due to limited space. For detailed architectural taxonomies, we recommend other specialized surveys [4], [8].

2.2 Robot Perception

2.2.1 Vision Encoders in VLA

(1) Convolutional Networks (CNNs). CNNs [10] remain indispensable in VLA due to their strong local feature extraction and translation equivariance, making them effective visual encoders in real-time and resource-constrained settings. Modern architectures such as ResNet and EfficientNet [11], [12] are widely adopted. CNNs commonly serve as visual backbones in end-to-end policies by encoding RGB or depth observations into compact features for downstream decision-making; representative systems such as Diffusion Policy [13] and SPECI [14] use ResNet-based encoders. CNNs also integrate naturally into hierarchical designs, where lightweight models handle high-frequency perception, as in HiRT [12], [15], which employs EfficientNet-B3. As world-model-based VLA frameworks grow in complexity, CNNs increasingly act as compact encoders for high-dimensional observations. For example, LUMOS [16] uses a CNN front-end to produce latent features consumed by RSSM [17] for prediction and planning.

(2) Vision Transformers (ViT). ViT [18] and its variants have become the dominant perception backbone in modern VLA systems. Their self-attention captures global context and long-range dependencies, and patch tokenization aligns visual inputs with Transformer-based language models, making ViT well suited for end-to-end VLA pipelines [19]. Contemporary VLA frameworks therefore rely heavily on large-scale pretrained ViT encoders, typically fine-tuned for stronger generalization and efficiency. ViT-based visual encoders in VLA generally follow four structural paradigms:

a) Language-Supervised Visual Encoders. Models such as CLIP [20] and SigLIP [21] learn vision features aligned to human semantics via contrastive learning from internet-scale image–text pairs. Adopting such encoders is now standard practice: for example, π_0 [22], RDT-1B [23], TriVLA [24], and ForceVLA [25] use SigLIP as their vision backbone, while many others rely on CLIP, e.g., DeeRVLA [26], RationalVLA [27], MinD [28]. Some works innovate in their use, for instance, OTTER [29] extracts features from the final layer of a frozen CLIP ViT to obtain strongly language-aligned visual representations.

b) Self-Supervised Visual Encoders. These models, exemplified by DINOv2 [30], avoid textual labels and learn

robust visual representations from large unlabeled corpora, capturing fine-grained geometry and spatial structure that make them particularly effective for contact-rich manipulation tasks requiring precise physical cues. For example, LexVLA [31] employs a frozen DINOv2 encoder and a lightweight adapter to map local visual features into sparse, language-aligned lexical representations.

c) Hybrid Architectures. To combine the semantic strengths of language-supervised encoders with the geometric precision of self-supervised ones, an increasingly common strategy is to adopt a hybrid approach. Recent VLA frameworks, including OpenVLA [32], OpenVLA-OFT [33], GraspVLA [34], UniVLA [35], and VLA-RL [36] often employ a SigLIP+DINOv2 hybrid to attain strong performance on both semantic understanding and geometric reasoning.

d) Vision-Language Models (VLMs). The most integrated paradigm directly adopts pretrained VLMs as high-level visual encoders, producing language-conditioned visual embeddings rather than raw pixel features. Examples include PaLI-X [37] in RT-H [38]; PaliGemma [39] in Hume [40] and Hi Robot [41]; Qwen-VL [42] in VTLA [43], CombatVLA [44], and OpenHelix [45], which leverage VLMs' fused vision-language context to provide higher-level inputs for policy and planning.

2.2.2 Language Encoders in VLA

Language instructions form the semantic core of VLA systems, defining task objectives and providing high-level context. The language encoder has evolved alongside advances in natural language processing, and in practice falls into three main categories:

(1) Transformer-Based Language Encoders. The earliest approach involves the use of standard Transformer-based language encoders. These VLA systems adopt text-only Transformers (e.g., BERT [46], T5 [47]) pretrained on large corpora to encode instructions, providing strong semantics as the entry point of the control stack. Classical examples include RDT-1B [23] with T5-XXL [47], RoboBERT [48] with BERT, and early partial implementations of Octo [49] that rely on such modules.

(2) Large Language Models (LLMs). With the rise of LLMs, VLA systems increasingly adopt billion-parameter models as their language backbone, leveraging their richer world knowledge and commonsense reasoning to interpret ambiguous and compositional instructions. Representative choices include Llama-family models (e.g., OpenVLA-OFT [33], VLA-RL [36] with Llama-2 [50] 7B; HiRT [15] with Llama-based InstructBLIP [51]), Gemma-family models [52] (e.g., π_0 [22]/ $\pi_{0.5}$ [53] with Gemma 2B [54]), and InternLM2 [55] (e.g., GraspVLA [34]).

(3) Vision-Language Models (VLMs). The recent trend is to adopt native VLMs, where the language module is no longer a standalone component but is jointly pre-trained with vision for end-to-end multimodal understanding. For example, several VLA systems explicitly adopt well-known vision-language models: DeeR-VLA [26] and RoboFlamingo [56] build on OpenFlamingo [57]; Diffusion-VLA [58] instead employs Qwen-VL [42] while MemoryVLA [59] is developed upon the 7B Prismatic VLM [60]. In contrast, Dexbotic [61] pretrains its own dedicated model, DexboticVLM, tailored to dexterous manipulation. Other

systems including InstructVLA [62], FlowVLA [63], and others, also utilize VLMs as their language encoders.

2.2.3 Proprioceptive Encoders in VLA

Proprioceptive inputs are provided by onboard sensors and typically include (i) joint states: per-joint position, velocity, and effort/torque; (ii) end-effector states: the 6-DoF pose (x, y, z , roll, pitch, yaw), optionally with linear/angular velocities, in the world or base frames [64], [65]; and (iii) gripper status: opening width/state and applied force. These data are low-dimensional, structured vectors.

Given the low-dimensional, structured nature of proprioception, MLPs are the standard, efficient encoders, whose outputs are fused with vision and language via concatenation or conditioning (e.g., FiLM [66]). Many VLA models follow this design. TriVLA [24] employs an embodiment-specific MLP, RDT-1B [23] encodes low-D robot states with an MLP, SPECI [14] trains an MLP from scratch on joint angles and gripper states, and systems such as OpenVLA-OFT [33] and the GR series [67], [68] similarly include MLP modules for proprioceptive fusion.

2.3 Robot Brain

The robot brain is the core of a VLA system, responsible for fusing multimodal representations from input modules, performing reasoning and planning, and ultimately generating action intentions. Current architectures primarily follow four mainstream technical directions:

(1) Transformer. The Transformer serves as a core VLA architecture by tokenizing vision, language, and proprioception inputs and using self-attention to fuse multimodal tokens and learn an end-to-end perception-to-action mapping. A Generalist Agent [69] demonstrates the capacity of a decoder-only Transformer to handle multiple modalities and tasks, and models such as VIMA [70] and GR-1/GR-2 [67], [68] further adopt Transformer-based generalist policies. Other approaches, such as SPECI [14], apply temporal Transformers across both high-level reasoning and low-level execution. Beyond Transformers, recent alternatives also emerge; RoboMamba [71] adapts the Mamba [72] architecture to VLA for more efficient long-sequence processing.

(2) Diffusion Transformer (DiT). Unlike Transformer-only policies that predict actions directly, this paradigm uses a diffusion model as the generative core, with a Transformer guiding the denoising process. Diffusion models are well suited for robot control because they model complex continuous distributions and produce smooth, natural motion trajectories. Diffusion Policy [13] provides early evidence of the effectiveness of denoising-based generation, helping establish diffusion as a strong policy-learning paradigm. More recent methods, such as RDT-1B [23] and TriVLA [24], integrate diffusion on top of Transformer backbones to map semantics to actions through multi-step denoising.

(3) Hybrid Architectures. These models pair Transformer-based semantic reasoning with a diffusion [73] or flow-matching [74] head for high-frequency, smooth control. π_0 [22] exemplifies this design by using a pretrained VLM as the Transformer backbone for perception and a separate Flow Matching head for action generation. Octo [49] and

ConRFT [75] follow a similar pattern, combining a Transformer backbone with a generative action head. Diffusion-VLA [58] injects LLM reasoning into the diffusion process to coordinate high-level planning with low-level execution. MinD [28] adopts a hierarchical hybrid structure, using distinct diffusion models for low-frequency video prediction and high-frequency action control.

(4) Vision-Language Models (VLMs). This paradigm treats a full pretrained vision-language model as the core robot brain, leveraging its perception, multimodal fusion, commonsense reasoning, and sequence modeling, while integrating robot-specific proprioception and action spaces on top. RT-2 [76] is a milestone in this direction, extending the VLM’s (i.e., PaLI-X [37]/PaLM-E [77]) output space to include action tokens, effectively creating an embodied agent. Nearly all current SOTA VLA models, including OpenVLA [32], $\pi_{0.5}$ [53], CoT-VLA [78], SafeVLA [79], DeeRVLA [26], GraspVLA [34], VTLA [43], UniVLA [35], VLA-RL [36], WorldVLA [80], TraceVLA [81], PointVLA [82], 3DVLA [83], and BridgeVLA [84] build their decision-making on strong pretrained VLMs. In hierarchical systems, VLMs often act as high-level planners or span both high- and low-level policies, as in A Dual Process VLA [85], Hi Robot [41], HAMSTER [86], and HiRT [15].

2.4 Robot Action

Robot action is the VLA system’s final execution interface, translating abstract decisions from the robot brain into concrete, low-level control commands. Its design directly determines action precision, smoothness, real-time performance, and generalization.

2.4.1 Action Representation

Action space representation defines the target language that the model predicts. Representing typically high-dimensional, continuous robot actions involves a key trade-off between performance and learnability.

(1) Discrete Spaces. Continuous controls are discretized into bins and cast as a next-token classification problem, naturally reusing Transformer stacks for sequence prediction. This is common in generalist Transformer agents (e.g., A Generalist Agent [69], VIMA [70], RT-H [38], SafeVLA [79]) and in many recent VLA systems (e.g., UniVLA [35], VLA-RL [36], WorldVLA [80], TraceVLA [81], CombatVLA [44]).

(2) Continuous Spaces. Actions are regressed directly in normalized continuous domains (e.g., joint angles, end-effector velocities), yielding smoother, higher-precision control at the cost of high demands on model learning ability. This aligns naturally with diffusion or flow-matching policies (e.g., Diffusion Policy [13], TriVLA [24], RDT-1B [23], π_0 [22]) and with continuous variants of prior discrete models (e.g., OpenVLA-OFT [33]). Other systems such as iRe-VLA [87], GraspVLA [34], and Hume [40] also adopt continuous control.

(3) Hybrid Spaces. To combine strengths, hybrids assign discrete and continuous encodings to different control facets: BridgeVLA [84] uses continuous translation with discretized rotation. HiRT [15] treats EE pose as continuous while gripper open/close is discrete. Hierarchical models often keep high-level skills discrete and low-level execution continuous (e.g., Hi Robot [41], HAMSTER [86], $\pi_{0.5}$ [53]).

2.4.2 Action Decoding

(1) Autoregressive Decoding. In autoregressive (AR) decoding, the policy emits actions step by step with causal masking, and each prediction conditions on all previously generated actions and observations, enabling modeling of long-range temporal dependencies. AR remains standard in early and many recent VLA models (e.g., A Generalist Agent [69], VIMA [70], RT-H [38], SafeVLA [79], GR-2 [68], 3DVLA [83], UniVLA [35], OpenVLA [32], TraceVLA [81], CombatVLA [44]).

(2) Non-Autoregressive Decoding. To reduce latency, non-AR decoders predict an action horizon in one or a few passes. One path replaces causal attention with bidirectional attention to infer all steps jointly (e.g., OpenVLA-OFT [33]). Another uses inherently non-AR generators such as diffusion or flow matching that iteratively denoise or transform the whole sequence in parallel (e.g., Diffusion Policy [13], TriVLA [24], RDT-1B [23], π_0 [22], RoboBERT [48], Hume [40], Deer-VLA [26]).

(3) Hybrid Decoding. A practical compromise is chunking: the policy operates autoregressively over coarse time (emitting chunks), but non-autoregressively within each chunk (parallel refinement), which improves both stability and throughput. A representative example is $\pi_{0.5}$ [53], which performs AR semantic decisions with parallel low-level chunk generation. CoT-VLA [78], UniVLA [35], and WorldVLA [80] follow the same design, which support long-horizon coherence with efficient local rollout.

3 EVOLUTION & MILESTONES

The evolution of Vision-Language-Action (VLA) models is driven by the need to overcome the brittleness of traditional modular pipelines and achieve the broad generalization seen in foundation models. This evolution reflects a steady shift from passive multimodal perception to active, embodied reasoning and control. An overview of VLA milestones is shown in Fig. 2 and Appendix Tab. S3.

From 2017 to 2019, the Vision-and-Language Navigation (VLN) benchmark [88] pioneers large-scale evaluation of agents aligning linguistic instructions with visual environments for physical navigation. EmbodiedQA [89] advances this direction by defining embodied intelligence through a closed perception-action loop, establishing an early theoretical foundation. Follow-up work such as BabyAI [90], RCM [91], and Point-Cloud EQA [92] further refine the paradigm by improving language-to-action learning and introducing early forms of 3D geometric reasoning.

The period from 2020 to 2021 marks a shift toward *long-horizon reasoning and language-conditioned embodied control*. ALFRED [93] introduces the first interactive benchmark combining high-level goals, step-by-step instructions, and object-environment interactions, establishing realistic long-horizon tasks. ALFWORLD [94] extends this direction by linking symbolic reasoning with visually grounded execution, and BEHAVIOR [95] standardizes long-horizon household evaluation in high-fidelity simulation. A pivotal milestone of this era is CLIPORT [96], which integrates pretrained visual representations into a language-conditioned policy, demonstrating that internet-scale knowledge enables zero-shot generalization in robotic manipulation.

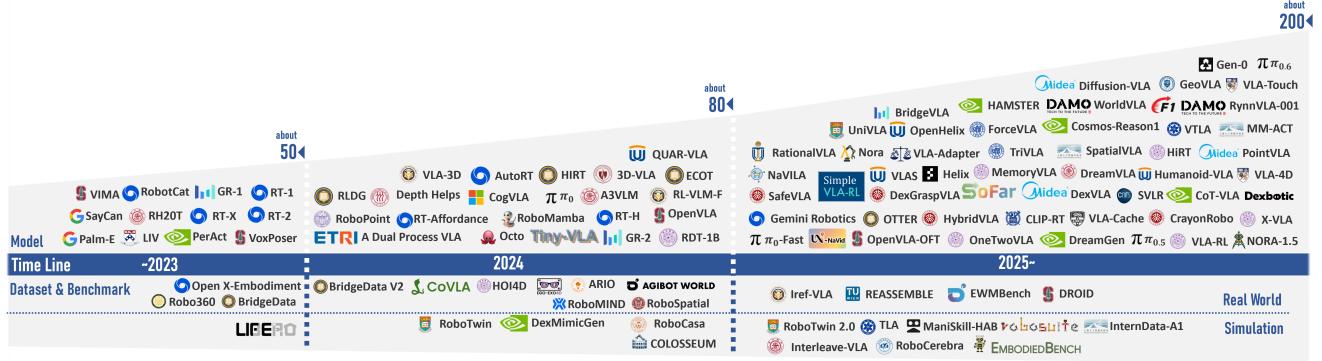


Fig. 2: The timeline of VLA models, datasets, and evaluation benchmarks from 2022 to 2025. The top row presents major VLA models introduced each year. The bottom row displays key datasets used to train and benchmarks to evaluate these models, grouped by release year.

Since 2022, VLA enters the era of *large models and generalized learning*. SayCan [97] is the first to introduce a hierarchical framework that separates LLM-based high-level planning from low-level skill execution, using affordance and value estimates from the robot to ground candidate subtasks and select feasible actions. Inner Monologue [98] for the first time embeds language models within continuous multimodal feedback loops, achieving self-reflection and dynamic behavioral adjustment. RT-1 and RT-2 [76], [99] realize end-to-end learning from vision and language to action via Transformer architectures, marking the birth of a truly unified VLA framework.

In 2023, multiple advances emerge, most notably in *unified multimodal backbones, generative action modeling, and cross-embodiment data scaling*. PaLM-E [77] embeds visual and state representations directly into pretrained LLMs, achieving for the first time a unified multimodal input space. The introduction of Diffusion Policy [13] applies generative diffusion models to action modeling, bringing greater stability and expressiveness to high-dimensional continuous control, and marking a key paradigm shift in policy generation for VLA. Open X-Embodiment [100] represents a meaningful milestone in robotic learning, providing large-scale and diverse cross-robot data with open access, and driving the field toward more general and powerful embodied models.

Building on the previous year's breakthroughs, 2024 broadens the frontier across *open-source scaling, generalist policies, flow/denoising action generation, web-scale video pre-training, and 3D world modeling*. Octo [49] establishes a generalist policy capable of cross-platform, multi-task control. OpenVLA [32] becomes the first fully open-source 7B VLA model, lowering the barrier for large-scale research and deployment. π_0 [22] is the first to combine pretrained VLMs with flow-matching action generation, setting a new architectural reference point for general and precise control. GR-2 [68] systematizes web-scale generative video pretraining for VLA, enabling broad generalization without proportional robot labels. 3D-VLA [83] marks a shift toward full 3D world modeling by coupling a generative 3D world model with VLA for plan-by-imagination.

By 2025, VLA research enters a stage of pluralistic evolution, where diverse embodiments, modalities, and learning paradigms co-evolve toward *general robotic intelligence*. Humanoid-VLA [101] and GR00T N1 [102] extend VLA

to full-body humanoid control. Another direction targets open-world autonomy, emphasizing deeper understanding and reasoning. PointVLA [82] injects point-cloud features without retraining the core model, enabling faithful 3D understanding for open-world settings. Cosmos-Reason1 [103] is the first to standardize physically grounded reasoning for VLAs, unifying ontologies and benchmarks into an open reasoning pipeline and shifting the field toward plug-and-play, physics-constrained reasoning. CoT-VLA [78] introduces the first explicit visual chain-of-thought, predicting subgoal images as intermediate reasoning before action generation. At the core, some models aim to unify prior advances by integrating hierarchy, reasoning, and control. $\pi_{0.5}$ [53] unifies high-level reasoning and low-level control via hierarchical Transformers, enabling long-horizon operation without target-specific robot data. LUMOS [16] integrates a learned world model with on-policy RL into a single system. VLA-RL [36] scales online RL to pretrained VLAs, addressing imitation learning's OOD limitations. GEN-0 [104] offers early evidence for scaling laws in robotics, showing that large-scale interaction data enables phase transitions in cross-embodiment generalization.

4 CHALLENGES & SOLUTIONS & FUTURE DIRECTIONS

Fig. 3 provides an overview of the five core challenges addressed in this section, along with their respective sub-challenges and the relevant papers involved.

4.1 Multi-Modal Alignment and Physical World Modeling

Fig. 4 illustrates the three levels of this challenge, which are elaborated in detail below.

4.1.1 The GAP between Semantics, Perception, and Physical Interaction

Vision-Language-Action (VLA) tasks center on three core components: vision for perceiving the world, language for conveying high-level instructions, and action for interacting with the physical environment. Together, they form an integrated embodied framework linking perception, reasoning, and execution. The central challenge is bridging the gap

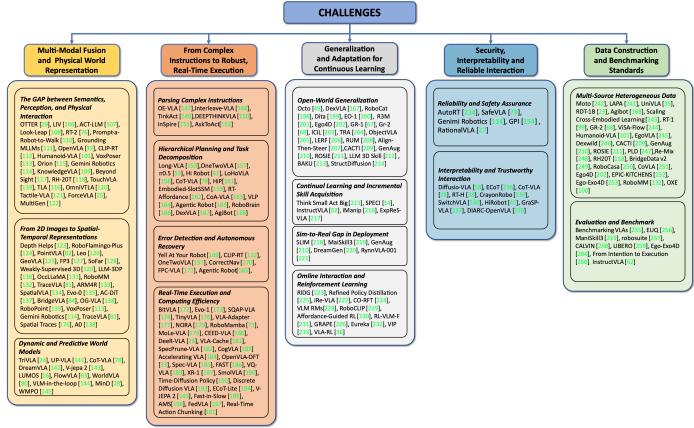


Fig. 3: **Taxonomy of VLA challenges**, encompassing 5 primary challenges and 15 sub-challenges, with representative works listed. Please zoom in for more details.

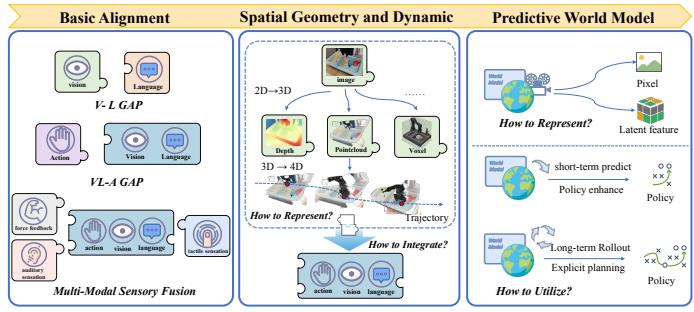


Fig. 4: **The challenge of Multi-Modal Alignment and Physical World Modeling**. First, Section 4.1.1 addresses the fundamental disalignment at the interface of information. Building upon this, Section 4.1.2 focuses on the construction of the world’s geometric and dynamic structure. Section 4.1.3 represents the highest level of understanding, as embodied in dynamic predictive capabilities.

between abstract semantics and grounded physical reality, which can be decomposed into three subproblems:

(1) Vision Language Gap. Vision provides high-dimensional perceptual input, while language offers abstract symbolic semantics. Establishing a precise mapping between these distinct modalities is essential for grounding visual understanding and goal reasoning in the physical world [45], [105]. Some approaches address this challenge by *enhancing visual representations* to make them more responsive to language conditioning. OTTER [29] introduces text-aware feature extraction that preserves semantics aligned with task descriptions, while LIV [106] employs a contrastive framework on robot-control data to construct a joint vision–language embedding space, enabling visual features to become inherently sensitive to linguistic cues. A recent paradigm bridges the vision–language gap via *symbolic reasoning* with natural language as an intermediate representation, powered by LLMs. ACT-LLM [107] translates visual observations into structured state descriptions for symbolic reasoning. Look Leap [108] pushes this further by generating full structured action plans, elevating vision–language alignment to a higher cognitive level and reframing the problem as one of reasoning.

(2) Vision-Language Action Gap. Although multimodal models achieve strong perception–semantics alignment, a

gap remains when grounding this understanding into physical action [109]. One direction is *end-to-end fine-tuning*, which reformulates control as sequence generation by discretizing the action space into tokens and fine-tuning a VLM to generate these action tokens in the same way it generates words. RT-2 [76] demonstrates the feasibility of this approach, and subsequent works such as Prompt-a-Robot-to-Walk [110], Grounding MLLMs in Actions [111], and Open-VLA [32] adopt similar paradigms. Another line of work introduces *shared intermediate representations* between language and action. CLIP-RT [112] extends vision–language alignment to action generation, and Humanoid-VLA [101] performs language–action pretraining to narrow the semantic–motor gap. VoxPoser [113] leverages LLM reasoning to produce intermediate programs and 3D affordance maps that ground perceptual semantics into spatial actions. Recent studies further mitigate this mismatch by introducing *hierarchical architectures* [109], [114], [115] that insert an explicit intermediate layer between language and action, where a VLM serves as a high-level planner and a separate low-level controller executes high-frequency motion.

(3) Multi-modal Sensory Fusion. As VLA systems evolve, perception extends beyond RGB images and language. For precise manipulation, vision and instruction alone are insufficient for accurate physical interaction and fine-grained control [116]. Incorporating additional modalities such as tactile, force, and audio sensing is therefore essential for achieving more reliable and comprehensive perception [117]–[119], yet it significantly increases the complexity of modality alignment and model optimization.

A common solution is to build *specialized encoders* for each sensory modality and align them with language using contrastive learning. TLA [116] integrates tactile perception to improve contact-rich manipulation, and OmniVLA [120] constructs a semantically aligned tactile encoder that links tactile feedback with linguistic concepts. After obtaining effective representations, the challenge shifts to *fusion*, ranging from deep fusion across the full pipeline, as in Tactile-VLA [121], to modular mixture-of-experts fusion that preserves VLM representations, as in ForceVLA [25]. Due to the high cost of collecting real multimodal data, *simulation-based generation* is emerging as a promising alternative. Multi-Gen [122] explores this direction by generating visual scenes in the simulator and synthesizing additional modalities such as audio to pretrain or enhance real-world policies.

4.1.2 From 2D Images to Spatial-Temporal Representations

Bridging the semantic–perceptual–physical gap requires spatial grounding, meaning that VLA models must accurately capture the 3D structure of the environment. Yet most pretrained VLMs are trained on 2D internet images, creating a core limitation: their reliance on RGB inputs restricts the spatial reasoning needed for real-world robotic operation [82]. Enabling a 2D-native model to acquire spatio-temporal understanding is therefore a central challenge.

(1) Constructing Spatio-Temporal Representations. Building spatio-temporal understanding begins with selecting a representation capable of expressing geometric structure and dynamics. Existing approaches primarily follow three directions. A straightforward option is to augment RGB

inputs with 2.5D *depth maps*, which provide per-pixel distance information and align naturally with 2D images. Depth Helps [123] uses depth as a supervision to learn spatial perception without real sensors, while RoboFlamingo-Plus [124] fuses preprocessed depth with RGB features to strengthen spatial awareness. These results show that even simple 2.5D cues can significantly enhance geometric reasoning. Then, *point clouds* preserve full 3D geometry and offer lossless 3D representation [125]. PointVLA [82] integrates point cloud inputs into pretrained VLA models to improve spatial reasoning without modifying the backbone. Later systems, such as An Embodied Generalist Agent in a 3D World [126] and GeoVLA [125], unify 2D and 3D modalities, while FP3 [127] rebuilds the perception–decision pipeline around point cloud representations under a pretraining–finetuning paradigm. Beyond pure geometry, other studies aim to infuse semantics into point clouds. SoFar [128] constructs semantic 3D scene graphs, Weakly-Supervised 3D Visual Grounding [129] transfers 2D–text alignment to 3D by leveraging CLIP, and LMM-3DP [130] fuses back-projected 2D semantic features with geometric point clouds to form unified semantic–geometric representations. To address the irregular structure of point clouds, other work discretizes 3D space into *voxels* or *occupancy grids*, enabling structured spatial reasoning. Oc-CLLaMA [131] assigns semantic labels to 3D voxels, while RoboMM [132] incorporates multi-view temporal cues to construct unified 3D occupancy grids. Finally, since real-world operation is dynamic, a static 3D snapshot is insufficient. ARM4R [133] captures spatio-temporal evolution by predicting the 4D *trajectory* of 3D point motion, extending static perception to a time-aware formulation [81].

(2) Architectural Integration. Once a spatio-temporal representation is chosen, the next challenge is incorporating geometric information into VLA models without disrupting pretrained alignment. A common strategy is augmentation and injection through *specialized adapters* that introduce 3D features while preserving the backbone’s integrity as much as possible. PointVLA [82] directly augments 2D models with point cloud inputs, while GeoVLA [125] processes 2D and 3D streams in parallel. SpatialVLA [134] projects 2D semantic features into 3D coordinates using positional encoding and spatial grids to form explicit space–action graphs. In contrast, implicit approaches avoid modifying the backbone by attaching external geometric modules, as in Evo-0 [135] with VGTT [136], or by using diffusion-based conditioning to model depth reliability, as in AC-DiT [137]. Another line of work circumvents direct 3D modeling by *reprojecting 3D data into the 2D domain*. BridgeVLA [84] renders point clouds into multi-view images, and OG-VLA [138] generates orthographic projections to recover 3D pose. Some systems predict in 2D and then lift results into 3D, such as A⁰ [138], which first predicts interaction points and trajectories in 2D and then lifts them into 3D via depth projection, and RoboPoint [139], which back-projects 2D keypoints into 3D to create structured action cues. These methods preserve the strengths of large-scale 2D pretraining while retaining essential 3D awareness. A third direction avoids explicit reconstruction altogether by relying on the *reasoning ability of large multimodal models*. VoxPoser [113] generates dense voxel-value maps from language-guided

code to directly impose linguistic constraints on spatial geometry, and Gemini Robotics [114] infers 3D structure through large-scale multimodal reasoning. Finally, to operationalize the 4D perspective within VLA, recent work *injects tracked motion as temporal context*. TraceVLA [81] overlays tracked keypoint trajectories as spatial memory, and Spatial Traces [140] fuses tracked points with depth maps to encode structure and motion within a unified input.

4.1.3 Dynamic and Predictive World Models

A truly embodied world representation cannot stop at static geometry or semantics, it must capture dynamics and causality, i.e., construct an internal, predictive world model capable of answering the fundamental question: if the agent executes an action, what happens next? Predictive world modeling forms the foundation for counterfactual reasoning, long-horizon planning, and physical understanding.

(1) Representation Space. A key design choice is how future states should be represented. One option is to predict directly in the observation space by generating future *pixel-level frames*, which provides a high-fidelity, human-interpretable imagination of future states. TriVLA [24] extends video diffusion models for multi-step visual forecasting, while UP-VLA [141] and CoT-VLA [78] generate key subgoal images that indicate the next salient task state. DreamVLA [142] enriches prediction with task-critical cues such as dynamic regions, depth, and affordances, and FlowVLA [63] introduces a visual chain-of-thought mechanism to synthesize physically consistent future scenes. WorldVLA [80] further models object motion, contact, and state transitions to simulate low-level physical evolution via learned world dynamics. A complementary approach is prediction in a *latent space*. This strategy first encodes high-dimensional visual observations into a compact, low-dimensional latent space and then learns a simpler model to predict the evolution of this latent state [143]. This is more computationally efficient and robust to irrelevant visual noise. For instance, VLM-in-the-Loop [144] explicitly leverages a pretrained latent world model to predict future latent states, while MinD [28] proposes a hierarchical world model that performs predictions in dynamic feature spaces at multiple levels of abstraction. WMPO [145] generates internally in latent space while aligning policy and optimization in pixel space.

(2) Utilization Paradigms. One paradigm is *policy enhancement*, where the world model is tightly integrated with the policy. Short-term future predictions serve as auxiliary inputs or auxiliary training signals [146], providing the policy with forward-looking intuition for more informed action selection. Most observation-space models, including TriVLA [24], CoT-VLA [78], and DreamVLA [142], follow this strategy by conditioning their action decoders on predicted future states. The second paradigm is *explicit planning*, in which the world model serves as a decoupled internal simulator. In this think-before-you-act framework, the agent performs multi-step rollouts of candidate action sequences within the model, evaluates their long-horizon outcomes, and chooses the best plan. This deliberative approach, used in systems such as LUMOS [16], VLM-in-the-Loop [144], and MinD [28], is particularly effective for tasks requiring long-term foresight and trade-offs.

4.1.4 Future Directions

Summary & Trends: Current VLA architectures struggle with two fundamental disconnects, which existing methods address via a patchwork strategy. (1) Regarding the Modality Disconnect, the prevailing trend relies on modular Late Fusion, where separate encoders process inputs in isolation before concatenation. (2) Regarding the Physical Disconnect, researchers currently introduce auxiliary modules or rely on state forecasting to approximate dynamics. However, these approaches remain superficial: late fusion limits deep cross-modal reasoning, while dynamic prediction often mimics physics without understanding causality.

Directions: To bridge these gaps, the field must simultaneously advance toward *Native Multimodal Architecture*. This means converting visual and physical data into tokens at the very beginning of training. By placing all modalities into the same language and shared space, the model does not need complex alignment steps. It can simply reason over all data types together, leading to a more natural and direct understanding of the physical world. An important next step is to develop a hybrid *Latent-Physics-Semantic World Model*. Such a model would internally represent 3D geometry, physical dynamics, semantic attributes and affordances. Given vision, optional depth/point-cloud/tactile input and a language instruction, the system encodes a unified world state, simulates candidate future states (e.g., object motion, contact, stability, affordance changes), and plans by reasoning jointly over semantics and physics. This integration grounds high-level semantic intent in physics-aware simulation, helping to close the gap between semantic understanding, perception, and physical interaction.

4.2 Instruction Following, Planning, and Robust Real-Time Execution

Fig. 5 illustrates the four levels of this challenge, which are elaborated in detail below.

4.2.1 Parsing Complex Instructions

Task instructions for VLA are often multimodal and underspecified, and failures in understanding propagate to perception, planning, and control. We highlight two primary sources of difficulty: (i) Open-ended, multimodal instruction forms. Instructions are no longer plain text, as they may mix language with images, scene cut-outs, internet photos, or hand-drawn sketches. (ii) Ambiguity and underspecification. Commands like “help me” or “clean this up” omit crucial task parameters (i.e., what, where, how, when).

(1) Open-Ended Instruction. To handle open-ended, mixed-modality prompts, recent methods attempt to *interleave images and text* into a single sequence, and use the same sequence modeling mechanism for understanding and control. OE-VLA [147] adopts a shared visual encoder for all images and a text tokenizer for all text, converting them into token streams that are strictly interleaved to preserve the original instruction order. Similarly, Interleave-VLA [148] introduces special tags to its tokenizer, allowing image feature vectors to be seamlessly inserted within a text sequence. These approaches enable the policy to understand non-text instructions and improve direct cross-modal grounding without relying on standardized phrasing.

(2) Ambiguous Instructions. Another line of work focuses on endowing the model with *deeper reasoning and interactive clarification capabilities*. When facing ambiguous commands, ThinkAct [149] infers and verifies the intended target via scene parsing and feedback, while Deep-ThinkVLA [150] resolves ambiguity with causal chain-of-thought and aligns subgoals with correct execution through outcome-driven RL. When spatial information is underspecified, InSpire [151] explicitly prompts the policy to answer “where is the target relative to the robot?” before acting, thereby auto-filling missing cues. Taking this a step further, AskToAct [152] trains an ambiguity-recognition module on synthetically incomplete queries and uses large-scale clarification dialogues to teach the agent to proactively request missing details when an instruction is underspecified.

4.2.2 Hierarchical Planning and Task Decomposition

While many VLA frameworks are optimized for short-horizon skills, executing long-horizon operations remains a largely unresolved challenge [153]. Agents must decompose high-level instructions into structured subgoals to act robustly. Pure end-to-end models, which directly map inputs to low-level actions without explicit intermediate reasoning, often struggle with multi-step planning and compositional tasks [154], [155]. To address this, hierarchical decomposition is a dominant paradigm [15], [40], [85], [86], [156]. Based on the type of intermediate representation they employ to bridge high-level intent and low-level control, current approaches can be broadly categorized into three families.

(1) Language-Driven Planning. These methods adopt a *modular hierarchical paradigm*, leveraging language to decompose tasks in semantic space. $\pi_{0.5}$ [53] embeds hierarchical reasoning within a single inference chain: the model first proposes explicit language-level sub-tasks from vision and instructions, then conditions continuous control on these sub-tasks. OneTwoVLA [157] performs structured textual reasoning at key decision points, generating scene descriptions, high-level plans, and next-step instructions, to decompose tasks within the semantic space. Hi Robot [41] employs a two-layer scheme where a VLM parses instructions into atomic sub-tasks, and a VLA controller handles low-level execution. Other methods use *end-to-end hierarchical paradigm*, like LoHoVLA [158], using a common VLM backbone to jointly produce language sub-steps and continuous actions, enabling long-horizon reasoning without a strict planner-executor split.

(2) Planning via Multimodal Intermediates. These methods perform planning via multimodal intermediates, using non-linguistic representations like visual goals or affordances as the stepping stones for decomposition. On the *vision-driven* side, CoT-VLA [78] employs pixel-level subgoal images as explicit intermediates [78], while Embodied-SlotSSM [159] employs slot-based [160] object-centric representations to create structured visual intermediates. HiP [161] further extends this idea with a three-tier pipeline in which an LLM generates abstract subgoals, a video diffusion model produces physically feasible visual trajectories, and an inverse dynamics model converts these trajectories into actions. On the *affordance-driven* side, RT-Affordance [162] plans tasks by decomposing complex robotic manipulation into man-

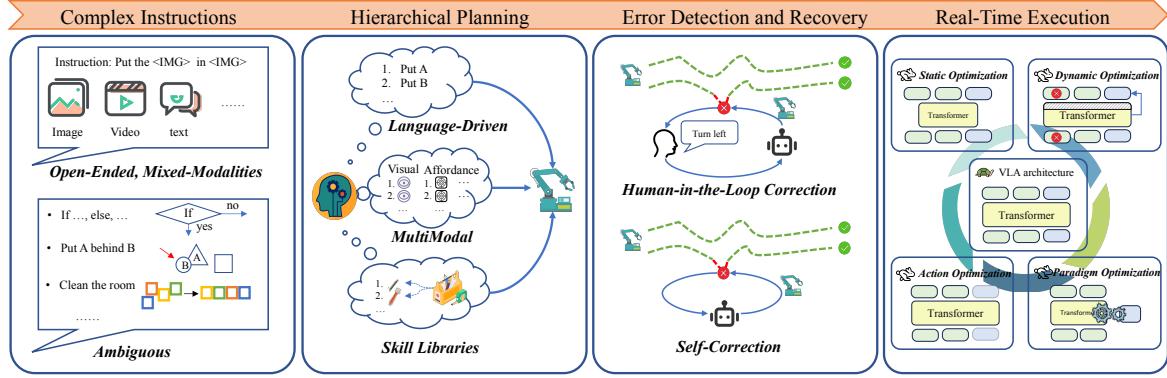


Fig. 5: **The Challenge of Instruction Following, Planning, and Robust Real-Time Execution.** The flow begins with Section 4.2.1, where the model understands what a human wants, even if the instructions are unclear or mixed with images. This understanding then moves to Section 4.2.2, where big goals are broken down into smaller, workable steps or plans. For the third step, Section 4.2.3, the robot carries out its plan, watches for problems, and fixes them if something goes wrong. Lastly, Section 4.2.4 is a rule for the whole process, requiring every step to happen quickly.

ageable affordance plans. CoA-VLA [163] internalizes an affordance chain at each step as an implicit planning signal.

(3) Compositional Planning with Skill Libraries. These methods decompose long-horizon tasks into reusable atomic skills and compose them into higher-level behaviors for efficient and interpretable task execution. For the *explicit skill* usage, VLP [164] builds a fine-grained library for data-efficient reuse of manipulation patterns. Agentic Robot [165] derives a short, semantically clear subgoal sequence from the library, decomposing a task into 2–5 verifiable atomic steps prior to execution. RoboBrain [166] also employs a hierarchical paradigm that expands human-understandable abstract instructions into executable atomic action sequences, achieving an intent–plan–action mapping through the joint learning of data and models. Other works explore the emergence of *implicit skills*. For instance, DexVLA [167] learns to automatically annotate semantic sub-steps within long-horizon action sequences through temporal alignment. AgiBot World [168] serves as a transition, using explicit skills during data collection but learning a policy that implicitly compresses high-dimensional control into semantic latent action tokens, enabling the emergence of composable behaviors.

4.2.3 Error Detection and Autonomous Recovery

Long-horizon VLA deployments are inherently vulnerable to execution interruptions, perception drift, and actuation failures. Without timely, on-policy correction, small mistakes can compound into cascading failures that derail the entire task. To address this, research efforts have largely followed two main lines of inquiry:

(1) Human-in-the-Loop Correction. These methods leverage a human user as an external source of intelligence to guide recovery. This can be *reactive*, where the human provides corrective signals during execution. For instance, Yell At Your Robot [169] integrates real-time human language feedback as corrective signals for immediate behavioral adjustment, while CLIP-RT [112] treats human language feedback as an ideal action template and embeds it into the decision process via similarity matching for efficient, retrain-free correction. This approach can also be *proactive*, where the agent solicits help when it detects ambiguity.

OneTwoVLA [157], for example, incorporates active human clarification as a key component, proactively querying for user input to resolve uncertainty before acting.

(2) Self-Correction. A more effective strategy is to enable the model to autonomously detect anomalous states and correct them. Specifically, CorrectNav [170] enables *self-recovery without extra modules* by iteratively collecting the model’s own error trajectories, automatically identifying deviations, and generating corrective actions and visual data to continuously fine-tune the model. Similarly, FPC-VLA [171] uses a VLM to assess the semantic appropriateness of key actions and, when necessary, generates natural language feedback with corrective directions. Agentic Robot [165] focuses on the architectural level, which achieves autonomous correction via a *standardized plan–act–verify closed loop*: a vision–language validator dynamically assesses subgoal completion and, upon failure, triggers predefined recovery strategies, effectively suppressing error accumulation.

4.2.4 Real-Time Execution and Computing Efficiency

The powerful capabilities of VLA come at the cost of substantial computational overhead. Yet, physical-world interaction is highly sensitive to latency, especially in complex and long-horizon tasks. Bridging the compute-latency gap between model capability and real-time performance is thus critical to the practical deployment of VLA systems. To address these issues, recent works focus on four directions:

(1) Static Optimization of Architecture. A line of work focuses on static architectural optimization, which reduces inherent computational complexity by refining the model’s structure. A common solution is *compression* and *quantization*. BitVLA [172] achieves ultra-low-precision efficiency via ternary 1-bit compression and distillation, while Evo-1 [173] offers a similar lightweight design with only 77M parameters. SQAP-VLA [174] introduces perceptual pruning strategies on the basis of quantization, and achieves a nearly two times inference speedup and half memory reduction. Besides, some methods directly adopt *lightweight backbones*, like NORA [175] and TinyVLA [176], while VLA-Adapter [177] introduces lightweight adapters to graft knowledge from a large model onto a smaller policy network. Other approaches fundamentally replace the computationally expensive Transformer attention mechanism with

linear attention. SARA-RT [178] converts high-cost Transformer policies into linear-attention variants to cut inference delay. RoboMamba [71] replaces the Transformer with the Mamba, attaining linear-time scaling and faster inference without explicit quantization or specialized accelerators.

(2) Dynamic Optimization of Decoding Process and Inference Strategies. Beyond static architectural changes, this line focuses on runtime adaptivity, dynamically adjusting compute budgets during decoding and inference based on task complexity, thereby reducing latency and computation while maintaining accuracy. One strategy is to create *dynamic inference paths*, which dynamically skip certain computation layers or terminate inference early at shallow depths, based on the complexity of the current input. For example, MoLe-VLA [179] leverages layer skipping to reduce FLOPs, while CEDD-VLA [180] and DeeR-VLA [26] design early exit mechanisms. Another is to perform dynamic token processing through *token pruning* or *caching*. VLA-Cache [181] designs adaptive caching strategies that treat static and dynamic tokens differently. SpecPrune-VLA [182] performs action-aware pruning conditioned on history and current observations. CogVLA [183] also reduces computation through instruction-driven visual token sparsification. Furthermore, methods employ *accelerated decoding* to overcome the sequential bottleneck of traditional approaches. For instance, Accelerating VLA [184] and OpenVLA-OFT [33] generate an entire action chunk in a single forward pass through parallel decoding. Spec-VLA [185] adopts speculative decoding to emit candidate action tokens in a single forward pass with relaxed acceptance.

(3) Optimization of Action Representation and Generation Paradigm. This type of method posits that the bottleneck in inference efficiency stems largely from how actions are represented and generated. By rethinking and optimizing action representations, efficiency can be fundamentally improved. One strategy is *efficient action tokenization*, which designs more compact and information-dense action tokens to reduce the number of prediction steps. For example, FAST [186] compresses action sequences to reduce training cost and wall time. XR-1 [187] leverages discrete visual-motor representations learned by VQ-VAE [188] to guide policy learning, while VQ-VLA [189] extends this idea by using a VQ-VAE tokenizer to compress long trajectories into a small set of discrete tokens. Another strategy is *asynchronous execution and inference*, where the system predicts the next action chunk while the current one is being executed, as seen in SmolVLA [190] and Real-Time Action Chunking [191]. A third strategy focuses on *accelerating diffusion policies* by reducing the number of required sampling iterations. Time-Diffusion Policy [192] replaces the traditional time-varying denoising process with a fixed, direction-consistent unified velocity field. Discrete Diffusion VLA [193] discretizes actions into tokens and employs masked diffusion with parallel prediction, alleviating the autoregressive decoding bottleneck.

(4) Optimization of Training Paradigm and System. This kind of work emphasizes the design of the training process and the implementation of the system to further reduce the inference overhead and improve the execution efficiency. A common principle among these approaches is to leverage additional knowledge or data during training

so the model can *take shortcuts at inference time*. For instance, ECoT-Lite [194] uses reasoning traces during training but completely bypasses explicit reasoning steps during inference. V-JEPA 2 [143] reduces planning overhead by predicting *compressed semantic representations* instead of raw pixels. Meanwhile, Fast-in-Slow [195] employs an elegant *dual-system architecture within a single model*, enabling tight coordination between slow, deliberate reasoning and fast, reactive execution. At the highest system level, some works elevate optimization to the *operating system or distributed learning*. For example, AMS [196] introduces OS-level action context caching and replay mechanisms, and FedVLA [197] explores efficient distributed training of VLA models under a federated learning framework.

4.2.5 Future Directions

Summary & Trends: To handle complex tasks, the community is currently divided into rigid hierarchical systems (using LLMs as high-level planners for code generation or sub-goal decomposition) for long-horizon reasoning, or massive end-to-end policies via instruction tuning for reactive skills. However, the former suffers from severe information loss between modules, while the latter lacks the reasoning capability for multi-stage correction, resulting in open-loop execution without introspection.

Directions: Future architectures must break this dichotomy by becoming *Adaptive*. Just like a human, the model should decide how much to think based on the task. For simple tasks like grabbing a cup, it should act instantly. For complex tasks like assembling furniture, it should automatically activate deeper reasoning skills to plan steps. To do this, one direction is to use *Unified Decision Tokens*. By treating seeing, thinking, and acting as a single stream of data, the model can naturally switch between fast action and deep thought without needing separate, rigid modules. This creates a true end-to-end unified mind that handles both simple reflexes and long-term planning. Beyond just acting efficiently, robots need to change how they understand their own actions. Today's robots are passive, i.e., they just follow instructions without asking why. Future VLA models must evolve toward *Self-Awareness*. The goal is an agent that not only knows what to do, but also understands why it is doing it. Models should shift from open-loop execution to closed-loop resilient autonomy, dynamically switching between replanning and reflex adjustment to autonomously recover from failures without intervention.

4.3 From Generalization to Continuous Adaptation

Fig. 6 illustrates the four levels of this challenge, which are elaborated in detail below.

4.3.1 Open-World Generalization

Despite strong cross-modal understanding and manipulation in closed settings, large VLA models often generalize poorly when deployed in open, dynamic real-world environments. Conventional imitation learning relies on large human-annotated datasets and fails to cover the long tail of real scenes. Therefore, achieving robust open-world generalization is a pivotal challenge.

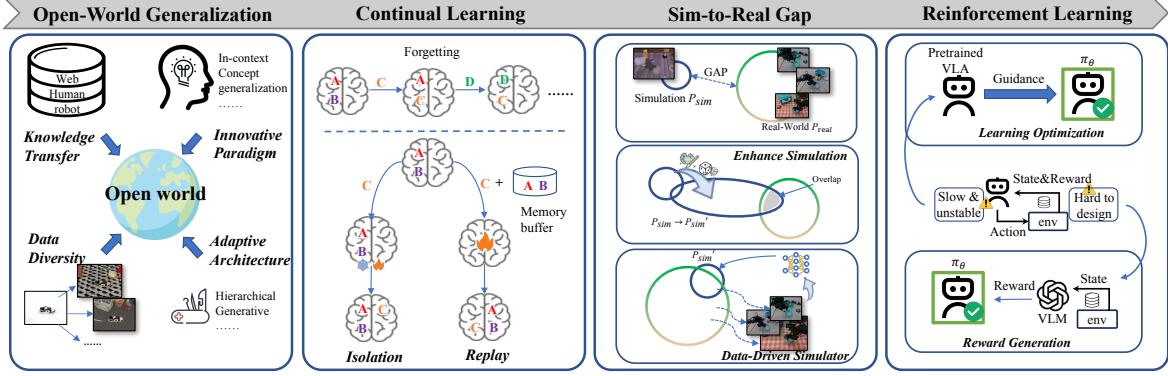


Fig. 6: **The Challenge of From Generalization to Continuous Adaptation.** This diagram illustrates how VLA models operate continuously in dynamic, open-world environments, highlighting four key enabling strategies. Section 4.3.1 represents the initial ability to perform well in settings not seen during training. Building on this, Section 4.3.2 focuses on how agents can continuously acquire new skills throughout their operational lifetime without forgetting old ones. Section 4.3.3 addresses the critical challenge of transferring learned policies from virtual environments to the physical world. Finally, Section 4.3.4 highlights how agents refine their behaviors and learn from real-time experience.

(1) Knowledge Transfer and Utilization. The most dominant approach posits that the key to generalization lies not in learning from scratch, but in effectively transferring vast prior knowledge from large-scale data sources. This is pursued in two main ways. *Multi-task/multi-robot pretraining* involves training on massive robotic datasets to learn a general, hardware-agnostic prior over behaviors. For example, Octo [49] pretrains a Transformer on about 800k robot trajectories to acquire general manipulation regularities and uses lightweight adapters for efficient fine-tuning, enabling rapid adaptation to new sensors and action spaces under limited data and compute. DexVLA [167] introduces billion-parameter diffusion action experts that pretrain across robot morphologies and adopts a three-stage curriculum to realize task-agnostic language-action mapping. RoboCat [198] pretrains on heterogeneous multi-robot data and continually improves on real trajectories for sustained task transfer. Dita [199] leverages the large OXE dataset [100] and diffusion Transformers to learn cross-environment behaviors, adapting with as few as 10 real demonstrations. EO-1 [200] further scales this paradigm by pretraining a shared backbone on the 1.5M-EO-Data dataset to achieve knowledge transfer and enhance open-world understanding. The second method is *internet/human video knowledge transfer*, which leverages data sources vastly larger than robotic datasets. Following CLIP [20], R3M [201] extends this paradigm to robotics by pretraining visual encoders on massive collections of human first-person videos (e.g., Ego4D [202]), thereby transferring general interaction knowledge into robotic policies. In addition, the GR series (e.g., GR-1 [67], GR-2 [68]) stands as a representative line of work in this direction that pre-train on massive human egocentric video datasets to transfer general physical and interaction knowledge into robotic policies.

(2) Paradigm-Level Innovations. Beyond knowledge transfer from pretrained models or web-scale data, a growing body of work explores how models learn, not just what they learn, which is a key to achieving robust generalization. For example, ICIL [203] follows the *in-context learning* paradigm that trains the model to infer tasks from a few demonstrations provided in the prompt at test time, enabling rapid, retrain-free adaptation. Another direction focuses on

emergent compositionality, where methods like TRA [204] use a temporal contrastive loss to imbue the learned representation space with a compositional structure, allowing the model to automatically combine learned skills into new tasks. A more profound shift is toward *conceptual generalization*, which moves beyond imitating actions to understanding semantic concepts. ObjectVLA [205] jointly trains on robot trajectories and box-labeled VL corpora to achieve zero-shot manipulation of unseen objects, while LERF [206] fuses CLIP with 3D NeRFs for natural-language localization and grasping of novel objects. Finally, to achieve robust deployment, new *adaptation paradigms* emerge. Align-Then-Steer [207] proposes a non-invasive adaptation method that steers a frozen VLA model's outputs using a lightweight, latent-space adapter. Robot Utility Models (RUM) [208] pair large-scale home demonstrations with multimodal LLM reasoning for runtime verification and automatic retries, achieving zero-shot deployment in new environments.

(3) Enhancing Data Diversity. Given the high cost of collecting real-world data, recent work expands the data distribution using generative models and semantic priors to build large-scale, more diverse training data at low robot cost. For *data augmentation*, CACTI [209] scales multi-task imitation by using Stable Diffusion for zero-shot inpainting of expert images to increase layout and appearance diversity without additional robot rollouts. GenAug [210] employs text-to-image synthesis conditioned on a few demonstrations and prompts to produce visually diverse yet functionally consistent scenes, improving robustness to unseen environment shifts. For *semantic augmentation*, ROSIE [211] distills knowledge from internet-scale VLMs into robot training, exposing policies to richer semantic combinations and task variants to strengthen open-set generalization.

(4) Adaptive Architectural Design. Beyond the above approaches, the design of the model architecture itself profoundly influences its generalization capability. Specifically, *hierarchical designs* enhance generalization by decomposing tasks into high-level planning and low-level execution. The high-level planner can leverage abstract knowledge learned from large-scale data, while the low-level executor focuses on acquiring reusable skills [212]. *Multimodal frameworks* that dynamically fuse multimodal sensor inputs can

significantly enhance robustness in complex environments, like BAKU [213]. Meanwhile, *generative diversity* methods like StructDiffusion [214] use language-guided diffusion to generate multiple physically plausible action structures instead of a single deterministic plan, improving robustness to unseen object sizes and shapes.

4.3.2 Continual Learning and Incremental Skill Acquisition

An embodied agent’s learning process should not end at deployment. It must continually acquire new skills throughout its lifetime to adapt to evolving environments and user needs. However, recent studies reveal a critical issue: as new tasks are learned, the parameters supporting previously acquired skills are often overwritten, leading to sharp performance regressions and the erosion of multimodal reasoning capabilities inherited from backbones [14], [62]. To solve this, existing efforts broadly follow two routes.

(1) Parameter Isolation and Expansion. These methods allocate dedicated parameter space for new skills or adopt modular designs that safeguard existing weights, thereby fundamentally preventing weight conflicts between old and new tasks and mitigating cross-task interference at its source. One prominent approach is *Prompt-Based and Codebook-Based Learning*, which encodes skill knowledge into a set of discrete, composable prompts or codebook entries. When acquiring a new skill, the system simply adds a new prompt or codebook entry without modifying existing components [14], [215]. The other approach uses *modular and expert-based architectures* to isolate knowledge. For example, InstructVLA [62] adopts a two-stage training paradigm and a Mixture-of-Experts architecture to intelligently route between reasoning and action modules, avoiding direct modification of its backbone. Similarly, the scalable PerceiverIO proposed in iManip [216] falls into this category by adding new, skill-specific weights while freezing old ones.

(2) Replay-based Knowledge Consolidation. Inspired by human review, these methods *rehearse a subset of past examples* while learning new tasks to reinforce retained knowledge. Since storing and replaying all historical data is impractical, the core challenge lies in intelligently selecting the most informative samples for replay. ExpReS-VLA [217] addresses this by introducing compressed experience replay to mitigate catastrophic forgetting in robotic VLA systems, while iManip [216] proposes a temporal replay strategy that avoids random sampling and instead replays critical frames during skill execution.

4.3.3 Sim-to-Real Gap in Deployment

The sim-to-real gap remains a core obstacle for deploying VLA policies, as discrepancies between simulated and real-world dynamics (e.g., friction, latency, actuation response) and perception (e.g., illumination, textures, sensor noise) severely degrade policy transfer despite the low-cost, large-scale data provided by simulators [218]. To address this challenge, researchers have explored a variety of strategies:

(1) Enhancing Simulation Fidelity and Robustness. The goal of this class of methods is to improve the direct transferability of policies, either by making the simulation environment more closely resemble the real world or by making the policy robust to the discrepancies between simulation and reality. A straightforward solution is to *enhance the*

visual fidelity of the simulator’s rendering. ManiSkill3 [219] leverages GPU-parallel rendering, domain randomization, and background composition to narrow the appearance gap and enable zero-shot transfer. Another alternative to improving the simulation is to make the policy more robust by *learning a stable intermediate representation*. SLIM [218], for instance, compresses high-dimensional RGB images into segmentation and depth maps, thereby filtering out task-irrelevant visual differences between sim and real.

(2) Data-driven Simulators. Recognizing that classical physics engines cannot fully capture real-world complexity, a complementary line sidesteps explicit sim modeling by learning from or generating experiences using real-world data. One direction is *generative augmentation* on real-world data, which attempts to expand a small set of real robot trajectories to enhance data diversity. For instance, GenAug [210] leverages web-scale image generative models to synthesize visually diverse but functionally consistent images from a few real robot demonstrations and semantic prompts, bypassing simulators entirely by exploiting the model’s prior over real-world visuals to generate highly realistic scenes. Another mainstream direction redefines physics-based simulation as *data-driven prediction*: it trains a powerful world model to learn physical dynamics and causal relationships directly from massive amounts of real-world data, such as DreamGen [220]. RynnVLA-001 [221] further advances this direction through large-scale video generation pretraining combined with human-centric trajectory perception modeling, enabling implicit transfer of human manipulation skills to robotic control.

4.3.4 Online Interaction and Reinforcement Learning

Imitation learning allows VLA models to quickly learn basic skills from offline data, but is limited by distributional shift and a performance ceiling capped by human demonstrators. Reinforcement Learning (RL) addresses these by enabling autonomous exploration, yet its application to large VLA models in high-dimensional continuous action spaces is hindered by low sample efficiency [222]–[225] and the difficulty of designing effective rewards [36], [226], [227]. To tackle these challenges, researchers integrate RL with VLA models’ strong priors, primarily through two directions:

(1) Optimizing the Learning Process. Rather than letting RL explore from scratch, this approach injects or distills the rich knowledge and structural priors already learned by VLA models into the RL policy, addressing the slow and unstable nature of RL training. For *knowledge transfer*, RLDG [223] first trains task-specialist RL policies, then distills their high-quality trajectories into a general VLA, improving precise control and generalization without fragile end-to-end RL fine-tuning. Refined Policy Distillation [225] adds a simple MSE constraint so that VLA action distributions guide the RL agent, maintaining stability under sparse rewards and viewpoint changes. iRe-VLA [222] alternates phases: it freezes the large backbone and trains a lightweight action head during RL for stability; it then unfreezes and fine-tunes with successful/expert trajectories under supervision to regain capacity. Beyond the above, some approaches *optimize the internal structure* of RL algorithms. For example, CORFT [224] designs a chunked temporal-difference learning mechanism that feeds entire action sequences into the critic

to predict multi-step returns, aligning with VLA’s chunked structure and significantly improving training stability and sample efficiency under limited data.

(2) Automating Reward Generation. Instead of costly hand-crafted rewards or labor-intensive preference labels, recent work leverages VLM/LLM perception and reasoning to automatically derive dense, high-quality rewards directly from observations and goals. One direction infers rewards through *perceptual alignment* by measuring similarity between the current visual state and the goal description in a shared embedding space. VLM-RMs [228] introduces this idea, and RoboCLIP [229] extends it to video trajectories by computing video–language similarity for sparse rewards. Affordance-Guided RL [230] converts VLM-predicted grasp points and target trajectories into continuous dense rewards that guide policy optimization. A second direction uses VLMs as critics to *rank trajectories or states* rather than relying on direct similarity scores. RL-VLM-F [231] employs GPT-4V to compare observation pairs and infer preferences for training a reward function without human labels, while GRAPE [226] decomposes tasks and generates stage-wise preferences for structured, multi-objective rewards. A third direction leverages LLMs’ zero-shot *code generation and high-level reasoning* to produce reward functions. Eureka [232] prompts an LLM with environment code and task specifications to generate executable rewards, VIP [233] views reward learning as implicit value optimization from video, and VLA-RL [36] fine-tunes a VLM into a structured process-reward model that transforms sparse feedback into next-action-token supervision.

4.3.5 Future Directions

Summary & Trends: To achieve generalization, the dominant approach currently hinges on Scaling Laws, i.e., aggregating massive, heterogeneous datasets to train large-scale transformers via passive imitation learning. While this has significantly improved task-level success rates on seen distributions, models remain hardware-dependent and temporally static. They are frozen after training, lacking the agency to actively explore or adapt to novel robot morphologies without extensive fine-tuning.

Directions: To realize “GPT moment” in embodied intelligence, the paradigm must shift from training fragmented, robot-specific policies toward developing *Morphology-Agnostic Representations*. By logically disentangling high-level semantic planning from low-level proprioceptive control, a unified brain can transfer manipulation skills across vastly different embodiments—from quadrupeds to humanoids—via lightweight, modular adapters. This would enable true *Zero-Shot Cross-Embodiment Transfer*, where a new robot is treated simply as a new peripheral for a universal policy. Furthermore, this generalization must be sustained through time via *Autonomous Open-Ended Evolution*. We envision a shift from static training sets to a self-reinforcing data engine, where agents exhibit intrinsic motivation to act as curious explorers. By combining self-supervised exploration with online reinforcement learning, future VLAs will transition from passive imitators to active learners, identifying their own knowledge gaps and generating high-quality training data in the wild. This creates a

virtuous closed loop of “*Deployment → Discovery → Evolution*,” allowing the system to continuously refine its world model and expand its capabilities without human.

4.4 Safety, Interpretability and Reliable Interaction

Fig. 7 illustrates the two levels of this challenge, which are elaborated in detail below.

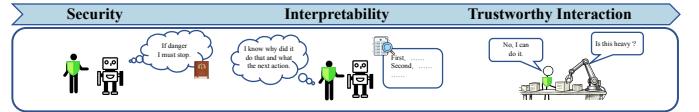


Fig. 7: The challenge of Safety, Interpretability and Reliable Interaction. This diagram shows how VLA systems build human trust, broken into three key layers. Section 4.4.1 is about making sure the robot is physically safe and works reliably. Moving up, Section 4.4.2 focuses on helping humans understand why the robot makes certain decisions, making the robot’s actions easy to understand and predict, leading to smooth collaboration.

4.4.1 Reliability and Safety Assurance

VLA models, particularly end-to-end deep learning systems, usually lack transparency in their decision-making and exhibit unpredictable behavior. When deployed in unstructured, human-shared physical environments, they may execute hazardous actions due to perception errors, generalization failures, or misinterpretation of instructions, potentially endangering humans, the environment, or themselves. Consequently, establishing a reliable and verifiable safety assurance mechanism is a critical prerequisite for the real-world deployment of VLA systems. To address this challenge, two directions are explored:

(1) Constraint-Based Safety Paradigms. This paradigm injects explicit rule systems, inside or outside the model, to hard-bound the action space and avoid unsafe behaviors. Specifically, applying *rule-based explicit constraints* is the most straightforward approach. AutoRT [234] introduces a robot constitution via structured prompting to encode multi-level constraints for behavior bounding in the wild. Alternatively, some works directly *internalize safety constraints* as an integral part of the model’s learning process. SafeVLA [79] explicitly models physically hazardous behaviors as a cost function within a constrained Markov decision process, where the training objective is to maximize task reward while ensuring the cumulative cost remains below a pre-defined safety threshold.

(2) Learning-based Alignment Paradigms. Since scenarios in the real world are highly complex and cannot be fully covered by a finite set of handcrafted rules, some methods aim to *internalize a human-aligned safety intuition and judgment*, enabling models to proactively detect and avoid risks. For example, Gemini Robotics [114] applies Constitutional AI post-training on safety data, ensuring that policies follow human-centric principles and thereby internalize safety intuition. Beyond passively adhering to predefined rules, the model must *actively assess the uncertainty and potential risks* of the current situation and adapt its behavior accordingly. GPI [154] integrates confidence estimation, probabilistic action generation, and language-guided backtracking to pause, seek help, or replan under uncertainty. Furthermore,

RationalVLA [27] introduces a learnable refusal token to reject unsafe/invalid commands, adding a rational safety layer between high-level semantics and low-level control.

4.4.2 Interpretability and Trustworthy Interaction

Most VLA models follow the end-to-end deep learning paradigm, which is inherently a black box and offers little mechanistic insight [235]. When a robot acts, its inability to explain its rationale to the user impedes debugging, erodes trust, and hinders efficient human-robot collaboration. Thus, a core challenge for VLA systems is to make decision logic more transparent and behavior more predictable. Research efforts are therefore shifting toward two aspects:

(1) Enhancing Process Interpretability. The aim is to expose the model’s abstract neural states as explicit, human-understandable intermediates at each step of the think-decide-act chain. *Chain-of-thought reasoning* is a well-known approach to enhancing interpretability. The intermediate reasoning process can be expressed either in linguistic form or in visual form. For the former, Diffusion-VLA [58] conditions a diffusion policy on natural-language reasoning, exposing step-wise intent. ECoT [236] outputs editable step-by-step rationales that users can correct via language. For the latter, CoT-VLA [78] adds visual subgoal images to render intermediate plans observable. Moreover, in hierarchical architectures, the *intermediate instructions generated by the high-level planner* inherently serve as a natural source of interpretability. For example, RT-H [38] separates language-action generation from execution, enabling self-explanation and language-level intervention. HiRobot [41] outputs readable low-level commands from a high-level planner, making task decomposition transparent. GraSP-VLA [237] explicitly converts visual inputs into symbolic states and performs planning in this symbolic space, making its intermediate process inherently interpretable. Besides, recent efforts aim to *decode the internal, hidden symbolic states* from trained, black-box VLA models. A representative work is DIARC-OpenVLA [238], which trains linear probes on hidden layers to explicitly map neural activations to symbolic states, providing a monitorable layer of decision transparency without altering the original model.

(2) Behavioral Predictability. Beyond explaining why a decision is made, it is equally important to design robot behaviors that are *inherently intuitive and aligned with human expectations*, thereby fostering trust directly through interaction. CrayonRobo [239] externalizes the model’s internal decision logic using structured, semantically explicit visual prompts, creating a shared, interpretable language that lets humans intuitively understand and even design the prompts for deeper collaboration. Another critical aspect is *predictable responses to dynamic instructions*. SwitchVLA [240] introduces structured task switching: upon mid-execution instruction changes, the agent rolls back conflicting actions before smoothly transitioning to the new goal, yielding natural, predictable behavior in open-ended interaction.

4.4.3 Future Directions

Summary & Trends: Currently, safety is predominantly handled by extrinsic guardrails (e.g., rule-based shields or constitution-based filtering like AutoRT) or post-hoc rationalization (prompting VLMs to caption their actions). While

providing a layer of protection, these reactive measures are separated from the policy’s core decision process, often failing to prevent intrinsic model hallucinations or confident but wrong actions in real-time.

Direction: To build truly trustworthy embodied agents, the field must evolve beyond imposing static safety rules toward cultivating *Intrinsic Uncertainty Awareness*. In unstructured open worlds, absolute safety cannot be guaranteed by pre-defined constraints alone. Instead, future VLA models require a System 2 reflective layer that actively estimates epistemic uncertainty, endowing the agent with a sense of doubt. This enables a paradigm shift from reactive emergency stops to *Proactive Risk Aversion*, where the agent autonomously pauses to solicit human clarification or replans when it detects ambiguity or potential hazard. Furthermore, trust relies on establishing *Shared Mental Model* through intervention-ready transparency. Interpretability should not be merely a post-hoc debugging tool, but an integral part of the execution loop. We envision agents that visualize their thought process, such as future trajectories, attention heatmaps, or subgoal decompositions, before physical action is taken. Crucially, this transparency must be actionable: it should empower users to not only anticipate robot behavior but also intervene effectively. By allowing humans to correct the robot’s reasoning chain via natural language or gestures, we can close the loop of *Interactive Safety*, ensuring that VLA systems are not just compliant, but genuinely aligned with human intent.

4.5 Data Construction and Benchmarking Standards

Fig. 8 illustrates the two levels of this challenge, which are elaborated in detail below.

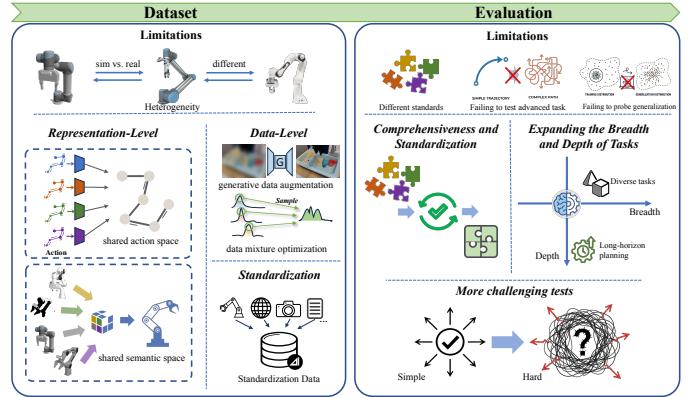


Fig. 8: The challenge of Data Construction and Benchmarking Standards. Section 4.5.1 addresses the critical bottleneck of acquiring and unifying diverse training resources to construct large-scale datasets. Section 4.5.2 focuses on the standardization and increasing complexity of assessment protocols.

4.5.1 Multi-Source Heterogeneous Data

The capabilities and generalization of VLA models are fundamentally constrained by the scale, diversity, and quality of their training data. However, acquiring and unifying high-quality, large-scale, and diverse data presents a formidable challenge, primarily due to the inherent heterogeneity of data sources (e.g., sim vs. real, different robot embodiments) and their respective control interfaces. To

address this, the research community systematically initiates explorations across three interconnected levels:

(1) Representation-Level Unification and Alignment. The core idea here is to model heterogeneous data within a shared, semantically consistent latent space, thereby eliminating heterogeneity at the cognitive level rather than directly handling raw discrepancies. This is achieved through two complementary strategies. The first aligns behaviors in a latent action space by *learning a unified discrete representation* that maps continuous, high-dimensional motions from different robots or human videos into semantically consistent action tokens. This filters out low-level control differences and aligns behaviors at a higher semantic level. LAPA [241], Moto [242], and UniVLA [35] learn such task-centric latent action representations through unsupervised or self-supervised video learning. A more holistic strategy constructs a *shared semantic space across all modalities and embodiments*, extending beyond action correspondence to unify perception, reasoning, and control. RDT-1B [23] and AgiBot World [168] map diverse robot actions into unified physical or latent vectors, while Scaling Cross-Embodied Learning [243] tokenizes heterogeneous visual and proprioceptive inputs for a shared Transformer to handle multiple morphologies. At the multimodal level, methods such as RT-1 [99], GR-2 [68], ViSA-Flow [244], and Humanoid-VLA [101] achieve consistent VLA grounding through unified tokenization, semantic alignment, or self-supervised learning. Human-to-robot transfer approaches, including EgoVLA [245] and DexWild [246], further align human and robot motion using MANO hand models and inverse kinematics, enabling cross-domain embodied transfer.

(2) Data-Level Augmentation and Optimization. Rather than altering the model’s latent space, this line of work directly operates on raw data. The first strategy, *generative data augmentation*, creates expanded data distributions using large pretrained generative models. This substantially increases visual diversity at low cost and improves robustness to appearance variations in heterogeneous real-world data. CACTI [209] and GenAug [210] augment robot data via inpainting or restyling, ROSIE [211] enriches data at the semantic level using VLM priors, and Models with Data Generation via Residual RL [247] generate additional samples through RL to further strengthen downstream VLA performance. The second strategy, *automated data mixture optimization*, focuses on making better use of existing heterogeneous datasets by treating data fusion as an optimization problem. Re-Mix [248] adjusts sampling weights of heterogeneous data subsets based on performance feedback, enabling the model to focus on informative samples and achieve efficient cross-domain fusion.

(3) Standardization and Benchmark Construction. This line of work reduces the heterogeneity of the data at the source by establishing standardized data collection protocols, synchronization mechanisms, and unified benchmarks. A major focus is *unified acquisition and synchronization* within individual datasets to ensure high quality and internal consistency. RH20T [118] enforces strict temporal alignment across multimodal sensors, and BridgeData V2 [249] organizes diverse data types into a standardized format. In simulation, RoboCasa [250] and CoVLA [251] provide large-scale, high-fidelity environments that act as standardized digital

laboratories. Another effort involves *collecting and aligning human-centric and multi-view data*, which is necessary for robots operating in human environments. Representative examples include Ego4D [202] and EPIC-KITCHENS [252], with Ego-Exo4D [253] further integrating first- and third-person viewpoints to support learning skilled activities from multiple perspectives. The broader ambition is *cross-domain standardization*, where heterogeneous datasets are aligned at scale to form unified fusion benchmarks. Open X-Embodiment (OXE) [100] marks a major milestone by aggregating dozens of datasets into a single benchmark for cross-embodiment generalization. RoboMM [132] advances it through a three-level semantic alignment framework that enables joint training across multiple datasets.

4.5.2 Evaluation Benchmarks

Standardized benchmarks play a pivotal role in embodied intelligence by establishing common evaluation protocols that enable fair comparison and reproducible research. However, as VLA models advance rapidly, the yardsticks used to measure them struggle to keep pace, revealing several critical limitations [254]. First, a lack of unified standards in metrics and experimental setups makes fair comparison difficult. Second, many existing benchmarks are limited to simple, short-horizon tasks, failing to test advanced cognitive reasoning. Third, they often lack a systematic way to probe frontier generalization capabilities. To address these gaps, the community actively develops a new generation of benchmarks and evaluations.

A primary direction of this effort is the pursuit of *comprehensiveness and standardization*. The work on Benchmarking VLAs [255] provides a blueprint by emphasizing unified I/O, metrics, and multi-robot coverage, shifting the focus from tasks to metrics. EUQ [256] introduces a human-assessed, multi-dimensional scoring system to capture process quality beyond binary success. At the infrastructure level, simulation platforms like ManiSkill3 [219] and robosuite [257] contribute standardized APIs and task suites, providing a reproducible foundation for fair and scalable evaluation. A second major direction focuses on *expanding the breadth and depth of tasks* to assess more complex capabilities. CALVIN [258] is designed to require the execution of long-horizon sequences of language-guided operations. LIBERO [259] is introduced as the first benchmark specifically for lifelong learning in robotics, featuring standardized metrics for knowledge transfer and forgetting. Furthermore, Ego-Exo4D [253] pioneers the synchronization of first- and third-person recordings for multi-perspective skill analysis. Finally, a third direction aims to design more challenging tests that focus on frontier *generalization and reasoning capabilities*. The From Intention to Execution [260] suite is introduced to probe the intention-execution gap and systematically covers challenges in object diversity, linguistic complexity, and visual-language reasoning. To specifically assess the abilities of instruction-tuned models, InstructVLA [62] releases the SimplerEnv-Instruct benchmark, a comprehensive suite of 80 zero-shot tasks featuring multilingual expressions, novel objects, and implicit intentions to evaluate contextual reasoning and generalization.

4.5.3 Future Directions

Summary & Trends: Driven by the pursuit of scaling laws, the field is currently preoccupied with aggregating massive, heterogeneous real-world datasets to fuel models. On the evaluation front, the standard remains simplistic, relying heavily on binary success rates in controlled settings. However, real-world collection is inherently unscalable and noisy, and binary metrics fail to capture the nuances of robustness, often masking critical failure modes.

Directions: To scale embodied intelligence, the field must transition towards a *Simulation-First, Failure-Centric Paradigm*. Relying solely on real-world data is unscalable; instead, we envision *Simulated Universes* acting as infinite data factories that generate diverse, labeled trajectories with perfect ground truth. The core challenge will be bridging the Sim-to-Real gap for perception and physics, allowing real-world data to serve efficiently as a high-quality alignment set to calibrate the simulator’s physics and rendering fidelity, rather than being the primary training source. Equally important is a shift in how we treat errors. Current pipelines often discard failed trajectories, wasting critical information. Future systems must *Turn Failure into Signal*, treating mistakes as gold mines for negative mining and contrastive learning. By explicitly training on what not to do and diagnosing why failures occur, agents can learn not only to avoid risks but also to autonomously recover from inevitable execution errors. Finally, evaluation must evolve from simple binary success rates to *Comprehensive Diagnostic Stress Testing*. Benchmarks should utilize high-fidelity simulation proxies to assess holistic capabilities—quantifying not only task completion but also safety margins, efficiency, and resilience to perturbations—thereby prioritizing robust adaptability over rote execution of memorized trajectories.

5 CONCLUSION

This survey presents a comprehensive anatomy of Vision-Language-Action (VLA) models, structured to guide readers from basic modules and historical milestones to the core challenges at the research frontier. We provide a detailed analysis of the five key problem areas: representation, execution, generalization, safety, and dataset evaluation, reviewing current solutions, and highlighting future opportunities for each. We hope that this work serves as a foundational roadmap, helping both newcomers and experienced researchers navigate and advance the rapidly evolving field of embodied intelligence.

REFERENCES

- [1] Y. Zhong *et al.*, “A survey on vision-language-action models: An action tokenization perspective,” *arXiv:2507.01925*, 2025.
- [2] W. Guan, Q. Hu, A. Li, and J. Cheng, “Efficient vision-language-action models for embodied manipulation: A systematic survey,” *arXiv:2510.17111*, 2025.
- [3] T.-Y. Xiang *et al.*, “Parallels between vla model post-training and human motor learning: Progress, challenges, and trends,” *arXiv:2506.20966*, 2025.
- [4] D. Zhang *et al.*, “Pure vision language action (vla) models: A comprehensive survey,” *arXiv:2509.19012*, 2025.
- [5] M. Ü. Din, W. Akram, L. S. Saoud, J. Rosell, and I. Hussain, “Vision language action models in robotic manipulation: A systematic review,” *arXiv:2507.10672*, 2025.
- [6] R. Sapkota, Y. Cao, K. I. Roumeliotis, and M. Karkee, “Vision-language-action models: Concepts, progress, applications and challenges,” *arXiv:2505.04769*, 2025.
- [7] R. Shao *et al.*, “Large vlm-based vision-language-action models for robotic manipulation: A survey,” *arXiv:2508.13073*, 2025.
- [8] K. Kawaharazuka, J. Oh, J. Yamada, I. Posner, and Y. Zhu, “Vision-language-action models for robotics: A review towards real-world applications,” *IEEE Access*, 2025.
- [9] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, “A survey on vision-language-action models for embodied ai,” *arXiv:2405.14093*, 2024.
- [10] K. O’shea and R. Nash, “An introduction to convolutional neural networks,” *arXiv:1511.08458*, 2015.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [12] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.
- [13] C. Chi *et al.*, “Diffusion policy: Visuomotor policy learning via action diffusion,” *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1684–1704, 2025.
- [14] J. Xu and X. Nie, “Speci: Skill prompts based hierarchical continual imitation learning for robot manipulation,” *arXiv:2504.15561*, 2025.
- [15] J. Zhang *et al.*, “Hirt: Enhancing robotic control with hierarchical robot transformers,” *arXiv:2410.05273*, 2024.
- [16] I. Nematollahi, B. DeMoss, A. L. Chandra, N. Hawes, W. Burgard, and I. Posner, “Lumos: Language-conditioned imitation learning with world models,” 2025.
- [17] D. Hafner *et al.*, “Learning latent dynamics for planning from pixels,” in *International conference on machine learning*. PMLR, 2019, pp. 2555–2565.
- [18] A. Dosovitskiy, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv:2010.11929*, 2020.
- [19] Y. Haruna, S. Qin, A. H. Adama Chukkol, A. A. Yusuf, I. Bello, and A. Lawan, “Exploring the synergies of hybrid convolutional neural network and vision transformer architectures for computer vision: A survey,” *Engineering Applications of Artificial Intelligence*, vol. 144, p. 110057, Mar. 2025.
- [20] A. Radford *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [21] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language image pre-training,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.
- [22] K. Black *et al.*, “π0: A visionlanguage-action flow model for general robot control, 2024a,” *arXiv:2410.24164*, 2024.
- [23] S. Liu *et al.*, “Rdt-1b: a diffusion foundation model for bimanual manipulation,” *arXiv:2410.07864*, 2024.
- [24] Z. Liu, Y. Gu, S. Zheng, X. Xue, and Y. Fu, “Trivila: A unified triple-system-based unified vision-language-action model for general robot control,” *arXiv:2507.01424*, 2025.
- [25] J. Yu *et al.*, “Forcevla: Enhancing vla models with a force-aware mofe for contact-rich manipulation,” *arXiv:2505.22159*, 2025.
- [26] Y. Yue *et al.*, “Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 56 619–56 643, 2024.
- [27] W. Song *et al.*, “Rationalvyla: A rational vision-language-action model with dual system,” *arXiv:2506.10826*, 2025.
- [28] L. Wang, R. Shelim, W. Saad, and N. Ramakrishnan, “Dmwm: Dual-mind world model with long-term imagination,” *arXiv:2502.07591*, 2025.
- [29] H. Huang *et al.*, “Otter: A vision-language-action model with text-aware visual feature extraction,” *arXiv:2503.03734*, 2025.
- [30] M. Oquab *et al.*, “Dinov2: Learning robust visual features without supervision,” *arXiv:2304.07193*, 2023.
- [31] Y. Li, Y. Wang, Y. Fu, D. Ru, Z. Zhang, and T. He, “Unified lexical representation for interpretable visual-language alignment,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 1141–1161, 2024.
- [32] M. J. Kim *et al.*, “Openvla: An open-source vision-language-action model,” *arXiv:2406.09246*, 2024.
- [33] M. J. Kim, C. Finn, and P. Liang, “Fine-tuning vision-language-action models: Optimizing speed and success,” *arXiv:2502.19645*, 2025.

- [34] S. Deng *et al.*, "Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data," *arXiv*:2505.03233, 2025.
- [35] Q. Bu *et al.*, "Univla: Learning to act anywhere with task-centric latent actions," *arXiv*:2505.06111, 2025.
- [36] G. Lu *et al.*, "Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning," *arXiv*:2505.18719, 2025.
- [37] X. Chen *et al.*, "Pali-x: On scaling up a multilingual vision and language model," *arXiv*:2305.18565, 2023.
- [38] S. Belkhale *et al.*, "Rt-h: Action hierarchies using language," *arXiv*:2403.01823, 2024.
- [39] L. Beyer *et al.*, "Paligemma: A versatile 3b vlm for transfer," *arXiv*:2407.07726, 2024.
- [40] H. Song *et al.*, "Hume: Introducing system-2 thinking in visual-language-action model," *arXiv*:2505.21432, 2025.
- [41] L. X. Shi *et al.*, "Hi robot: Open-ended instruction following with hierarchical vision-language-action models," *arXiv*:2502.19417, 2025.
- [42] X. Chu *et al.*, "Qwen look again: Guiding vision-language reasoning models to re-attention visual information," *arXiv*:2505.23558, 2025.
- [43] C. Zhang, P. Hao, X. Cao, X. Hao, S. Cui, and S. Wang, "Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation," *arXiv*:2505.09577, 2025.
- [44] P. Chen *et al.*, "Combatvla: An efficient vision-language-action model for combat tasks in 3d action role-playing games," *arXiv*:2503.09527, 2025.
- [45] C. Cui *et al.*, "Openhelix: A short survey, empirical analysis, and open-source dual-system vla model for robotic manipulation," *arXiv*:2505.03912, 2025.
- [46] J. Devlin, M.-W. Chang, K. Lee, and K. 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)*, 2019, pp. 4171–4186.
- [47] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.
- [48] S. Wang, S. Liu, W. Wang, J. Shan, and B. Fang, "RoboBERT: An end-to-end multimodal robotic manipulation model," *arXiv*:2502.07837, 2025.
- [49] O. M. Team *et al.*, "Octo: An open-source generalist robot policy," *arXiv*:2405.12213, 2024.
- [50] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv*:2307.09288, 2023.
- [51] W. Dai *et al.*, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 49 250–49 267, 2023.
- [52] G. Team *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv*:2403.08295, 2024.
- [53] Z. Zhou, Y. Zhu, J. Wen, C. Shen, and Y. Xu, "Vision-language-action model with open-world embodied reasoning from pre-trained knowledge," *arXiv*:2505.21906, 2025.
- [54] G. Team *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv*:2408.00118, 2024.
- [55] Z. Cai *et al.*, "Internlm2 technical report," *arXiv*:2403.17297, 2024.
- [56] X. Li *et al.*, "Vision-language foundation models as effective robot imitators," *arXiv*:2311.01378, 2023.
- [57] A. Awadalla *et al.*, "Openflamingo: An open-source framework for training large autoregressive vision-language models," *arXiv*:2308.01390, 2023.
- [58] J. Wen *et al.*, "Diffusion-vla: Generalizable and interpretable robot foundation model via self-generated reasoning," *arXiv*:2412.03293, 2024.
- [59] H. Shi *et al.*, "Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation," *arXiv*:2508.19236, 2025.
- [60] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh, "Prismatic vlms: Investigating the design space of visually-conditioned language models," in *Forty-first International Conference on Machine Learning*, 2024.
- [61] B. Xie *et al.*, "Dexbotic: Open-source vision-language-action toolbox," *arXiv*:2510.23511, 2025.
- [62] S. Yang *et al.*, "Instructvla: Vision-language-action instruction tuning from understanding to manipulation," *arXiv*:2507.17520, 2025.
- [63] Z. Zhong *et al.*, "Flowvla: Thinking in motion with a visual chain of thought," *arXiv*, pp. arXiv–2508, 2025.
- [64] D. Lim, M.-J. Kim, J. Cha, D. Kim, and J. Park, "Proprioceptive external torque learning for floating base robot and its applications to humanoid locomotion," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 8510–8517.
- [65] Z. Zhang *et al.*, "Ta-vla: Elucidating the design space of torque-aware vision-language-action models," *arXiv*:2509.07962, 2025.
- [66] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [67] H. Wu *et al.*, "Unleashing large-scale video generative pre-training for visual robot manipulation," *arXiv*:2312.13139, 2023.
- [68] C.-L. Cheang *et al.*, "Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation," *arXiv*:2410.06158, 2024.
- [69] S. Reed *et al.*, "A generalist agent," *arXiv*:2205.06175, 2022.
- [70] Y. Jiang *et al.*, "Vima: General robot manipulation with multi-modal prompts," *arXiv*:2210.03094, vol. 2, no. 3, p. 6, 2022.
- [71] J. Liu *et al.*, "Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 40 085–40 110, 2024.
- [72] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *ArXiv*, vol. abs/2312.00752, 2023.
- [73] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [74] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv*:2210.02747, 2022.
- [75] Y. Chen, S. Tian, S. Liu, Y. Zhou, H. Li, and D. Zhao, "Conrft: A reinforced fine-tuning method for vla models via consistency policy," *arXiv*:2502.05450, 2025.
- [76] B. Zitkovich *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.
- [77] D. Driess *et al.*, "Palm-e: An embodied multimodal language model," 2023.
- [78] Q. Zhao *et al.*, "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1702–1713.
- [79] B. Zhang *et al.*, "Safevla: Towards safety alignment of vision-language-action model via safe reinforcement learning," *arXiv*, pp. arXiv–2503, 2025.
- [80] J. Cen *et al.*, "Worldvla: Towards autoregressive action world model," *arXiv*:2506.21539, 2025.
- [81] R. Zheng *et al.*, "Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies," *arXiv*:2412.10345, 2024.
- [82] C. Li, J. Wen, Y. Peng, Y. Peng, F. Feng, and Y. Zhu, "Pointvla: Injecting the 3d world into vision-language-action models," *arXiv*:2503.07511, 2025.
- [83] H. Zhen *et al.*, "3d-vla: A 3d vision-language-action generative world model," *arXiv*:2403.09631, 2024.
- [84] P. Li *et al.*, "Bridgevla: Input-output alignment for efficient 3d manipulation learning with vision-language models," *arXiv*:2506.07961, 2025.
- [85] B. Han, J. Kim, and J. Jang, "A dual process vla: Efficient robotic manipulation leveraging vlm," *arXiv*:2410.15549, 2024.
- [86] Y. Li *et al.*, "Hamster: Hierarchical action models for open-world robot manipulation," *arXiv*:2502.05485, 2025.
- [87] Y. Guo *et al.*, "Improving vision-language-action model with online reinforcement learning," *arXiv*:2501.16664, 2025.
- [88] P. Anderson *et al.*, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3674–3683.
- [89] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1–10.

- [90] M. Chevalier-Boisvert *et al.*, "Babyai: A platform to study the sample efficiency of grounded language learning," *arXiv:1810.08272*, 2018.
- [91] X. Wang *et al.*, "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6629–6638.
- [92] E. Wijmans *et al.*, "Embodied question answering in photorealistic environments with point cloud perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6659–6668.
- [93] M. Shridhar *et al.*, "Alfred: A benchmark for interpreting grounded instructions for everyday tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10740–10749.
- [94] M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler, and M. Hausknecht, "Alfworld: Aligning text and embodied environments for interactive learning," *arXiv:2010.03768*, 2020.
- [95] S. Srivastava *et al.*, "Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments," in *Conference on robot learning*. PMLR, 2022, pp. 477–490.
- [96] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on robot learning*. PMLR, 2022, pp. 894–906.
- [97] M. Ahn *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv:2204.01691*, 2022.
- [98] W. Huang *et al.*, "Inner monologue: Embodied reasoning through planning with language models," *arXiv:2207.05608*, 2022.
- [99] A. Brohan *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv:2212.06817*, 2022.
- [100] Q. Vuong *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," in *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023.
- [101] P. Ding *et al.*, "Humanoid-vla: Towards universal humanoid control with visual integration," *arXiv:2502.14795*, 2025.
- [102] J. Björck *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv:2503.14734*, 2025.
- [103] A. Azzolini *et al.*, "Cosmos-reason1: From physical common sense to embodied reasoning," *arXiv:2503.15558*, 2025.
- [104] G. A. Team, "Gen-0: Embodied foundation models that scale with physical interaction," *Generalist AI Blog*, 2025, <https://generalistai.com/blog/preview-uqlxvb-bb.html>.
- [105] W. Li *et al.*, "Semanticvila: Semantic-aligned sparsification and enhancement for efficient robotic manipulation," *arXiv:2511.10518*, 2025.
- [106] Y. J. Ma, V. Kumar, A. Zhang, O. Bastani, and D. Jayaraman, "Liv: Language-image representations and rewards for robotic control," in *International Conference on Machine Learning*. PMLR, 2023, pp. 23 301–23 320.
- [107] J. Bi, L. B. Wen, Z. Liu, and C. Xu, "Actllm: Action consistency tuned large language model," *arXiv:2506.21250*, 2025.
- [108] Y. Hu, F. Lin, T. Zhang, L. Yi, and Y. Gao, "Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning," *arXiv:2311.17842*, 2023.
- [109] D. Driess *et al.*, "Knowledge insulating vision-language-action models: Train fast, run fast, generalize better," *arXiv:2505.23705*, 2025.
- [110] Y.-J. Wang, B. Zhang, J. Chen, and K. Sreenath, "Prompt a robot to walk with large language models," in *2024 IEEE 63rd Conference on Decision and Control (CDC)*. IEEE, 2024, pp. 1531–1538.
- [111] A. Szot, B. Mazoure, H. Agrawal, R. D. Hjelm, Z. Kira, and A. Toshov, "Grounding multimodal large language models in actions," *Advances in Neural Information Processing Systems*, vol. 37, pp. 20 198–20 224, 2024.
- [112] G.-C. Kang, J. Kim, K. Shim, J. K. Lee, and B.-T. Zhang, "Cliprt: Learning language-conditioned robotic policies from natural language supervision," *arXiv:2411.00508*, 2024.
- [113] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," *arXiv:2307.05973*, 2023.
- [114] G. R. Team *et al.*, "Gemini robotics: Bringing ai into the physical world," *arXiv:2503.20020*, 2025.
- [115] H. Fu *et al.*, "Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation," *arXiv:2503.19755*, 2025.
- [116] P. Hao *et al.*, "Tla: Tactile-language-action model for contact-rich manipulation," *arXiv:2503.08548*, 2025.
- [117] J. Jones, O. Mees, C. Sferrazza, K. Stachowicz, P. Abbeel, and S. Levine, "Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding," *arXiv:2501.04693*, 2025.
- [118] H.-S. Fang *et al.*, "Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot," *arXiv:2307.00595*, 2023.
- [119] J. Bi, K. Y. Ma, C. Hao, M. Z. Shou, and H. Soh, "Vla-touch: Enhancing vision-language-action models with dual-level tactile feedback," *arXiv:2507.17294*, 2025.
- [120] Z. Cheng *et al.*, "Omnivila: Vision-tactile-language-action model with semantic-aligned tactile sensing," *arXiv:2508.08706*, 2025.
- [121] J. Huang, S. Wang, F. Lin, Y. Hu, C. Wen, and Y. Gao, "Tactile-vla: Unlocking vision-language-action model's physical knowledge for tactile generalization," *arXiv:2507.09160*, 2025.
- [122] R. Wang *et al.*, "The sound of simulation: Learning multimodal sim-to-real robot policies with generative audio," in *9th Annual Conference on Robot Learning*, 2025.
- [123] X. Pang *et al.*, "Depth helps: Improving pre-trained rgb-based policy with depth information injection. in 2024 ieee," in *RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 7251–7256.
- [124] S. Wang, "Roboflamingo-plus: Fusion of depth and rgb perception with vision-language models for enhanced robotic manipulation," *arXiv:2503.19510*, 2025.
- [125] L. Sun, B. Xie, Y. Liu, H. Shi, T. Wang, and J. Cao, "Geovla: Empowering 3d representations in vision-language-action models," *arXiv:2508.09071*, 2025.
- [126] J. Huang *et al.*, "An embodied generalist agent in 3d world," *arXiv:2311.12871*, 2023.
- [127] R. Yang, G. Chen, C. Wen, and Y. Gao, "Fp3: A 3d foundation policy for robotic manipulation," *arXiv:2503.08950*, 2025.
- [128] Z. Qi *et al.*, "Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation," *arXiv:2502.13143*, 2025.
- [129] X. Xu *et al.*, "Weakly-supervised 3d visual grounding based on visual language alignment," *IEEE Transactions on Multimedia*, 2025.
- [130] Y. Li, G. Yan, A. Macaluso, M. Ji, X. Zou, and X. Wang, "Integrating lmm planners and 3d skill policies for generalizable manipulation," *arXiv:2501.18733*, 2025.
- [131] J. Wei, S. Yuan, P. Li, Q. Hu, Z. Gan, and W. Ding, "Occlama: An occupancy-language-action generative world model for autonomous driving," *arXiv:2409.03272*, 2024.
- [132] F. Yan *et al.*, "Robomm: All-in-one multimodal large model for robotic manipulation," *arXiv:2412.07215*, 2024.
- [133] D. Niu *et al.*, "Pre-training auto-regressive robotic models with 4d representations," *arXiv:2502.13142*, 2025.
- [134] D. Qu *et al.*, "Spatialvla: Exploring spatial representations for visual-language-action model," *arXiv:2501.15830*, 2025.
- [135] T. Lin *et al.*, "Evo-0: Vision-language-action model with implicit spatial understanding," *arXiv:2507.00416*, 2025.
- [136] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5294–5306.
- [137] S. Chen *et al.*, "Ac-dit: Adaptive coordination diffusion transformer for mobile manipulation," *arXiv:2507.01961*, 2025.
- [138] R. Xu *et al.*, "A0: An affordance-aware hierarchical model for general robotic manipulation," *arXiv:2504.12636*, 2025.
- [139] W. Yuan *et al.*, "Robopoint: A vision-language model for spatial affordance prediction for robotics," *arXiv:2406.10721*, 2024.
- [140] M. A. Patratskiy, A. K. Kovalev, and A. I. Panov, "Spatial traces: Enhancing vla models with spatial-temporal understanding," *arXiv:2508.09032*, 2025.
- [141] J. Zhang, Y. Guo, Y. Hu, X. Chen, X. Zhu, and J. Chen, "Up-vla: A unified understanding and prediction model for embodied agent," *arXiv:2501.18867*, 2025.
- [142] W. Zhang *et al.*, "Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge," *arXiv:2507.04447*, 2025.
- [143] M. Assran *et al.*, "V-jpa 2: Self-supervised video models enable understanding, prediction and planning," *arXiv:2506.09985*, 2025.
- [144] Y. Wu, R. Tian, G. Swamy, and A. Bajcsy, "From foresight to forethought: Vlm-in-the-loop policy steering via latent alignment," *arXiv:2502.01828*, 2025.
- [145] F. Zhu, Z. Yan, Z. Hong, Q. Shou, X. Ma, and S. Guo, "Wmpo: World model-based policy optimization for vision-language-action models," *arXiv:2511.09515*, 2025.

- [146] Q. Lv *et al.*, "F1: A vision-language-action model bridging understanding and generation to actions," *arXiv:2509.06951*, 2025.
- [147] W. Zhao, G. Li, Z. Gong, P. Ding, H. Zhao, and D. Wang, "Unveiling the potential of vision-language-action models with open-ended multimodal instructions," *arXiv:2505.11214*, 2025.
- [148] C. Fan *et al.*, "Interleave-vla: Enhancing robot manipulation with interleaved image-text instructions," *arXiv:2505.02152*, 2025.
- [149] C.-P. Huang, Y.-H. Wu, M.-H. Chen, Y.-C. F. Wang, and F.-E. Yang, "Thinkact: Vision-language-action reasoning via reinforced visual latent planning," *arXiv:2507.16815*, 2025.
- [150] C. Yin *et al.*, "Deepthinkvla: Enhancing reasoning capability of vision-language-action models," 2025.
- [151] J. Zhang *et al.*, "Inspire: Vision-language-action models with intrinsic spatial reasoning," *arXiv:2505.13888*, 2025.
- [152] X. Zhang *et al.*, "Asktoact: Enhancing llms tool use via self-correcting clarification," *arXiv:2503.01940*, 2025.
- [153] Y. Fan *et al.*, "Long-vla: Unleashing long-horizon capability of vision language action model for robot manipulation," *arXiv:2508.19958*, 2025.
- [154] S. Kanta *et al.*, "Toward general physical intelligence for resilient agile manufacturing automation," *arXiv:2508.11960*, 2025.
- [155] C.-P. Huang, Y.-H. Wu, M.-H. Chen, Y.-C. F. Wang, and F.-E. Yang, "Thinkact: Vision-language-action reasoning via reinforced visual latent planning," *arXiv:2507.16815*, 2025.
- [156] A. Abdolmaleki *et al.*, "Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer," *arXiv:2510.03342*, 2025.
- [157] F. Lin, R. Nai, Y. Hu, J. You, J. Zhao, and Y. Gao, "Onetwovla: A unified vision-language-action model with adaptive reasoning," *arXiv:2505.11917*, 2025.
- [158] Y. Yang, J. Sun, S. Kou, Y. Wang, and Z. Deng, "Lohovla: A unified vision-language-action model for long-horizon embodied tasks," *arXiv:2506.00411*, 2025.
- [159] N. Chung *et al.*, "Rethinking progression of memory state in robotic manipulation: An object-centric perspective," *arXiv:2511.11478*, 2025.
- [160] J. Jiang, F. Deng, G. Singh, M. Lee, and S. Ahn, "Slot state space models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 11 602–11 633, 2024.
- [161] A. Ajay *et al.*, "Compositional foundation models for hierarchical planning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 22 304–22 325, 2023.
- [162] S. Nasiriany *et al.*, "Rt-affordance: Affordances are versatile intermediate representations for robot manipulation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 8249–8257.
- [163] J. Li *et al.*, "Coa-vla: Improving vision-language-action models via visual-text chain-of-affordance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 9759–9769.
- [164] D. Li *et al.*, "An atomic skill library construction method for data-efficient embodied manipulation," *arXiv:2501.15068*, 2025.
- [165] Z. Yang *et al.*, "Agentic robot: A brain-inspired framework for vision-language-action models in embodied agents," *arXiv:2505.23450*, 2025.
- [166] Y. Ji *et al.*, "Robobrain: A unified brain model for robotic manipulation from abstract to concrete," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1724–1734.
- [167] J. Wen, Y. Zhu, J. Li, Z. Tang, C. Shen, and F. Feng, "Dexvla: Vision-language model with plug-in diffusion expert for general robot control," *arXiv:2502.05855*, 2025.
- [168] Q. Bu *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," *arXiv:2503.06669*, 2025.
- [169] L. X. Shi *et al.*, "Yell at your robot: Improving on-the-fly from language corrections," *arXiv:2403.12910*, 2024.
- [170] Z. Yu *et al.*, "Correctnav: Self-correction flywheel empowers vision-language-action navigation model," *arXiv:2508.10416*, 2025.
- [171] Y. Yang *et al.*, "Fpc-vla: A vision-language-action framework with a supervisor for failure prediction and correction," *arXiv:2509.04018*, 2025.
- [172] H. Wang, C. Xiong, R. Wang, and X. Chen, "Bitvla: 1-bit vision-language-action models for robotics manipulation," *arXiv:2506.07530*, 2025.
- [173] T. Lin *et al.*, "Evo-1: Lightweight vision-language-action model with preserved semantic alignment," *arXiv:2511.04555*, 2025.
- [174] H. Fang, Y. Liu, Y. Du, L. Du, and H. Yang, "Sqap-vla: A synergistic quantization-aware pruning framework for high-performance vision-language-action models," *arXiv:2509.09090*, 2025.
- [175] C.-Y. Hung *et al.*, "Nora: A small open-sourced generalist vision language action model for embodied tasks," *arXiv:2504.19854*, 2025.
- [176] J. Wen *et al.*, "Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation," *IEEE Robotics and Automation Letters*, 2025.
- [177] Y. Wang *et al.*, "Vla-adapter: An effective paradigm for tiny-scale vision-language-action model," *arXiv:2509.09372*, 2025.
- [178] I. Leal *et al.*, "Sara-rt: Scaling up robotics transformers with self-adaptive robust attention," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6920–6927.
- [179] R. Zhang *et al.*, "Mole-vla: Dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation," *arXiv:2503.20384*, 2025.
- [180] W. Song *et al.*, "Ceed-vla: Consistency vision-language-action model with early-exit decoding," *arXiv:2506.13725*, 2025.
- [181] S. Xu, Y. Wang, C. Xia, D. Zhu, T. Huang, and C. Xu, "Vla-cache: Towards efficient vision-language-action model via adaptive token caching in robotic manipulation," *arXiv:2502.02175*, 2025.
- [182] H. Wang, J. Xu, J. Pan, Y. Zhou, and G. Dai, "Specprune-vla: Accelerating vision-language-action models via action-aware self-speculative pruning," *arXiv:2509.05614*, 2025.
- [183] W. Li, R. Zhang, R. Shao, J. He, and L. Nie, "Cogvla: Cognition-aligned vision-language-action model via instruction-driven routing & sparsification," *arXiv:2508.21046*, 2025.
- [184] W. Song *et al.*, "Accelerating vision-language-action model integrated with action chunking via parallel decoding," *arXiv:2503.02310*, 2025.
- [185] S. Wang *et al.*, "Spec-vla: speculative decoding for vision-language-action models with relaxed acceptance," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 26 916–26 928.
- [186] K. Pertsch *et al.*, "Fast: Efficient action tokenization for vision-language-action models," *arXiv:2501.09747*, 2025.
- [187] S. Fan *et al.*, "Xr-1: Towards versatile vision-language-action models via learning unified vision-motion representations," *arXiv:2511.02776*, 2025.
- [188] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.
- [189] Y. Wang, H. Zhu, M. Liu, J. Yang, H.-S. Fang, and T. He, "Vqvla: Improving vision-language-action models via scaling vector-quantized action tokenizers," *arXiv:2507.01016*, 2025.
- [190] M. Shukor *et al.*, "Smolvla: A vision-language-action model for affordable and efficient robotics," *arXiv:2506.01844*, 2025.
- [191] K. Black, M. Y. Galliker, and S. Levine, "Real-time execution of action chunking flow policies," *arXiv:2506.07339*, 2025.
- [192] Y. Niu, S. Zhou, Y. Li, Y. Den, and L. Wang, "Time-unified diffusion policy with action discrimination for robotic manipulation," *arXiv:2506.09422*, 2025.
- [193] Z. Liang *et al.*, "Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies," *arXiv:2508.20072*, 2025.
- [194] W. Chen *et al.*, "Training strategies for efficient embodied reasoning," *arXiv:2505.08243*, 2025.
- [195] H. Chen *et al.*, "Fast-in-slow: A dual-system foundation model unifying fast manipulation within slow reasoning," *arXiv:2506.01953*, 2025.
- [196] W. Zheng, B. Li, B. Xu, E. Feng, J. Gu, and H. Chen, "Leveraging os-level primitives for robotic action management," *arXiv:2508.10259*, 2025.
- [197] C. Miao *et al.*, "Fedvla: Federated vision-language-action learning with dual gating mixture-of-experts for robotic manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 6904–6913.
- [198] K. Bousmalis *et al.*, "Robocat: A self-improving generalist agent for robotic manipulation," *arXiv:2306.11706*, 2023.
- [199] Z. Hou *et al.*, "Diffusion transformer policy," *arXiv:2410.15959*, 2024.
- [200] D. Qu *et al.*, "Embodiedonevision: Interleaved vision-text-action pretraining for general robot control," *arXiv*, pp. arXiv–2508, 2025.

- [201] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," *arXiv:2203.12601*, 2022.
- [202] K. Grauman *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*, 2022, pp. 18 995–19 012.
- [203] L. Fu *et al.*, "In-context imitation learning via next-token prediction," *arXiv:2408.15980*, 2024.
- [204] V. Myers, B. C. Zheng, A. Dragan, K. Fang, and S. Levine, "Temporal representation alignment: Successor features enable emergent compositionality in robot instruction following," *arXiv:2502.05454*, 2025.
- [205] M. Zhu *et al.*, "Objectvla: End-to-end open-world object manipulation without demonstration," *arXiv:2502.19250*, 2025.
- [206] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 19 729–19 739.
- [207] Y. Zhang *et al.*, "Align-then-steer: Adapting the vision-language action models through unified latent guidance," *arXiv:2509.02055*, 2025.
- [208] H. Etukuru *et al.*, "Robot utility models: General policies for zero-shot deployment in new environments," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 8275–8283.
- [209] Z. Mandi, H. Bharadhwaj, V. Moens, S. Song, A. Rajeswaran, and V. Kumar, "Cacti: 256 a framework for scalable multi-task multi-scene visual imitation learning. 257," *arXiv:2212.05711*, vol. 258, 2022.
- [210] Z. Chen, S. Kiami, A. Gupta, and V. Kumar, "Genaug: Retargeting behaviors to unseen situations via generative augmentation," *arXiv:2302.06671*, 2023.
- [211] T. Yu *et al.*, "Scaling robot learning with semantically imagined experience," *arXiv:2302.11550*, 2023.
- [212] Y. Li, G. Yan, A. Macaluso, M. Ji, X. Zou, and X. Wang, "Integrating lmm planners and 3d skill policies for generalizable manipulation," *arXiv:2501.18733*, 2025.
- [213] S. Haldar, Z. Peng, and L. Pinto, "Baku: An efficient transformer for multi-task policy learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 141 208–141 239, 2024.
- [214] W. Liu, Y. Du, T. Hermans, S. Chernova, and C. Paxton, "Structdiffusion: Language-guided creation of physically-valid structures using unseen objects," *arXiv:2211.04604*, 2022.
- [215] Y. Yao *et al.*, "Think small, act big: Primitive prompt learning for lifelong robot manipulation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 573–22 583.
- [216] Z. Zheng *et al.*, "imanip: Skill-incremental learning for robotic manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 13 890–13 900.
- [217] S. N. Syed, Y. Ahuja, A. Jakobsson, and J. Ichnowski, "Expres-vla: Specializing vision-language-action models through experience replay and retrieval," *arXiv:2511.06202*, 2025.
- [218] H. Zhang *et al.*, "Slim: Sim-to-real legged instructive manipulation via long-horizon visuomotor learning," *arXiv:2501.09905*.
- [219] S. Tao *et al.*, "Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai," *arXiv:2410.00425*.
- [220] J. Jang *et al.*, "Dreamgen: Unlocking generalization in robot learning through video world models," *arXiv:2505.12705*, 2025.
- [221] Y. Jiang *et al.*, "Rynnvla-001: Using human demonstrations to improve robot manipulation," *arXiv:2509.15212*, 2025.
- [222] Y. Guo *et al.*, "Improving vision-language-action model with online reinforcement learning," *arXiv:2501.16664*, 2025.
- [223] C. Xu, Q. Li, J. Luo, and S. Levine, "Rldg: Robotic generalist policy distillation via reinforcement learning," *arXiv:2412.09858*, 2024.
- [224] D. Huang, Z. Fang, T. Zhang, Y. Li, L. Zhao, and C. Xia, "Co-rft: Efficient fine-tuning of vision-language-action models through chunked offline reinforcement learning," *arXiv:2508.02219*, 2025.
- [225] T. Jülg, W. Burgard, and F. Walter, "Refined policy distillation: From vla generalists to rl experts," *arXiv:2503.05833*, 2025.
- [226] Z. Zhang *et al.*, "Grape: Generalizing robot policy via preference alignment," *arXiv:2411.19309*, 2024.
- [227] C.-Y. Hung *et al.*, "Nora-1.5: A vision-language-action model trained using world model- and action-based preference rewards," *arXiv:2511.14659*, 2025.
- [228] J. Rocamonde, V. Montesinos, E. Nava, E. Perez, and D. Lindner, "Vision-language models are zero-shot reward models for reinforcement learning," *arXiv:2310.12921*, 2023.
- [229] S. Sontakke *et al.*, "Roboclip: One demonstration is enough to learn robot policies," *Advances in Neural Information Processing Systems*, vol. 36, pp. 55 681–55 693, 2023.
- [230] O. Y. Lee, A. Xie, K. Fang, K. Pertsch, and C. Finn, "Affordance-guided reinforcement learning via visual prompting," *arXiv:2407.10341*, 2024.
- [231] Y. Wang *et al.*, "Rl-vlm-f: Reinforcement learning from vision language foundation model feedback," *arXiv:2402.03681*, 2024.
- [232] Y. J. Ma *et al.*, "Eureka: Human-level reward design via coding large language models," *arXiv:2310.12931*, 2023.
- [233] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "Vip: Towards universal visual reward and representation via value-implicit pre-training," *arXiv:2210.00030*, 2022.
- [234] M. Ahn *et al.*, "Autort: Embodied foundation models for large scale orchestration of robotic agents," *arXiv:2401.12963*, 2024.
- [235] B. Häon, K. Stocking, I. Chuang, and C. Tomlin, "Mechanistic interpretability for steering vision-language-action models," *arXiv:2509.00328*, 2025.
- [236] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine, "Robotic control via embodied chain-of-thought reasoning," *arXiv:2407.08693*, 2024.
- [237] M. Neau, Z. Falomir, P. E. Santos, A.-G. Bosser, and C. Buche, "Grasp-vla: Graph-based symbolic action representation for long-horizon planning with vla policies," *arXiv:2511.04357*, 2025.
- [238] H. Lu, H. Li, P. S. Shahani, S. Herbers, and M. Scheutz, "Probing a vision-language-action model for symbolic states and integration into a cognitive architecture," *arXiv:2502.04558*, 2025.
- [239] X. Li *et al.*, "Object-centric prompt-driven vision-language-action model for robotic manipulation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27 638–27 648.
- [240] M. Li *et al.*, "Switchvla: Execution-aware task switching for vision-language-action models," *arXiv:2506.03574*, 2025.
- [241] S. Ye *et al.*, "Latent action pretraining from videos," *arXiv:2410.11758*, 2024.
- [242] Y. Chen *et al.*, "Moto: Latent motion token as the bridging language for learning robot manipulation from videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 19 752–19 763.
- [243] R. Doshi, H. Walke, O. Mees, S. Dasari, and S. Levine, "Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation," *arXiv:2408.11812*, 2024.
- [244] C. Chen, Q. Yang, X. Xu, N. Fazeli, and O. Andersson, "Visa-flow: Accelerating robot skill learning via large-scale video semantic action flow," *arXiv:2505.01288*, 2025.
- [245] R. Yang *et al.*, "Egovla: Learning vision-language-action models from egocentric human videos," *arXiv:2507.12440*, 2025.
- [246] T. Tao, M. K. Srirama, J. J. Liu, K. Shaw, and D. Pathak, "Dexwild: Dexterous human interactions for in-the-wild robot policies," *arXiv:2505.07813*, 2025.
- [247] W. Xiao *et al.*, "Self-improving vision-language-action models with data generation via residual rl," *arXiv:2511.00091*, 2025.
- [248] J. Hejna, C. Bhateja, Y. Jiang, K. Pertsch, and D. Sadigh, "Remix: Optimizing data mixtures for large scale imitation learning," *arXiv:2408.14037*, 2024.
- [249] H. R. Walke *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning*. PMLR, 2023, pp. 1723–1736.
- [250] S. Nasiriany *et al.*, "Robocasa: Large-scale simulation of everyday tasks for generalist robots," *arXiv:2406.02523*, 2024.
- [251] H. Arai *et al.*, "Covla: Comprehensive vision-language-action dataset for autonomous driving," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 1933–1943.
- [252] D. Damen *et al.*, "The epic-kitchens dataset: Collection, challenges and baselines," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4125–4141, 2020.
- [253] K. Grauman *et al.*, "Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 383–19 400.
- [254] M. U. Din, W. Akram, L. S. Saoud, J. Rosell, and I. Hussain, "Vision language action models in robotic manipulation: A systematic review," *arXiv:2507.10672*, 2025.

- [255] P. Guruprasad, H. Sikka, J. Song, Y. Wang, and P. P. Liang, "Benchmarking vision, language, & action models on robotic learning tasks," *arXiv:2411.05821*, 2024.
- [256] P. Valle, C. Lu, S. Ali, and A. Arrieta, "Evaluating uncertainty and quality of visual language action-enabled robots," *arXiv:2507.17049*, 2025.
- [257] Y. Zhu *et al.*, "robosuite: A modular simulation framework and benchmark for robot learning," *arXiv:2009.12293*, 2020.
- [258] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7327–7334, 2022.
- [259] B. Liu *et al.*, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44776–44791, 2023.
- [260] I. Fang, J. Zhang, S. Tong, and C. Feng, "From intention to execution: Probing the generalization boundaries of vision-language-action models," *arXiv:2506.09930*, 2025.
- [261] Z. Zhou *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*, 2025, pp. 5377–5395.
- [262] T. Shen *et al.*, "The journey/dao/tao of embodied intelligence: From large models to foundation intelligence and parallel intelligence," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 6, pp. 1313–1316, 2024.
- [263] Q. Li *et al.*, "Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation," *arXiv:2411.19650*, 2024.
- [264] S. Li, Y. Gao, D. Sadigh, and S. Song, "Unified video action model," *arXiv:2503.00200*, 2025.
- [265] S. Ye *et al.*, "Latent action pretraining from videos," *arXiv:2410.11758*, 2024.
- [266] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv:1707.06347*, 2017.
- [267] J. Xiao *et al.*, "World-env: Leveraging world model as a virtual environment for vla post-training," *arXiv:2509.24948*, 2025.
- [268] H. Zhang, S. Zhang, J. Jin, Q. Zeng, R. Li, and D. Wang, "Robustvla: Robustness-aware reinforcement post-training for vision-language-action models," *arXiv:2511.01331*, 2025.
- [269] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. Pmlr, 2018, pp. 1861–1870.
- [270] J. Luo *et al.*, "Serl: A software suite for sample-efficient robotic reinforcement learning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 16 961–16 969.
- [271] J. Luo, C. Xu, J. Wu, and S. Levine, "Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning," *Science Robotics*, vol. 10, no. 105, p. eads5033, 2025.
- [272] P. Intelligence, " $\pi_{0,6}^*$: A vla that learns from experience," <https://pi.website/blog/pistar06>, accessed: 2025-02-18,2025.
- [273] H. Zhang, N. Zantout, P. Kachana, Z. Wu, J. Zhang, and W. Wang, "Vla-3d: A dataset for 3d semantic scene understanding and navigation," *arXiv:2411.03540*, 2024.
- [274] A. Khazatsky *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv:2403.12945*, 2024.
- [275] T. Perrett *et al.*, "Hd-epic: A highly-detailed egocentric video dataset," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 23 901–23 913.
- [276] Y. Liu *et al.*, "Hoi4d: A 4d egocentric dataset for category-level human-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 013–21 022.
- [277] A. Padmakumar *et al.*, "Teach: Task-driven embodied agents that chat," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2017–2025.
- [278] L. Yu, X. Chen, G. Gkioxari, M. Bansal, T. L. Berg, and D. Batra, "Multi-target embodied question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6309–6318.
- [279] B. Jia, T. Lei, S.-C. Zhu, and S. Huang, "Egotaskqa: Understanding human tasks in egocentric videos," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3343–3360, 2022.
- [280] Z. Cheng *et al.*, "Embodiedeval: Evaluate multimodal llms as embodied agents," *arXiv:2501.11858*, 2025.
- [281] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.
- [282] T. Mu *et al.*, "Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations," *arXiv:2107.14483*, 2021.
- [283] J. Gu *et al.*, "Maniskill2: A unified benchmark for generalizable manipulation skills," *arXiv:2302.04659*, 2023.
- [284] A. Shukla, S. Tao, and H. Su, "Maniskill-hab: A benchmark for low-level manipulation in home rearrangement tasks," *arXiv:2412.13211*, 2024.
- [285] A. Mandlekar *et al.*, "What matters in learning from offline human demonstrations for robot manipulation," *arXiv:2108.03298*, 2021.
- [286] L. Zheng, F. Yan, F. Liu, C. Feng, Z. Kang, and L. Ma, "Robocas: A benchmark for robotic manipulation in complex object arrangement scenarios," *arXiv:2407.06951*, 2024.
- [287] R. Yang *et al.*, "Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents," *arXiv:2502.09560*, 2025.
- [288] H. Yue *et al.*, "Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models," *arXiv:2505.09694*, 2025.
- [289] T. Chen *et al.*, "Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation," *arXiv:2506.18088*, 2025.

A APPENDIX

A.1 Applications

The true measure of Vision-Language-Action (VLA) models lies in their ability to solve real-world problems. By integrating perception, reasoning, and control, these models are uniquely equipped to translate abstract human intent into grounded, executable actions, bridging the long-standing gap between high-level cognition and low-level robotics. Leveraging large-scale pretraining, VLA-driven robots demonstrate unprecedented capabilities in generalization and adaptation, surpassing conventional modular pipelines in both autonomy and efficiency. This section surveys their transformative impact across two primary domains: household robotics and industrial automation.

A.1.1 Embodied Manipulation and Household Robotics

The unstructured, dynamic, and human-centric nature of household environments makes them a major proving ground for VLA models. Unlike structured factory settings, homes require robots to understand natural language, handle a vast diversity of unseen objects, and perform complex, long-horizon tasks.

VLA models excel in this domain precisely because their core architecture is well-suited to these challenges. Their ability to leverage internet-scale knowledge allows them to recognize and interact with a near-infinite variety of household items without task-specific training (e.g., SayCan [97], RT-1 [99], RT-2 [76]). The evolution of benchmarks from ALFRED [93] to real-world validations like ChatVLA [261] confirms their robustness in sequential reasoning. Furthermore, the hierarchical reasoning inherent in many VLA systems (e.g., Helix [262]) enables the decomposition of vague commands like “clean the kitchen” into concrete, executable subtasks.

Looking ahead, the next frontier for household VLA systems lies in achieving true personalization and collaborative intelligence. Future models must move beyond simply executing one-off commands and learn to understand a user’s long-term preferences, habits, and implicit intentions. This requires a deeper integration of interactive learning, where robots can learn from real-time verbal feedback, ask clarifying questions when faced with ambiguity, and even proactively suggest actions based on learned routines.

Furthermore, to become truly ubiquitous, these systems must operate on low-power, on-device hardware, necessitating breakthroughs in model efficiency and compression. The ultimate goal is to transform household robots from simple instruction-followers into proactive, adaptive, and truly personalized domestic assistants.

A.1.2 Industrial and Field Robotics

Following their success in household scenarios, VLA models now extend to industrial domains, where they promise to bring unprecedented flexibility to manufacturing, logistics, and field operations. Industrial environments, however, impose far stricter demands on precision, reliability, and safety. The evolution of VLA models for industrial use is therefore characterized by a clear focus on enhancing their physical intelligence and robustness.

This evolution is proceeding along three major directions: (1) Incorporating physical perception via tactile and force sensors (e.g., Tactile-VLA [121], VTVA [43]); (2) Developing industrial-grade Reasoning for complex processes (e.g., ForceVLA [25], CogACT [263]); (3) Ensuring Safety and Reliability through mechanisms like safe reinforcement learning (e.g., SafeVLA [79]).

The future of VLA in industry hinges on bridging the gap between flexible intelligence and the rigorous demands of production environments. The next wave of innovation will likely focus on certification-ready safety and formal verification, moving beyond empirical safety to provide provable guarantees on robot behavior. Another critical direction is zero-shot adaptation to new tasks and parts in highly customized manufacturing, where reprogramming a robot for every new product is economically infeasible. This requires VLA models to learn from CAD files, technical manuals, and video demonstrations of human workers. Finally, the integration of VLA into multi-agent systems enables fleets of robots to collaboratively perform complex assembly or logistics tasks, coordinated by a central language-based understanding of the overall production goal.

A.2 Basic Modules

A.2.1 Training Strategy

Current VLA training follows three largely complementary routes that are often combined in practice:

(1) Behavioral Cloning (BC). In current VLA research, BC is the dominant paradigm: it formulates control as supervised imitation, learning a mapping from multimodal observations (i.e., vision, language, proprioception) to expert actions by minimizing the prediction demonstration discrepancy. In practice, BC underpins a broad spectrum of VLA systems across architectures, from diffusion-based controllers to multimodal Transformer generalists (e.g., Diffusion Policy [13], TriVLA [24], VIMA [70], Octo [49], RDT-1B [23], RT-H [38], Hi Robot [41], GR-2 [68], 3D-VLA [83], RoboMM [132]). Beyond flat policies, BC is also employed as a pre- or post-training stage in hierarchical pipelines, e.g., adapting continuous control in $\pi_{0.5}$ [53] and driving the fast S-Sys1 executor in Dual-Process VLA [85].

(2) Predictive Modeling. Instead of imitating actions, predictive modeling learns to anticipate the world in future observations or latent dynamics, providing powerful self-supervised signals that internalize physics and causality. World models exemplify this idea: WorldVLA [80], LUMOS [16], and UVA [264] train with predictive and self-supervised objectives, enabling effective learning from unstructured data and strong performance on complex, long-horizon robotic tasks. Additionally, other approaches learn discrete latent actions from unlabeled video as in self-supervised prediction like world modeling (e.g., LAPA [265]).

(3) Reinforcement Learning (RL). RL moves beyond demonstrations to optimize policies through interaction and reward feedback, and in VLA it is often built upon BC-pretrained backbones to enhance robustness and long-horizon performance. On-policy methods (e.g., PPO [266]) update from freshly collected rollouts (e.g., LUMOS [16], VLA-RL [36], RoboCLIP [229], World-Env [267],

RobustVLA [268], EUREKA [232], Refined Policy Distillation [225]); off-policy methods (e.g., SAC [269]) exploit replay for sample efficiency (e.g., ConRFT [75], SERL [270], RL-VLM-F [231]), with HIL-SERL [271] further leveraging human demos and online corrections. Beyond these on- and off-policy paradigms, the latest $\pi_{0.6}^*$ [272] introduces RE-CAP (RL with Experience and Corrections via Advantage-Conditioned Policies), a scalable RL framework for large VLA models that incorporates advantage-conditioned policy extraction into flow-matching/diffusion-based VLAs, enabling stable and scalable training without relying on complex RL objectives such as PPO.

A.2.2 Datasets

Recent progress in embodied intelligence is driven by a shift toward data-centric development, where the scale, diversity, and quality of training data largely determine a VLA model’s generalization and robustness. VLA datasets form a diverse and evolving ecosystem, each providing complementary supervision signals for different aspects of embodied reasoning and control. Tab. S1 is an overview of representative embodied datasets. This section categorizes major datasets by their core properties and primary research roles.

(1) Simulation-Centered Datasets. These datasets are collected in controlled virtual environments, which support large-scale, safe, and reproducible data generation with full access to state information. This makes them well suited for studying high-level reasoning and long-horizon planning. ALFRED [93] provides expert demonstrations for 25,000 language-grounded household tasks in AI2-THOR and emphasizes long-horizon compositionality. LIBERO [259] targets lifelong robot learning by offering procedurally varied tasks that evaluate incremental skill acquisition and retention. Recent datasets such as VLA-3D [273] incorporate detailed 3D scene representations paired with language instructions to support the development of 3D-aware vision–language–action models.

(2) Real-World Robotic Manipulation Datasets. These datasets are collected from real robotic systems and capture the full complexity of real-world sensing, dynamics, and environmental variability. They are essential for training policies that remain robust under uncertainty and can generalize to unstructured settings. BridgeData V2 [249] provides large-scale multi-task demonstrations collected across institutions using a standardized single-arm platform and serves as a central resource for generalist manipulation learning. DROID [274] expands task and environment diversity by offering more than 350K in-the-wild trajectories gathered from 50+ real environments with a low-cost mobile manipulator. AgiBot World [168] further increases scale with a million-level corpus spanning broad task and object variations to support large VLA model training. Open X-Embodiment (OXE) [100] aggregates over 60 datasets across 22 robot embodiments and currently represents the most comprehensive resource for studying cross-morphology transfer and the emergence of generalist policies.

(3) Human-Centric and Egocentric Datasets. These datasets capture data from a first-person human perspective. Although they typically do not include robotic action labels,

they are crucial for grounding perception in human experience and for learning to infer human intent. The primary approach involves large-scale, egocentric video collection. Ego4D [202] provides thousands of hours of egocentric video that support pretraining visual representations for human-object interaction, which can be effectively transferred to robotic policies. More specialized datasets further enhance this capability. HD-EPIC [275] offers detailed, annotated egocentric recordings of unscripted kitchen activities, and HOI4D [276] captures 4D human interactions with diverse objects, enabling fine-grained modeling of interaction dynamics. TEACH [277] shifts the focus to instruction following by collecting dialogue-driven task execution data, making it a valuable resource for training agents that can collaborate with humans and resolve ambiguities through communication.

(4) Embodied Visual Question Answering Datasets. Embodied VQA datasets pair visual scenes with language-based question-answer supervision and are increasingly used to train VLA models that require semantic alignment and environment-level reasoning. MT-EQA [278] provides 19,287 QA pairs for multi-target embodied question answering, requiring agents to navigate 3D indoor environments and infer object attributes through active exploration. Ego-TaskQA [279] expands cognitive scope with 368K generated questions refined into 40K high-quality pairs covering description, prediction, explanation, and counterfactual reasoning. EmbodiedEval [280] further broadens task diversity with 328 embodied tasks across 125 scenes, including Attribute QA on object and scene properties and Spatial QA that evaluates spatial reasoning through interaction and observation.

A.2.3 Evaluation

Standardized benchmarks are central to embodied intelligence research because they define common evaluation protocols that support fair comparison, systematic diagnosis of model limitations, and reproducible experimentation. The current VLA benchmark ecosystem is diverse, with platforms tailored to assess different dimensions of embodied competence, from basic skills to advanced cognitive abilities. Tab. S2 is an overview of representative embodied benchmarks. This section reviews major benchmarks and categorizes them by the primary capabilities they aim to evaluate.

(1) Language-Conditioned Manipulation. This category evaluates a model’s ability to follow natural language instructions and produce precise manipulation actions. RL-Bench [281] provides over 100 language-annotated tasks with motion-planned demonstrations and serves as a standard benchmark for imitation and reinforcement learning. The ManiSkill series (ManiSkill [282], ManiSkill2 [283], ManiSkill-HAB [284]) offers large-scale simulation environments designed to assess multi-task manipulation and policy generalization, with ManiSkill-HAB providing high-fidelity home-environment tasks. RoboMimic [285] evaluates offline learning methods using human demonstrations and highlights key challenges in leveraging human-generated data for manipulation policies.

(2) Long-Horizon and Interactive Task Completion. This category evaluates tasks that require sequential reasoning,

memory, and sustained interaction with the environment or a human user. ALFRED [93] assesses long-horizon compositional household tasks involving irreversible state changes, which challenge planning, memory, and instruction following. CALVIN [258] links language commands with continuous control and evaluates an agent’s ability to execute long sequences of language-guided operations in unseen environments while maintaining state and performing sequential reasoning. TEACH [277] advances toward interactive task execution by introducing dialogue-based instruction following, where the agent must seek clarification and recover from errors through natural language communication.

(3) Advanced Cognitive Capabilities. This category includes benchmarks designed to probe higher-level cognitive functions beyond basic instruction following, such as lifelong learning and physical reasoning. LIBERO [259] quantifies lifelong learning dynamics through forward and backward transfer metrics that measure how an agent acquires

new skills and retains prior ones across a task sequence. RoboCAS [286] probes embodied cognition in cluttered and physically unstable scenes, exposing the limitations of current models in physical reasoning, spatial understanding, and robust interaction with unpredictable environments.

(4) Evaluation of Embodied Foundation Models. This category shifts the evaluation focus from single-task agents to the holistic and emergent capabilities of large pretrained multimodal systems. EmbodiedBench [287] evaluates multimodal large language models such as GPT-4o across high-level semantic planning and low-level physical control to diagnose their end-to-end embodied competence. EWM-Bench [288] measures the physical realism of generative world models by assessing the motion and semantic consistency of their predicted futures. RoboTwin [289] targets cross-robot generalization and evaluates policies on dual-arm collaborative tasks, emphasizing their ability to transfer from large-scale synthetic data.

TABLE S1: An overview of representative embodied datasets. We exhibit different facets of these datasets, including embodiment, perspective, episodes, scenes, tasks&skills, and collection. More details are discussed in Section A2.2.

Name (Year)	Embodiment	Perspective	Episodes	Scenes	Tasks & Skills	Collection
Simulation-Centered Datasets						
ALFRED [93] (2020)	Simulated human agent	First-person	8,055 expert demonstrations	~120 indoor scenes 4 simulated domains	8 composite household activities	Simulation (AI2-THOR)
LIBERO [259] (2022)	Simulated robot arm	First-person	~6,500		130 skills	Simulation (Robosuite)
VLA-3D [273] (2024)	Virtual agent in 3D scenes	Third-person	9.7M referential pairs	11.5k reconstructed 3D rooms	Spatial navigation & grounding	Simulation (Matterport3D / ScanNet)
Real-World Robotic Manipulation Datasets						
BridgeData V2 [249] (2023)	Robot arm (WidowX)	Mixed (first- & third-person)	60,096 trajectories	24 real environments	13 core manipulation skills	Real robot (VR teleoperation + scripted)
DROID [274] (2024)	Robot arm (Franka Emika Panda)	Mixed (wrist & external cameras)	~76k (\approx 350 hours)	564 distinct real scenes	86 tasks	Real robot (VR teleoperation by 50 operators)
Open X-Embodiment [100] (2023)	22 robot types	Mixed (first- & third-person)	1M+ trajectories	160k+ unified scenes	527 skills	Web-scale aggregation of real-robot data
AgiBot World [168] (2024)	Dual-arm humanoid robot fleet	First-person	1M+ trajectories	5 domains (home, retail, office, restaurant, industry)	217 tasks	Real robot (multi-robot facility)
Human-Centric and Egocentric Datasets						
Ego4D [202] (2021)	Human	First-person	~3,700 hours (~1M clips)	74 locations across 9 countries	Multi-activity	Real human egocentric video
TEACH [277] (2021)	Human commander + embodied agent	Mixed (first- & third-person)	~3k dialog-based episodes	~200 simulated homes	17 composite household tasks	Human teleoperation in simulation
HOI4D [276] (2022)	Human	First-person	~4,000 sequences	610 indoor scenes	54 tasks across all 16 categories	Head-mounted dual RGB-D
HD-EPIC [275] (2025)	Human	First-person	~4,881 object itineraries	9 Real kitchen scenes	–	Wearable sensors (Project Aria glasses)
Embodied Visual Question Answering Datasets						
MT-EQA [278] (2019)	–	First-person	~19,287 QA pairs	588 environments	61 unique object in 8 unique room	Simulation (House3D)
EgoTaskQA [279] (2022)	Human	First-person	~40K QA pairs	Kitchen	48 relationships and 14 object attributes	Head-mounted egocentric RGB video
EmbodiedEval [280] (2025)	–	First-person	328 tasks	125 unique scenes	Navigation, Spatial, Attribute, ...	–

TABLE S2: An overview of representative embodied benchmarks. We exhibit different facets of these benchmarks, including task type, evaluation metrics, and environment/platform. More details are discussed in Section A2.3.

Name (Year)	Task Type	Evaluation Metric	Environment / Platform
Language-Conditioned Manipulation & Control			
RLBench [281] (2020)	Multi-task tabletop manipulation	Success rate	PyRep / CoppeliaSim
ManiSkill Series [282]–[284]	Multi-task object-centric manipulation	Success / completion rate (per task)	SAPIEN (ManiSkill), Habitat-based (ManiSkill-HAB)
RoboMimic [285] (2021)	Multi-stage robot manipulation	Success rate	MuJoCo
Long-Horizon and Interactive Task Completion			
ALFRED [93] (2020)	Vision-language instruction following	Success rate, Goal-Condition Success	ALFRED simulator
CALVIN [258] (2022)	Language-guided multi-step manipulation	Success rate, zero-shot generalization	Simulated tabletop (4 scenes)
TEACh [277] (2021)	Dialog-driven embodied task completion	Success rate, EDH / TfD / TATC	AI2-THOR
Advanced Cognitive Capabilities			
LIBERO [259] (2023)	Continual multi-task manipulation	Success rate, Fwd/Bwd Transfer, AUC	Robosuite
RoboCAS [286] (2024)	Multi-object arrangement & long-horizon manipulation	Success under spatial/clearance constraints	Custom arrangement scenes (SAPIEN)
Evaluation of Embodied Foundation Models			
EmbodiedBench [287] (2025)	Vision-driven embodied agent evaluation	Success rate, Subgoal success rate	AI2-THOR, Habitat 2.0, CoppeliaSim
EWM Bench [288] (2025)	World-model evaluation	Scene consistency, motion correctness, semantic alignment	Synthetic + real embodied datasets
RoboTwin [289] (2025)	Multi-robot imitation, cross-embodiment manipulation	Success rate, sim ↔ real transfer rate, latency	Isaac Gym / PyBullet

TABLE S3: An overview of VLA milestones. We exhibit different facets of these methods, including robot perception, brain, action, training strategy, primary dataset, and evaluation, which corresponding to the subsections in Section 3.

Name	Perception (Visual/Language)	Brain	Action	Training	Primary Dataset	Evaluation
By 2021						
EmbodiedQA [89]	CNN/LSTM	LSTM+FNN	Discrete(Autoregressive)	BC	EQA dataset	EQA v1
VNL [88]	ResNet-152 / LSTM	LSTM	Discrete(Autoregressive)	BC	R2R	R2R
RCM [91]	ResNet-152/LSTM	LSTM	Discrete(Autoregressive)	BC + RL	R2R	R2R
Point-Cloud EQA [92]	PointNet++ & ResNet50/LSTM	RRN+GRU-RNN	Discrete (Sequentia)	BC	MP3D-EQA	Matterport3D
ALFWorld [94]	Mask R-CNN / Seq2Seq	Seq2Seq	Discrete(Autoregressive)	BC	TextWorld	ALFRED benchmark
CLIPort [96]	ResNet-50 / Transformer	FCN + Affordances	Discrete(Autoregressive)	BC	Ravens	Ravens
2022						
SayCan [97]	Resnet-18 / LLM	LLM	Discrete(Autoregressive)	BC + RL	-	-
Inner Monologue [98]	LLM	LLM	Discrete(Autoregressive)	BC	Everyday Robots, Ravens	Ravens & Ravens
RT-1 [99]	EfficientNet-B3 / USE	Transformer	Discrete(Autoregressive)	BC	Self-collected	Self-built benchmark
RT-2 [76]	PaLI-X + PaLM-E	VLM	Discrete(Autoregressive)	BC + co-finetuning	WebLI + RT-1	Real-world Generalization Benchmark
2023						
PaLM-E [77]	ViT / PaLM	VLM	Discrete(Autoregressive)	Multimodal SFT	WebLI, ...	OK-VQA, ...
Diffusion Policy [13]	ResNet-18	Transformer/DiT	Continuous(DDPM)	BC	Human demonstration data	Robomimic, Push-T, ...
2024						
3D-VLA [83]	VLM	3D-LLM	Continuous(trajec- tory seg- ment)	co-finetuning	OXE, RLBench, ...	RLBench, RoboVQA, ...
Octo [49]	CNN / T5	Transformer	Continuous(DDPM)	BC	OXE	Policy general- ization, SR
OpenVLA [32]	SigLip + Dino	Transformer	Discrete(Autoregressive)	BC	OXE	Libero
GR-2 [68]	VQGAN / CLIP	Transformer	Continuous	Predictive modeling	HowTo100M, Ego4D, ...	CALVIN, ...
π_0 [22]	VLM	Transformer	Continuous(Flow Matching)	BC	OXE, Bridge v2, ...	-
2025						
Humanoid-VLA [101]	Transformer	Transformer	Continuous(Autoregressive)	BC	Humanoid-S, AMASS	HumanML3D, Humanoid-S, ...
GR00T N1 [102]	Eagle-2 VLM	VLM + DiT	Continuous(Flow Matching)	BC	GR00T N1 dataset, OXE, ...	GR-1 Tabletop Tasks, ...
PointVLA [82]	CNN (3D) / VLM	VLM	Continuous(DDPM)	BC	Self-collected data	RoboTwin
CoT-VLA [78]	Transformer	LLM	Discrete	BC	OXE	Libero, Brige v2, ...
$\pi_{0.5}$ [53]	VLM	Transformer	Hybrid(Flow Matching)	BC + Predictive modeling	OXE	Real family scenes
LUMOS [16]	CNN / Sentence-BERT	Goal-conditioned actor-critic	Continuous/	BC	-	MLPerf Training Benchmarks
VLA-RL [36]	Siglip + Dino / Llama-2	Transformer	Discrete(Autoregressive)	RL	Online-collected	LIBERO
Cosmos-R1 [103]	ViT/LLM	LLM	Discrete(Autoregressive)	BC + RL	BridgeData v2, RoboVQA, ...	RoboFail, AgiBot, ...