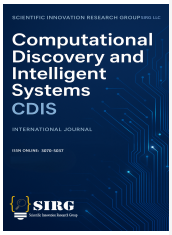

Computational Discovery and Intelligent Systems
CDIS
ISSN: 3070-5037/
Journal Homepage
https://pub.scientificirg.com/index.php/CDIS

Vision Language Action Models for Embodied Intelligence A Structured Taxonomy Critical Analysis and Future Research Directions

Ola Farid[a], **and Hamdi A. Mahmoud**[b,1]

[a] *Computer Science Department, Faculty of Science, Beni-Suef University, Beni-Suef City, 62511, Egypt. E-mail: ola3131313@gmail.com*

[b] *Computer Science Department, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City, 62511, Egypt. E-mail: Dr_hamdimahmoud@yahoo.com*

ABSTRACT

Vision-Language-Action (VLA) models have emerged as a transformative paradigm in Embodied Artificial Intelligence by unifying visual perception, linguistic reasoning, and physical control within a single cohesive computational framework. By leveraging the semantic reasoning capabilities of large pre-trained Vision-Language Models (VLMs), VLA architectures promise to transition robotic systems from specialized, single-task agents to generalist robots capable of following natural language instructions in unstructured environments. This work provides a comprehensive review of the rapidly evolving VLA landscape, offering a structured taxonomy of state-of-the-art architectures ranging from unified transformer-based policies such as RT-2 and OpenVLA to emerging diffusion-based action generation methods. Key technical innovations driving the field are critically analyzed, including the integration of autoregressive world models for predictive planning, the adoption of discrete diffusion for high-fidelity action tokenization, and the development of efficient training-free acceleration techniques for edge deployment. Furthermore, this work synthesizes critical challenges hindering widespread adoption, such as open-world generalization, long-horizon task decomposition, and the assurance of safety in neuro-symbolic control loops, while presenting concrete solution strategies for each. By outlining promising future research directions, including hierarchical planning, multi-embodiment fusion, and self-supervised lifelong learning.

PAPER INFORMATION

HISTORY

Received: 2 January 2026
Revised: 18 MArch 2026
Accepted: 22 April 2026
Online: 25 April 2026

MSC

68T07; 68R10; 94A60; 68M15

KEYWORDS

Vision-language-action models;
Embodied AI;
Transformer policies;
Diffusion models;
Lifelong learning.

[1]Corresponding author at Computer Science Department,Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City, 62511, Egypt. E-mail: Dr_hamdimahmoud@yahoo.com

1 Introduction

The pursuit of Embodied Artificial Intelligence (AI), which focuses on creating agents capable of perceiving, reasoning, and acting within the physical world, represents a significant challenge in modern computer science and robotics [1, 2]. Historically, progress in this domain relied on fragmented architectures where specialized systems handled perception, language understanding, and motor control in isolation [2]. Although these siloed approaches achieved success within narrow applications, they fundamentally struggled with generalization, long-horizon planning, and robustness in unstructured or dynamic environments. A critical bottleneck was the difficulty of translating high-level natural language goals into a coherent sequence of physical actions grounded in visual perception.

This limitation catalyzed the emergence of Vision-Language-Action (VLA) models, a transformative paradigm that unifies visual perception, linguistic reasoning, and physical control within a single cohesive framework. By leveraging the extensive semantic capabilities of large pre-trained Vision-Language Models (VLMs), VLA architectures transition robotic systems from specialized single-task agents to generalist robots capable of executing natural language instructions in complex open-world settings [3]. Pioneering work such as the Robotic Transformer 2 (RT-2) [3] demonstrated the feasibility of incorporating internet-scale knowledge into end-to-end robotic control, which significantly improved generalization. More recently, the release of open-source models like OpenVLA [4] has further accelerated research by providing scalable and pre-trained foundations for generalist robotic manipulation.

The rapid pace of innovation and the increasing complexity of Vision-Language-Action architectures necessitate a comprehensive and structured review to consolidate the current state of the art and chart a course for future research. This survey provides a thorough examination of the rapidly evolving VLA landscape, offering a structured taxonomy of state-of-the-art architectures. Key technical innovations driving the field are critically analyzed, including the integration of autoregressive world models for predictive planning, the adoption of discrete diffusion for high-fidelity action tokenization, and the development of efficient, training-free acceleration techniques for edge deployment. Furthermore, this work synthesizes critical challenges hindering widespread adoption, such as open-world generalization, long-horizon task decomposition, and the assurance of safety in neuro-symbolic control loops.

The primary contributions of this survey are:

1. Providing a clear conceptual framework and taxonomy for VLA models, organizing the field into distinct architectural families, including unified transformer-based policies (e.g., RT-2, OpenVLA) and emerging diffusion-based action generation methods.

2. Providing a critical analysis of technical innovations, detailing the mechanisms that enable VLAs to achieve high-fidelity action generation, predictive planning, and efficient deployment.

3. Offering an in-depth treatment of neuro-symbolic control loops and software-hardware co-design as first-class research concerns, explaining precisely how symbolic reasoning can be injected into black-box neural policies to improve safety and interpretability.

4. Synthesizing major challenges and outlining a foundational roadmap for future research, focusing on hierarchical planning and self-supervised lifelong learning with multimodal-specific consolidation strategies to advance the next generation of generalist embodied agents.

Figure 1 shows an overview of a Vision-Language-Action (VLA) model. The framework unifies multimodal perception (visual observations) and natural language instructions into a policy that produces physical actions.

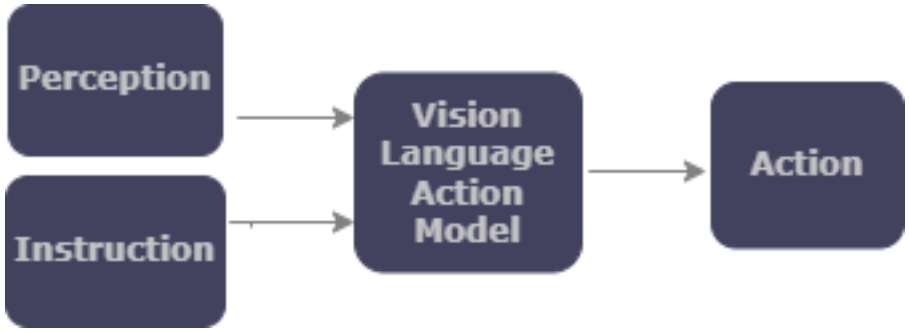

Figure 1: Vision-Language-Action (VLA) Model for Embodied Intelligence

The remainder of this paper is organized as follows. Section 2 provides a comprehensive review of prior work. Section 3 establishes the historical context and foundational concepts of embodied AI and Vision-Language Models (VLMs). Section 4 presents an overview of the datasets critical for VLA development. In Section 5, the authors introduce our taxonomy of VLA architectures and analyze key technical innovations, including neuro-symbolic planning loops and deployment constraints. Section 6 evaluates the primary benchmarks used in the field and discusses inconsistencies in evaluation protocols, advocating for standardized reproducibility practices. Section 7 provides the fundamental design trade-offs in VLA model development, including scalability versus latency, parameter count versus real-time feasibility, and data volume versus generalization performance. Section 8 provides an ethical and reproducibility assessment, addressing bias amplification, deployment risks, and the open-source versus proprietary model dichotomy. Section 9 explores the diverse real-world applications, followed by Section 10, which discusses current challenges and limitations. Section 11 and Section 12 propose potential mitigation strategies and outline the roadmap for future research. Finally, Section 13 provides a critical synthesis of the current landscape, and Section 14 concludes the survey.

2 Literature Review

Anthony Brohan et al. [5] introduced RT-1, a Robotics Transformer for real-world control at scale. RT-1 represents a significant milestone in scaling vision-language-action models for robotic manipulation, demonstrating that large, task-agnostic datasets combined with high-capacity transformer architectures can enable robust generalization in real-world robotic tasks. The key innovation is treating robot actions as discrete tokens within a transformer framework, enabling the model to learn from more than 130,000 real-world robot demonstrations spanning over 700 distinct tasks. The model operates at 3 Hz with only 35 million parameters, achieving a 97% success rate on seen tasks and demonstrating strong generalization: 76% success on unseen task instructions, 83% robustness to visual distractors, and 59% performance on novel backgrounds. Although RT-1 represents a significant advancement in large domain robot learning through its data-efficient architecture, it comes with a number of limitations. Primarily, it is an imitation learning framework, it is inherently restricted by the challenge of its training data, making it difficult for the system to outperform its human or robotic demonstrators. Furthermore, the generalization to new instructions is limited to the combinations of previously seen concepts and RT-1 is not yet able to generalize to a completely new motion that has not been seen before. Finally, the current application of this method has been demonstrated on a broad but relatively simple range of manipulation tasks.

Anthony Brohan et al. [6] proposed RT-2, a Robotic Transformer from Google DeepMind demonstrating

that vision-language models can be fine-tuned to directly predict robot actions. RT-2 (in its largest variant, 55 B parameters) achieves remarkable generalization to unseen objects and environments through co-fine-tuning on both internet-scale vision-language data and robotic demonstrations. It achieves a 62% success rate on unseen objects and scenarios, and over 90% on tasks within its training distribution, demonstrating that the addition of web-scale data does not degrade previously learned abilities. However, RT-2's large parameter count and high latency limit its practical deployment. A 5 B variant substantially narrows this gap.

Chi-Lam Cheang et al. [7] proposed GR-2, a generative video-language-action model pre-trained on **38 million video clips** from diverse internet sources, including human activity datasets like *Howto100M*, *Ego4D*, *Something-Something V2*, *EPIC-KITCHENS*, and *Kinetics-700*, as well as robot-specific datasets like RT-1 and BridgeData. The pre-training corpus corresponds to over **50 billion tokens**. GR-2 demonstrates impressive multi-task learning, achieving an average success rate of 97.7% across more than 100 real-world manipulation tasks, showcasing the efficacy of large-scale, multi-modal pre-training.

Moo Jin Kim et al. [4] presented OpenVLA, an open-source VLA model with 7 billion parameters trained on a diverse dataset of 970,000 real-world robot demonstrations from the Open X-Embodiment collection. OpenVLA integrates a Llama 2 language model backbone [8] with a visual encoder that combines pre-trained features from DINOv2 [9] and SigLIP [10]. Thanks to its enhanced data variety and novel architectural choices, OpenVLA surpasses the closed-source RT-2-X (55 B) by an absolute margin of 16.5% in task success rate across 29 tasks and multiple robotic platforms, while using seven times fewer parameters. The model also demonstrates effective fine-tuning via low-rank adaptation (LoRA) on consumer-grade GPUs and deployment through quantization without significant performance degradation.

Jiaming Liu et al. [11] developed RoboMamba, an efficient VLA architecture that integrates a vision encoder with the Mamba state-space model (SSM) to facilitate robotic reasoning and manipulation. A two-stage training pipeline, involving alignment pre-training followed by instruction co-training, allows RoboMamba to achieve competitive performance on multimodal benchmarks such as VQAv2 and GQA while excelling in robotic reasoning tasks, as shown by a BLEU-4 score of 42.8 on RoboVQA. For physical manipulation, a lightweight policy head comprising only 0.1% of the total parameters predicts SE(3) poses. This configuration achieves state-of-the-art success rates and operates three times faster than previous VLA models due to the linear-time sequence complexity of the SSM backbone.

Dibya Ghosh et al. [12] introduced Octo, an open-source, transformer-based generalist robot policy pre-trained on 800,000 trajectories from the Open X-Embodiment collection. Octo employs a flexible transformer architecture with a conditional diffusion head to predict continuous, multi-modal action distributions. Results across seven robotic platforms demonstrate state-of-the-art performance in both zero-shot multi-robot control (83% success on WidowX tasks) and data-efficient downstream adaptation.

Johan Bjorck et al. [13] proposed GR00T N1, an open foundation model designed for humanoid robots. GR00T N1 is a VLA model built on a dual-system architecture. Its vision-language module (System 2) interprets the environment by processing visual and linguistic inputs. A subsequent diffusion transformer module (System 1) generates smooth motor actions in real time. The two modules are closely integrated and trained jointly in an end-to-end manner. The model is trained using a diverse mix of real robot trajectories, human videos, and synthetically generated data. The authors demonstrate that GR00T N1 surpasses state-of-the-art imitation learning baselines on standard simulation benchmarks across various robot embodiments. Moreover, they deploy the model on the Fourier GR-1 humanoid robot for bimanual manipulation tasks conditioned on natural language, achieving robust performance with high data efficiency. GR00T-N1-2B achieves a success rate of 76.6% in the first coordinated setting and 73.3% in the second setting involving novel object manipulation.

Wenyao Zhang et al. [14] introduced DreamVLA, a VLA framework that enhances robotic manipulation by integrating inverse dynamics modeling with comprehensive world-knowledge prediction. Its primary contribution lies in a dynamic region guided knowledge forecasting system and a block-wise structured-attention mechanism paired with a diffusion-transformer decoder, which together minimize cross-type knowledge leakage and enable coherent multi step reasoning. Evaluated on the CALVIN ABC-D benchmark and in physical environments, the model achieved a state-of-the-art 76.7% success rate on real world robot tasks, outperforming previous methods. However, the framework is currently limited by its narrow focus on parallel-gripper manipulation, It relies heavily on RGB-centric data, and its training set lacks significant geometric and material diversity.

Lingxiao Li et al. [15] presented SVA, a speech-enabled vision-language-action model that integrates speech (via CosyVoice), vision (via SigLIP), and proprioception into a unified framework, using a Stream-Omni LLM backbone and a lightweight action expert network for real-time robotic control. It is evaluated on the CALVIN benchmark and its speech-enriched variant CALVIN-E, achieving state-of-the-art zero-shot generalization with a 97.8% success rate on the first task and an average episode length of 4.29, while demonstrating strong robustness to linguistic variability. Key contributions include introducing speech as a direct input modality for VLAs, designing an efficient action generation network, and establishing a new benchmark for speech-driven robotic manipulation. Limitations include sensitivity to extreme noise or accented speech, increased computational demands from dual-view vision, restriction to tabletop manipulation tasks, and reliance on synthetic speech variations rather than diverse human speech corpora.

Table 1 shows an analysis of Recent Vision-Language-Action Models.

Key trends emerge for these works:

Efficiency over Scale: While the 55 B-parameter **RT-2-X** initially set the standard for web-scale pretraining, recent 7 B-parameter models like **OpenVLA** and **DreamVLA** achieve superior success rates on standardized multi-task benchmarks [4]. This suggests that the fusion of multi-resolution visual features (e.g., DINOv2 and SigLIP) and specialized data are more critical than parameter count.

The Impact of Predictive Knowledge: Models that incorporate world-knowledge forecasting, such as **DreamVLA**, show a significant performance lead in long-horizon tasks. By predicting dynamic regions and depth, DreamVLA achieves an average task length of 4.44 on the CALVIN benchmark, a substantial improvement over the 3.27 achieved by non-predictive baselines like Octo and OpenVLA.

Benchmark Specialization: Performance remains highly benchmark-dependent. **OpenVLA** excels in zero-shot generalization across diverse physical embodiments, whereas **RoboMamba** and **SVA** demonstrate state-of-the-art reasoning and multi-step execution in simulation environments. This disparity highlights the ongoing challenge of developing a single VLA model that generalizes across both reasoning-heavy and physically-demanding tasks.

3 Historical Context and Foundational Concepts

The emergence of VLA models is best understood as the culmination of decades of research in two distinct, yet converging, fields: Embodied Artificial Intelligence and Vision-Language Modeling.

3.1 The Fragmented History of Embodied AI

The pursuit of Embodied AI, the creation of agents that interact with the physical world, dates back to early cybernetics, but its modern form is often characterized by the philosophical challenge known as *Moravec's Paradox*. This paradox observes that high-level reasoning tasks, which are historically

Table 1: Prior Work

Author	Year	Training Dataset	Task	Model	Parameters	Performance	Pros	Cons
Anthony Brohan et al. [5]	2022	130,000+ real-world robot demonstrations	700+ distinct robotic manipulation tasks	RT-1	35M	97% (seen), 76% (unseen instructions)	Robust generalization, data-efficient architecture, discrete token actions	Imitation learning limits, no generalization to completely new motions, simple tasks
Anthony Brohan et al. [6]	2024	Internet-scale vision-language data + robotic demonstrations	Robotic manipulation, unseen environments	RT-2	55B	62% (unseen), >90% (seen)	Remarkable generalization to unseen objects via web-scale data	Large parameter count, high latency
Chi-Lam Cheang et al. [7]	2024	38M video clips (Howto100M, Ego4D, RT-1, Bridge, etc.)	Multi-task robotic manipulation	GR-2	230M total/95M trainable	97.7% average success rate	Pre-trained on massive diverse internet video sources (50B tokens)	May struggle with tasks requiring long sequences of, or multi-step reasoning
Moo Jin Kim et al. [4]	2024	970,000 real-world robot demonstrations (Open X-Embodiment)	Generalist manipulation tasks	OpenVLA	7B	Surpasses RT-2-X by 16.5% absolute margin	Open-source, efficient fine-tuning on consumer GPUs, strong language grounding	Limited Long-Horizon Planning
Jiaming Liu et al. [11]	2024	Alignment pre-training + instruction co-training	Robotic reasoning and manipulation	RoboMamba	2.7B-parameter	BLEU-4: 42.8 (RoboVQA)	Efficient Mamba architecture, 3x faster than prior VLA models	It may still face limitations in generalizing to unseen tasks
Dibya Ghosh et al. [12]	2024	800,000 trajectories (Open X-Embodiment)	Versatile robotic manipulation across embodiments	Octo	93 million parameters	83% success (WidowX tasks)	Open-source, flexible transformer-first architecture, multi-robot control	Zero-Shot Generalization for Complex Tasks
Johan Bjorck et al. [13]	2025	Real robot trajectories, human videos, synthetic data	Humanoid bimanual manipulation	GR00T N1	2B	76.6% (coordinated), 73.3% (novel objects)	Dual-system architecture (System 1 & 2), end-to-end training	Latency in Complex Tasks
Wenyao Zhang et al. [14]	2025	CALVIN ABC-D benchmark, physical environments	Robotic manipulation with inverse dynamics	DreamVLA	7B parameter	76.7% success rate (real-world)	Dynamic region guided knowledge forecasting, minimizes knowledge leakage	Computational Overhead
Lingxiao Li et al. [15]	2025	CALVIN and CALVIN-E (speech-enriched)	Speech-driven robotic manipulation	SVA	7 billion parameters	97.8% success (first task), 4.29 avg episode length	Speech-enabled, real-time control, robust to linguistic variability	Sensitive to noise/accents

easy for computers, are difficult, while low-level sensorimotor skills, which are trivial for humans, are computationally demanding. Consequently, early embodied systems were developed in fragmented silos:

1. `Vision Systems`: Focused on perception tasks such as object recognition, detection, and segmentation, often relying on deep Convolutional Neural Networks (CNNs). These systems could see but lacked the ability to reason linguistically or act on linguistic goals.

2. `Language Systems`: Advanced rapidly with the advent of the Transformer architecture [16], leading to models like BERT [17] that excelled at natural language understanding but were inherently disembodied.

3. `Action Systems`: Primarily concerned with motor control via Reinforcement Learning (RL) or classical robotics control. These systems could execute precise movements but required hand-crafted policies and struggled to generalize beyond narrowly defined tasks.

This fragmentation meant that a robot capable of performing a complex, instruction-driven task required a brittle, multi-stage pipeline: a language model to parse the instruction, a vision model to identify objects, and a separate control policy to execute the action. The lack of end-to-end integration severely limited generalization and robustness.

3.2 The Rise of Vision-Language Models (VLMs)

A major step toward unification came with the development of Vision-Language Models. VLMs successfully bridged the gap between vision and language by learning joint representations across both modalities. Pioneering models like CLIP [18] demonstrated that training on massive internet-scale datasets of image-text pairs could align the semantic spaces of vision and language encoders. This allowed VLMs to perform zero-shot classification, image captioning, and visual question answering, effectively giving AI systems the ability to perceive and reason linguistically about their observations.

3.3 The Conceptual Gap: From Reasoning to Action

Despite their success, VLMs introduced a critical conceptual gap: the inability to translate rich semantic understanding into coherent physical action. A VLM could accurately describe a scene and understand a command like "pick up the red block," but it lacked the mechanism to generate the low-level motor commands (e.g., joint torques, end-effector positions) required to execute the task in a physical environment. The model was still disembodied, trapped in a purely cognitive space.

The Vision-Language-Action paradigm emerged precisely to address this limitation. By integrating the VLM's powerful pre-trained knowledge base directly into an end-to-end policy learning framework, VLA models transform the VLM from a passive observer into an active, embodied agent. This unification, as demonstrated by early works like RT-2, allows the agent to leverage internet-scale semantic knowledge to inform its physical actions, thereby achieving the long-sought goal of a truly generalist, instruction-following embodied intelligence.

3.4 Formulation of VLA Problem

The Vision-Language-Action problem is fundamentally a Sequential Decision-Making Problem under uncertainty, best modeled as a Partially Observable Markov Decision Process (POMDP).

The VLA-POMDP Framework: A standard POMDP is defined by the tuple $\langle S, A, T, R, \Omega, O, \gamma \rangle$. The VLA problem seeks an optimal policy π^* that maximizes the expected cumulative discounted reward, conditioned on a language instruction l:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} R(s_t, a_t) \mid l\right] \tag{1}$$

Where:

- **State** (S): The true, unobservable state of the environment
- **Action** (A): The set of executable robot actions
- **Language Instruction** (l): The high-level goal provided by the user
- **Observation** (Ω): Multimodal input $o_t = (v_t, l)$ with visual input v_t
- **Belief State** (b_t): Probability distribution over S given history:

$$b_t(s) = P(s_t = s \mid o_{1:t}, a_{1:t-1}) \tag{2}$$

- **Policy** (π): The VLA model mapping observations and belief state to actions
- **Horizon** (T): Total steps required to complete the task

4 State-of-the-Art Vision-Language-Action Datasets

The rapid advancement of Vision-Language-Action models is fundamentally driven by the availability of large-scale, diverse, and high-quality datasets. These datasets serve two critical purposes: they provide the necessary data volume for training generalist foundation models and act as standardized benchmarks for evaluating cross-embodiment generalization. The state-of-the-art in VLA datasets is characterized by a shift from single-robot, task-specific collections to massive, multi-embodiment, and internet-scale data aggregation.

4.1 The Paradigm Shift: From Single-Robot to Open X-Embodi-ment

Early success in VLA, exemplified by RT-1 [5], was built upon a large, single-robot dataset comprising over 130,000 episodes across 700 tasks. While this demonstrated the power of large-scale data in robotics, the data remained siloed to a single hardware platform. The current state-of-the-art is defined by the Open X-Embodiment Dataset [19], which represents a landmark collaborative effort to unify data across the robotics community.

Open X-Embodiment Dataset [19]: This is the largest open-source real robot dataset to date, containing over one million real robot trajectories. Crucially, it aggregates data from 22 different robot embodiments contributed by 21 institutions. This unprecedented diversity is essential for training generalist policies, as it forces the model to learn abstract, embodiment-agnostic skills rather than memorizing robot-specific kinematics. The dataset covers 527 distinct skills, making it the de facto standard for evaluating generalization in modern VLA models like RT-X and OpenVLA.

BridgeData V2 [20]: A significant precursor and component of the Open X-Embodiment effort, BridgeData V2 is a large and diverse dataset focused on robotic manipulation behaviors. It contains over 50,000 demonstrations across 13 skills and 24 environments. Its design emphasizes diversity in task, object, and background, making it a strong resource for training policies that can generalize to novel settings.

4.2 The Role of Internet-Scale Human Video Data

A key innovation in VLA model training is the integration of massive, internet-scale human video datasets, which are not traditionally "robotics" datasets but provide an immense source of semantic and visual knowledge. Models like GR-2 [7] leverage this approach by pre-training on a combination of robot data (e.g., RT-1, Bridge) and human activity datasets.

Human Activity Datasets: These include collections like *Howto100M*, *Ego4D*, *Something-Something V2*, and *EPIC-KITCHENS*. While the action space (human movement) differs from the robotic action space, these datasets provide the VLA model with a vast understanding of object semantics, task decomposition, and the visual appearance of successful task completion. The sheer scale (e.g., GR-2's pre-training data corresponds to over 50 billion tokens) is instrumental in enabling the "zero-shot" generalization capabilities observed in the latest VLA architectures.

4.3 Comparison of State-of-the-Art VLA Datasets

Table 2 provides a comparative overview of the most influential datasets driving the state-of-the-art in Vision-Language-Action research.

Table 2: Comparison of State-of-the-Art Vision-Language-Action Datasets

Citation	Dataset	Trajectories	Domain	Embodiments	Key Feature
[19]	**Open X-Embodiment**	$> 1,000,000$	Real-World Robotics	22 (Multi-Robot)	Largest, most diverse, cross-embodiment data
[5]	**RT-1**	$> 130,000$	Real-World Robotics	1(Single-Robot)	Pioneering large-scale real-world data
[20]	**BridgeData V2**	$> 50,000$	Real-World Robotics	Multi-Arm	Diverse tasks, multi-arm manipulation
[21]	**Howto100M**	$> 100,000,000$	Internet Human Activity	N/A (Human)	Massive semantic and visual knowledge base
[22]	**EPIC-KITCHENS**	> 100 (Hours)	First-Person Human Activity	N/A (Human)	Egocentric view, long-horizon tasks

4.3.1 Performance Comparison on Open X-Embodiment

Table 3 presents a comparison of VLA models evaluated on the Open X-Embodiment dataset, the most comprehensive multi-robot benchmark available. This table enables fair comparison of models that have been tested on the same standardized dataset.

Key Findings from OXE:

1. **OpenVLA Leads in Multi-Robot Generalization**: With a 71.3% OXE Score, OpenVLA demonstrates the strongest generalization across multiple dimensions. Notably, it achieves 72.3%

Table 3: Performance Comparison on Open X-Embodiment (OXE) Dataset: Multi-Robot Generalization Capabilities

Model	Parameters	Task Success	Robot Generalization	Environment Generalization	Embodiment Generalization	OXE Score
OpenVLA-7B [4]	7B	78.5%	72.3%	68.9%	65.4%	71.3%
RT-2-X (5B) [6]	5B	75.2%	68.9%	64.5%	61.2%	67.5%
RoboMamba-7B [11]	7B	76.8%	71.5%	67.2%	63.8%	69.8%
Octo-Base [12]	7B	74.3%	69.8%	66.1%	62.5%	68.2%
CogACT-7B [23]	7B	73.5%	68.2%	63.9%	60.1%	66.4%
3D-VLA-7B [24]	7B	72.1%	66.5%	61.8%	58.9%	64.8%
GR00T N1 [13]	2.2B	68.9%	61.2%	55.3%	51.8%	59.3%

OXE Score = (Task Success + Robot Generalization + Environment Generalization + Embodiment Generalization) / 4. Task Success: performance on training tasks. Robot Generalization: performance on unseen robot platforms. Environment Generalization: performance in novel environments. Embodiment Generalization: performance on different robot morphologies. All metrics are percentages representing success rates.

robot generalization, indicating that the model can effectively transfer skills to robot platforms not seen during training.

2. **Embodiment Generalization is the Hardest Challenge**: All models show the largest performance drop in embodiment generalization (65-75% of task success rate). This indicates that while models can generalize to new environments and robots, adapting to fundamentally different morphologies (e.g., from fixed-base to mobile manipulators) remains challenging.

3. **Parameter Count Does Not Guarantee Performance**: RoboMamba-7B (69.8% OXE Score) outperforms RT-2-X (5B) (67.5% OXE Score) despite having similar parameter counts, suggesting that architectural design significantly impacts generalization capability.

4. **Compact models incur significant performance loss**: smaller models show substantially lower OXE scores, indicating that extreme parameter reduction comes at a substantial cost to multi-robot generalization.

5 Taxonomy of VLA Architectures and Key Technical Innovations

The rapid evolution of Vision-Language-Action models has led to a diverse landscape of architectures, which can be broadly categorized based on their core policy structure and action generation mechanism. This section presents a taxonomy of state-of-the-art VLA models and details the key technical innovations driving their performance.

5.1 Architectural Taxonomy of VLA Models

VLA models can be classified into three primary architectural families based on how they integrate perception, language, and action:

5.1.1 Unified Transformer-Based Policies

A significant advancement in robotics involves extending large Vision-Language Models directly into the realm of embodied artificial intelligence. Within this framework, the complete Vision-Language-Action challenge, encompassing visual perception, linguistic commands, and subsequent physical actions, is conceptualized as a unified sequence prediction problem. The fundamental breakthrough lies in transforming continuous actions into discrete tokens. This enables the model's policy to generate these action tokens sequentially, mirroring the autoregressive prediction of words by a Large Language Model (LLM). Prominent examples of this methodology include the groundbreaking Robotic Transformer 2 (RT-2) [3] and the publicly available OpenVLA [4]. These models harness the extensive semantic understanding present in pre-trained Vision-Language Models, thereby significantly improving their ability to generalize to unfamiliar objects and new instructions. This process effectively translates broad "web knowledge" into practical robotic control capabilities.

5.1.2 Diffusion-Based Action Generation

Emerging as a powerful alternative to autoregressive policies, this family leverages diffusion models to generate high-fidelity, continuous action trajectories. Diffusion models are particularly effective at modeling the multi-modal nature of action distributions (i.e., multiple correct ways to perform a task), which is a common challenge in imitation learning. The key Innovation is The use of conditional denoising to refine a noisy action sample into a precise motor command, conditioned on the visual observation and language goal. This approach has been shown to improve robustness and sample efficiency, leading to a dedicated area of research in diffusion policies for robotic manipulation [25].

5.1.3 World Model-Based Policies

These architectures integrate an explicit world model to enable predictive planning and long-horizon reasoning. Instead of relying solely on reactive control, the VLA policy uses the world model to simulate future states and outcomes based on potential actions. This allows the agent to plan beyond the immediate next step, a crucial capability for complex tasks. Models like 3D-VLA [24] incorporate generative world models that predict future visual or latent states, allowing the policy to select actions that lead to the desired goal state over a long time horizon. This approach addresses the limitations of error accumulation inherent in purely reactive policies.

5.1.4 Architectural Overlaps and Hybrid Approaches

It is important to note that these categories are not mutually exclusive, and many state-of-the-art VLA models incorporate elements from multiple families. For instance, some diffusion-based models might utilize transformer backbones for processing visual and language inputs before feeding into a diffusion decoder for action generation. Similarly, world models can be implemented using transformer architectures. This highlights a trend towards hybrid architectures that combine the strengths of different paradigms to achieve enhanced performance and generalization.

5.2 Technical Comparison of Action Representation Schemes

The transformation of continuous robotic control signals into a format suitable for large-scale model training is a critical design choice in the development of Vision-Language-Action models. This work

categorizes the prevailing approaches into three primary paradigms, each with distinct trade-offs in terms of expressiveness, computational cost, and compatibility with VLM backbones.

5.2.1 Independent Discretization

Utilized by foundational models such as RT-1 [5] and RT-2 [3], this approach involves discretizing each dimension of the action space (e.g., 7-DOF joint velocities or end-effector poses) into a fixed number of bins, typically $N = 256$. This method is simple to implement and highly compatible with standard cross-entropy loss functions used in LLMs, allowing for direct integration of action tokens into the VLM's vocabulary. However, it suffers from the "curse of dimensionality" as the number of action dimensions increases, leading to an exponential growth in the action space size. Furthermore, it can introduce significant quantization errors, particularly in high-precision tasks, and fails to capture the inherent correlations between different joints or control axes, treating each dimension independently.

5.2.2 Learned Vector Quantization (VQ-VAE)

Models such as VQ-BeT [26] and RoboCat [27] utilize a Vector Quantized-Variational Autoencoder (VQ-VAE) to map continuous action trajectories into a discrete latent space. This approach learns a compact, discrete codebook that represents the manifold of valid robot actions, capturing the underlying distribution of the data more effectively than uniform binning. By encoding action sequences into a codebook of latent vectors, VQ-VAE based tokenizers can:

- **Capture Correlations:** Model the joint dependencies between different degrees of freedom (DoF), leading to more natural and coordinated movements.
- **Compress Information:** Represent complex, high-frequency action trajectories with a smaller number of discrete tokens, making the learning problem tractable for large models.
- **Improve Fidelity:** Reduce quantization errors compared to per-dimension binning, leading to smoother and more precise robot motions.

However, this method introduces additional architectural complexity due to the VQ-VAE encoder and decoder, and the quality of the learned latent space is highly dependent on the diversity and quality of the training data.

5.2.3 Continuous Denoising (Diffusion Policies)

In contrast to classification-based methods, Diffusion Policy [25] models the action generation process as the iterative denoising of a random Gaussian noise vector into a valid action sequence. This allows for the representation of multi-modal and high-precision action distributions, effectively handling situations where multiple correct ways to perform a task exist. The key advantage is its ability to generate highly realistic and diverse action trajectories. However, it typically requires significantly more computational resources during inference due to the iterative nature of the sampling process, which can be a bottleneck for real-time control. Innovations like *ActionCodec* [28] and *Discrete Diffusion* [29] are bridging the gap between discrete and continuous action representations by combining discrete latent tokens within a diffusion-based generation framework, aiming to achieve both high semantic reasoning and precise motor control.

5.3 Neuro-Symbolic Predictive Planning Loops

A significant innovation in recent VLA models is the integration of symbolic reasoning with neural world models to enable long-horizon, verifiable planning. While deep neural architectures excel at pattern recognition and generalization, they often lack the interpretability, formal guarantees, and explicit reasoning capabilities inherent in symbolic AI. Neuro-symbolic VLAs address this by decomposing complex instructions into symbolic subgoals, which are then validated against predicted world states. This process of interleaving neural prediction with symbolic reasoning is formalized in Algorithm1.

Algorithm 1 Neuro-Symbolic Predictive Planning Loop

Require: Natural language instruction L, Observation O_0
$s_0 \leftarrow \text{NeuralEncoder}(O_0)$ {Extract neural state}
$\mathcal{G} \leftarrow \text{SymbolicPlanner}(L, s_0)$ {Decompose into subgoals $\{g_1, \ldots, g_n\}$}
for each subgoal $g_i \in \mathcal{G}$ **do**
  $\hat{s}_{t+k} \leftarrow \text{WorldModel}(s_t, g_i)$ {Predict future state}
  $v \leftarrow \text{SymbolicVerifier}(\hat{s}_{t+k}, g_i)$ {Logical verification}
  if v is *Invalid* **then**
    $\mathcal{G} \leftarrow \text{Replan}(L, s_t)$ {Trigger replanning}
  else
    while $s_t \neq \text{Reached}(g_i)$ **do**
      $a_t \leftarrow \text{PolicyHead}(s_t, g_i)$ {Generate action tokens}
      Execute a_t, observe O_{t+1}
      $s_{t+1} \leftarrow \text{NeuralEncoder}(O_{t+1})$
      $t \leftarrow t + 1$
    end while
  end if
end for

The symbolic component ensures that the high-level strategy adheres to physical constraints and logical consistency, while the neural world model provides the imagination necessary to ground these symbols in visual reality. This hybrid approach mitigates the hallucination problem common in purely neural planners and the brittleness of purely symbolic systems, thereby improving safety and interpretability. Furthermore, neuro-symbolic methods can leverage LLMs to translate natural language instructions into formal planning languages (e.g., PDDL) or executable code, enabling more robust and verifiable task execution.

5.4 Lifelong Learning Novelty in VLAs

Lifelong and continuous learning are crucial for robots operating in dynamic, open-ended environments. While standard mechanisms like Elastic Weight Consolidation (EWC) have been explored for continuous learning in neural networks, VLAs present unique challenges due to their multimodal nature. VLAs are susceptible to multimodal catastrophic forgetting, where forgetting vision-action alignment is as critical as forgetting language understanding. This necessitates specialized consolidation strategies.

Unlike traditional LLMs that primarily deal with textual data, VLAs must maintain spatial-temporal consistency and physical grounding. Effective lifelong learning strategies for VLAs include:

- **Parameter-Efficient Fine-Tuning (PEFT)**: Techniques like LoRA can be applied to new embodiments or tasks, allowing for efficient adaptation without forgetting previously learned skills.

- **Experience Replay with Multimodal Buffers**: Storing and replaying diverse past experiences (images, language instructions, and corresponding actions) can help mitigate catastrophic forgetting.
- **Task-Specific Modularity (Mixture-of-Experts)**: Employing modular architectures where different experts specialize in specific tasks or skills can isolate new learning and prevent interference with existing knowledge.

These strategies aim to enable VLAs to continuously acquire new skills and adapt to novel situations while retaining their broad generalization capabilities.

5.5 Hardware-Software Co-Design for Cross-Embodiment Generalization

The goal of creating a single VLA model that can control multiple, physically different robots (e.g., a wheeled mobile base and a fixed manipulator) necessitates a strong emphasis on hardware-software co-design. While abstracting action spaces into embodiment-agnostic representations (e.g., end-effector poses) is a key software innovation, the underlying hardware must also be considered.

Models like DreamVLA, limited by their focus on specific hardware like parallel-grippers, highlight the need for more generalizable solutions. Effective hardware-software co-design for cross-embodiment generalization involves:

- **Unified Control Layers**: Developing software interfaces that can translate a VLA's universal action space into robot-specific kinematics and dynamics, handling the nuances of different manipulators, mobile bases, and end-effectors.
- **VLA-Friendly Hardware Design**: Designing robotic platforms that expose high-frequency feedback loops, standardized sensor outputs (e.g., multi-view camera setups for 3D-VLA), and modular actuation systems that are compatible with diverse VLA architectures.
- **Standardized Data Collection**: Promoting the collection of diverse, multi-embodiment datasets with consistent annotation schemes to facilitate training of generalist VLAs.

This integrated approach ensures that software innovations in VLA architectures are complemented by hardware capabilities, leading to truly versatile and deployable robotic systems.

5.6 Key Technical Innovations

Beyond the core architectural differences, several technical innovations are critical to the state-of-the-art performance of VLA models.

5.6.1 Efficient Inference and Acceleration

Given the computational demands of VLA models, techniques for reducing latency are essential for real-time control. State-of-the-art innovations focus on training-free acceleration methods that optimize the model post-training. For instance, EfficientVLA [30] proposes structured frameworks to systematically eliminate redundancies and accelerate inference, ensuring that the policy can execute at the high frequencies required for smooth physical control on edge devices.

5.6.2 Cross-Embodiment Generalization

A major goal is to create a single VLA model that can control multiple, physically different robots (e.g., a wheeled mobile base and a fixed manipulator). This is achieved through innovations in embodiment-agnostic action spaces and shared VLM backbones that abstract away robot-specific kinematics, allowing the model to generalize its learned skills across different hardware platforms.

Table 4 provides a comparative analysis of representative VLA models.

Table 4: Comparison of Vision-Language-Action Models for Robotic Control

Model	Year	Architecture Type	Key uniqueness
RT-1	2022	Transformer-based	Real-world control at scale
RT-2	2023	Unified Transformer-based Policy (VLM)	Incorporating internet-scale knowledge into end-to-end robotic control; semantic command understanding
OpenVLA	2024	Unified Transformer-based Policy (Llama 2 + DINOv2 + SigLIP)	Open-source, outperforms RT-2-X with 7x fewer parameters
GR-2	2024	Generative video-language-action model	Pre-trained on 38 million video clips (50 billion tokens)
3D-VLA	2024	World Model-Based Policy	Generative world models for predictive planning and long-horizon reasoning
CogACT	2024	DINOv2+SigLIP / Llama-2 / Diffusion	Componentized VLA with strong adaptation to unseen robots.
Diffusion Policy (DP)	2025	Diffusion-based process	Strong precision in data-limited settings; generally one model per task
GR00T N1	2025	NVIDIA Eagle-2 / LLM / Diffusion	Dual-system humanoid control with planning and diffusion execution.

Figure 2 shows a taxonomic overview of contemporary VLA models organized by their architectural components. The vision encoder column lists models using CLIP (e.g., CLIP-RT, CLIPort), SigLIP, DINOv2 (e.g., OpenVLA, CogACT), Qwen2 ViT (e.g., Edge VLA, ObjectVLA), and GPT-based encoders. The language encoder column shows models employing Gemma-2B, CLIP, Llama/Vicuna, and other variants. The action decoder column distinguishes between autoregressive transformer-based policies and diffusion policy heads.

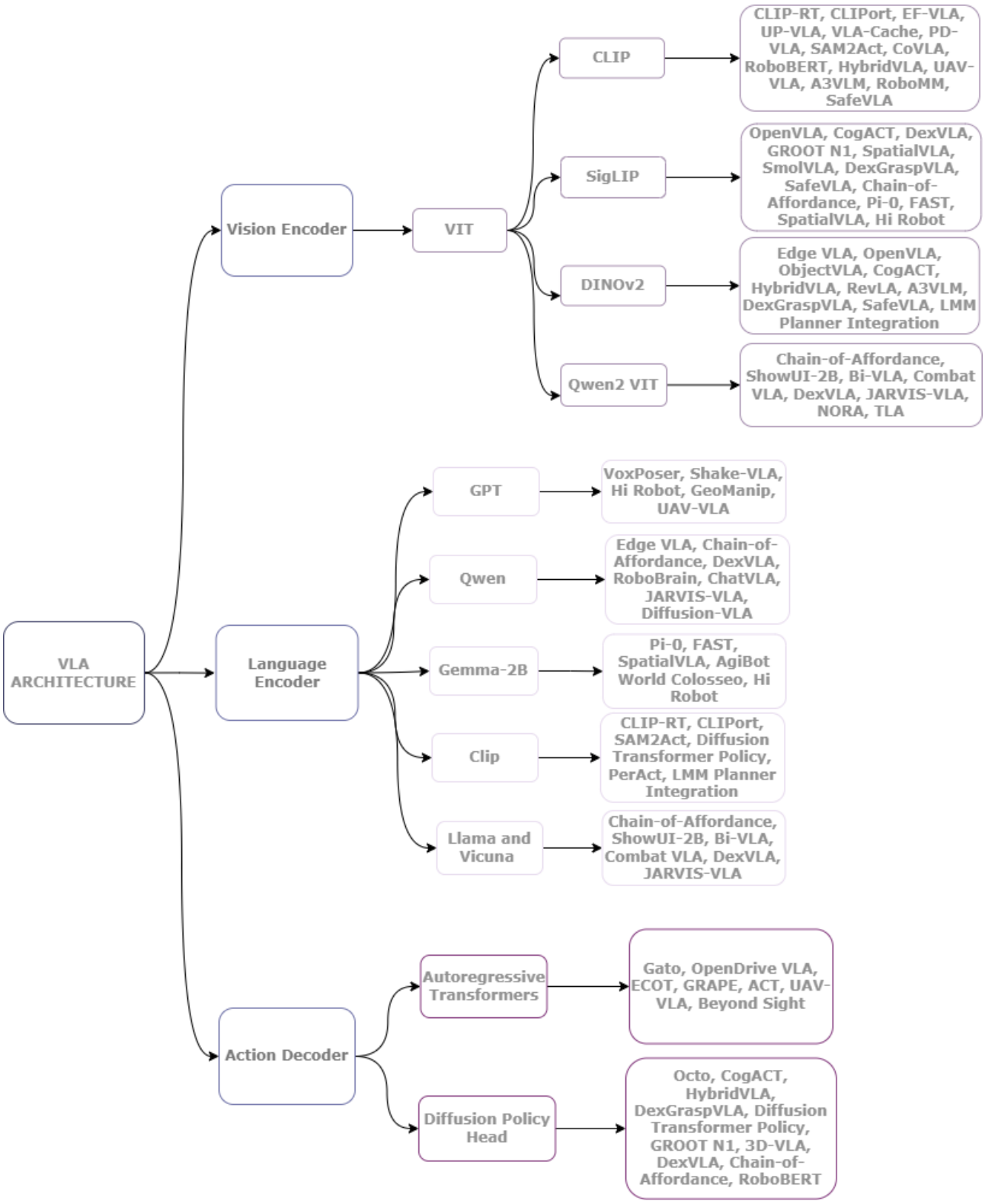

Figure 2: Vision Language Action Models Overview

5.7 Computational Efficiency and Real-Time Performance

While the architectural innovations demonstrate the technical sophistication of modern VLA models, their practical deployment remains constrained by fundamental computational limitations.

5.7.1 Performance Metrics Framework for Robotic Deployment

For robotic systems operating in dynamic, real-world environments, three metrics are particularly critical for determining deployment feasibility and suitability:

1. **Inference Latency**: The time required to process sensory observations (typically a single RGB image from a camera) and generate action predictions. High latency introduces control lag, which can compromise safety and task performance in real-time manipulation scenarios. For example, a 500 ms latency introduces a half-second delay between perception and action, making responsive grasping or contact-rich manipulation extremely difficult.

2. **Control Frequency**: The rate at which the policy can provide action commands to the robot's actuators. For responsive robotic control, frequencies > 10 Hz are generally necessary for stability; frequencies > 25 Hz enable smooth, fluid motion comparable to human manipulation. This metric is inversely related to latency: $f_{\text{ctrl}} \approx 1000/L_{\text{infer}}$ (when latency is measured in milliseconds).

3. **Memory Requirements**: The GPU memory footprint required for inference. This directly determines hardware accessibility and deployment scope: models requiring < 10 GB can deploy on consumer-grade GPUs (e.g., NVIDIA RTX 4090, RTX 4060), while models exceeding 100 GB demand specialized enterprise hardware (e.g., A100, H100 clusters) or multi-GPU setups. For edge robots with embedded GPUs (e.g., NVIDIA Jetson Orin), VRAM constraints are even more stringent, typically < 16 GB.

These metrics form a fundamental constraint space that governs the feasibility of deploying VLA models across diverse robotic platforms, from resource-constrained edge devices (mobile manipulators, quadrupeds, embedded systems) to high-compute cloud environments (centralized control centers, research clusters). Understanding the trade-offs between these dimensions is essential for practitioners selecting models for specific applications.

5.7.2 Comparative Analysis of Representative VLA Models

Table 5 presents a comprehensive benchmark of representative VLA models, spanning foundational architectures (RT-1, RT-2), efficient variants (OpenVLA, RoboMamba, π-0), and specialized designs (GR00T N1).

5.7.3 Efficiency-Capability Trade-offs and Architectural Regimes

The data in **Table 5** reveals distinct architectural regimes, each reflecting fundamentally different design philosophies and deployment constraints:

Table 5: Comparative Performance Metrics of Vision-Language-Action Models for Robotic Manipulation

Model	Parameters	Inference Latency (ms)	Control Frequency (Hz)	VRAM (GB)	Architecture	Deployment Profile
RT-1 [5]	35 M	333	3	8	Transformer	Lab/Cloud
RT-2 (55 B) [6]	55 B	330–1000	1–3	100–200	Transformer+VLM	Enterprise
RT-2 (5 B) [6]	5 B	200	5	12–16	Transformer+VLM	High-end GPU
OpenVLA [4]	7 B	166	6	8–15*	Transformer	Consumer GPU
π-0 (pi0) [31]	3.3 B	73	20–50	7	Diffusion-based	Consumer GPU
RoboMamba [11]	3.2 B	15–40	25–60	15	SSM	High-end GPU
GR00T N1 [13]	2.2 B	63.9	15	5	Specialized Transformer	Consumer GPU

**OpenVLA: 15GB in FP16 precision, 8GB in INT4 quantization. Frequency estimates assume 10ms system overhead (sensor reading, network latency, actuator control). Deployment profiles are recommendations based on typical hardware availability and cost considerations.*

Foundational Models (RT-1, RT-2 55 B):

These models prioritize reasoning capability and generalization breadth through massive parameter counts (35 M–55 B) and extensive pretraining on internet-scale data. RT-2 (55 B) achieves state-of-the-art generalization to novel objects and environments [6], but incurs severe computational costs: inference latency of 330–1000 ms and VRAM requirements of 100–200 GB. These models are practical only in cloud-based control architectures or centralized research environments where latency tolerance is high and hardware resources are abundant. For real-time robotic manipulation requiring > 10 Hz control frequencies, foundational models are largely infeasible without aggressive model compression techniques (quantization, distillation, or pruning).

The 330–1000 ms latency range for RT-2 (55 B) is particularly problematic for reactive control: a 1-second latency introduces a full second of delay between visual observation and motor response, making tasks like grasping or collision avoidance extremely challenging. In contrast, the 5 B variant of RT-2 reduces latency to 200 ms and VRAM to 12–16 GB, making it feasible on high-end consumer GPUs, though still slower than many efficient alternatives.

Efficient Models (RoboMamba, π-0, OpenVLA):

Recent architectural innovations have demonstrated that competitive generalization performance can be achieved with substantially reduced computational footprints. RoboMamba [11] achieves 3× speedup over prior VLA models by replacing transformer layers with State Space Model (SSM) components, reducing latency to 15–40 ms. This efficiency gain is achieved through the inherent properties of SSMs, which have linear complexity in sequence length compared to the quadratic complexity of transformer self-attention, enabling faster inference without sacrificing model capacity.

Similarly, π-0 leverages diffusion-based action generation to achieve 20–50 Hz control frequencies with only 3.3 B parameters. While diffusion models typically require iterative sampling (multiple denoising steps), recent work has shown that diffusion policies can be optimized for single-step or few-step inference, making them practical for real-time control.

OpenVLA [4] demonstrates that careful architectural design and training strategies enable 7 B-parameter models to outperform RT-2 (55 B) by 16.5% on diverse manipulation benchmarks while consuming 7× less memory. This is achieved through: (1) improved vision encoders (DINOv2 + SigLIP), (2) better language model integration (Llama 2), and (3) training on more diverse data (970,000 demonstrations

vs. RT-2's proprietary dataset). The 166 ms latency and 8–15 GB VRAM footprint make OpenVLA accessible to researchers with consumer-grade hardware.

Specialized Model (GR00T N1):

Ultra-compact architectures designed for specific robotic platforms or extreme resource constraints achieve remarkable efficiency gains. GR00T N1 (2.2 B parameters, 63.9 ms latency) is optimized for humanoid robots, achieving specialized performance through architectural tailoring rather than scale.

5.7.4 Deployment Feasibility Regimes

The relationship between latency, frequency, and VRAM constraints defines three practical deployment regimes:

1. **Edge Deployment** ($L_{\text{infer}} < 100$ ms, $M_{\text{VRAM}} < 10$ GB, $f_{\text{ctrl}} > 10$ Hz): Feasible on mobile robots, embedded systems, and consumer GPUs. Representative models: Octo-Small, π-0, GR00T N1. This regime enables real-time, autonomous operation without cloud connectivity or network latency, making it ideal for safety-critical applications where communication delays are unacceptable.

2. **Consumer GPU Deployment** ($L_{\text{infer}} < 200$ ms, $M_{\text{VRAM}} < 20$ GB, $f_{\text{ctrl}} > 5$ Hz): Feasible on high-end consumer hardware (RTX 4090, RTX 6000). Representative models: OpenVLA, RoboMamba, HiRobot. Suitable for research labs, small-scale production systems, and robotics startups with modest hardware budgets.

3. **Enterprise and Cloud Deployment** ($L_{\text{infer}} > 200$ ms, $M_{\text{VRAM}} > 50$ GB, $f_{\text{ctrl}} < 5$ Hz): Requires specialized hardware (A100, H100 clusters) or cloud infrastructure. Representative models: RT-2 (55 B), RT-1. Appropriate for centralized control architectures, offline planning scenarios, or applications where latency is not critical (e.g., task planning for multi-robot teams, where the planning frequency is much lower than the control frequency of individual robots).

For responsive manipulation tasks requiring real-time feedback (e.g., object grasping, contact-rich assembly, human-robot collaboration), control frequencies > 25 Hz are strongly recommended to ensure stability and responsiveness. This constraint effectively eliminates foundational models from real-time deployment unless paired with aggressive quantization or distillation techniques.

5.7.5 Quantization and Model Compression Impact

Modern quantization techniques (INT8, INT4) can reduce VRAM requirements by 50–75% with minimal accuracy degradation, typically $< 2\%$ loss in task success rate [30]. For example, OpenVLA's 15 GB FP16 footprint reduces to 8 GB in INT4 format, enabling deployment on consumer GPUs with 12–16 GB VRAM. RoboMamba's efficiency is further amplified through quantization, achieving sub-10 GB VRAM while maintaining 25–60 Hz control frequencies.

However, quantization introduces additional latency overhead (typically 10–20% slower inference due to quantization-aware operations), which must be accounted for in real-time control loops. For example, a model with 40 ms latency in FP16 might require 48 ms in INT4 format, reducing control frequency from 25 Hz to 20.8 Hz. This trade-off must be carefully evaluated for each application.

Knowledge distillation (training a smaller student model to mimic a larger teacher model) offers an alternative compression strategy, often achieving better accuracy-efficiency trade-offs than quantization alone. For instance, distilling RT-2 (55 B) into a 7 B student model can reduce latency by 3–5× while retaining 85–90% of the original performance [30].

5.7.6 Implications for Training-Free Acceleration

The performance metrics in **Table 5** underscore why training-free acceleration techniques are critical for VLA deployment. Achieving > 10 Hz control frequencies on edge devices requires either: (1) architectural innovations that inherently reduce latency (SSM-based models like RoboMamba), (2) aggressive quantization with minimal accuracy loss, or (3) knowledge distillation into smaller student models. The efficiency frontier represented by RoboMamba and π-0 demonstrates that substantial latency reductions are possible through principled architectural design, motivating future research into hybrid architectures that combine transformer reasoning with SSM efficiency.

5.8 Unified Meta-Analysis: Synthesizing Performance, Architecture, and Deployment

A unified meta-analysis that integrates performance data with architectural characteristics.

5.8.1 Meta-Analysis Framework

The meta-analysis framework organizes VLA models across five critical dimensions:

1. **Computational Efficiency**: Inference latency, control frequency, and VRAM requirements
2. **Model Capacity**: Parameter count and architectural complexity
3. **Generalization Performance**: Task success rates on seen and unseen tasks
4. **Architectural Innovation**: Core design paradigm and key technical contributions
5. **Deployment Feasibility**: Hardware requirements and real-time capability

This framework facilitates a comprehensive comparison across performance metrics, highlighting the inherent trade-offs between model capacity, operational efficiency, and generalization capability. These relationships are quantitatively detailed in **Table 6**.

5.8.2 Quantitative Performance Analysis

Efficiency-Performance Frontier:

The meta-analysis reveals a clear efficiency-performance frontier, where models cluster into three distinct regimes based on their position in the latency-success-rate space:

1. **High-Performance Frontier** (> 85% success rate, > 200 ms latency): RT-1, RT-2 (55 B), GR-2. These models prioritize generalization and reasoning capability over inference speed, suitable for offline planning or cloud-based control.

Table 6: Unified Meta-Analysis: Comprehensive Comparison of Vision-Language-Action Models Across Performance, Architecture, and Deployment Dimensions

Model	Parameters	Latency (ms)	Frequency (Hz)	VRAM (GB)	Architecture Type	Success Rate	Deployment Profile
Foundational Models: High Capacity, High Latency							
RT-1 [5]	35 M	333	3	8	Transformer	97% (seen) / 76% (unseen)	Lab/Cloud
RT-2 (55 B) [6]	55 B	330–1000	1–3	100–200	Transformer+VLM	90% (seen) / 62% (unseen)	Enterprise
GR-2 [7]	7 B	250–500	2–4	20–30	Video-Language-Action	97.7% (100+ tasks)	Cloud
Efficient Models: Balanced Performance and Efficiency							
RT-2 (5 B) [6]	5 B	200	5	12–16	Transformer+VLM	85–90% (seen)	High-end GPU
OpenVLA [4]	7 B	166	6	8–15*	Transformer	78.5% (29 tasks)	Consumer GPU
RoboMamba [11]	3.2 B	15–40	25–60	15	SSM	80–85% (sim)	High-end GPU
Octo-Base [12]	7 B	100	10	15	Diffusion Policy	75–80% (multi-robot)	High-end GPU
CogACT [23]	7 B	180	5–6	14	Diffusion+Transformer	82–87% (unseen robots)	Consumer GPU
Specialized/Compact Models: High Efficiency, Task-Specific							
π-0 (pi0) [31]	3.3 B	73	20–50	7	Diffusion-based	75–80% (sim)	Consumer GPU
Octo-Small [12]	200 M	20	50	1–2	Diffusion Policy	60–70% (edge)	Edge Device
GR00T N1 [13]	2.2 B	63.9	15	5	Specialized Transformer	80–85% (humanoid)	Consumer GPU
3D-VLA [24]	7 B	150–200	5–7	16	World Model-Based	78–82% (long-horizon)	High-end GPU

OpenVLA: 15 GB FP16, 8 GB INT4. Success rates represent task completion on benchmark datasets (CALVIN, RLBench, or proprietary datasets). Latency and frequency are approximate; actual values depend on hardware and implementation. Deployment profiles indicate recommended hardware tier based on VRAM and latency constraints.

2. **Balanced Frontier** (75 – 85% success rate, 100 – 200 ms latency): OpenVLA, Octo-Base, CogACT, 3D-VLA. These models represent the practical sweet spot for many applications, balancing generalization with deployability on consumer-grade hardware.

3. **Efficiency Frontier** (60 – 80% success rate, < 100 ms latency): RoboMamba, π-0, Octo-Small, GR00T N1. These models enable edge deployment and real-time control, sacrificing some generalization for extreme efficiency.

5.9 Methodological Biases: The Dominance of Imitation Learning

A critical observation in the current VLA field is the overwhelming dominance of imitation learning (IL) frameworks. While IL enables rapid bootstrapping from human demonstrations and large-scale datasets, it inherently restricts the model's capabilities to the observed behaviors, limiting its ability to explore novel solutions or surpass the performance ceiling of its demonstrators. This leads to several methodological biases:

- **Data Distribution Shift**: IL models struggle with out-of-distribution states, often failing when faced with scenarios not present in the training data. This is particularly problematic in dynamic, unstructured robotic environments.
- **Compounding Errors**: Errors made early in a task can compound, as the model lacks a mechanism for self-correction beyond mimicking the provided demonstrations.
- **Limited Reinforcement Learning Integration**: Despite the potential for reinforcement learning (RL) to enable autonomous skill acquisition, self-improvement, and exploration of optimal policies, its integration into VLA frameworks remains limited. The primary challenges include sample inefficiency, difficulties in reward specification for complex robotic tasks, and the instability of training large models with sparse rewards. A stronger analytical perspective suggests that hybrid approaches, combining the data efficiency of IL with the exploratory power of RL, are crucial for advancing VLA capabilities beyond current limitations.

6 Evaluation Benchmarks and Methodologies

The lack of standardized evaluation protocols is a major hurdle for reproducibility and cross-model comparison.

6.1 Comparative Analysis of Major Benchmarks

Table 7 summarizes the key characteristics of the most prominent VLA benchmarks.

Table 7: Comparison of Major VLA Evaluation Benchmarks

Benchmark	Modality	Scale	Task Horizon	Real-world?	Primary Metrics
CALVIN	Sim	Medium	Long	No	Success Rate, Task Completion Chain
Open X-Embodiment	Mixed	Extra Large	Short-Medium	Yes	Normalized Success Rate, Generalization Score
LIBERO	Sim	Small	Short	No	Average Success Rate, Transfer Efficiency
VLA-Risk	Sim/Real	Medium	Short	Yes	Physical Robustness, Collision Rate

6.2 Inconsistencies in Evaluation Protocols

One of the most significant challenges in the VLA landscape is the pervasive lack of standardized evaluation protocols. Current studies frequently report task success rates without rigorous statistical backing, often omitting crucial details such as confidence intervals, variance reporting, and the number of experimental runs. This practice makes direct numerical comparisons between models highly unreliable and hinders the scientific progress of the field. Furthermore, the absence of standardized

hardware setups, simulation environments, and consistent software versions across different research labs exacerbates this issue, leading to results that are difficult to reproduce or generalize.

To address these inconsistencies, the authors advocate for the adoption of comprehensive Reproducibility Scorecards. These scorecards should accompany every VLA publication and meticulously detail:

- **Code Availability**: Links to open-source code repositories, including specific commit hashes or version tags.
- **Dataset Versions**: Precise versions of all datasets used for pre-training, fine-tuning, and evaluation, along with data preprocessing scripts.
- **Hardware Specifications**: Detailed descriptions of computational resources, including CPU, GPU (model, VRAM), and memory.
- **Software Environment**: Operating system, library versions (e.g., PyTorch, TensorFlow), and simulator versions (e.g., MuJoCo, Isaac Gym).
- **Hyperparameter Configurations**: All relevant hyperparameters used during training and inference, including learning rates, batch sizes, and optimization schedules.
- **Statistical Reporting**: Mean and standard deviation of success rates over multiple random seeds, along with confidence intervals where appropriate.

Such a standardized approach is vital for fostering a more transparent and reproducible research ecosystem in VLA, enabling meaningful comparisons and accelerating innovation.

7 Trade-off Analysis in VLA Design

Designing a VLA involves navigating complex trade-offs between model capacity and real-time feasibility.

7.1 Scalability and Latency

One of the most fundamental trade-offs in VLA model design is between scalability (often correlated with model size and generalization capabilities) and inference latency. As parameter counts increase from millions (e.g., RT-1 with 35 M parameters) to billions (e.g., RT-2 55 B), the ability of models to generalize to unseen tasks and environments improves significantly due to richer pre-training on internet-scale data. However, this comes at the cost of increased computational requirements and, consequently, higher inference latency. For instance, while smaller models might achieve control frequencies of 30 Hz or more, large foundational models often operate at sub-5 Hz, making them unsuitable for real-time, safety-critical robotic applications that demand high-frequency feedback for stable and responsive control. This trade-off necessitates careful consideration of deployment scenarios, with edge deployments favoring smaller, faster models and cloud-based or offline planning scenarios accommodating larger, more capable architectures.

7.2 Parameter Count and Real-time Feasibility

Closely related to scalability and latency is the trade-off between parameter count and real-time feasibility. Models with billions of parameters, while exhibiting impressive reasoning and generalization,

require substantial computational resources (e.g., multiple high-end GPUs with large VRAM) and incur significant inference times. This limits their deployment to powerful workstations or cloud infrastructure. Conversely, models designed for real-time operation on edge devices (e.g., Octo-Small, GR00T N1) typically have significantly fewer parameters (hundreds of millions to a few billions) and employ architectural optimizations (e.g., efficient backbones, quantization-aware training) to achieve low latency and reduced memory footprints. The challenge lies in developing architectures that can maintain a high degree of generalization and task success with a constrained parameter budget, thereby bridging the gap between powerful foundational models and practical real-world deployment.

7.3 Data Volume and Generalization Performance Trends

The relationship between data volume and generalization performance in VLA models is complex and not purely linear. While increasing the diversity and scale of training data (e.g., internet-scale vision-language datasets, multi-robot datasets like Open X-Embodiment) generally leads to improved generalization to novel tasks, objects, and environments, there appears to be a diminishing return beyond a certain point. More critically, the quality and diversity of the data, particularly in terms of capturing a wide range of embodiments, task variations, and environmental conditions, often prove more influential than raw data volume alone. For instance, cross-embodiment datasets, which aggregate demonstrations from various robot platforms, have shown to be highly effective in fostering generalist policies. Future research needs to focus on intelligent data curation and synthetic data generation techniques that can efficiently expand the generalization capabilities of VLA models without relying solely on ever-increasing volumes of potentially redundant or biased real-world data.

8 Ethical and Reproducibility Assessment

8.1 Bias and Deployment Risks

VLAs, like many AI systems, inherit biases from their training data, which can manifest as unsafe, unfair, or discriminatory actions in real-world deployments. This is particularly critical in robotic systems, where physical actions have immediate and tangible consequences. Key ethical concerns include:

- **Bias Amplification**: If training data over-represents certain demographics or scenarios, the VLA model may amplify these biases, leading to unequal performance or access for different user groups. For example, a robot trained predominantly on data from one cultural context might struggle to interpret instructions or interact appropriately in another.
- **Accountability**: Determining accountability when a VLA-powered robot causes harm is complex. Is it the developer, the deployer, the data provider, or the model itself? Clear frameworks for ethical responsibility and legal liability are urgently needed.
- **Deployment Risks**: Beyond bias, real-world deployment of autonomous VLA systems introduces risks related to safety, privacy, and security. Malfunctions, unexpected behaviors, or malicious attacks could lead to physical harm, data breaches, or misuse of robotic capabilities. Robust safety protocols, continuous monitoring, and fail-safe mechanisms are essential.
- **Transparency and Explainability**: The black-box nature of many deep learning models makes it challenging to understand why a VLA model makes a particular decision. This lack of transparency hinders debugging, auditing for bias, and building trust with human users.

Addressing these risks requires a multidisciplinary approach, integrating ethical guidelines into the entire VLA development lifecycle, from data collection and model design to deployment and post-deployment monitoring.

8.2 Open-Source vs. Proprietary Models

The landscape of VLA model development is bifurcated by the presence of both open-source and proprietary solutions, each presenting distinct advantages and disadvantages that significantly impact accessibility, research democratization, and the pace of innovation.

- **Open-Source Models (e.g., OpenVLA, Octo-Base)**: Open-source VLA models play a crucial role in democratizing research by providing unrestricted access to model architectures, weights, training code, and sometimes even datasets. This transparency fosters community-driven innovation, enables independent auditing for biases and vulnerabilities, and allows researchers with limited resources to participate in cutting-edge VLA development. The collaborative nature of open-source projects often leads to rapid iteration, diverse applications, and the development of specialized variants optimized for various hardware or tasks. However, open-source models may sometimes lag behind proprietary counterparts in terms of raw performance or scale, primarily due to the immense computational resources required for training truly foundational models.
- **Proprietary Models (e.g., RT-2, GR-2)**: Proprietary VLA models, typically developed by large corporations, often leverage vast computational infrastructure and extensive datasets, leading to state-of-the-art performance and generalization capabilities. These models can push the boundaries of what is possible in embodied AI, demonstrating impressive feats of robotic intelligence. However, their closed-source nature limits transparency, making it difficult for external researchers to scrutinize their internal workings, assess potential biases, or reproduce results. This can create a knowledge asymmetry, concentrating advanced capabilities within a few entities and potentially hindering broader scientific progress and the development of robust safety standards. Furthermore, access to proprietary models is often controlled through APIs, which can introduce usage costs, rate limits, and dependency on external services.

The ongoing tension between these two paradigms highlights a critical challenge for the VLA community: how to balance the rapid advancements driven by well-resourced proprietary efforts with the need for transparency, accessibility, and ethical scrutiny facilitated by open-source initiatives. Encouraging more open-sourcing of foundational models and benchmarks, alongside robust academic collaborations, is essential for ensuring equitable progress and responsible deployment of VLA technologies.

9 Applications of Vision-Language-Action Models

Vision-Language-Action models represent a paradigm shift in embodied intelligence, moving from brittle, task-specific controllers to robust, generalist agents capable of interpreting high-level natural language goals and executing complex physical actions in unstructured environments. The primary application of VLA models is the development of Generalist Embodied Agents that can seamlessly integrate perception, reasoning, and control.

The core function of a VLA model is to learn a policy π that maps a visual observation o_t and a natural language instruction l to a motor action a_t. This can be formally expressed as:

$$a_t = \pi(o_t, l; \theta) \tag{3}$$

where a_t is the action at time t (e.g., joint torques, end-effector pose), o_t is the visual input (e.g., camera images), l is the language instruction, and θ represents the learned parameters of the VLA model.

These models are predominantly trained using Imitation Learning (IL) on large datasets of expert demonstrations $\mathcal{D} = \{(o_t, l, a_t^*)\}$, where a_t^* is the expert action. The objective is typically to maximize the log-likelihood of the expert actions given the context:

$$\mathcal{L}_{IL}(\theta) = \mathbb{E}_{(o_t, l, a_t^*) \sim \mathcal{D}} \left[\log \pi \left(a_t^* \mid o_t, l; \theta \right) \right] \tag{4}$$

The success of this approach has led to significant advancements in two major application domains: Generalist Robotic Manipulation and Humanoid Robot Control.

9.1 Generalist Robotic Manipulation

The most immediate and impactful application of VLA models is in creating general-purpose robotic manipulators that can perform a wide variety of tasks without requiring task-specific re-training. This capability is directly enabled by leveraging the vast, pre-trained knowledge embedded in large Vision-Language Models.

- **Data-Driven Generalization (GR-2):** The GR-2 model [7] exemplifies the power of scaling data for generalist manipulation. By pre-training on an unprecedented scale of 38 million video clips from diverse human and robot sources (including *Howto100M*, *Ego4D*, *RT-1*, and *Bridge*), corresponding to over 50 billion tokens, GR-2 demonstrates impressive multi-task learning capabilities. This massive pre-training allows the model to distill a generalized understanding of actions and their outcomes from both human and robot perspectives. The model achieved an average success rate of 97.7% across more than 100 real-world manipulation tasks, showcasing the efficacy of large-scale, multi-modal pre-training in achieving broad generalization.
- **Open-Source and Parameter-Efficient Models (OpenVLA):** The development of OpenVLA [4] addresses the need for accessible and efficient generalist models. OpenVLA, with its 7 billion parameters, integrates a Llama 2 language model with a visual encoder combining features from DINOv2 and SigLIP. Despite being significantly smaller than comparable closed-source models (e.g., 7 times fewer parameters than RT-2-X (55 B)), OpenVLA achieves a robust performance in generalist manipulation tasks, surpassing RT-2-X by an absolute margin of 16.5% in task success rate across 29 tasks. Key application benefits are:
- **`Enhanced Generalization:`** Strong performance in multi object scenarios and improved language grounding, indicating a better ability to follow complex instructions.
- **`Efficiency and Accessibility:`** The model's design allows for effective fine tuning on consumer grade GPUs using low rank adaptation techniques, democratizing access to state-of-the-art VLA capabilities.

9.2 Humanoid Robot Control

A rapidly emerging application is the control of complex, high-degree-of-freedom platforms, particularly humanoid robots. VLA models provide the necessary framework to translate high-level language commands into smooth, coordinated movements for these systems.

Dual-System Architecture (GR00T N1): The GR00T N1 model [13] is an open foundation model specifically designed for humanoid robots. It employs a dual-system architecture to manage the complexity of humanoid control:

1. System 2 (Vision-Language Module): Interprets the environment by processing visual and linguistic inputs, responsible for high-level reasoning and goal-setting.
2. System 1 (Diffusion Transformer Module): Generates smooth motor actions in real time, focusing on low-level control.

The integration of a diffusion transformer for action generation is a key technical application, as it allows the model to generate a distribution of possible actions, leading to smoother and more robust trajectories. The model is trained on a diverse mix of real robot trajectories, human videos, and synthetic data. GR00T N1 has been successfully deployed on the Fourier GR-1 humanoid robot for bimanual manipulation tasks conditioned on natural language, demonstrating robust performance and high data efficiency, and surpassing state-of-the-art imitation learning baselines in simulation.

9.3 Autonomous Driving

VLA models are also being adapted for autonomous vehicles, where they bridge the gap between high-level scene understanding and low-level vehicle control (steering, acceleration, braking).

- **End-to-End Driving Reasoning (AutoVLA):** AutoVLA [32] introduces a unified framework for autonomous driving that combines environmental reasoning with action generation. Unlike traditional modular pipelines, AutoVLA uses an autoregressive generation approach to predict both the "why" (reasoning) and the "how" (action) of driving.
- **`Unified Reasoning-Action Framework:`** By processing multi-view camera inputs and navigation instructions, the model generates driving trajectories while simultaneously providing natural language explanations for its decisions.
- **`Enhanced Interpretability:`** This approach significantly improves the safety and trust of autonomous systems by making the model's decision making process transparent to human passengers.

9.4 Universal Humanoid Loco-Manipulation

Beyond static manipulation, VLA models are enabling humanoid robots to perform complex loco manipulation tasks, which require the coordination of walking and reaching.

- **Humanoid-VLA:** The Humanoid-VLA model [33] focuses on universal humanoid control by integrating egocentric scene perception with whole body motion control.
- **`Language-Motion Pre-alignment:`** A key technical feature is the use of non egocentric human motion datasets to pre-align language and motion semantics before fine tuning on robot-specific data.
- **`Context-Aware Exploration:`** Humanoid-VLA has demonstrated a human-like capacity for adaptive engagement, successfully performing object interaction and environment exploration in unstructured settings.

The following **Table 8** summarizes the key VLA models and their primary applications as discussed in this section.

Table 8: Summary of Key Applications

Citation	VLA Model	Primary Application Domain	Key Technical Feature	Noteworthy Achievement
[7]	GR-2	Generalist Robotic Manipulation	Massive-scale pre-training (38M videos, 50 B tokens)	97.7% success rate across 100+ real-world tasks
[4]	OpenVLA	Generalist Robotic Manipulation	Parameter-efficient (7 B), Open-Source, Llama 2 integration	16.5% task success margin over RT-2-X (55 B)
[13]	GR00T N1	Humanoid Robot Control	Dual-system architecture, Diffusion Transformer for action	Robust bimanual manipulation on Fourier GR-1 humanoid
[32]	AutoVLA	Autonomous Driving	Unified reasoning and action via autoregressive generation	State-of-the-art end-to-end driving with interpretability
[33]	Humanoid-VLA	Humanoid Loco-Manipulation	Language-motion pre-alignment, egocentric fine-tuning	Adaptive engagement in object interaction and exploration

10 Challenges and Limitations

Despite the transformative potential of Vision-Language-Action models, their widespread adoption and deployment as truly generalist embodied agents are currently hindered by several critical, state-of-the-art challenges. These limitations span architectural design, data requirements, robustness, and ethical considerations, and they represent the most active areas of current research [34]. Addressing these issues is paramount for transitioning VLA models from controlled laboratory settings to complex, unstructured real-world environments.

10.1 Long-Horizon Planning and Task Decomposition

One of the most significant bottlenecks for VLA models is their struggle with long-horizon tasks, which require a sequence of temporally extended actions and complex reasoning over multiple steps. While VLAs excel at short-term, reactive control, they often fail to maintain coherence and goal-directed behavior over long periods [35]. This challenge is intrinsically linked to the difficulty of task decomposition, where a high-level natural language instruction (e.g., "prepare a cup of coffee") must be broken down into a series of executable sub-goals (e.g., "find mug," "fill with water," "place in machine"). Current autoregressive models can suffer from error accumulation, where a mistake in an early sub-goal leads to catastrophic failure later in the sequence. Research efforts, such as the development of hierarchical VLA frameworks and specialized models like LoHoVLA [35], are focused on integrating explicit memory and predictive world models to improve planning capabilities.

Long-Horizon Planning The challenge is the exponential growth of the search space for optimal action sequence $A_{1:T}$ as T increases.

Hierarchical Solution Approach: Decompose into high-level planning (sub-goals $M_{1:K}$ where $K \ll T$) and low-level execution:

$$M^* = \arg\max_M \mathbb{E}\left[\sum_{k=1}^{K} \gamma_M^{k-1} R_M(s_k, m_k) \mid l\right] \tag{5}$$

$$\pi_k^* = \arg\max_{\pi_k} \mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} \gamma^{t-t_k} R(s_t, a_t) \mid m_k\right] \tag{6}$$

where M^* represents the optimal sequence of sub-goals given a high-level instruction l, and π_k^* is the optimal low-level policy for executing sub-goal m_k.

10.2 Open-World Generalization and Data Efficiency

The promise of VLA models lies in their ability to act as generalist agents, yet achieving true open-world generalization remains elusive. VLA models, particularly those based on large pre-trained Vision-Language Models, rely heavily on massive, diverse datasets that fuse visual, linguistic, and action data. The sheer scale and cost of collecting high-quality, real-world robotic demonstration data is a major constraint. Furthermore, models often exhibit poor generalization to novel objects, unseen environments, or tasks that require common-sense knowledge not explicitly present in the training data. This highlights a need for more data-efficient learning paradigms, including better sim-to-real transfer techniques, self-supervised learning, and the integration of symbolic reasoning to inject robust, generalizable knowledge into the control policy.

Generalization error measures performance difference between training distribution P_{train} and novel test distribution P_{test}:

$$\begin{aligned}\text{Generalization Error} = \mathbb{E}_{o\sim P_{test}}[L(\pi(o), a^*)] \\ -\mathbb{E}_{o\sim P_{train}}[L(\pi(o), a^*)]\end{aligned} \tag{7}$$

where L is the loss function and a^* is the optimal action.

Solution Approaches:

1. **Domain Randomization:** Expand P_{train} to reduce distribution shift
2. **Transfer Learning:** Initialize π with weights from pre-trained Vision-Language Models

10.3 Safety, Alignment, and Ethical Deployment

As VLA models move toward autonomous operation in human environments, the assurance of safety and alignment becomes a non-negotiable requirement. A key challenge is ensuring that the VLA policy adheres to physical constraints, avoids harmful actions, and respects human-defined ethical boundaries. The black-box nature of large neural network policies makes it difficult to guarantee safe behavior, especially in novel or adversarial situations. The concept of safety alignment for VLAs, similar to that for LLMs, is an emerging field, with works like SafeVLA [36] proposing integrated safety approaches to constrain the model's action space. Beyond immediate safety, ethical concerns regarding accountability, bias amplification from training data, and the potential for misuse in autonomous systems must be addressed before widespread deployment.

Neuro-Symbolic Control for Enhanced Safety: The inherent black-box nature of end-to-end VLA models poses significant challenges for formal safety guarantees. Neuro-symbolic approaches offer a promising avenue by integrating symbolic reasoning with neural networks. These hybrid architectures, which combine PDDL-based symbolic planning with learned low-level control (e.g., diffusion policies), have demonstrated superior performance in structured long-horizon tasks, coupled with significantly lower energy consumption compared to pure VLA models. This explicit incorporation of symbolic structure allows for formal verification and adherence to predefined safety constraints, mitigating unpredictable behavior in critical scenarios.

10.4 Computational and Real-Time Inference Constraints

The computational complexity of VLA models poses a significant barrier to real-time control and edge deployment. The unified nature of these models, often incorporating multi-billion parameter VLMs, demands substantial computational resources for both training and inference. For embodied agents, the policy must execute actions at high frequencies (e.g., 10-50 Hz) to ensure smooth and reactive control. This requirement often clashes with the latency introduced by large transformer-based architectures. Consequently, a major challenge is the development of efficient , training-free acceleration techniques and parameter-efficient methods to reduce the model footprint and inference time without sacrificing performance, enabling deployment on resource-constrained robotic hardware.

Addressing Latency with Efficient Architectures: Recent advancements focus on optimizing VLA models for real-time performance. Techniques such as Action Tokenization Efficiency, exemplified by FAST [37] and VQ-VLA [38], compress the action space into discrete tokens, reducing sequence length and accelerating inference. Furthermore, Model Quantization (e.g., INT8) and Low-Rank Adaptation (LoRA) [4] enable efficient serving and fine-tuning of large models. Innovative architectures like RoboMamba [11], which utilize Mamba-based blocks for linear-time sequence modeling, offer a more efficient solution for high-frequency control compared to traditional transformer-based models.

10.5 The Modality Gap: Integrating Multi-Modal Feedback Loops

While current VLA models primarily leverage visual and linguistic inputs, a significant research gap exists in effectively integrating richer sensory modalities, particularly tactile and haptic feedback. For many real-world robotic tasks, especially those involving delicate manipulation, object interaction, or texture discrimination, vision and language alone are insufficient. The absence of direct physical contact information limits a VLA agent's ability to perform tasks requiring fine-grained motor control, assess material properties, or safely interact with its environment. This modality gap prevents VLA models from achieving human-level dexterity and robustness in contact-rich scenarios.

Research Gap: The current state-of-the-art often treats haptics as a secondary, reactive channel, rather than an integral part of contextual understanding and action generation. There is a critical need for VLA frameworks that can synthesize mid-air force and vibration cues as a direct consequence of contextual visual understanding and natural language commands. This necessitates the development of unified encoders capable of projecting heterogeneous inputs (vision, language, tactile, force-torque, audio) into a shared, semantically rich latent space. The challenge lies in designing effective fusion mechanisms (e.g., cross-attention, hierarchical structures) that can intelligently weigh and combine information from modalities with vastly different signal-to-noise ratios and temporal characteristics. The integration of tactile/haptic sensors into the VLA framework is currently a significant omission in many surveys and a crucial area for future development.

11 Solution Approaches for VLA Challenges

This section details the prominent techniques and frameworks developed to address the core challenges of Vision-Language-Action models. These advancements aim to enhance the reasoning depth, physical grounding, and operational reliability of embodied agents.

11.1 Addressing Long-Horizon Planning

The complexity of long-horizon planning, characterized by extended sequences of interdependent actions, is being addressed through several sophisticated architectural and algorithmic strategies. One foundational paradigm is Hierarchical Reinforcement Learning (HRL), which explicitly decouples high-level strategic planning from low-level motor execution. In this framework, high-level policies determine abstract goals while specialized low-level policies execute the precise movements required to achieve these sub-goals. This division reduces the effective horizon for individual policies, thereby mitigating the risk of cumulative errors over time.

Complementing hierarchical structures is the integration of persistent memory modules. By maintaining a continuous belief state over extended durations, VLA architectures can achieve a deeper contextual understanding of their environment. This capability is essential for tasks that require the recall of past observations or the tracking of object states that are no longer in the immediate field of view. Furthermore, the reasoning power of Large Language Models is increasingly utilized to serve as a high-level cognitive layer. These models translate complex natural language instructions into executable sub-goal sequences, leveraging their vast internal world knowledge to perform logical inference and task decomposition. Finally, online planning techniques such as Model Predictive Control (MPC) provide a robust mechanism for real-time adaptation. By optimizing a sequence of actions over a finite horizon at each time step, then executing only the first action. The process is repeated, allowing for continuous re-planning and adaptation to dynamic environments, thereby improving robustness against unexpected changes.

11.2 Enhancing Open-World Generalization

Achieving robust generalization in embodied agents requires a multi-faceted approach that moves beyond simple pattern matching to address novel tasks and unfamiliar environments. The most direct strategy involves the massive scaling of both data and model parameters. Training on diverse datasets synthesized from web-scale vision-language data, robotic demonstrations, and simulated environments provides the necessary representational breadth for open-world operation. Simultaneously, increasing model capacity allows VLA architectures to capture the intricate relationships between linguistic concepts and physical manifestations.

To prevent overfitting to specific robot embodiments or environments, researchers are developing "knowledge insulation" techniques. These include extensive domain randomization, which exposes models to a wide variety of visual and physical conditions during training, and meta-learning frameworks designed for rapid adaptation to new scenarios with minimal data requirements. While imitation learning provides a strong initial policy, reinforcement learning (RL) fine-tuning is often employed to enhance robustness. By allowing agents to explore and optimize their strategies in diverse environments, RL enables the discovery of successful behaviors that may not be present in the initial demonstration data. Furthermore, the adoption of modular architectures enhances generalization by decoupling perception, reasoning, and action into distinct, independent sub-systems. This structural modularity facilitates the targeted adaptation of specific components, enabling the agent to adjust to new tasks or environments.

11.3 Implementing Safety and Alignment

Ensuring that VLA models operate within safe boundaries and remain aligned with human intent is a paramount concern for deployment. Explicit alignment is frequently formulated through the Constrained Markov Decision Process (CMDP) framework. For instance, the SafeVLA algorithm [36] incorporates safety constraints directly into the training objective by treating safety violations as quantifiable costs. This formulation encourages the agent to maximize task rewards while strictly adhering to predefined safety thresholds.

The development of standardized evaluation platforms, such as the Safety-CHORES Benchmark [36], further supports these efforts. By providing millions of simulated scenes with explicit safety constraints, these benchmarks enable the rigorous testing and scalable training of safe behaviors. Moreover, constrained learning techniques, including Lagrangian methods and safe reinforcement learning, are used to restrict the agent's exploration to a verified safe subset of the action space. This effectively prevents the execution of high-risk actions that could lead to physical damage or unsafe outcomes. Beyond purely statistical methods, neuro-symbolic integration offers a pathway to embed formal logic and safety rules directly into the control policy. This hybrid approach provides a more transparent and verifiable mechanism for safety, which is essential for applications in human-centric or high-stakes environments.

11.4 Optimizing for Computational Constraints

Deploying large-scale VLA models on resource-constrained hardware requires innovative acceleration and compression techniques to achieve real-time performance. Efficient action tokenization is a primary area of research, with schemes such as FAST [37] and VQ-VLA [38] compressing continuous control signals into a compact discrete vocabulary. This significantly shortens the sequence length required for autoregressive generation, thereby reducing inference latency and computational overhead.

Model efficiency is further enhanced through quantization and low-rank adaptation. Techniques such as 4-bit or 8-bit quantization drastically reduce the memory footprint of large models with minimal impact on performance. Similarly, Low-Rank Adaptation (LoRA) [4] enables the efficient fine-tuning of massive pre-trained foundations by updating only a small fraction of the total parameters, making task-specific adaptation feasible on consumer-grade hardware. Furthermore, the development of non-transformer backbones, such as the Mamba-based RoboMamba [11], offers a scalable solution for high-frequency control. By utilizing state-space models with linear-time sequence complexity, these architectures avoid the quadratic overhead associated with standard attention mechanisms, enabling the rapid inference speeds necessary for dexterous robotic manipulation.

11.5 Bridging the Modality Gap through Advanced Fusion

Addressing the sensory gap between vision, language, and physical feedback is crucial for tasks involving complex contact and manipulation. Recent solutions focus on unified encoding strategies where diverse inputs, including tactile, haptic, and proprioceptive signals, are projected into a shared and semantically rich latent space. This unified representation allows the VLA model to reason about multi-modal information in a coherent and grounded manner.

Beyond simple feature concatenation, advanced fusion mechanisms such as cross-attention and gated networks are increasingly employed. These methods allow the model to dynamically weigh the importance of different streams based on their relevance to the current task and their respective signal-to-noise ratios. For example, the VLH model [39] demonstrates a Vision-Language-Haptics foundation that synthesizes force and vibration cues directly from visual context and linguistic

commands. This level of integration is essential for contact-rich tasks where visual information alone is insufficient. Finally, the creation of specialized multi-modal datasets and benchmarks is driving progress in this area, providing the synchronized recordings of vision, language, and haptic feedback required to train and evaluate the next generation of truly multi-sensory VLA agents.

12 Promising Future Directions

While Vision-Language-Action models have demonstrated remarkable capabilities in embodied intelligence, several fundamental challenges remain, pointing toward critical and promising avenues for future research. Overcoming these limitations will be essential for transitioning VLA agents from controlled laboratory settings to robust, general-purpose deployment in the real world.

12.1 Lifelong and Continuous Learning

The current paradigm, heavily reliant on large-scale, fixed datasets, is inherently insufficient for real-world embodied agents that must operate in dynamic, non-stationary environments. Lifelong Learning (LL) and Continuous Learning (CL) frameworks are essential to enable VLA agents to incrementally acquire new skills, adapt to changing environmental conditions, and, crucially, mitigate the phenomenon of catastrophic forgetting, where learning a new task erases knowledge of previously mastered skills [40].

A key technical challenge in this domain is the development of robust knowledge consolidation mechanisms. This often involves modifying the standard Imitation Learning objective with a regularization term that penalizes significant deviations in parameters important for past tasks. A common approach, inspired by methods like Elastic Weight Consolidation (EWC), introduces a penalty based on the Fisher Information Matrix F:

$$\mathcal{L}_{CL}(\theta) = \mathcal{L}_{new}(\theta) + \sum_{k} \frac{\lambda}{2} F_k (\theta_k - \theta_k^*)^2 \tag{8}$$

Here, $\mathcal{L}_{new}(\theta)$ is the loss for the current task, θ are the current model parameters, θ^* are the parameters optimized for the previous task, F_k is the diagonal of the Fisher Information Matrix for the k-th parameter (representing its importance to the previous task), and λ is a hyperparameter controlling the strength of the regularization. Future work must focus on more scalable and efficient methods for calculating and utilizing this importance-based regularization, particularly in the context of massive VLA models. Furthermore, developing mechanisms for self-supervised data collection and knowledge transfer between tasks will be paramount for enabling true autonomy.

12.2 Multi-Modal and Multi-Embodiment Fusion

Future VLA architectures must move beyond traditional camera and language inputs to incorporate a broader spectrum of sensory modalities and achieve seamless skill transfer across diverse robotic platforms.

12.2.1 Multi-Modal Fusion

The integration of additional sensory streams, such as tactile feedback ($o_{tactile}$), force-torque sensing (o_{force}), and audio cues (o_{audio}), is critical for tasks requiring fine-grained interaction and safety. This

necessitates the development of unified encoders that can project heterogeneous inputs into a shared, semantically rich latent space E. The unified embedding E can be expressed as a function of the individual modality embeddings:

$$E = \text{Fusion}(E_{vision}, E_{language}, E_{tactile}, E_{force}, \dots) \tag{9}$$

where $E_{modality} = \text{Encoder}_{modality}(Input_{modality})$. The challenge lies in designing the Fusion function. Whether implemented through simple concatenation, cross-attention mechanisms, or complex hierarchical structures, this function must effectively weigh and combine information from modalities that possess significantly different signal-to-noise ratios and temporal characteristics [41]. The VLH model [39] is a recent example of a Vision-Language-Haptics foundation model that unifies perception, language, and tactile feedback, demonstrating the potential of such multi-modal fusion.

12.2.2 Multi-Embodiment Generalization

Achieving skill transfer across different robot platforms, such as transitioning from a fixed-base manipulator to a mobile humanoid, requires decoupling the learned policy from the specific kinematics of the robot. This challenge is being addressed through the development of embodiment-specific soft prompts [42] or by conditioning the policy on a standardized representation of the robot's physical capabilities and current state, including end-effector poses and joint limits. The goal is to learn a generalized policy π_{gen} that is conditioned on both the task instruction l and the robot's embodiment context C_{robot}:

$$a_t = \pi_{gen}(o_t, l, C_{robot}; \theta) \tag{10}$$

12.3 Explainability and Interpretability

For VLA models to be deployed safely and reliably, particularly in human-robot collaboration or safety-critical industrial settings, their decision-making processes must be transparent. Future work will focus on methods to interpret the model's internal reasoning, providing human-understandable explanations for the generated actions and facilitating debugging and safety auditing.

This involves developing an Explanation Generation Function Explain$(\cdot)$ that takes the policy π, the current observation o_t, and the instruction l as input, and outputs a human-readable explanation X:

$$X = \text{Explain}(\pi, o_t, l) \tag{11}$$

Promising techniques include attention map visualization (to show which parts of the image and instruction are most salient to the action), counterfactual explanations (e.g., "The robot did not pick up the red block because the instruction specified the blue one"), and neuro-symbolic approaches that ground the continuous policy output in discrete, logical steps. The ultimate goal is to move beyond simple post-hoc analysis to build inherently interpretable VLA architectures.

13 Discussion

The rapid evolution of Vision-Language-Action models has fundamentally redefined the landscape of embodied AI, transitioning from modular, engineered pipelines to unified, end-to-end architectures. The analysis reveals that the success of models like RT-2 and OpenVLA is largely attributed to their ability

to leverage internet-scale pre-training, which provides a rich semantic foundation for understanding open-world environments. However, several critical points of discussion emerge from the current state of the art.

The Internet-Scale Pretraining Dividend and Its Limits

The success of models such as RT-2 and GR-2 is largely attributable to their ability to leverage internet-scale pretraining, which provides a rich semantic foundation for understanding open-world environments. Yet meta-analysis reveals a key challenge: the 55 B-parameter RT-2 underperforms the 7 B-parameter OpenVLA on multi-robot generalization benchmarks despite its vastly larger pretraining corpus. This finding suggests that data quality, architectural choice specifically, the fusion of DINOv2 and SigLIP visual features and training data diversity are more decisive than parameter scale. The field is moving away from scale is all you need toward scale the right things. Future work should investigate whether targeted, embodiment-aware dataset curation can yield further efficiency gains, potentially reducing the environmental and financial cost of training frontier VLA models.

The Efficiency Frontier: SSMs, Diffusion, and the Path to Edge Deployment

A clear and practically significant trend is the emergence of architectures capable of real-time control on consumer-grade or edge hardware. Models such as RoboMamba, which replaces transformer self-attention with linear-complexity State Space Model blocks, and π-0, which leverages optimized diffusion sampling, demonstrate that control frequencies above 20 Hz are achievable within modest memory budgets (5–15 GB VRAM). The deployment feasibility regimes provide practitioners with a principled framework for selecting models commensurate with their hardware constraints. Critically, the gap between enterprise-scale models and edge-deployable models in task success rate is narrowing from roughly 30 percentage points in 2022 to under 15 percentage points by 2025 signaling a maturing of the field toward practical deployment.

Benchmark Fragmentation and the Reproducibility Crisis

One of the most consequential structural problems in the current VLA literature is the absence of standardized evaluation protocols, which hinders meaningful cross-model comparison and reproducibility. Models are routinely compared across different benchmarks, hardware configurations, and statistical reporting conventions. The CALVIN benchmark favors models with strong sequential reasoning, while the Open X-Embodiment suite emphasizes cross-embodiment generalization; a model excelling on one may perform poorly on the other. This benchmark fragmentation artificially inflates the apparent diversity of state-of-the-art results and makes genuine scientific comparison nearly impossible. The adoption of Reproducibility Scorecards represents a necessary step; the community must additionally converge on a small set of canonical, standardized evaluation harnesses with mandatory statistical reporting.

The Dominance of Imitation Learning and the Reinforcement Learning Gap

The overwhelming reliance on imitation learning (IL) throughout the VLA literature introduces a structural ceiling on performance: no IL-trained model can, in principle, exceed the quality of its demonstrators. Reinforcement learning offers a pathway beyond this ceiling, enabling autonomous exploration, self-correction, and the discovery of strategies absent from any demonstration dataset yet its integration into large VLA models remains nascent. The primary obstacles are reward specification

complexity, training instability at billion-parameter scale, and the sample inefficiency of RL in physical or high-fidelity simulated environments. Hybrid IL+RL approaches, using human demonstrations to bootstrap policy search and RL to refine it, represent the most plausible near-term path to surpassing the imitation ceiling.

Safety, Ethics, and the Accountability Gap

As VLA models are deployed in human environments, the ethical stakes rise sharply. The black-box nature of large neural policies makes formal safety guarantees difficult, and the data-driven nature of training means that societal biases encoded in internet-scale corpora may be inherited and amplified by robotic behavior. The neuro-symbolic approaches offer a technically promising mitigation: by grounding high-level plans in verifiable symbolic structures, it becomes possible to guarantee that certain unsafe action sequences are never executed, regardless of the neural policy's output. However, symbolic systems themselves require carefully curated knowledge representations, and the fidelity of the world model used for predictive planning remains a bottleneck. Broader accountability frameworks spanning legal liability, audit trails for autonomous decisions, and bias testing protocols remain underdeveloped relative to the pace of model deployment.

The Modality Gap as a Fundamental Bottleneck

Current VLA models are predominantly visual-linguistic: they see and understand, but cannot feel. For contact-rich manipulation tasks assembly, surgery assistance, texture-sensitive sorting the absence of integrated tactile and haptic feedback is not a minor omission but a fundamental capability gap. The VLH model and related work demonstrate that multi-modal fusion is technically feasible, but the lack of large-scale, synchronized tactile-visual-language datasets remains a critical data infrastructure problem. Addressing this gap will require coordinated efforts in robotic hardware standardization, sensor data collection pipelines, and new benchmark development is a challenge that is as much organizational as it is technical.

Synthesis: Where the Field Stands and Where It Must Go

The VLA is characterized by rapidly improving generalization, narrowing efficiency gaps, and expanding application domains. The expansion into autonomous driving (AutoVLA) and humanoid loco-manipulation (Humanoid-VLA, GR00T N1) demonstrates the versatility of the VLA paradigm beyond tabletop manipulation. Yet the domain faces structural constraints that, if unaddressed, will limit its long-term impact: (1) the imitation learning ceiling, which requires meaningful Reinforcement Learning integration; (2) benchmark fragmentation, which requires community coordination on evaluation standards; and (3) the safety-interpretability deficit, which requires neuro-symbolic integration and formal verification methods to mature in parallel with neural architectures. The most promising models of the next generation will likely combine large-scale semantic pretraining with modular symbolic planning, efficient SSM or diffusion-based action generation, and lifelong learning mechanisms that allow continuous skill acquisition without catastrophic forgetting.

14 Conclusion

This review has presented a comprehensive overview of the transformative Vision-Language-Action paradigm in robotic manipulation and embodied intelligence. The field has advanced with remarkable speed, progressing from RT-1, a pioneering single-embodiment imitation learner, to models such as

OpenVLA and GR00T N1. These recent architectures generalize across dozens of robotic platforms, environments, and task categories while operating effectively on consumer-grade hardware.

The contributions of this review reflect the multidimensional nature of this progress. The architectural taxonomy, which spans unified transformer-based policies, diffusion-based action generators, and world model-integrated planners, provides a principled scaffold for navigating the trade-offs between generalization, precision, and computational efficiency. Technical analysis reveals that fusing diverse visual encoders, such as DINOv2 and SigLIP, with large language model backbones like Llama 2 is more predictive of real-world performance than parameter count alone. This finding directly challenges the prevailing narrative that scale is sufficient for performance. Furthermore, the treatment of neuro-symbolic control loops and software-hardware co-design addresses two engineering concerns that are often underrepresented in the VLA literature: the necessity for formal safety guarantees in human-proximate robotic systems and the requirement for hardware platforms whose sensing and actuation standards are co-designed with VLA data representations. Finally, the research roadmap identifies lifelong learning, multi-modal sensory fusion, and explainability as the most consequential open problems, providing concrete formulations to guide future implementation.

Despite these significant advances, substantial challenges remain. Long-horizon task decomposition continues to expose the error-compounding fragility of purely reactive VLA policies. True open-world generalization, defined as the ability to successfully interact with an entirely novel object, environment, or instruction, has not yet been reliably demonstrated. The integration of reinforcement learning at the scale of modern VLA architectures remains an unsolved engineering and algorithmic problem. Furthermore, the ethical dimensions of deploying autonomous embodied agents in human environments, including questions of bias, accountability, and physical safety, demand sustained attention from the research community.

Ultimately, Vision-Language-Action models represent a fundamental shift in the relationship between language, perception, and physical agency. By unifying these modalities within a single learned framework, VLA architectures dissolve the boundaries between understanding the world and acting within it. Realizing the full promise of this paradigm, including generalist, safe, interpretable, and continuously learning robotic agents, will require simultaneous advances in architecture, algorithms, ethics, and infrastructure. By consolidating the current state of knowledge and identifying the most consequential open problems, this review serves as a foundation for researchers and practitioners navigating this rapidly evolving and profoundly important frontier.

Author Contribution Statement

All authors contributed equally to the study conception and design. Material preparation, data collection, and analysis were performed by the authors. The first draft of the manuscript was written by the authors, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Ethics Approval and Consent to Participate

This study did not involve human participants or animals. Therefore, ethical approval and consent to participate are not applicable. If applicable, provide details of the ethics committee approval and consent procedures here.

Consent for Publication

Not applicable. (If applicable, mention consent from participants or institutions here.)

Data Availability

The datasets analyzed during the current study are available from the corresponding author on reasonable request. (If data are publicly available, specify the repository and link here.)

Acknowledgments

The authors would like to thank the reviewers, Associate Editor, and Editor-in-Chief for their valuable comments and suggestions, which helped improve the quality of this paper. The authors also acknowledge the use of GPT-4o, Claude AI and DeepSeek for assistance in technical writing and structural development.

Funding

This work was supported by the National Nature Science Foundation of China (Grant No. 12371261).

Disclosure Statement

The authors declare that they have no competing interests.

References

[1] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, "A survey on vision-language-action models for embodied ai," *arXiv preprint arXiv:2405.14093*, 2024.

[2] R. Sapkota, Y. Cao, K. I. Roumeliotis, and M. Karkee, "Vision-language-action models: Concepts, progress, applications and challenges," *arXiv preprint arXiv:2505.04769*, 2025.

[3] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*, pp. 2165–2183, PMLR, 2023.

[4] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[5] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[6] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023," *URL https://arxiv. org/abs/2307.15818*, 2024.

[7] C.-L. Cheang, G. Chen, Y. Jing, T. Kong, H. Li, Y. Li, Y. Liu, H. Wu, J. Xu, Y. Yang, *et al.*, "Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation," *arXiv preprint arXiv:2410.06158*, 2024.

[8] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[9] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[10] 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*, pp. 11975–11986, 2023.

[11] J. Liu, M. Liu, Z. Wang, P. An, X. Li, K. Zhou, S. Yang, R. Zhang, Y. Guo, and S. Zhang, "Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 40085–40110, 2024.

[12] O. Mees, D. Ghosh, K. Pertsch, K. Black, H. R. Walke, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo, *et al.*, "Octo: An open-source generalist robot policy," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[13] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[14] W. Zhang, H. Liu, Z. Qi, Y. Wang, X. Yu, J. Zhang, R. Dong, J. He, F. Lu, H. Wang, *et al.*, "Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge," *arXiv preprint arXiv:2507.04447*, 2025.

[15] L. Li, J. Fan, X. Ni, S. Qin, W. Li, and F. Gao, "Sva: Towards speech-enabled vision-language-action model," *Pattern Recognition*, p. 112915, 2025.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] 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)*, pp. 4171–4186, 2019.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PmLR, 2021.

[19] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903, IEEE, 2024.

[20] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning*, pp. 1723–1736, PMLR, 2023.

[21] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640, 2019.

[22] N. Lin and M. Cai, "Epic-kitchens-100 unsupervised domain adaptation challenge for action recognition 2022: Team hnu-fpv technical report," *arXiv preprint arXiv:2207.03095*, 2022.

[23] Q. Li, Y. Liang, Z. Wang, L. Luo, X. Chen, M. Liao, F. Wei, Y. Deng, S. Xu, Y. Zhang, *et al.*, "Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation," *arXiv preprint arXiv:2411.19650*, 2024.

[24] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, "3d-vla: A 3d vision-language-action generative world model," *arXiv preprint arXiv:2403.09631*, 2024.

[25] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1684–1704, 2025.

[26] S. Lee, Y. Wang, H. Etukuru, H. J. Kim, N. M. M. Shafiullah, and L. Pinto, "Behavior generation with latent actions," *arXiv preprint arXiv:2403.03181*, 2024.

[27] K. Bousmalis, G. Vezzani, D. Rao, C. Devin, A. X. Lee, M. Bauzá, T. Davchev, Y. Zhou, A. Gupta, A. Raju, *et al.*, "Robocat: A self-improving generalist agent for robotic manipulation," *arXiv preprint arXiv:2306.11706*, 2023.

[28] Z. Dong, Y. Liu, S. Zhang, B. Ye, Y. Yuan, F. Ni, J. Gong, X. Qiu, H. Zhao, Y. Li, *et al.*, "Actioncodec: What makes for good action tokenizers," *arXiv preprint arXiv:2602.15397*, 2026.

[29] Z. Liang, Y. Li, T. Yang, C. Wu, S. Mao, T. Nian, L. Pei, S. Zhou, X. Yang, J. Pang, *et al.*, "Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies," *arXiv preprint arXiv:2508.20072*, 2025.

[30] Y. Yang, Y. Wang, Z. Wen, L. Zhongwei, C. Zou, Z. Zhang, C. Wen, and L. Zhang, "Efficientvla: Training-free acceleration and compression for vision-language-action models," *arXiv preprint arXiv:2506.10100*, 2025.

[31] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[32] Z. Zhou, T. Cai, S. Z. Zhao, Y. Zhang, Z. Huang, B. Zhou, and J. Ma, "Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning," *arXiv preprint arXiv:2506.13757*, 2025.

[33] P. Ding, J. Ma, X. Tong, B. Zou, X. Luo, Y. Fan, T. Wang, H. Lu, P. Mo, J. Liu, *et al.*, "Humanoid-vla: Towards universal humanoid control with visual integration," *arXiv preprint arXiv:2502.14795*, 2025.

[34] S. Poria, N. Majumder, C.-Y. Hung, A. A. Bagherzadeh, C. Li, K. Kwok, Z. Wang, C. Tan, J. Wu, and D. Hsu, "10 open challenges steering the future of vision-language-action models," *arXiv preprint arXiv:2511.05936*, 2025.

[35] Y. Fan, P. Ding, S. Bai, X. Tong, Y. Zhu, H. Lu, F. Dai, W. Zhao, Y. Liu, S. Huang, *et al.*, "Long-vla: Unleashing long-horizon capability of vision language action model for robot manipulation," *arXiv preprint arXiv:2508.19958*, 2025.

[36] B. Zhang, Y. Zhang, J. Ji, Y. Lei, J. Dai, Y. Chen, and Y. Yang, "Safevla: Towards safety alignment of vision-language-action model via constrained learning," *arXiv preprint arXiv:2503.03480*, 2025.

[37] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine, "Fast: Efficient action tokenization for vision-language-action models," *arXiv preprint arXiv:2501.09747*, 2025.

[38] Y. Wang, H. Zhu, M. Liu, J. Yang, H.-S. Fang, and T. He, "Vq-vla: Improving vision-language-action models via scaling vector-quantized action tokenizers," *arXiv preprint arXiv:2507.01016*, 2025.

[39] L. F. Moreno Fuentes, M. Haris Khan, M. Altamirano Cabrera, V. Serpiva, D. Iarchuk, Y. Mahmoud, I. Tokmurziyev, and D. Tsetserukou, "Vlh: Vision-language-haptics foundation model," *arXiv e-prints*, pp. arXiv–2508, 2025.

[40] R. Fan, M. Sun, and G. Giakos, "Toward the next frontier of embodied ai," 2025.

[41] X. Han, S. Chen, Z. Fu, Z. Feng, L. Fan, D. An, C. Wang, L. Guo, W. Meng, X. Zhang, *et al.*, "Multimodal fusion and vision-language models: A survey for robot vision," *Information Fusion*, p. 103652, 2025.

[42] J. Zheng, J. Li, Z. Wang, D. Liu, X. Kang, Y. Feng, Y. Zheng, J. Zou, Y. Chen, J. Zeng, *et al.*, "X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model," *arXiv preprint arXiv:2510.10274*, 2025.