FROM DATA TO DYNAMICS: A COMPREHENSIVE SURVEY OF DATASETS AND TRAINING PARADIGMS FOR EMBODIED WORLD MODELS

A PREPRINT

Junjie Liu[*]
School of Computer Science
Xi'an Shiyou University
Xi'an 710065, China
202215050307@stumail.xsyu.edu.cn

Elias D. Striatum
Department of Electrical Engineering
Mount-Sheikh University
Santa Narimana, Levand
stariate@ee.mount-sheikh.edu

February 27, 2026

ABSTRACT

Embodied AI has recently moved from narrow task-specific control to broader Vision-Language-Action and world-model-based systems. This survey focuses on the data and training side of that transition. We organize the literature with three axes: **data provenance** (passive video, embodied interaction, synthetic/simulated), **state representation** (pixel, latent, object-centric), and **inference/training paradigm** (autoregressive, diffusion/flow-based, reinforcement-style post-training). We then analyze how large cross-platform robot datasets (e.g., Open X-Embodiment **?**), world-model architectures (e.g., π_0 **?**), and simulator/data-engine pipelines interact in modern embodied learning. A comparative summary table is provided to distinguish **algorithm-oriented** and **data-oriented** studies within a unified survey structure. Finally, we discuss open problems in physical consistency, long-horizon decision quality, robustness under distribution shift, and evaluation protocols for real-world deployment.

1 Introduction

This survey synthesizes recent progress in embodied AI with a focus on model design, data foundations, and training paradigms, and provides a structured comparison of algorithm-oriented and data-oriented studies in Section **??**.

1.1 The Frontier of Physical Intelligence

The trajectory of Artificial Intelligence in the mid-2020s is defined by a stark dichotomy: the mastery of "Digital Intelligence" versus the nascent struggle for "Physical Intelligence." While Large Language Models (LLMs) have achieved near-human or superhuman proficiency in symbolic reasoning, coding, and creative generation—effectively solving the "Digital" domain—embodied agents remain fundamentally constrained by Moravec's Paradox. This paradox, formulated in the 1980s, posits that high-level reasoning requires relatively little computation, whereas low-level sensorimotor skills (perception, mobility, manipulation) require enormous computational resources and evolutionary prior. As of 2025, artificial agents can pass the Turing Test, score in the 99th percentile of the LSAT, and generate award-winning art, yet they struggle to fold a shirt, clear a cluttered table, or navigate a crowded sidewalk with the fluidity and robustness of a biological organism.

The central hypothesis driving the current research landscape is that **World Models** constitute the critical missing link in bridging this gap. Just as LLMs internalized the statistical structure of language through internet-scale text corpora, World Models aim to learn the "syntax of reality"—the underlying laws of physics, causal relationships, object

[*] Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

permanence, collision dynamics, and material properties—from massive multimodal sensory streams **????**. A robust world model serves as an internal simulator, allowing an agent to predict the consequences of its actions without the risks and costs of real-world trial and error.

The imperative for World Models arises from the fundamental limitations of end-to-end model-free Reinforcement Learning (RL). While model-free methods have solved specific tasks like Go, Dota 2, or StarCraft, they suffer from extreme sample inefficiency, often requiring millions or billions of interaction cycles to converge. In the digital realm of video games, this simulation time is cheap. In the physical world, such data collection is prohibitively expensive, dangerous, and unscalable. Robots break, batteries die, and environments reset slowly. World Models address this by enabling **Model-Based Reinforcement Learning (MBRL)**, where the agent "dreams" potential futures, evaluates counterfactuals, and refines policies in a learned latent space before committing to expensive real-world actions. This capability transforms the agent from a reactive system to a predictive one, capable of planning over long horizons and generalizing to unseen environments **??**.

1.2 Historical Context: From Data Scarcity to the Data Engine Era

The evolution of Embodied AI can be characterized by the transition from the "Data Scarcity" era to the "Data Engine" era. Historically, robotic learning was bottlenecked by the difficulty of acquiring high-quality interaction data. Unlike computer vision or NLP, where datasets like ImageNet or CommonCrawl provided billions of static examples, robotics lacked a comparable resource due to the physical risks and logistical complexities of teleoperation. Early approaches relied on small-scale, lab-controlled datasets or low-fidelity simulations, resulting in policies that failed to generalize beyond their training distribution (the "Reality Gap").

However, the period from 2024 to 2026 marked a paradigm shift driven by two converging trends: the rise of **Embodied Data Engines** and the scaling of **Generative Video Priors**. Initiatives such as the Open X-Embodiment collaboration **?** and the DROID project **?** demonstrated that diverse, cross-institutional robot data could be aggregated to train generalist policies. These projects fundamentally altered the data landscape, shifting from static repositories to dynamic "engines" where active learning loops continuously refine the model's understanding of rare events. The collection of real-world interaction data has scaled from thousands of trajectories to millions, enabled by lower-cost teleoperation hardware and fleet learning.

Simultaneously, the explosion of generative video models (e.g., Sora, Gen-3) revealed that internet-scale video data contains implicit physical knowledge. The field is now pivoting toward **Physicalization**—the process of grounding these generative priors in rigorous physical dynamics. Systems like π_0 **?**, $\pi_{0.5}$ **?**, and $\pi^*_{0.6}$ **?** exemplify this shift, treating video not merely as pixels to be predicted but as a rich source of diverse physical interactions that can be distilled into actionable control policies via techniques like Flow Matching and reinforcement learning **?**. This "Video-as-Policy" approach leverages the vast visual diversity of the web to handle open-world object recognition, while using robot interaction data to ground the dynamics in physical reality.

1.3 Formalism: World Models as Actionable Dynamic Systems

To rigorously define the scope of this survey, we must distinguish between *Generative Video Models*, which prioritize visual fidelity for human consumption, and *Embodied World Models*, which prioritize physical consistency and action-conditionability for control. Formally, an Embodied World Model is a Partially Observable Markov Decision Process (POMDP) characterized by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, P, R, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and $\mathcal{O}$ is the observation space.

The core function of the World Model is to approximate the transition dynamics of the environment. We define the **State Transition Model** as:

$$s_{t+1} \sim P_\theta(s_{t+1}|s_t, a_t, z_t) \tag{1}$$

where s_t denotes the compact latent state at time t, a_t represents the agent's action (proprioception and control signals), and z_t captures stochastic environmental uncertainty. Unlike pure video prediction, this transition must be *action-conditioned*; the model must predict how the state changes *because of* the agent's intervention, distinguishing causality from correlation **?**. A model that predicts a cup moving without an applied force is a hallucination; a model that predicts a cup moving *given* a pushing action is a world model.

Complementing the transition model is the **Observation Model** (or Decoder), which reconstructs high-dimensional sensory data (pixels, point clouds) from the latent state:

$$o_{t+1} \sim P_\phi(o_{t+1}|s_{t+1}) \tag{2}$$

While early approaches focused on pixel-space prediction, modern architectures increasingly operate in latent spaces (e.g., VQ-VAE codebooks or diffusion latents) to avoid the computational cost of rendering high-resolution images during planning. The objective is to minimize the divergence between the predicted future state distribution and the actual distribution observed from data, often formulated as minimizing a variational lower bound (ELBO) or a flow-matching objective **??**.

Furthermore, in goal-conditioned settings, we often introduce a reward model $r_t \sim P_\psi(r_t|s_t)$ or a value function $v_t = V(s_t)$ to guide the planning process towards desirable states **?**. The agent's policy $\pi(a_t|s_t)$ is then optimized to maximize cumulative reward within this "dreamed" environment:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau\sim P_\theta,\pi}\left[\sum_{t=0}^{H} \gamma^t r_t\right] \tag{3}$$

1.4 Paradigm Shifts: The Rise of Actionable Generative Models

The convergence of generative modeling and robotics has catalyzed a new class of algorithms that blur the line between video generation and robot control. We identify three distinct phases in this evolution:

1.4.1 Phase I: Passive Observation (2018–2023)

Early World Models like the original "World Models" paper **?** and Dreamer variants focused on learning dynamics from low-dimensional inputs or simple simulated environments (e.g., Atari, DeepMind Control Suite). These models were primarily "spectators," predicting future frames based on past frames without a deep understanding of complex, contact-rich physical interactions. They excelled in game environments where physics is deterministic and discrete, but struggled with the visual complexity, contact discontinuities, and high-frequency dynamics of the real world.

1.4.2 Phase II: Text-to-Video and the "Physics Hallucination" Problem (2024)

The advent of large-scale diffusion models (Sora, Kling, Gen-3) enabled the generation of photorealistic video from text prompts. These models demonstrated an unprecedented grasp of visual semantics, lighting, and texture. However, they suffered from "Physics Hallucination"—objects would morph, disappear, or violate conservation laws (e.g., mass, momentum). A glass might merge into the table, or a hand might pass through a solid object. While visually impressive, they lacked the *controllability* required for robotics. A robot cannot plan a grasp if the object teleports or changes shape in the predicted future. This highlighted the necessity of **Action-Conditioning** and **Physical Grounding ?**.

1.4.3 Phase III: Actionable World Models and Flow Matching (2025–Present)

The current state-of-the-art, represented by architectures like BridgeV2W **?** and π_0 **?**, explicitly integrates robot actions into the generative process. These models utilize **Flow Matching**—a continuous-time generalization of diffusion—to model the vector field of state transitions. By conditioning generation on precise proprioceptive history and future action tokens, these systems can predict the outcome of a robot's interaction with deformable objects (e.g., pressing a sponge, folding a cloth) with sufficient fidelity to serve as a policy. Furthermore, techniques like **Frequency-space Action Sequence Tokenization (FAST) ?** have emerged to handle the high-frequency nature of robotic control (typically 50Hz–100Hz), compressing continuous action trajectories into discrete tokens compatible with transformer-based world models.

1.5 The Cross-Embodiment Challenge

A unique challenge in Embodied AI, distinct from NLP or Computer Vision, is the heterogeneity of hardware. "Physical Intelligence" must generalize across morphologies: from a 7-DoF Franka Emika arm to a bimanual Aloha setup, a quadrupedal Spot robot, or a humanoid figure. A World Model trained on a single robot is of limited utility. The 2026 frontier focuses on **Cross-Embodiment Generalization**—learning a universal physics representation that can be fine-tuned or prompted for specific kinematic chains. This requires handling diverse action spaces (joint angles vs. end-effector poses) and observation spaces (wrist cameras vs. head-mounted depth sensors). Models like $\pi_{0.5}$ **?** approach this by co-training on heterogeneous datasets, learning to map diverse embodiment data into a shared latent action space. Moreover, multi-agent scenarios **?** introduce further complexity, requiring models to anticipate the intentions and dynamics of other actors in shared environments.

1.6 Evaluation: Beyond Visual Fidelity

Evaluating World Models for embodied agents presents a distinct set of challenges compared to evaluating generative video. Standard computer vision metrics like Fréchet Inception Distance (FID) or FVD (Fréchet Video Distance) measure visual realism but fail to capture physical correctness. A video of a robot hand passing through a mug might have a low FID (it looks realistic) but represents a catastrophic physics violation. Consequently, the field is moving toward **Task-Centric Evaluation**—measuring the success rate of a policy trained inside the world model when deployed in the real world (Sim-to-Real). New metrics focus on "Physical Consistency," "Action-Controllability," and "Counterfactual Accuracy" **??**.

1.7 Contributions and Scope

This survey provides a comprehensive analysis of the "World Model" paradigm in Embodied AI, synthesizing developments from robotics, computer vision, and generative modeling. Our contributions are four-fold:

1. **Unified Taxonomy**: We propose a structured classification framework categorizing models by *Data Provenance* (Passive Video vs. Active Interaction), *Representation* (Pixel vs. Latent vs. Object-Centric), and *Inference Mechanism* (Autoregressive vs. Diffusion vs. Flow), providing clarity in a crowded landscape (Section **??**).
2. **The Data Engine Analysis**: We detail the architecture of modern Embodied Data Engines, analyzing how massive datasets like Open X-Embodiment are curated, stored, and retrieved to fuel world model training, addressing the "Physical Grounding Gap" **?**.
3. **Emerging Training Paradigms**: We rigorously review cutting-edge training methodologies, specifically contrasting traditional next-token prediction with **Flow Matching** and **Reinforcement Learning with Verifiable Rewards (RLVR)** for physical alignment **?**.
4. **Infinite Simulation**: We explore the role of procedural generation and "Infinite Photorealistic Worlds" (e.g., Infinigen **?**) in bridging the sim-to-real gap, allowing agents to learn in simulation and zero-shot transfer to reality.

This survey serves as a technical roadmap for researchers aiming to build the "GPT for Robotics"—a general-purpose world model that grounds digital intelligence in the physical world. By systematically reviewing the convergence of large-scale generative models with embodied control, we illuminate the path toward agents that not only see the world but understand its underlying physics well enough to act upon it.

2 Taxonomy and Problem Formulation

The rapid evolution of embodied AI and world models in the 2025–2026 landscape demands a structured classification framework that moves beyond traditional boundaries of model architecture. While early taxonomies focused primarily on the dichotomy between model-free and model-based reinforcement learning, the current era is defined by the convergence of large-scale generative models, massive heterogeneous datasets, and diverse training objectives. To systematize this complex landscape, we propose a comprehensive taxonomy centered on three primary axes: **(1) Data Provenance and Construction**, which delineates the source and nature of the training signal; **(2) Representation Modality**, which distinguishes how the world state is encoded and predicted; and **(3) Inference Paradigm**, which categorizes the temporal generation mechanism. This framework, illustrated in Table **??**, provides a unified lens to analyze systems ranging from foundational video generators like NVIDIA Cosmos to latent dynamics models like Meta's V-JEPA 2 and TARS Robotics' AWE2.0.

2.1 Axis 1: Data Provenance and Construction

The first axis categorizes world models based on the origin and active nature of their training data, reflecting a fundamental shift from data scarcity to data abundance through synthesis and internet-scale mining. **Passive Observation** models are trained primarily on large-scale, non-interactive video data sourced from the internet (e.g., YouTube, Ego4D). These systems, such as the foundation of NVIDIA Cosmos **?** and Genie **?**, excel at learning general physical priors and visual dynamics but often lack explicit action conditioning. The key technical challenge in this category is the "physicalization" of video—filtering for physical consistency, causal logic, and temporal coherence to ensure the model learns a valid simulator of reality rather than a mere sequence of pixels **?**. In contrast, **Embodied Interaction** models are grounded in active robot interaction data, capturing precise action-effect correlations from distributed teleoperation (DROID **?**), human-centric wearables (TARS WIYH **?**), and large-scale aggregation initiatives (Open X-Embodiment **?**). While these datasets provide high-fidelity proprioceptive and tactile signals, they are constrained by the high cost of physical collection. Finally, **Generative Synthesis** represents the frontier of "world model as data engine," where data is produced by high-fidelity simulators or the world models themselves. This category encompasses heterogeneous parallel simulation platforms like ManiSkill3 **?**, digital twins such as RoboGSim **?**, and approaches like GigaWorld **?** where synthetic experiences are generated to bootstrap policy learning.

2.2 Axis 2: Representation Modality

The second axis distinguishes models by their internal state representation, reflecting the inherent trade-off between visual fidelity and the abstract compactness required for robust control. **Pixel-Space Generative Models** predict future states directly in the high-dimensional pixel space, often treating video generation as the primary objective. Architectures like Vid2World **?** and Navigation World Models (NWM) **?** leverage diffusion backbones to produce interpretative rollouts, though they require significant computational resources for high-dimensional prediction. **Latent-Space Dynamics Models** operate in a compressed, abstract space, focusing on semantic and dynamic consistency rather than pixel-level perfection. This approach, typified by Joint Embedding Predictive Architectures (V-JEPA 2 **?**) and Latent Action Models (Genie), prioritizes sample efficiency and long-horizon planning stability by ignoring high-frequency visual noise that is irrelevant to task success. **Object-Centric and Structured Models** explicitly factorize the world state into objects and interaction primitives (e.g., FIOC-WM **?**). These approaches leverage inductive biases to improve combinatorial reasoning and generalization in complex, multi-object environments, enabling more precise manipulation and causal reasoning.

2.3 Axis 3: Inference and Training Paradigm

The third axis categorizes the temporal generation mechanism and the optimization objectives used to align the world model with physical reality. **Autoregressive and Sequential** models generate future states step-by-step, conditional on the entire past history $x_{<t}$. Representative architectures like RSSM-based DreamerV3 **?** and Transformer-based Genie are inherently causal and suitable for real-time control, though they can suffer from error accumulation over long horizons. **Global Diffusion and Flow Matching** models represent a significant shift toward parallelized or iterative refinement mechanisms. Examples include π_0 **?**, which utilizes Flow Matching to achieve 50Hz continuous control, and Sora-inspired physics simulators. These methods typically achieve higher temporal coherence and are more robust to the multi-modal nature of human-like behavior. Finally, **Verifiable Reward Optimization (RLVR)** represents an emerging paradigm where models are optimized via reinforcement learning against verifiable physical or logical constraints, rather than just likelihood maximization. This post-training alignment, as seen in RLVR-World **?**, ensures that the world model's predictions remain physically plausible and logically consistent during long-horizon rollouts.

Primary Axis	**Category**	**Representative Works**
Data Provenance	Passive Observation	Cosmos **?**, Genie **?**, Ego4D **?**.
	Embodied Interaction	RT-X **?**, DROID **?**, TARS AWE2.0 **?**, HumanPlus.
	Generative Synthesis	ManiSkill3 **?**, RoboGSim **?**, GigaWorld **?**, DrEureka.
Representation	Pixel-Space	Vid2World **?**, NWM **?**, DriveDreamer.
	Latent-Space	V-JEPA 2 **?**, DreamerV3 **?**, LWM.
	Object-Centric	FIOC-WM **?**, ManiGaussian.
Inference	Autoregressive	RSSM, Transformer-based (Genie **?**).
	Global/Flow	π_0 **?**, Video Diffusion (Sora).
	RL-Optimized	RLVR-World **?**, CoT-Reasoning.

Table 1: Taxonomy of World Models for Embodied AI (2025–2026 Landscape)

This taxonomy serves as the structural backbone for the subsequent sections, where we delve into the specific data engines (Section **??**), training methodologies (Section **??**), and simulation platforms (Section **??**) that embody these classifications.

3 Comprehensive Literature Review

This section provides a structured synthesis of representative studies in embodied AI. For clarity, Section **??** further organizes the literature into algorithm-oriented and data-oriented papers.

The landscape of Embodied AI has undergone a seismic and rapid transformation over the past four years, evolving from disjointed, modular pipelines of perception and planning to unified, end-to-end Vision-Language-Action (VLA) architectures. This paradigm shift has been driven by the convergence of large-scale pre-training, multimodal integration, and a renewed, critical focus on physical grounding. This section provides an extensive and detailed synthesis of the literature, tracing the trajectory from the early, text-centric LLM-based planners of 2022 to the high-fidelity, world-model-augmented, and physically aware systems of 2026. We categorize this vast body of work into five emerging and distinct technical themes: spatial and physical grounding, cognitive reasoning and planning, world model integration, computational efficiency, and robust generalization.

3.1 The Evolution of Embodied Intelligence (2022–2026)

The progression of embodied agents can be characterized by a distinct shift from "semantic reasoning" in abstract spaces to "sensorimotor mastery" in the physical world. This evolution reflects a deepening understanding of what is required to translate internet-scale knowledge into actionable robotic skills.

3.1.1 2022–2023: The Era of LLMs as Planners

In the nascent stages of large model adoption in robotics, research primarily focused on leveraging the semantic reasoning capabilities of text-only Large Language Models (LLMs) to decompose abstract, high-level tasks into executable primitives. Huang et al. (2022) **?** demonstrated that LLMs could perform zero-shot planning for embodied agents, provided the output space was strictly constrained to valid, pre-defined actions. This "Planner-Actor" paradigm was further formalized by Wu et al. (2023) **?** with the *Plan, Eliminate, and Track (PET)* framework. PET utilized LLMs not just for generating steps, but for actively filtering irrelevant objects from the perception pipeline and tracking subtask completion, essentially acting as a high-level cognitive controller over a fixed low-level policy.

Similarly, *LLM-Planner* **?** and *Voyager* **?** showcased the power of LLMs in open-ended, game-like environments such as Minecraft. *Voyager*, in particular, introduced an automatic curriculum and an iterative prompting mechanism that allowed the agent to write and refine its own code skills, effectively "learning" by expanding a library of executable primitives. Nottingham et al. (2023) **?** explored the concept of "embodied decision making" using language-guided world modeling, asking whether agents could "dream" of pixelated outcomes to inform their text-based plans. Zhang et al. (2024) **?** extended this to multi-agent cooperation with *CoELA*, where LLMs facilitated communication and joint planning. However, these systems universally suffered from a "grounding gap"—the fundamental disconnect between high-level semantic plans (e.g., "pick up the apple") and the low-level physical execution (e.g., torque control, grasp synthesis). They relied heavily on hand-coded primitives or separate, often brittle, low-level controllers, limiting their applicability to complex, contact-rich real-world tasks **??**.

3.1.2 2024: The Rise of Unified VLA Models

The introduction of Vision-Language-Action (VLA) models marked a radical departure from modular pipelines toward end-to-end learning. By tokenizing robot actions and training them alongside text and images, models began to treat motor control as a sequence modeling problem. Zhang et al. (2025) **?** provided a comprehensive survey of "Pure" VLA models. Cui et al. (2025) **?** consolidated this end-to-end learning framework, while Xiang et al. (2025) **?** and Li et al. (2025) **?** demonstrated its efficacy in fine-grained control tasks. This unified approach allowed models to inherit the vast semantic knowledge of their VLM backbones while learning the "syntax" of physical action.

Sapkota et al. (2026) **?** further categorized these developments, noting the emergence of "generalist agents" capable of cross-task transfer. The *VLA-Adapter* **?** represented a key architectural innovation, introducing efficient adaptation mechanisms to fine-tune massive pre-trained models for robotic tasks without catastrophic forgetting or prohibitive compute costs. Similarly, *VITA-VLA* **?** explored efficient "teaching" methods, distilling action knowledge into VLMs via expert action models. Kawaharazuka et al. (2025) **?** reviewed this transition towards real-world applications, emphasizing that while VLAs solved the "interface" problem between language and action, they introduced new challenges in data efficiency and precision.

3.1.3 2025–2026: Specialization and World Model Integration

As the field matured into 2025 and 2026, the research focus shifted from "can it act?" to "how well, how fast, and how reliably can it act?" The limitations of pure imitation learning—specifically its struggle with distribution shifts and long-horizon compounding errors—became apparent. Li et al. (2025) **?** and Ding et al. (2025) **?** formalized the role of *World Models*—internal simulators that predict future states to guide decision-making. These surveys established a taxonomy for world models in embodied AI, distinguishing between decision-coupled and general-purpose simulators.

Systems like *GigaBrain-0* **?** and π_0 **?** began to integrate advanced generative techniques like flow matching and diffusion processes to handle the multimodal, continuous nature of action distributions. The π_0 model, for instance, proposed a flow matching architecture on top of a VLM to inherit internet-scale semantics while outputting continuous actions, demonstrating capabilities in diverse tasks like laundry folding and box assembly. Simultaneously, the "Pragmatic" turn described by Wu et al. (2026) **?** with *LingBot-VLA* emphasized reliability, pushing for training on massive datasets (20,000+ hours) to achieve industrial-grade robustness. This era is defined by a move towards specialized architectures that address specific bottlenecks: *NORA* **?** for efficiency, *WholeBodyVLA* **?** for humanoid control, and *ObjectVLA* **?** for open-world object generalization.

3.2 Theme I: Spatial and Physical Grounding

A critical bottleneck for early VLAs was their reliance on 2D image inputs, which often discarded the precise 3D geometry and physical forces required for delicate or dynamic manipulation. Recent work has aggressively addressed this "dimensionality deficit" by incorporating depth, point clouds, and tactile feedback.

3.2.1 From 2D to 3D Representations

To overcome the inherent limitations of 2D vision, Turgunbaev et al. (2025) **?** and Din et al. (2025) **?** argue that robust perception requires full 3D understanding. Sun et al. (2025) **?** introduced *GeoVLA*. This framework integrates 3D information by converting depth maps into point clouds and employing a customized Point Embedding Network. These geometric embeddings are fused with vision-language features, empowering the model with a richer spatial understanding that 2D-only models lack. Similarly, *PointVLA* **?** injects the 3D world directly into the VLA by processing point cloud data, allowing for more precise manipulation in cluttered or depth-complex environments.

Li et al. (2025) **?** proposed *3DS-VLA*, a 3D spatial-aware model designed for robust multi-task manipulation, further cementing the importance of explicit 3D encoders. Zhou et al. (2025) **?** extended this grounding to the temporal dimension with *4D-VLA*. By utilizing sequential RGB-D inputs, 4D-VLA captures spatiotemporal dynamics and aligns the coordinate systems of the robot and the scene, addressing the "coordinate system chaos" that often plagues pre-training. For applications where full 3D perception is expensive, *OG-VLA* **?** offers an elegant alternative. It utilizes orthographic projections to canonicalize input observations, generating images that encode the next position and orientation of the end-effector in a view-invariant manner. *SpatialVLA* **?** and *QDepth-VLA* **?** explore auxiliary supervision signals. *SpatialVLA* introduces Ego3D Position Encoding and Adaptive Action Grids, while *QDepth-VLA* augments the VLA with a quantized depth prediction task, forcing the model to learn structural representations of the scene implicitly. Patratskiy et al. (2025) **?** further emphasized this with "Spatial Traces." Berg et al. (2025) **?** complement this physical grounding with deep semantic grounding, ensuring objects are understood not just by location but by function.

3.2.2 Incorporating Force and Tactile Feedback

Beyond geometry, successful manipulation—especially in contact-rich tasks—requires sensing and modulating force. *ForceVLA* **?** treats 6-axis force-torque data not as a secondary signal, but as a "first-class" modality. It introduces *FVLMoE*, a force-aware Mixture-of-Experts fusion module that dynamically integrates force feedback with visual-language embeddings, achieving significant success in tasks like plug insertion. *Tactile-VLA* **?** unlocks the physical knowledge embedded in VLMs. By connecting the VLM to a hybrid position-force controller and a tactile reasoning module, it enables the robot to adapt its strategy based on tactile feedback, achieving zero-shot generalization in contact-rich scenarios. *VLA-Touch* **?** adopts a dual-level approach. Dolgopolyi et al. (2025) **?** further advance this by bridging visual and tactile modalities into a unified representation space. A tactile-language model provides semantic feedback for high-level planning, while a diffusion-based controller refines actions based on tactile signals.

Zhang et al. (2026) introduced the *CompliantVLA-adaptor* **?**, which represents a step towards safety and compliance. This model interprets task contexts to adapt the stiffness and damping parameters of a variable impedance controller, ensuring safe interaction. Finally, *TA-VLA* **?** systematically elucidates the design space for torque-aware architectures, finding that predicting torque as an auxiliary output—rather than just using it as input—encourages the model to build a physically grounded internal representation of interaction dynamics.

3.3 Theme II: Cognitive Reasoning and Planning

While foundational VLAs are adept at reactive control, they often lack the "System 2" reasoning capabilities required for multi-step problem solving, ambiguity resolution, and error recovery. The integration of Chain-of-Thought (CoT) reasoning into robotic policies has become a major research vector in 2025-2026.

3.3.1 Chain-of-Thought in Robotics

Zhong et al. (2026) proposed *ACoT-VLA* **?**, which formulates reasoning as a structured sequence of "coarse action intents." The model features an Explicit Action Reasoner (EAR) that proposes reference trajectories and an Implicit Action Reasoner (IAR) that extracts latent action priors, co-forming a chain of thought that guides the final policy. This aligns with *CoT-VLA* **?**, which employs visual chain-of-thought reasoning to better interpret visual cues. *GraphCoT-VLA* **?** takes a structured approach, constructing a real-time 3D Pose-Object graph. Xue et al. (2025) **?** introduce *LeVerb*, which explicitly leverages verbalization to ground complex planning steps.

Wu et al. (2026) **?** introduced a novel runtime verification mechanism with *Do What You Say*. This framework steers the VLA by verifying that the generated action sequences actually align with the model's own intermediate textual plan, effectively using the plan as a ground-truth constraint to filter hallucinated actions. *ManualVLA* **?** explicitly trains models to generate "manuals"—intermediate steps consisting of images, position prompts, and text—before executing actions. *VLA-R1* **?** integrates Reinforcement Learning from Verifiable Rewards (RLVR) to reinforce this reasoning quality, ensuring that the chain-of-thought is not just plausible but executable. *Reasoning-VLA* **?** applies this paradigm to autonomous driving, using learnable action queries initialized from ground-truth trajectories to generate continuous action paths.

3.3.2 Self-Reflection and Counterfactuals

Advanced reasoning also involves knowing what *not* to do and learning from hypothetical failures. *Counterfactual VLA* **?** enables agents to perform "self-reflective" reasoning. It generates time-segmented meta-actions and then performs counterfactual reasoning to simulate potential outcomes, identifying unsafe behaviors and correcting the plan before execution. Hsieh et al. (2025) **?** focused on the critical safety capability of "rejecting the impossible," teaching VLAs to identify and refuse unfeasible instructions.

VLA-Reasoner **?** empowers models with Monte Carlo Tree Search (MCTS) to explore future states at test time. By rolling out possible action trajectories and evaluating them with a world model, it scales compute during inference to improve decision quality. *Latent-CoT* **?** pushes this reasoning into a latent space, interleaving action-proposal tokens with world-model tokens to reason about future outcomes without the computational overhead of decoding natural language. *d-VLA* **?** integrates a multimodal chain-of-thought directly into a diffusion objective, optimizing perception, reasoning, and action jointly.

Recent architectures have further formalized this cognitive cycle. Yin et al. (2025) **?** enhanced the reasoning depth of VLAs with *DeepThinkVLA*, utilizing a recursive module for iterative state evaluation in long-horizon tasks. In parallel, Zhong et al. (2025) **?** introduced *FlowVLA*, which grounds Chain-of-Thought reasoning directly into visual motion flows, effectively bridging the gap between abstract semantic planning and continuous control trajectories. Song et al. (2025) **?** complemented this with *RationalVLA*, a dual-system architecture that generates explicit rationales for action selection, enabling robust error recovery and interpretability.

3.4 Theme III: World Models as Predictive Simulators

The integration of World Models—systems that predict future sensory data conditioned on actions—has fundamentally changed how agents learn and plan. They serve as internal simulators that allow agents to "experience" the future without acting in the real world.

3.4.1 Video Generation for Control

Generative video models have evolved from passive media creators to active control interfaces. *DriveVLA-W0* **?** utilizes world modeling to predict future images in autonomous driving scenarios, using this dense supervision to amplify data scaling laws. *WristWorld* **?** addresses the specific data scarcity of wrist-view cameras. It acts as a 4D world model that generates temporally coherent wrist-view videos solely from anchor views, bridging the gap between abundant third-person data and the fine-grained first-person data needed for manipulation.

BridgeV2W **?** introduces "embodiment masks" to align video generation with robot kinematics. By rendering the robot's URDF into the video generation process, it ensures that the predicted futures are physically consistent with the robot's capabilities. *Dream-VL* and *Dream-VLA* **?** leverage diffusion backbones to "dream" potential futures, achieving state-of-the-art performance on benchmarks like LIBERO. *WorldVLA* **?** and *RynnVLA-002* **?** propose unified architectures where the world model and the policy are jointly trained, allowing the action model to aid visual generation and vice versa. *GigaWorld-0* **?** scales this up, serving as a massive data engine that generates diverse, controllable, and physically plausible video data for training VLA models.

3.4.2 Reinforcement Learning with World Models

World models provide a safe, resettable sandbox for Reinforcement Learning (RL), mitigating the sample inefficiency and safety risks of real-world training. *WMPO* **?** introduces World-Model-based Policy Optimization, a framework for on-policy VLA RL that uses pixel-based predictions to align "imagined" trajectories with VLA features. *VLA-RFT* **?** uses a data-driven world simulator to provide dense, trajectory-level rewards for reinforcement fine-tuning, achieving strong robustness with minimal samples.

World-Env **?** constructs a virtual environment using a video-based world simulator and a VLM-guided reflector for reward generation, enabling effective post-training without physical interaction. *IRL-VLA* **?** applies this to autonomous driving, building a reward world model via inverse reinforcement learning. *SimpleVLA-RL* **?** and *RLinf-VLA* **?** provide scalable frameworks for this paradigm, implementing efficient resource allocation strategies to handle the heavy computational load of rendering, training, and inference in parallel. *Latent Action Pretraining* **?** utilizes world modeling to learn latent action representations from unlabeled video data, enabling unsupervised skill acquisition.

3.5 World Model Integration in VLA Architectures

The integration of world models into Vision-Language-Action (VLA) architectures has emerged as a transformative paradigm in 2025-2026, shifting the focus from purely reactive policies to predictive agents capable of simulating future outcomes. This integration addresses the critical "horizon" limitation of standard imitation learning by equipping agents with an internal simulator of physical dynamics.

3.5.1 Unified World Model Architectures

Recent works have proposed unified architectures that seamlessly blend perception, prediction, and action. Bi et al. (2025) introduced *Motus* **?**, a unified latent action world model that employs a Mixture-of-Transformer (MoT) to integrate understanding, video generation, and action experts. By leveraging optical flow to learn latent actions, Motus achieves superior performance in both simulation and real-world tasks. Lillemark et al. (2026) advanced this by proposing *Flow Equivariant World Models* **?**, which unify self-motion and object motion as Lie group flows. This equivariance provides stable latent world representations over long horizons, significantly outperforming diffusion-based baselines in partially observed environments. Magne et al. (2026) scaled this concept to gaming with *NitroGen* **?**, a foundation model trained on 40,000 hours of gameplay that exhibits strong cross-game generalization.

3.5.2 Video Generation as World Simulation

Generative video models are increasingly serving as the "visual cortex" of world models. Liao et al. (2025) presented the *Genie Envisioner* **?**, a platform that maps latent representations to executable actions via flow matching, supported by a large-scale instruction-conditioned video diffusion model. Wan et al. (2025) introduced *WorldAgen* **?**, which unifies state-action prediction with test-time world model training. In the domain of flow-based representations, Guo et al. (2026) proposed *FlowDreamer* **?**, an RGB-D world model that captures motion dynamics through flow fields. Wang et al. (2023) **?** also contributed to this landscape with the *Genie* interactive world model, emphasizing the generative aspects of environment simulation.

3.5.3 Predictive Modeling and Unified Prediction

Beyond generation, world models are being optimized for explicit prediction tasks. Han et al. (2025) developed *Percept-WAM* **?**, a World-Awareness-Action Model that integrates 2D/3D perception into tokens, enabling robust spatial grounding in autonomous driving. Xu et al. (2025) introduced *WAM-Diff* **?**, a masked diffusion framework that iteratively refines discrete sequence predictions for trajectory generation. Zhang et al. (2025) proposed *UP-VLA* **?**, a unified model trained with both understanding and future prediction objectives, bridging the gap between high-level semantics and low-level physical dynamics.

The scope of world modeling has also expanded to encompass comprehensive knowledge integration. Zhang et al. (2025) **?** developed *DreamVLA*, which integrates a vast repository of world knowledge into the generative process, allowing the agent to simulate diverse physical interactions and outcomes with high fidelity. This aligns with the capabilities of *GigaBrain-0* **?**, which leverages massive-scale pre-training to power a world-model-centric VLA capable of generalist behavior.

3.5.4 Mechanistic Integration and Steering

Finally, understanding *how* these world models influence VLA behavior is crucial. Haon et al. (2025) **?** introduced a framework for mechanistic interpretability, identifying sparse semantic directions within VLA representations that causally link to action selection. This allows for direct steering of the model's behavior at inference time. Wang et al. (2026) **?** provided a mechanistic taxonomy of video generation as world models, distinguishing between state construction and dynamics modeling to advance robust world simulators.

3.6 Theme IV: Efficiency and Real-Time Inference

As VLA models scale to billions of parameters, inference latency becomes a critical bottleneck. Closed-loop robotic control often requires low-latency updates, which is challenging for massive transformer models.

3.6.1 Architectural Optimization and Pruning

To address the computational cost, *EfficientVLA* **?** introduces a training-free acceleration framework that prunes redundant layers in the language module and optimizes visual token selection. *HyperVLA* **?** proposes a hypernetwork-based architecture that generates a small, task-specific policy at inference time, drastically reducing the number of active parameters while retaining the capacity of the large model. *SpecPrune-VLA* **?** leverages the temporal redundancy of robot actions, employing action-aware speculative pruning to reduce visual tokens based on global history.

SQAP-VLA **?** synergistically combines quantization and pruning, proposing quantization-aware pruning criteria to achieve nearly 2x speedups without performance loss. *MoLe-VLA* **?** draws inspiration from the "Shallow Brain Hypothesis," using a Mixture-of-Layers strategy to dynamically skip layers based on the complexity of the current state. *AC*^{2}*-VLA* **?** introduces Action-Context-aware Adaptive Computation. Focusing on optimization dynamics, Guo et al. (2025) **?** and Zhao et al. (2025) **?** propose methods for making VLA training and inference more effective through improved loss landscapes.

3.6.2 Tokenization and Decoding Strategies

Efficiently representing the action space is equally important. *FAST* **?** introduces Frequency-space Action Sequence Tokenization, using the Discrete Cosine Transform (DCT) to compress high-frequency action sequences into fewer tokens. *VQ-VLA* **?** scales vector-quantized action tokenizers. Hancock et al. (2025) **?** systematically analyze action representations, advocating for tokenization schemes that preserve kinematic fidelity. *Spec-VLA* **?** and *CEED-VLA* **?** adapt speculative decoding and early-exit strategies from the LLM domain to VLA policies, relaxing convergence conditions for easier steps to accelerate generation. *PD-VLA* **?** integrates parallel decoding with action chunking, allowing the model to generate multiple action steps simultaneously. *NinA* **?** replaces diffusion decoders with Normalizing Flows, enabling one-shot sampling that is significantly faster than iterative denoising.

3.7 Efficient Vision-Language-Action Models

The deployment of Vision-Language-Action (VLA) models on embodied systems—ranging from mobile manipulators to humanoid robots—faces severe computational constraints. While foundation models offer strong generalization, large parameter scales can incur latencies that violate real-time control needs in dynamic manipulation. Consequently, a dedicated wave of research in 2025 has focused on "VLA efficiency," developing specialized architectures and acceleration protocols that exploit spatiotemporal redundancy in robotic tasks.

3.7.1 Efficient Architectures for Edge Deployment

Deploying VLAs on the edge requires aggressive optimization of the architecture itself. Budzianowski et al. (2025) introduced *EdgeVLA* **?**, a framework tailored for resource-constrained inference. The paper reports substantial latency reduction through non-autoregressive design choices and selective quantization. Complementing this, Wen et al. (2025)

proposed *TinyVLA* **?**, showing that compact distilled models can preserve useful manipulation ability at lower compute cost. To scale capacity without proportional inference-cost growth, Du et al. (2025) developed *HiMoE-VLA* **?**. This hierarchical Mixture-of-Experts architecture activates only a subset of experts for each token, improving compute efficiency while retaining specialization.

Addressing the need for accessible robotics, Shukor et al. (2025) **?** proposed *SmolVLA*, demonstrating that highly optimized, smaller-scale architectures can achieve competitive performance through high-quality data curation, effectively democratizing VLA deployment. This push for efficiency is mirrored in architectural innovations like *HyperVLA* **?**, which utilizes hypernetworks to generate lightweight, task-specific policies on the fly, and *MoLe-VLA* **?**, which implements a layer-skipping Mixture-of-Experts strategy to dynamically adjust computational expenditure based on task complexity.

3.7.2 Temporal Acceleration and Token Caching

Beyond static compression, researchers have leveraged the high temporal correlation inherent in robotic control loops. Liu et al. (2025) introduced *TTF-VLA* **?**, which employs Temporal Token Fusion. Recognizing that consecutive frames in high-frequency control often contain limited semantic change, TTF-VLA merges redundant visual tokens across timesteps and processes scene “deltas” more efficiently. Similarly, Xu et al. (2025) proposed *VLA-Cache* **?**, an adaptive token-caching mechanism analogous to KV-caching in LLMs but specialized for action sequences. Their reported results show meaningful latency reduction with limited task-performance degradation.

3.7.3 Lightweight Adaptation and Discrete Representations

Efficient training paradigms are as critical as inference acceleration. Goyal et al. (2025) established *VLA-0* **?**, a rigorous baseline that treats actions simply as text tokens without specialized heads. VLA-0 demonstrates that stripping away complex auxiliary losses often yields superior performance, serving as a minimalist standard for benchmarking efficiency claims. For adapting pre-trained VLMs to robotic tasks, Li et al. (2025) introduced *ControlVLA* **?**, which employs a ControlNet-style architecture with zero-initialized projection layers. This supports few-shot adaptation with reduced forgetting of VLM priors. Li et al. (2025) also proposed *HAMSTER* **?**, a hierarchical framework that separates high-level semantic planning (predicting coarse 2D paths) from low-level motor control. By offloading complex reasoning to a lower-frequency planner, HAMSTER keeps high-frequency motor control lightweight and responsive. Optimizing the action representation itself, Liang et al. (2025) explored *Discrete Diffusion VLA* **?**, which formulates action generation as a discrete diffusion process to break the autoregressive bottleneck. Liang et al. (2025) also introduced *PixelVLA* **?**, which operates directly in pixel space or discrete latent codes to bypass complex kinematic state estimators. Finally, efficient multi-modal integration has emerged as a key theme. Wei et al. (2025) presented *Audio-VLA* **?**, efficiently adding contact audio perception to improve interaction robustness, while Tan et al. (2025) developed *Interactive VLA* (RIPT-VLA) **?**, a post-training paradigm that optimizes policies using sparse binary rewards from user interactions, streamlining the adaptation loop.

Expanding the sensory horizon, Zhang et al. (2025) **?** introduced *VTLA* (Vision-Tactile-Language-Action), a multimodal framework that deeply fuses tactile feedback with visual-linguistic streams to handle occlusion and precise manipulation. Wei et al. (2025) **?** further augmented this sensory suite with *Audio-VLA*, incorporating contact-rich audio signals to improve interaction robustness in environments where visual feedback is insufficient.

3.8 Theme V: Generalization and Robustness

The ultimate goal of VLA research is "Generalist" behavior—the ability to operate robustly across diverse environments, embodiments, and tasks without extensive retraining.

3.8.1 Cross-Embodiment and Cross-Task Transfer

X-VLA **?** tackles the challenge of diverse robot morphologies by using soft prompts to adapt a single transformer backbone to multiple embodiments. *MergeVLA* **?** explores model merging techniques, attempting to combine VLA experts trained on disjoint skills into a single generalist agent by mitigating parameter interference. *Galaxea* **?** provides a massive open-world dataset and a dual-system framework (G0) to support this cross-embodiment learning. *ObjectVLA* **?** focuses on object-level generalization, leveraging vision-language data to manipulate unseen objects without requiring specific demonstrations. *MimicDreamer* **?** bridges the human-robot domain gap by aligning human videos with robot kinematics, creating a scalable source of pre-training data. *Scalable VLA Pretraining* **?** similarly leverages large-scale human hand activity videos. Yang et al. (2025) **?** push beyond standard VLA architectures to handle extreme embodiment shifts, while Ye et al. (2025) **?** focus on efficient learning strategies for rapid adaptation. In the realm of open-world interaction, Zhou et al. (2025) **?** presented *ChatVLA-2*, which leverages its dialogue capabilities to facilitate open-world embodied reasoning, allowing the agent to generalize to new tasks and embodiments through interactive feedback loops. For long-horizon tasks, Fan et al. (2025) **?** introduced *Long-VLA*, specifically designed to maintain coherent behavior across extended manipulation sequences. Complementing this, *Interleave-VLA* **?** enhances instruction following by processing interleaved image-text prompts, bridging the gap between human demonstration formats and robotic execution.

3.8.2 Robustness to Perturbations and Distribution Shift

Despite high benchmark scores, VLAs can be brittle in deployment. *Eva-VLA* **?** and *LIBERO-Plus* **?** provide rigorous stress tests, revealing that models often fail under modest viewpoint changes, visual distractors, or instruction perturbations. *VLA-Pilot* **?** introduces an inference-time policy steering method to improve zero-shot deployment robustness without fine-tuning. *ReconVLA* **?** addresses the "dispersed attention" problem. Bendikas et al. (2025) **?** similarly propose focusing mechanisms to stabilize attention. For deployment adaptation, Hancock et al. (2025) **?** introduce techniques for run-time calibration.

VLA^2 **?** leverages external retrieval modules to handle unseen concepts, boosting success rates on hard-generalization tasks. Pugacheva et al. (2025) **?** systematically analyzed the impact of irrelevant context in language instructions ("Bring the apple, not the sofa"), highlighting the need for robust instruction filtering. *Align-Then-Steer* **?** introduces a framework to adapt VLAs to new domains by aligning action spaces via a unified latent space. *FPC-VLA* **?** integrates a supervisor for failure prediction and correction, enhancing reliability in open-ended tasks.

3.9 Vision-Language-Action Models for Autonomous Driving

The paradigm of Vision-Language-Action (VLA) models has recently permeated the domain of autonomous driving, marking a departure from traditional modular stacks of perception, prediction, and planning. While VLA architectures for robotic manipulation focus on dexterity and object interaction, driving-specific VLAs must contend with high-speed dynamics, multi-agent interactions, and strict safety-critical latency requirements. The literature in 2025 and 2026 reflects a concerted effort to adapt large-scale pre-trained models to these constraints, leveraging the semantic understanding of VLMs to handle long-tail corner cases and ambiguous traffic rules.

3.9.1 Benchmarks and Reasoning-Enhanced Architectures

The shift towards end-to-end learning in driving has necessitated new evaluation methodologies. Hao et al. (2025) **?** introduced the *DriveAction* benchmark, which serves as a standardized testbed for assessing VLA agents. Unlike traditional metrics focused solely on collision rates or displacement error, *DriveAction* evaluates the semantic alignment of agent behaviors with language instructions and traffic laws, highlighting the unique value proposition of VLA models in interpreting complex scene contexts.

To address the "black box" nature of end-to-end driving, recent works have integrated explicit reasoning modules. Yuan et al. (2025) **?** proposed *AutoDrive-R2*, a hierarchical VLA that employs a "reasoning-through-planning" mechanism. By explicitly generating textual explanations and strategic plans before outputting low-level control commands, *AutoDrive-R2* achieves a form of interpretable "System 2" driving, allowing for better error analysis and trust calibration. Complementing this reasoning focus, Guo et al. (2025) **?** developed *VDRive*, which combines VLA pre-training with reinforcement learning. *VDRive* specifically targets the alignment of VLA outputs with safety constraints, demonstrating that RL fine-tuning can suppress the hallucinated or unsafe behaviors sometimes exhibited by generative foundation models, thus marrying the generalization of VLAs with the robustness of control theory.

3.9.2 World Models as Driving Simulators

The scale of data required for robust autonomous driving has driven the adoption of World Models as engines for simulation and prediction. Wang et al. (2025) **?** explored the utility of generative world models in synthesizing rare and dangerous driving scenarios. Their work shows that agents trained on "dreamt" data covering adverse weather and adversarial traffic patterns generalize significantly better to real-world anomalies, effectively hallucinating safety-critical training data that is too dangerous to collect physically.

Jiang et al. (2025) **?** further advanced this domain by investigating the fidelity of world model representations. Their research, "Better Driving Understanding," posits that high-fidelity video prediction alone is insufficient for driving; instead, world models must capture the latent causal structure of the environment. By enforcing consistency in the physics and intent of traffic agents within the world model, they enable VLA agents to perform long-horizon planning with greater reliability. These contributions collectively signal a move towards "Generative Autonomy," where driving policies are distilled from foundational world models that continuously simulate and refine the agent's understanding of the road.

3.10 VLA Architecture Variants and Specializations

As the core principles of Vision-Language-Action models stabilize, research in late 2025 and early 2026 has diverged into specialized architectural variants designed to address specific limitations in spatial reasoning, cognitive depth, and domain adaptability. This section surveys these emerging variants, which move beyond the generalist paradigm to optimize for specific axes of performance.

3.10.1 Spatial and 3D Grounding

While standard VLAs operate primarily in 2D pixel space, accurate manipulation requires precise 3D understanding. Li et al. (2025) **?** introduced *Spatial VLA*, which explicitly encodes geometric constraints into the token sequence, allowing for finer control in cluttered environments. Extending this to full volumetric understanding, Bhat et al. (2025) **?** proposed diverse 3D VLA approaches that consume point cloud data directly, bypassing the information loss inherent in 2D projections. Feng et al. (2025) **?** further refined this with a spatial-aware VLA architecture that fuses depth information with semantic tokens to enhance relative positioning accuracy.

3.10.2 Reasoning and Cognitive Architectures

To bridge the gap between high-level planning and low-level control, recent works have embedded structured reasoning processes directly into the VLA backbone. Li et al. (2025) **?** developed *Chain-of-Action VLA* (CoA-VLA), which generates intermediate action-reasoning traces before committing to physical motor commands. Similarly, Song et al. (2025) **?** introduced *RationalVLA*, which explicitly models the rationale behind action selection to improve interpretability and error recovery. Yin et al. (2025) **?** pushed this further with *DeepThinkVLA*, integrating a deep reasoning module that performs iterative state evaluation for long-horizon tasks.

3.10.3 Multi-Modal and Hybrid Approaches

Hybrid architectures seek to combine the strengths of different control paradigms. Fang et al. (2025) **?** proposed *DualVLA*, a dual-stream architecture that processes high-frequency motor control separately from low-frequency semantic planning. This concept is paralleled by Liu et al. (2025) **?**, who explored a *HybridVLA* framework that fuses discrete and continuous action representations. Jin et al. (2025) **?** introduced a Dual-Actor VLA system, employing specialized actor heads for coarse and fine manipulation phases to decouple gross motion from precise interaction.

3.10.4 Domain-Specific Adaptations

Generalist models often struggle with the nuances of specific high-stakes domains, leading to the rise of task-specialized VLAs. Deng et al. (2025) **?** optimized *GraspVLA* specifically for the physics of grasping diverse objects, achieving higher stability than general-purpose baselines. In healthcare, Li et al. (2025) **?** tailored *RoboNurse-VLA* to handle the delicate interactions and safety constraints required for patient care. For aerial robotics, Lykov et al. (2025) **?** presented *CognitiveDrone*, adapting VLA principles to the 6-DOF dynamics of drone flight. In the high-speed domain, Serpiva et al. (2025) **?** developed *RaceVLA*, optimizing latency and lookahead for autonomous racing. Li et al. (2025) **?** proposed *JARVIS-VLA*, a specialized assistant architecture designed for complex, multi-stage household management tasks. For humanoid control, Ding et al. (2025) **?** introduced *Humanoid-VLA*, achieving universal full-body control with visual integration, while Jiang et al. (2025) **?** developed *WholeBodyVLA* for coordinated loco-manipulation. Chen et al. (2026) **?** extended VLA to gaming with *CombatVLA*, demonstrating efficient control in 3D action role-playing environments.

3.10.5 Interactive and Adaptive Models

The final frontier involves models that can communicate and adapt during deployment. Zhou et al. (2025) **?** pioneered *ChatVLA*, enabling users to adjust robot behavior in real-time through natural language dialogue, followed by the enhanced *ChatVLA-2* **?** which supports multi-turn context correction. Li et al. (2025) **?** introduced *SwitchVLA*, a dynamic architecture that can switch between specialist modules based on task requirements. The concept of omni-modal integration was explored by Hirose et al. (2025) **?** and Guo et al. (2025) **?** with their respective *OmniVLA* implementations, both aiming to unify audio, vision, and text into a cohesive sensorimotor policy. Jang et al. (2025) **?** focused on context-aware adaptation with *ContextVLA*, while Koo et al. (2025) **?** and Dey et al. (2025) **?** proposed *RetoVLA* and *ReVLA* respectively, emphasizing retrieval-augmented mechanisms to adapt to novel environments by referencing past experiences.

4 The Data Engine: Fueling World Models

The transition from specialized, task-specific robotic control to generalist Embodied AI has been propelled by a fundamental shift in data strategy: the move from small-scale, domain-specific datasets to massive, heterogeneous **Data Engines**. Just as Large Language Models (LLMs) rely on internet-scale text to learn the statistical structure of language, World Models for Embodied AI require a continuous, diverse stream of physical interaction data to ground their understanding of dynamics, physics, and causality. Recent comprehensive surveys **?? ????** underscore this paradigm shift. The core premise is that a robust world model must not only predict future pixels but also understand the underlying causal mechanisms **?** and physical laws governing the environment **?**.

Historically, robotic datasets were constrained by the “Moravec’s Paradox” of data collection: high-level semantic data (images, text) was abundant, but low-level physical interaction data (forces, torques, contact dynamics) was scarce. Early datasets like ImageNet provided semantic labels but no physical grounding. The robotics community responded with datasets like YCB-Video and GraspNet, which offered physical object models but lacked the diversity of real-world environments. The 2020s saw the rise of “Action-Labelled Video” (e.g., Something-Something), yet these still lacked the precise proprioceptive state information required for control. The breakthrough came with the Open X-Embodiment (OXE) initiative **?**, which aggregated data across embodiments, effectively creating the “ImageNet of Robotics.” However, OXE was largely a retrospective aggregation. The new era of “Data Engines” is characterized by *proactive*, *adversarial*, and *scalable* data generation loops.

This section systematically reviews the three critical fuel sources for these models: Passive Video Data (for learning physics and semantics), Embodied Interaction Data (for learning control and causal intervention), and Synthetic Simulation Data (for scaling and infinite diversity). We further analyze the emerging paradigm of “Adversarial Data Collection” and the trade-offs between teleoperation and wearable data acquisition, culminating in a discussion of the unified “Data Engine” loop that powers state-of-the-art models like π_0 **?** and NVIDIA Cosmos **?**.

4.1 Passive Video Datasets: The Foundation of Physical Priors

Passive video data serves as the bedrock for training foundational world models. Unlike robotic interaction data, which is expensive, slow, and dangerous to collect, passive video is abundant, diverse, and covers a vast range of real-world scenarios. Models pre-trained on this data inherit “internet-scale semantic knowledge,” enabling them to understand object properties, human interactions, and physical dynamics before they ever control a robot.

4.1.1 Scaling Trends in Multimodal Robot Pre-training

Recent work consistently reports scaling behavior with both model size and data scale, while also emphasizing that *data quality* and interaction density are critical, not just raw volume **??**. In practice, this has shifted embodied learning from a “collect more video” strategy to a “collect more physically informative video” strategy.

In this survey, we therefore treat scaling observations as directional evidence: larger and more diverse corpora generally help, but curation quality, temporal consistency, and action relevance strongly affect downstream control performance.

4.1.2 Physicalization of Internet Data

The core challenge with utilizing internet video for robotic learning is that it is often “disembodied”—curated for human consumption (e.g., vlogs, tutorials) rather than for robotic control. The **Physicalization** of internet data involves algorithmic filtering and annotation to make raw video streams useful for training world models.

Cosmos Filtration Metrics NVIDIA’s Cosmos pipeline implements a rigorous three-stage filtration process to ensure that only physically meaningful data enters the training stream, rejecting content that lacks clear causal dynamics:

1. **Motion Saliency and Optical Flow**: The system computes the dense optical flow field $\vec{v}(x, y, t)$ for every frame using a lightweight flow network. A clip is retained only if the mean flow magnitude $|\vec{v}|$ exceeds a dynamic threshold δ in regions of interest. This step aggressively filters out "talking heads," static slide presentations, and screen recordings, ensuring the model focuses on object motion and deformation. The threshold δ is adaptive, scaling with the resolution and frame rate of the source video to maintain physical consistency.
2. **Contact Probability Estimation**: A specialized interaction detector network, pre-trained on hand-object interaction datasets (e.g., 100DOH), predicts the probability $P(\text{contact})$ between agent effectors (hands, tools) and environmental objects. Only clips where $P > 0.85$ are retained. This focuses the model’s capacity on the mechanics of manipulation—grasping, pushing, cutting, and assembly—rather than passive observation

of scenery. This filter is crucial for learning the *contact dynamics* that are often glossed over in visual representations.

3. **Viewpoint Stability and Ego-Motion**: Using RANSAC-based background matching and simultaneous localization and mapping (SLAM) techniques, the pipeline estimates camera ego-motion. It filters out video with erratic, non-physical camera cuts or heavy video editing effects that violate the "single-world" temporal coherence assumption essential for learning consistent physics. Stable ego-motion allows the world model to disentangle camera movement from object movement, a prerequisite for accurate state estimation.

Overall, the key message from this line of work is that filtration is not a minor preprocessing step; it is a core part of the training signal design for embodied world models.

4.1.3 Bridging the View Gap: Egocentric Vision and Embodiment Masks

The domain gap between third-person internet video (e.g., YouTube cooking tutorials) and the first-person view of a robot is a primary hurdle. Third-person video provides global context but obscures fine-grained manipulation details due to occlusion and distance. Egocentric video, captured from head-mounted cameras (e.g., GoPro, Aria glasses), bridges this gap. Datasets like **Ego4D ?** (3,670 hours) and **EgoVid-5M ?** provide massive repositories of egocentric human behavior. These datasets capture hand-object interactions from the perspective of the agent, providing a direct mapping to the visual inputs of a humanoid or manipulator. The tasks covered range from household chores (cooking, cleaning) to skilled labor (carpentry, gardening), offering a rich source of procedural knowledge.

Furthermore, **BridgeV2W ?** demonstrates the power of bridging video generation models to embodied control using **Embodiment Masks**. By aligning coordinate-space actions with pixel-space videos from internet sources, BridgeV2W effectively "hallucinates" the robot's embodiment into the video. The process involves rendering a kinematic model of the robot (e.g., a URDF of a Franka arm) over the video frames, aligned with the estimated 3D pose of the human hand. This turns passive observation into "pseudo-active" data, allowing world models to learn proprioceptive feedback loops from video data that originally lacked any robot presence. This technique effectively multiplies the available embodied training data by orders of magnitude.

4.2 Language-Guided and Open-Ended Environments

Beyond physical manipulation, the "Data Engine" must also encompass high-level semantic reasoning and long-horizon planning. Open-ended environments provide a unique source of data for training agents to reason about abstract goals.

4.2.1 The Minecraft Crucible: Voyager and MineDojo

Minecraft has emerged as a fertile ground for training open-ended agents due to its infinite procedural generation and rich semantic hierarchy. **MineDojo ?** aggregates internet-scale knowledge (YouTube videos, Wiki pages, Reddit threads) to train agents capable of thousands of tasks. It provides a massive database of time-aligned video and text, allowing agents to learn the correlation between language instructions (e.g., "build a nether portal") and complex sequences of actions. Building on this, **Voyager ?** utilizes GPT-4 to iteratively write and refine code for an embodied agent, creating a self-improving curriculum. This approach generates a distinct type of data: *programmatic action traces* aligned with high-level reasoning. The "pixelated sheep" dream of embodied agents **?** is realized here, where the world model learns not just physics, but the *logic* of crafting and survival.

4.2.2 Instruction Following and Dialogue

Data for instruction following is critical for human-robot interaction. The **TEACh** dataset **?** and **DialFRED ?** provide rich dialogues paired with embodied tasks, allowing agents to learn from feedback. These datasets capture the back-and-forth of clarification: "Pick up the blue cup." "Do you mean the one on the left?" "Yes." This interactive data is essential for training agents that can resolve ambiguity in the real world. **LLM-Planner ?** and **Language Models as Zero-Shot Planners ?** demonstrate how to leverage the reasoning capabilities of LLMs to generate high-level plans that can be executed by low-level policies. Frameworks like **JARVIS ?** and **CoELA ?** extend this to multi-agent cooperation, generating data where agents must communicate to solve problems. These datasets are essential for training the "System 2" reasoning components of world models. Further research into multi-agent embodied systems **???** and human-agent collaboration **? ?** highlights the growing need for datasets that capture the social dynamics of interaction, not just physical manipulation.

4.3 Embodied Interaction Data: Learning Causal Intervention

While passive video teaches “what happens,” embodied interaction data teaches “how to make it happen.” This data consists of trajectories $\tau = (o_t, a_t, r_t, o_{t+1})$—observations, actions, rewards, and next observations—collected from robots interacting with the physical world. This high-cost, high-value data is the "gold standard" for training the policy head of Vision-Language-Action (VLA) models.

4.3.1 The Open X-Embodiment (OXE) Initiative and Standardization

The Open X-Embodiment (OXE) initiative **?** represents a watershed moment in robotic learning, aggregating data from 22 different robot embodiments across 34 research labs to provide over 1 million trajectories of real-world interaction. Prior to OXE, robotic datasets were fragmented, with incompatible action spaces (e.g., joint position vs. end-effector velocity) and observation formats. OXE standardized data formats using the **Reinforcement Learning Datasets (RLDS)** schema, which defines a unified protocol for serializing multimodal trajectories. This standardization allowed for the training of generalist policies like RT-X and π_0 **?**. By learning a “Universal Action Space” across morphologies (Franka, UR5, WidowX), these models achieved a 2x improvement in zero-shot generalization compared to single-robot training. Recent extensions like *Interleave-VLA* **?** have further enriched this data by re-annotating 210,000 episodes with interleaved image-text instructions, enabling fine-grained language control.

4.3.2 Distributed Data Collection: DROID

To go beyond aggregated historical data, the **Distributed Robot Interaction Dataset (DROID) ?** established a massive, consistent dataset for single-arm manipulation. Unlike OXE, which aggregated disparate legacy datasets, DROID was collected proactively using a standardized protocol across multiple institutions.

DROID Collection Protocol To ensure consistency, DROID emphasizes common hardware/protocol templates across sites:

- **Embodiment consistency**: standardized arm-centric collection setup to reduce embodiment-induced variance.
- **Multi-view sensing**: synchronized camera views and calibration-aware recording pipelines.
- **In-the-wild diversity**: broad environment and task diversity to improve out-of-distribution robustness.

This standardization minimizes the "embodiment gap" during training, allowing models to focus purely on learning environmental dynamics rather than compensating for sensor calibration errors. DROID also emphasizes “diversity” in object instances, ensuring that the model learns robust visual representations that generalize to unseen objects.

4.3.3 High-Intensity Humanoid Data: AgiBot and WholeBodyVLA

For humanoid robots, the data requirements scale with the complexity of the morphology. AgiBot pioneered high-intensity data collection for bimanual humanoids. *WholeBodyVLA* **?** was trained on 10,000 hours of data from the AgiBot X2 humanoid, focusing on whole-body control.

Hardware Shadowing and Retargeting AgiBot utilizes an “Efficient Human Shadowing” system where a human wearer’s motion (captured via IMUs and vision) is retargeted to the humanoid’s 40-DOF kinematics in real-time. This captures rich loco-manipulation knowledge—such as advancing, turning, and squatting while grasping—that is inherently missing from stationary arm datasets. The system uses a whole-body inverse kinematics (IK) solver to map the human’s task-space actions to the robot’s joint-space configuration, ensuring the generated motions are feasible and stable. Similarly, *LingBot-VLA* **?** expanded this to 20,000 hours across 9 dual-arm configurations, targeting the “pragmatic” coordination gap. The **Galaxea** dataset **?** further extends this by providing open-world data for mobile manipulation, pairing precise subtask-level language annotations with consistent robotic embodiment to facilitate hierarchical planning.

4.3.4 Specialized Data for Advanced VLA Architectures

As VLA architectures evolve, so do their data requirements. **PointVLA ?** introduces 3D point cloud data into the VLA training loop, addressing the limitations of 2D vision in handling depth ambiguity. Point clouds provide explicit geometry, crucial for grasping irregular objects. **DynamicVLA ?** focuses on dynamic object manipulation, collecting data on moving targets to train predictive control policies. This dataset includes scenarios like catching thrown objects or manipulating tools on a moving conveyor belt, challenging the temporal resolution of standard VLAs. **CompliantVLA**

? incorporates force/torque feedback data, enabling the learning of variable impedance control for contact-rich tasks. By recording the interaction forces during successful demonstrations (e.g., wiping a surface, inserting a peg), the model learns to modulate its stiffness, achieving safer and more robust manipulation. Furthermore, research into the architecture itself, such as **VLM4VLA ?** and **ACoT-VLA ?**, highlights the need for data that supports "Action Chain-of-Thought" reasoning, where intermediate rationales are annotated alongside the final action.

4.3.5 Quantifying Data Diversity and Coverage

A critical unresolved challenge is defining a rigorous metric for "data diversity." Mere volume (terabytes) does not equate to information content. Recent works have proposed **Task-Space Entropy** as a diversity metric, estimating the coverage of the state-action space covered by a dataset. For instance, the **Galaxea** dataset utilizes a hierarchical taxonomy of 500+ atomic skills (e.g., "grasp-mug," "open-fridge") to ensure uniform distribution across semantic categories. In contrast, uncurated datasets often suffer from a "long-tail" distribution, where common actions (walking, pick-and-place) are overrepresented while critical safety-critical behaviors (recovery from falls, emergency stops) are rare. **Eva-VLA ?** introduces a benchmark for evaluating robustness under physical variations, implicitly measuring the diversity of the training data's domain randomization. Future Data Engines must incorporate active learning loops that explicitly maximize this diversity metric, querying for data in under-represented regions of the state space.

4.3.6 Data Efficiency and Advanced Representations

As the parameter counts of World Models escalate, the efficiency of the "Data Engine" becomes a governing constraint. A comprehensive 2026 survey by Yu et al. **?** establishes a unified taxonomy for "Efficient Vision-Language-Action Models," dissecting the pipeline into efficient model design, training, and, crucially, *efficient data collection*. Yu et al. highlight that current datasets often suffer from high redundancy and low "causal density." They advocate for active learning frameworks where the robot explicitly seeks novel interactions—a strategy that moves beyond the passive ingestion of internet scale data. This analysis suggests that the next leap in performance will come not just from more data, but from *better* data that targets the sparse regions of the agent's state-space.

This shift mirrors the maturation of data strategies in other high-stakes scientific domains. The "Consortium Model" championed by the AACR Project GENIE **?** in oncology offers a compelling blueprint for Embodied AI. Pugh et al. **?** demonstrated that aggregating standardized, clinical-grade data from disparate institutions—while maintaining rigorous metadata standards—could unlock insights unavailable to any single entity. In robotics, this "AACR approach" validates the trajectory of projects like Open X-Embodiment. It emphasizes that the true value of a dataset lies in its schema consistency and the harmonization of heterogeneous sources (e.g., aligning action spaces across Franka and UR5 robots), rather than raw byte count. The lesson from **?** is that data utility scales with standardization, not just accumulation.

Complementing these organizational strategies are technical innovations in data representation and augmentation. To robustify World Models against viewpoint shifts and sensor noise, Gadre et al. **?** propose *Continuous Scene Representations* (CSRs). Unlike traditional discrete augmentations (e.g., random crops or rotations) which introduce artifacts, CSRs encode the 3D scene as an implicit neural function. This allows the data pipeline to generate infinite, consistent views of a scene, effectively performing "geometry-aware" augmentation. For a world model, this means training on trajectories that are not just replayed, but re-rendered from novel perspectives, enforcing 3D consistency in the learned latent space. Gadre et al. **?** show that agents trained with CSRs exhibit superior generalization in navigation and manipulation tasks, as they learn the underlying continuous geometry of the world rather than overfitting to pixel-level patterns.

4.4 Synthetic Simulation Data: The Infinite Afterburner

As real-world collection hits scaling limits (cost, hardware wear, safety), synthetic data from high-fidelity simulators has become the "Data Afterburner." Simulators allow for the generation of infinite, labeled data for corner cases that are dangerous or rare in the real world.

4.4.1 ManiSkill3: GPU-Parallelized Physics

ManiSkill3 **?** represents the state-of-the-art in parallelized physical simulation, utilizing the SAPIEN engine to achieve 30,000+ FPS on a single NVIDIA H100 GPU. Unlike traditional CPU-based simulators (Gazebo, MuJoCo), ManiSkill3 runs the entire physics pipeline on the GPU, eliminating the PCIe bottleneck. This architecture enables massive parallelization: thousands of environments can run simultaneously, generating billions of interaction steps per hour. Frameworks like *RLinf-VLA* **?** utilize these simulators to achieve 97.66% success across 25 ManiSkill tasks, leveraging "hybrid fine-grained pipeline allocation" to speed up training by nearly 2x. This speed enables the training of policies via Reinforcement Learning (RL) with billions of interaction steps, a scale impossible in the physical world. ManiSkill3 also supports diverse rendering pipelines, including ray-tracing, to minimize the visual sim-to-real gap.

4.4.2 RoboCasa and Generalizable Home Tasks

RoboCasa ? addresses the semantic poverty of earlier simulators. Built on the MuJoCo engine, it provides over 100 photorealistic kitchen environments and thousands of interactable objects sourced from high-quality asset stores. It focuses on "Everyday Tasks" (e.g., loading a dishwasher, making coffee, organizing cabinets), providing the semantic diversity needed to train VLA agents for domestic deployment. The benchmark revealed that photorealistic assets are critical for Sim-to-Real transfer; models trained on lower-fidelity assets (e.g., convex hulls, simple textures) suffer from severe "vision-gap" failures during real-world execution. RoboCasa mitigates this by using generative AI to texture assets, ensuring high visual fidelity. It also procedurally generates scene layouts, ensuring that the agent cannot simply memorize the map of a single kitchen.

4.4.3 Urban and Large-Scale Environments

Moving beyond the household, **Urban Generative Intelligence (UGI) ?** provides a platform for embodied agents in city-scale environments, integrating urban knowledge graphs with simulation. This is complemented by **MultiPLY ?**, which focuses on multi-sensory object-centric data in 3D worlds, and **LEO ?**, a generalist agent trained in diverse 3D environments. **PhyScene ?** takes this further by synthesizing physically interactable 3D scenes, allowing agents to manipulate the structural elements of the environment itself. These large-scale simulators are crucial for training agents that must navigate and interact with the complex, dynamic topology of the real world **? ?**. They provide the "macro" context (navigation, social norms) that complements the "micro" context (manipulation) of RoboCasa.

4.4.4 Sim-to-Real Transfer Challenges

Despite the advances in fidelity, the Sim-to-Real gap remains a formidable barrier. The **Physical Grounding Gap ?** refers to the discrepancy between simulated contact dynamics (often simplified as rigid body impulses) and real-world compliance (deformation, friction, stiction). Recent approaches like **VLA-RFT ?** and **World-Env ?** attempt to bridge this by using real-world data to train a "Neural Simulator" or World Model, which then serves as the training environment for the policy. This hybrid approach—training in a neural simulation grounded in real data—offers a promising path to reliable transfer. By fine-tuning the simulator's physics parameters to match real-world observations (System Identification), or by learning a residual policy that corrects for simulation errors, researchers are steadily closing this gap.

4.5 Adversarial Data Collection and Failure Mining

A key innovation in recent data engines is the concept of **Adversarial Data Collection**, championed by AgiBot and the developers of $\pi^*_{0.6}$ **?**. Instead of collecting only successful trajectories, this approach explicitly seeks out failure modes and "edge cases" to robustify the world model.

The RECAP Method The RECAP (RL with Experience and Corrections via Advantage-conditioned Policies) method **?** formalizes this loop. It incorporates expert teleoperated interventions provided *during* autonomous execution. When the policy's uncertainty metric $\mathcal{U}(s)$ exceeds a threshold, control is handed over to a human operator who corrects the behavior. These correction trajectories are heavily weighted during retraining. By focusing data collection on the decision boundaries where the model is most likely to fail, RECAP effectively turns deployment failures into high-value training signals. This methodology has been shown to double task throughput and halve failure rates in complex, long-horizon tasks like laundry folding and espresso making. It represents a shift from "static dataset" to "active learning," where the model itself directs the data collection process.

Test-Time Adaptation Complementing adversarial collection is **Test-Time Adaptation** (TTA). Systems like **TT-VLA ?** and **VLA-Reasoner ?** utilize online reinforcement learning or Monte Carlo Tree Search (MCTS) during deployment. By simulating potential future outcomes using the internal world model, these agents can reject unsafe actions before execution. This "thinking fast and slow" paradigm allows the agent to adapt to novel perturbations (e.g., a person bumping into the robot) without requiring full model retraining. It essentially treats the deployment phase as a continuous data collection and learning opportunity.

Dataset	Year	Type	Scale	Key Features	Ref.
Open X-Embodiment	2023	Real	1M+ traj.	Aggregation of 22+ embodiments.	**?**
DROID	2024	Real	76k eps.	Consistent Franka hardware setup.	**?**
AgiBot-G1	2025	Real	20k+ hours	Bimanual humanoid coordination.	**?**
LingBot	2026	Real	20k hours	Dual-arm pragmatic manipulation.	**?**
Galaxea	2025	Real	Open-World	Mobile manipulation w/ subtasks.	**?**
Ego4D	2022	Video	3,670 hours	Massive egocentric human behavior.	**?**
Cosmos-Video	2025	Video	20M hours	Physicalized internet footage.	**?**
ManiSkill3	2025	Sim	Infinite	GPU-parallelized physics steps.	**?**
RoboCasa	2024	Sim	100+ Envs	Photorealistic home manipulation.	**?**
BridgeV2W	2026	Hybrid	Varies	Embodiment masks for video gen.	**?**
TEACh	2022	Dialog	3k dialogues	Instruction following w/ clarification.	**?**
MineDojo	2022	Open	Internet	Massive Minecraft tasks.	**?**

Table 2: Key Datasets and Benchmarks for Embodied AI World Models (2022–2026). The "Type" column distinguishes between Real robot data, Passive Video, Simulation, and Hybrid/Dialogue formats.

4.6 Data Governance, Privacy, and Safety

As Data Engines ingest petabytes of real-world video, governance becomes paramount. **Ego4D** implements strict privacy protocols, blurring faces and personally identifiable information (PII) to comply with GDPR and CCPA. However, for world models to understand human emotion and intent, some facial cues are necessary, creating a trade-off between privacy and performance. In the domain of humanoid robotics, “Safety Filters” are critical. Datasets like AgiBot-G1 must be scrubbed of unsafe or harmful behaviors to prevent the policy from learning dangerous actions (e.g., swinging tools near humans). The concept of “Constitutional AI” is being adapted for robotics to ensure that the data engine itself adheres to safety norms, filtering out trajectories that violate predefined constraints.

4.7 Summary: The Data Engine Flywheel

The “Data Engine” of 2026 operates as a virtuous cycle: massive video data (π_0 **?**) builds physical intuition, high-quality robot interaction teaches specific skills (DROID **?**), online corrections ($\pi^*_{0.6}$ **?**) close the performance gap, and synthetic expansion (ManiSkill3 **?**) multiplies these scenarios into millions of variations. This “flywheel” ensures that generalist agents can transition from episodic task-solvers to truly autonomous physical companions. The integration of these diverse data streams—video, real, and sim—into a unified training curriculum remains the defining challenge and opportunity for the field. Future work must address the “Physical Grounding Gap” **?** and develop more robust metrics for “Data Diversity” to ensure that this engine continues to scale effectively. The ultimate goal is a self-sustaining ecosystem where agents learn from every interaction, whether physical, simulated, or observed, continuously refining their world models in pursuit of Artificial General Intelligence.

5 Training Methodologies for World Models

The training landscape for World Models in Embodied Artificial Intelligence has undergone a seismic shift, transitioning from simple Imitation Learning (IL) on domain-specific datasets to a sophisticated, multi-stage paradigm reminiscent of Large Language Model (LLM) training pipelines. This evolution, often referred to as the “2026 Shift,” is characterized by the adoption of massive-scale self-supervised pre-training, generative dynamics modeling via Flow Matching, and rigorous post-training alignment using Verifiable Rewards (RLVR). Modern training methodologies now address the tripartite challenge of: **(1) Scalable Representation Learning**, compressing high-dimensional sensory streams (O_t) into actionable latent states; **(2) Generative Dynamics Modeling**, predicting future world states (S_{t+1}) conditioned on agent interventions; and **(3) Policy Optimization**, selecting optimal actions (A_t) through verifiable reasoning chains rather than memorized heuristics.

This section provides a comprehensive treatise on these methodologies, detailing the mathematical foundations of Flow Matching, the architectural transition from UNets to Diffusion Transformers (DiT), and the emergence of System-2 reasoning in robotic control. We structure this analysis into four primary phases: Pre-training (Representation), Generative Modeling (Dynamics), Policy Optimization (Control), and Post-Training (Alignment).

5.1 Phase I: Foundational Pre-training Objectives

Pre-training establishes the fundamental semantic and physical priors required for generalist operation. By reconciling the distinct statistical modalities of vision (continuous, high-dimensional), language (discrete, semantic), and action (causal, precise), pre-training enables Vision-Language-Action (VLA) models to inherit internet-scale knowledge while grounding it in physical reality.

5.1.1 Autoregressive Next-Token Prediction (NTP)

The dominant paradigm for early VLAs, inherited directly from LLMs, is Autoregressive Next-Token Prediction (NTP). This approach serializes the multimodal embodied experience into a unified sequence of discrete tokens. Given a trajectory history $\tau_{<t} = (o_{<t}, l, a_{<t})$, where o represents observations, l language instructions, and a actions, the model parameters θ are optimized to maximize the conditional log-likelihood of the next token x_t. The general objective is defined as:

$$\mathcal{L}_{\text{NTP}}(\theta) = -\mathbb{E}_{\tau \sim \mathcal{D}} \left[\sum_{t=1}^{T} \log P_\theta(x_t | x_{<t}, \mathcal{C}) \right] \tag{4}$$

where $\mathcal{C}$ represents context or conditioning information. While effective for semantic generalization, NTP suffers from distinct limitations in the embodied domain:

1. **Discretization Error**: Continuous action spaces must be quantized, typically into 256 or 1024 bins, leading to precision loss in high-frequency control tasks. This "quantization noise" can destabilize contact-rich manipulation.
2. **Error Accumulation**: The autoregressive nature leads to compounding errors during long-horizon rollouts, a phenomenon known as the "delusion" problem. If the model samples a slightly off-distribution action at t, the input at $t + 1$ shifts further from the training distribution, leading to catastrophic divergence.
3. **Modality Mismatch**: Forcing continuous visual dynamics into discrete token sequences often results in inefficient utilization of model capacity. Vision tokens (e.g., from VQ-GAN) often capture high-frequency textures irrelevant to control, while missing subtle geometric cues.

Case Study: Early VLA Data Mixture Early large-scale VLA training pipelines exemplified the NTP approach by mixing web-scale vision-language data with robot interaction data **?**. A common idea is co-training/co-fine-tuning with an explicit balance term λ:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{robot}} + \lambda \mathcal{L}_{\text{web}} \tag{5}$$

This formulation captures the key tradeoff between retaining broad semantic priors and preserving action-valid control behavior.

5.1.2 Masked Auto-Encoding and In-Context Learning

Beyond NTP, Masked Auto-Encoders (MAE) have proven effective for visual representation learning. In the embodied context, VideoMAEs are trained to reconstruct masked patches of video trajectories. The loss function operates in pixel space or latent space:

$$\mathcal{L}_{\text{MAE}} = \sum_{i \in \mathcal{M}} ||\text{Patch}_i - \text{Reconstruct}(\text{VisiblePatches})||^2 \tag{6}$$

where $\mathcal{M}$ is the set of masked indices. This forces the model to learn spatiotemporal continuity and object permanence. When combined with language conditioning, this forms the basis of "In-Context" robot learning, where a model can infer a task from a few demonstration frames provided in the prompt context, without weight updates.

5.1.3 Joint Embedding and Latent Dynamics (JEPA)

To address the inefficiencies of pixel-space prediction, Joint Embedding Predictive Architectures (JEPA) have emerged as a powerful alternative. Unlike generative models that reconstruct input pixels, JEPA-based world models, such as DreamVLA **?**, optimize a feature-prediction objective in latent space. The objective is to minimize the distance between the predicted representation of a future state and the actual representation computed by a target encoder:

$$\mathcal{L}_{\text{JEPA}}(\theta, \phi) = \mathbb{E}_x \left[\|\text{sg}(E_\phi(x_{t+k})) - P_\theta(E_\phi(x_t), a_{t:t+k})\|_2^2 \right] \tag{7}$$

where E_ϕ is the encoder (often updated via Exponential Moving Average, sg denotes stop-gradient), and P_θ is the predictor. This formulation forces the model to capture semantic dynamics—such as object permanence and contact events—while discarding high-frequency visual noise that is irrelevant to control. PointVLA **?** and 4D-VLA **?** extend this to 3D representations, using spatiotemporal masking to align visual features with geometric structure. The key advantage is sample efficiency: JEPA models can learn robust dynamics from significantly fewer frames than pixel-reconstruction models, as they do not waste capacity on modeling stochastic textures.

5.2 Phase II: Generative Action and Dynamics (Flow Matching)

The most significant development in 2025-2026 has been the transition from Denoising Diffusion Probabilistic Models (DDPM) to Flow Matching (FM) for action generation and dynamics modeling. This shift addresses the critical latency and stability bottlenecks inherent in stochastic diffusion processes.

5.2.1 Diffusion Policies and Score Matching

Prior to the adoption of Flow Matching, Diffusion Policies represented the state-of-the-art. These models treat action generation as a conditional denoising process. The forward process $q(x_t|x_0)$ adds Gaussian noise to the data, while the reverse process $p_\theta(x_{t-1}|x_t)$ learns to recover the clean signal. The standard objective, often referred to as "supervision by prediction," minimizes the noise prediction error. The forward diffusion process is typically described by a Stochastic Differential Equation (SDE):

$$dx = f(x, t)dt + g(t)dw \tag{8}$$

where w is a standard Wiener process. The reverse-time SDE, which generates samples from the data distribution, is given by:

$$dx = [f(x, t) - g^2(t)\nabla_x \log p_t(x)]dt + g(t)d\bar{w} \tag{9}$$

The term $\nabla_x \log p_t(x)$ is the score function. A neural network $s_\theta(x, t)$ is trained to approximate this score via Denoising Score Matching:

$$\mathcal{L}_{\text{DSM}} = \mathbb{E}_{x_0, \epsilon \sim \mathcal{N}(0,I), t \sim \mathcal{U}(0,1)} \left[\|\epsilon - \epsilon_\theta(x_t, t, \mathcal{C})\|^2 \right] \tag{10}$$

where $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$. For continuous control, the Denoising Diffusion Implicit Model (DDIM) formulation allows for deterministic sampling by setting the stochastic noise term to zero during generation. However, inference remains computationally expensive, often requiring dozens of denoising steps (NFE: Number of Function Evaluations) to produce a clean action chunk. NVIDIA Cosmos **?** and GigaWorld **?** utilized this framework to model complex multi-modal distributions, effectively capturing the multi-modality of human behavior (e.g., the multiple valid ways to grasp a mug). To improve inference speed, techniques like Consistency Distillation (CD) are employed, where a student model learns to map any point on the trajectory directly to the origin x_0, theoretically enabling 1-step generation, though often at the cost of mode-coverage diversity.

5.2.2 Independent Conditional Flow Matching (ICFM)

Flow Matching (FM) simplifies the generative process by regressing a vector field that deterministically transports a prior distribution p_0 (e.g., standard Gaussian) to the data distribution p_1 via an Ordinary Differential Equation (ODE). The core innovation in models like π_0 **?** is the use of **Independent Conditional Flow Matching (ICFM)**.

Unlike Diffusion, which essentially simulates a stochastic process that can be curvy and chaotic, ICFM constructs a probability path $p_t(x)$ by marginalizing over conditional paths defined between specific data points $x_1 \sim p_{\text{data}}$ and noise samples $x_0 \sim p_{\text{prior}}$. The conditional flow $u_t(x|x_1)$ is defined to follow a straight path between noise and data:

$$x_t = tx_1 + (1 - (1 - \sigma_{\min})t)x_0 \tag{11}$$

Here, $t \in [0, 1]$ acts as the interpolation parameter. Differentiating with respect to time t yields the conditional vector field, which is simply the velocity of the particle moving along this straight line:

$$u_t(x|x_1) = \frac{x_1 - (1 - \sigma_{\min})x_0}{1 - (1 - \sigma_{\min})t} \tag{12}$$

The Flow Matching objective $\mathcal{L}_{\text{CFM}}$ trains a neural network $v_\theta(x, t)$ to approximate this vector field by minimizing the expected mean squared error over the path:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],x_0\sim p_0,x_1\sim p_1}\left[\|v_\theta(x_t, t, \mathcal{C}) - u_t(x_t|x_1)\|^2\right] \tag{13}$$

Crucially, because the target vector field $u_t(x|x_1)$ is linear and available in closed form, training is significantly more stable than score matching, which requires estimating the gradient of the log-density (a quantity that can be singular at low noise levels).

5.2.3 Comparison: ODE vs. SDE for Control

The deterministic nature of the Flow Matching ODE offers distinct advantages for control:

- **Consistency**: Given the same starting noise x_0, the generated trajectory is always identical. This is crucial for verifying safety and stability.
- **Efficiency**: ODE solvers can take larger steps than SDE solvers. π_0 typically uses 10-step Euler integration, achieving 50Hz control, whereas Diffusion often requires 50-100 steps.
- **Optimal Transport**: The straight paths learned by ICFM approximate the Optimal Transport plan, minimizing the "kinetic energy" of the transformation, which correlates with easier learning dynamics.

5.2.4 Optimal Transport and Rectified Flows

To further straighten the paths and reduce the transport cost, **Optimal Transport (OT) Conditional Flow Matching** couples x_0 and x_1 such that the total displacement is minimized. Instead of pairing random noise with random data, min-batch OT pairs samples (x_0^i, x_1^j) to minimize $\sum \|x_0^i - x_1^j\|^2$. This results in trajectories that do not cross, facilitating easier learning for the network. Rectified Flow is a special case of this, where the model is reflowed: the generated data from a trained model is used as the x_0 for a second stage of training, recursively straightening the flow. π_0 utilizes a variation of these techniques to ensure that the learned vector field is smooth, allowing for large step sizes during inference.

During inference, actions are generated by solving the ODE:

$$\frac{dx_t}{dt} = v_\theta(x_t, t, \mathcal{C}) \tag{14}$$

using numerical integrators like Euler (1 step) or Heun (2 steps). This formulation allows π_0 to generate high-frequency, dexterous control signals (up to 50Hz) with fewer function evaluations than comparable diffusion policies, which is critical for real-time reactivity in dynamic environments.

5.3 Phase III: Architecture Evolution

The shift in generative objectives has been paralleled by an architectural evolution. While earlier diffusion policies relied on 1D UNets with FiLM conditioning, the field has coalesced around the **Diffusion Transformer (DiT)** as a scalable backbone for both vision and action generation.

5.3.1 Scalability of DiT

The DiT architecture treats all inputs—noisy action chunks, observation history, and language embeddings—as a unified sequence of tokens. This allows the model to leverage the efficient attention mechanisms of Transformers while maintaining the generative capabilities of diffusion/flow models. Unlike UNets, which have rigid downsampling/upsampling structures, DiTs scale predictably with compute and data. The core operation is the self-attention mechanism, which allows every token to attend to every other token:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{15}$$

In the context of π_0, the VLA backbone (e.g., PaliGemma or SigLIP) acts as a massive conditional encoder, feeding semantic features into a lightweight Flow Matching head implemented as a DiT. This decoupling allows the heavy "thinking" components (vision and language processing) to scale independently of the high-frequency "acting" components. The DiT blocks typically incorporate Adaptive Layer Norm (adaLN) to inject the timestep t and context embeddings c directly into the normalization layers, modulating the features based on the noise level.

5.3.2 Action Tokenization: FAST and VQ-VAE

A critical and often overlooked component of training is action tokenization. Standard approaches typically bin continuous actions into discrete integers, which destroys high-frequency information. **Frequency-space Action Sequence Tokenization (FAST) ?** addresses this by applying a Discrete Cosine Transform (DCT) to action chunks before quantization. The DCT concentrates the signal energy into low-frequency components:

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right] \tag{16}$$

This transformation allows the model to prioritize the structural aspects of motion (low frequencies) while selectively preserving the high-frequency details required for precise impacts and contact dynamics. The resulting coefficients are then quantized. FAST demonstrates that this domain-specific tokenization allows autoregressive models to match the precision of diffusion models on tasks like peg-in-hole insertion. Alternative approaches like **Residual VQ-VAE** (RVQ) use a hierarchy of codebooks to quantize continuous vectors with increasing precision. However, these often suffer from codebook collapse. Techniques like Finite Scalar Quantization (FSQ) alleviate this by projecting vectors into a fixed grid, removing the need for a learned codebook entirely and simplifying training stability. Recent advancements have further refined discrete action spaces. **VQ-VLA ?** introduces vector-quantized action tokenizers that learn a compact codebook of motion primitives. Complementing this, **NinA ?** leverages normalizing flows to map complex action distributions into a latent Gaussian space, enabling VLA models to output continuous, multi-modal control signals without the precision loss inherent in binning.

5.3.3 Training Efficiency and Architectural Compression

As VLA models scale to billions of parameters, training efficiency becomes paramount. **EfficientVLA ?** proposes a training-free acceleration framework that identifies and prunes redundant tokens in the vision encoder during inference, achieving a 2.5x speedup with negligible performance degradation. On the training side, **MoLe-VLA ?** introduces a layer-skipping mechanism that dynamically routes computation through essential transformer blocks based on the complexity of the current observation, effectively reducing the FLOPs required for learning simple pick-and-place behaviors while retaining capacity for dexterous manipulation.

5.4 Phase IV: Policy Optimization (RLVR)

While Imitation Learning (IL) provides a strong initialization, it is fundamentally limited by the quality of the demonstration data and the "distribution shift" problem. To surpass demonstration quality, many pipelines transition to Reinforcement Learning (RL). A representative trend is **Reinforcement Learning with Verifiable Rewards (RLVR)**.

5.4.1 The Necessity of Verification

In traditional RL, reward functions are often dense, shaped, and difficult to specify (e.g., distance to target + orientation error + torque penalty). These shaped rewards often lead to "reward hacking," where the agent exploits the function without solving the task. RLVR relies on sparse, binary outcome signals that can be programmatically verified (e.g., "is the block in the bin?", "is the door open?"). This aligns the optimization landscape with the true task objective rather than a proxy. The objective function maximizes the expected sum of verifiable rewards:

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T} \gamma^t R_{\text{verify}}(s_t) \right] \tag{17}$$

5.4.2 Group Relative Policy Optimization (GRPO)

Applying RLVR to large VLA models introduces significant computational challenges. Standard Proximal Policy Optimization (PPO) requires maintaining a value network (Critic) and computing per-token generalized advantage estimates (GAE). This doubles the memory footprint. **Group Relative Policy Optimization (GRPO)**, adapted from the reasoning domain (e.g., DeepSeek-R1), eliminates the need for a critic network. Instead, it samples a group of outputs $\{o_1, o_2, ..., o_G\}$ for the same input and computes the advantage of each output relative to the group average:

$$A_i = \frac{R_i - \text{mean}(\{R_1, ..., R_G\})}{\text{std}(\{R_1, ..., R_G\}) + \epsilon} \tag{18}$$

The policy is then updated to maximize the surrogate objective:

$$\mathcal{L}_{\text{GRPO}} = \frac{1}{G} \sum_{i=1}^{G} \min \left(\frac{\pi_\theta(o_i|q)}{\pi_{\text{old}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_\theta(o_i|q)}{\pi_{\text{old}}(o_i|q)}, 1-\epsilon, 1+\epsilon \right) A_i \right) - \beta D_{KL}(\pi_\theta || \pi_{\text{ref}}) \tag{19}$$

For robotics, this allows for efficient, parallelized exploration in simulation or diverse real-world batches, as seen in VLA-R1 **?** and VLA-RL **?**. The absence of a critic network makes GRPO particularly suitable for fine-tuning massive VLA backbones where memory is the primary constraint.

5.4.3 Outcome vs. Process Supervision

In reasoning LLMs, a distinction is made between Outcome-Supervised Reward Models (ORMs) and Process-Supervised Reward Models (PRMs). In robotics, this analogy holds:

- **ORM**: Reward is given only at the end of the episode (Success/Fail). This is standard RLVR.
- **PRM**: Reward is given at key checkpoints (e.g., "grasped object", "lifted object", "aligned with bin"). PRMs provide denser signal but require more sophisticated verification logic (e.g., using vision-based success detectors or tactile feedback). Expanding the scope of reward modeling, **ReWorld ?** moves beyond scalar rewards to multi-dimensional reward modeling. By decomposing the reward signal into components—safety, efficiency, and task success—ReWorld enables the synthesis of nuanced behaviors that prioritize different objectives dynamically.

Recent work suggests that PRMs are essential for long-horizon tasks, as the credit assignment problem becomes intractable with pure ORMs.

5.5 Advanced RL Fine-tuning Strategies

Beyond the foundational objectives of RLVR, recent advancements have introduced specialized fine-tuning methodologies designed to address the unique challenges of embodied learning, specifically sample efficiency, consistency, and optimization stability.

5.5.1 Exploration and Sample Efficiency

Efficient exploration remains a primary bottleneck for online robotic learning. **?** introduces exploration-driven training paradigms that intrinsically motivate agents to visit novel state-space regions, significantly reducing the data requirements for generalization. Complementing this, **?** proposes sample-efficient methods that leverage prioritized experience replay to maximize the utility of limited real-world interactions. STARE-VLA **?** integrates structured exploration into the pre-training objective, allowing the model to hypothesize and test physical interactions. Furthermore, **?** presents on-the-fly adaptation techniques, enabling policies to adjust to novel dynamics or kinematic constraints during deployment without catastrophic forgetting. For test-time adaptation, *EVOLVE-VLA* **?** introduces environment feedback during inference, allowing policies to self-improve without retraining by leveraging world model predictions to guide exploration.

5.5.2 Consistency and Refinement

Ensuring temporal coherence in generated actions is critical for smooth control. **?** addresses the issue of improving VLA consistency by enforcing trajectory-level constraints rather than per-step optimization. **?** proposes a unified training approach that jointly optimizes for semantic understanding and kinematic precision, ensuring that high-level plans translate effectively into low-level motor commands. Similarly, **?** introduces refined VLA training protocols that iteratively distill successful trajectories back into the policy, smoothing out sub-optimal actions derived from noisy demonstrations.

5.5.3 Reward and Optimization

Novel optimization landscapes have been explored to stabilize training. **?** introduces a leave-one-out training strategy to identify and mitigate the impact of conflicting demonstrations in large-scale datasets. **?** focuses on VLA optimization techniques that balance the competing gradients from vision and action heads, preventing modality collapse. Additionally, **?** presents methods for enhancing VLA performance through auxiliary reward shaping, guiding the agent towards robust manipulation strategies even in sparse-reward environments. Addressing the brittleness of policies under perturbation, **RobustVLA ?** incorporates a robustness-aware post-training phase where the agent is trained against an adversarial dynamics model, ensuring stability across varying friction and mass distributions. To mitigate the "deadly triad" of offline RL in flow-based models, **?** introduces a **Balancing Signal** technique that adaptively weighs the contribution of conservative value estimation against the generative likelihood, preventing the policy from exploiting out-of-distribution actions.

5.5.4 Fine-tuning Strategies

General purpose fine-tuning has also evolved. **?** provides a comprehensive analysis of fine-tuning methods for foundation models in robotics, highlighting the efficacy of parameter-efficient techniques. **?** further explores learning strategies that decouple the policy's high-level reasoning from its low-level execution, allowing for modular fine-tuning of specific skill primitives.

5.6 Phase V: Alignment and Post-Training

Post-training alignment is essential to ensure safety, robustness, and adherence to user intent. This phase fine-tunes the pre-trained base model using high-quality, targeted data.

5.6.1 Direct Preference Optimization (DPO)

Direct Preference Optimization (DPO) has become the standard for aligning VLA models without the instability of RL training. Given a dataset of preference pairs (y_w, y_l) where y_w is the preferred trajectory (e.g., safer, smoother, or successful) and y_l is the dispreferred one, DPO optimizes the policy π_θ directly. It leverages an analytical mapping between the reward function and the optimal policy to bypass reward modeling:

$$r^*(x,y) = \beta \log \frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x) \tag{20}$$

Substituting this into the Bradley-Terry preference model yields the DPO loss:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right] \tag{21}$$

In the context of embodiment, preferences can be derived from human feedback (RbHF - Reinforcement Learning from Human Feedback), success detectors, or even energy efficiency metrics. NORA-1.5 **?** utilizes this to align agents towards "socially compliant" behaviors, such as avoiding sudden movements near humans or handling fragile objects with care.

5.6.2 Sim-to-Real Distillation (VLA-RFT)

A major bottleneck in robotic learning is the scarcity of real-world interaction data. **VLA-RFT (Reinforced Fine-Tuning) ?** leverages high-fidelity simulators (like Isaac Lab or ManiSkill) to generate massive amounts of successful trajectories, which are then distilled into the VLA. The process involves:

1. **Teacher Training**: Train a specialist expert π_E in simulation using RL with privileged state information (e.g., exact object poses, friction coefficients).
2. **Data Generation**: Roll out π_E to collect a dataset $\mathcal{D}_{\text{sim}}$. This dataset contains millions of successful trajectories.
3. **Student Distillation**: Fine-tune the generalist VLA π_θ on $\mathcal{D}_{\text{sim}}$ mixed with a regularization buffer of real-world data $\mathcal{D}_{\text{real}}$.

The loss function for distillation is typically a combination of behavioral cloning on the teacher's actions and a feature-matching loss to align the visual representations:

$$\mathcal{L}_{\text{distill}} = \lambda_1 \mathcal{L}_{\text{BC}}(\pi_\theta, \pi_E) + \lambda_2 \|\Phi(\text{sim\_img}) - \Phi(\text{real\_img})\|^2 \tag{22}$$

This approach effectively transfers the "reasoning" capabilities learned in simulation (where physics is consistent but appearance is synthetic) to the real world, utilizing the VLA's visual backbone to bridge the domain gap. Domain Randomization (varying textures, lighting, and physics params in sim) is crucial here to prevent the student from overfitting to simulation artifacts.

5.6.3 Interactive and Human-Centric Alignment

Beyond static preferences, **Interactive Post-Training ?** enables VLAs to query human supervisors for feedback only when model uncertainty exceeds a threshold, creating an active learning loop that maximizes data efficiency. This human-in-the-loop paradigm aligns with the **Parallels with Human Motor Learning ?** perspective, which argues that VLA post-training should mirror the stages of human skill acquisition—progressing from coarse imitation to fine-grained motor refinement through self-correction.

5.7 Phase VI: System-2 Reasoning and Test-Time Compute

The frontier of VLA training lies in enabling "System-2" thinking—deliberative, slow reasoning that occurs before action execution. This contrasts with the "System-1" fast, reactive policies learned via standard IL.

5.7.1 Chain-of-Thought (CoT) for Action

Models like ACoT-VLA **?** explicitly generate an intermediate reasoning trace z_{thought} before predicting low-level actions a_t. This effectively factorizes the policy:

$$P(a_t|o_t) = \sum_z P(a_t|z_{\text{thought}}, o_t) P(z_{\text{thought}}|o_t) \tag{23}$$

These thoughts can represent sub-goals ("grasp the handle"), spatial constraints ("avoid the obstacle on the left"), or physics predictions ("the object is heavy, lift slowly"). Training involves supervising these thought traces using data distilled from large reasoning models (e.g., GPT-4o) or generated via self-correction loops. During inference, the model can generate multiple thought chains and select the most consistent one (Self-Consistency).

5.7.2 Test-Time Scaling in Robotics

Analogous to the scaling of reasoning in LLMs (e.g., o1/R1), recent work explores scaling test-time compute for robotics. By allowing the world model to perform multiple rollouts and evaluating them with a value function or verifiable reward model (MCTS-style search), the agent can improve its success rate without changing its weights. The effective inference compute C_{infer} becomes a new scaling dimension:

$$\text{Performance} \propto f(N_{\text{param}}, D_{\text{train}}, C_{\text{infer}}) \tag{24}$$

VLA-Reasoner **?** demonstrates that allocating compute to tree-search planning significantly outperforms greedy decoding in long-horizon manipulation tasks. The model builds a search tree where nodes are states and edges are actions, using a learned World Model to expand the tree and a Value Function to prune unpromising branches. This "thinking time" allows the robot to simulate the consequences of its actions before committing to them in the real world, crucial for safety-critical applications.

5.8 Summary of Training Evolution

The trajectory of training methodologies indicates a convergence of generative modeling and reinforcement learning. The "2026 Shift" establishes a standardized recipe: pre-train on internet-scale data with NTP/JEPA, fine-tune dynamics and control with Flow Matching and DiT, and align with RLVR and DPO. This consolidated pipeline provides the robust foundation necessary for the next generation of Generalist Embodied Agents, moving from simple mimicry to verifiable, reasoned action.

6 Simulation, Digital Twins, and Synthetic Environments

The paradigm of Embodied AI is undergoing a fundamental shift from learning from static, finite datasets to learning in dynamic, infinite, and interactive environments. While real-world data remains the gold standard for fidelity, it is fundamentally constrained by the high cost of collection, safety risks associated with trial-and-error learning, and the inability to scale to long-tail scenarios **??**. As Vision-Language-Action (VLA) models scale to billions of parameters **??**, and initiatives like Open X-Embodiment **?** consolidate diverse datasets to build generalist policies **?**, their hunger for data outpaces the capability of physical robot farms. Consequently, advanced simulation platforms, high-fidelity digital twins, and procedurally generated "Infinite Worlds" have emerged as critical pillars for training generalist world models **??**. This transition mirrors the evolution in Large Language Models (LLMs), where synthetic data is increasingly used to refine reasoning; in robotics, simulation *is* the engine of synthetic experience.

The evolution of robotic simulation can be traced through three distinct generations. The first generation (e.g., Gazebo, PyBullet) focused on kinematic accuracy and rigid body dynamics but lacked visual fidelity. The second generation (e.g., AI2-THOR, Habitat **?**, iGibson) prioritized visual realism through 3D scanning and ray-tracing. We are now entering the third generation: **Neural and Generative Simulation**. This era is characterized by differentiable physics, neural rendering (NeRF/3DGS), and procedurally generated content that scales with compute **?**.

In this section, we survey the state-of-the-art in this third generation of simulation. We categorize these developments into four key pillars:

1. **Procedural Content Generation (PCG)**: Algorithms for creating unbounded, diverse environments (Section **??**).
2. **Digital Twins and Neural Rendering**: Techniques for creating photorealistic replicas of the real world with explicit physics (Section **??**).
3. **Sim-to-Real Mathematical Formalisms**: The theoretical underpinnings of Domain Randomization and System Identification (Section **??**).
4. **Neural Simulators**: The emerging use of video generation models as learned physics engines (Section **??**).

6.1 Procedural Generation and Infinite Worlds

The primary bottleneck in scaling embodied agents is not model architecture, but the scarcity of diverse 3D environments. Traditional simulators rely on manually designed assets or fixed 3D scans, which limits the agent's exposure to the "long tail" of real-world variations. The concept of "Infinite Worlds" addresses this by leveraging Procedural Content Generation (PCG) to create unbounded training environments, enabling open-ended learning **??**.

6.1.1 Algorithmic Foundations of PCG

Procedural generation in robotics draws heavily from computer graphics but imposes stricter constraints on physical plausibility and semantic consistency.

- **Wave Function Collapse (WFC)**: A constraint satisfaction algorithm that generates consistent tile-based layouts. In robotic simulation, WFC ensures that generated floorplans obey architectural rules (e.g., connectivity, accessibility), preventing dead ends that would confuse navigation agents **?**.
- **L-Systems and Shape Grammars**: Employed to synthesize diverse organic structures and architectural elements. By varying production rules, a simulator can generate infinite variations of trees or buildings without storing meshes.

- **Fractal Noise Synthesis**: Essential for terrain generation, creating realistic heightmaps for outdoor navigation benchmarks utilized by legged robots **?**.

Infinigen ? exemplifies the state-of-the-art in this domain. Unlike prior systems, Infinigen synthesizes geometry and textures from scratch using randomized mathematical rules. This "asset-free" approach provides "pixel-perfect" ground truth for auxiliary objectives (depth, optical flow, segmentation) and prevents VLA models from overfitting to specific object instances **?**.

6.1.2 Generative Asset Synthesis

Integration of Generative AI into simulation pipelines is accelerating. Models like Shap-E and Point-E allow simulators to conjure 3D assets from text descriptions on the fly:

$$\text{Asset}(c) = G_{\text{3D}}(T(c)) \tag{25}$$

where c is a semantic category and G_{3D} is a text-to-3D generator. This enables the creation of specific objects needed for instruction following tasks, such as "pick up the cat-shaped mug" **?**.

The generative capabilities underpinning these assets trace back to foundational work in generative information extraction **?**, which has evolved into sophisticated systems capable of achieving human parity in content-grounded generation **?**. These advances allow simulators to populate worlds not just with objects, but with semantically rich, interactive content.

6.1.3 City-Scale and Social Simulation

The scope of simulation is expanding to city-scale digital twins. **Urban Generative Intelligence (UGI) ?** represents a leap in this direction. UGI integrates Large Language Models (LLMs) to populate environments with reactive social agents **?**. In this framework, non-player characters (NPCs) are governed by LLM agents with explicit goals and personalities **?**, enabling robots to learn social navigation and intent inference in safety-critical urban scenarios.

6.2 Digital Twins and Neural Rendering

Digital Twins bridge the gap between procedural diversity and real-world fidelity **?**. The field is witnessing a rapid migration from mesh-based rasterization to neural rendering techniques.

6.2.1 The Rise of Neural Radiance Fields (NeRFs)

Neural Radiance Fields (NeRFs) revolutionized scene reconstruction by encoding volume density σ and color $\mathbf{c}$ in the weights of a Multi-Layer Perceptron (MLP):

$$F_{\Theta}(\mathbf{x}, \mathbf{d}) \to (\mathbf{c}, \sigma) \tag{26}$$

While NeRFs capture complex, non-Lambertian surfaces better than traditional scanning, they suffer from slow inference speeds and implicit geometry, making physics integration difficult **?**.

6.2.2 3D Gaussian Splatting (3DGS): The New Standard

3D Gaussian Splatting (3DGS) has emerged as a strong alternative for robotic simulation. 3DGS represents a scene as a set of anisotropic 3D Gaussians, offering explicit geometry for physics (collision detection via BVH) and real-time differentiable rendering.

RoboGSim ? utilizes 3DGS to enable real-time, photorealistic reconstruction of real-world environments. By leveraging differentiability, RoboGSim allows for scene editing—removing or rearranging objects by manipulating Gaussians—turning static scans into dynamic training grounds.

6.2.3 Neural Physics and Differentiable Simulation

Differentiable physics engines (e.g., Brax, DiffTaichi) allow gradients to propagate from reward signals back to policy or system parameters **?**. When combined with neural rendering, this enables end-to-end optimization of the entire pipeline, from pixels to torques **?**.

6.3 Sim-to-Real Transfer: Mathematical Formalisms

The ultimate test of simulation is the successful transfer of learned policies to the physical world, governed by the trade-off between **Domain Randomization (DR)** and **System Identification (SysID)**.

6.3.1 Domain Randomization (DR)

DR creates a distribution of simulated environments to ensure policy invariance. Formally, we optimize policy π_θ over a distribution of environments $\mathcal{E}_\phi$:

$$\max_\theta \mathbb{E}_{\phi \sim P(\Phi)} \left[\mathbb{E}_{\tau \sim \pi_\theta(\cdot|\mathcal{E}_\phi)}[R(\tau)] \right] \tag{27}$$

Automatic Domain Randomization (ADR) techniques adaptively update $P(\Phi)$ to maintain a curriculum. Recent embodied RL/VLA studies increasingly combine randomization and reward design to improve transfer **?**.

6.3.2 System Identification (SysID)

SysID optimizes simulation parameters ϕ to match real-world observations y_{real}:

$$\min_\phi \|y_{\text{sim}}(\phi, \pi) - y_{\text{real}}\|_2^2 \tag{28}$$

Recent advances allow this optimization via gradient descent, enabling world models to learn specific robot dynamics (e.g., motor backlash) directly from interaction data **?**.

6.3.3 Appearance Adaptation

To bridge the visual gap, approaches use **Domain-Invariant Feature Learning**, training perception backbones to map both domains to a common latent space using contrastive losses or masked auto-encoding, minimizing Maximum Mean Discrepancy (MMD):

$$\mathcal{L}_{\text{adapt}} = \text{MMD}(f(X_{\text{sim}}), f(X_{\text{real}})) \tag{29}$$

6.3.4 Cross-Embodiment and High-Speed Adaptation

Transferring policies to novel robot morphologies remains a distinct challenge. Research indicates that bringing foundation models like RT-1-X to new kinematics (e.g., SCARA robots) often necessitates targeted fine-tuning strategies **?**. Furthermore, closing the sim-to-real gap for dynamic tasks requires improving the speed and accuracy of predictive learning models **?** to handle real-time constraints.

6.4 Neural Simulators: The Video Generation Shift

A paradigm shift in 2025/2026 is the move to **Neural Simulators**—generative video models that predict future frames x_{t+1} given current frame x_t and action a_t:

$$x_{t+1} \sim P_\psi(x_{t+1}|x_t, a_t, c) \tag{30}$$

These models offer universal physics (modeling fluids, soft bodies) and open-set generalization **?**.

BridgeV2W ? bridges video generation to embodied control via *embodiment masks*, projecting the robot's kinematic chain into the video generator to enforce physical consistency. **DreamVLA ?** utilizes diffusion backbones to predict future outcomes for long-horizon planning, allowing agents to perform MCTS in latent space.

Challenges include temporal consistency (drift in long rollouts), controllability, and lack of ground-truth state access **?**.

Beyond pure video prediction, the integration of reasoning and control is paramount. New frameworks leverage Vision Language World Models (VLWM) to perform explicit planning with reasoning **?**. To support efficient execution, techniques like RT-Trajectory **?** utilize coarse trajectory sketches to guide generalization, while SARA-RT **?** employs self-adaptive attention to scale performance. Crucially, recent work on "knowledge insulating" architectures **?** suggests that separating high-level semantic knowledge from low-level control dynamics can significantly improve generalization and training efficiency in these advanced VLA systems.

6.5 Standardized Benchmarks and Evaluation

Rigorous benchmarks are essential for evaluating generalist agents.

LIBERO ? evaluates knowledge transfer. **CALVIN ?** benchmarks long-sequence instruction following. **SimplerEnv ?** addresses predictive evaluation. **RoboCasa ?** and **Behavior-1K ?** push semantic complexity. Minimizing the **Sim-to-Real Gap** remains the central objective:

$$\text{Gap} = \frac{\text{SR}_{\text{sim}} - \text{SR}_{\text{real}}}{\text{SR}_{\text{sim}}} \tag{31}$$

Benchmark	Primary Focus	Key Features	Notable Works
LIBERO	Lifelong Learning	Knowledge transfer across 90 tasks; tests catastrophic forgetting.	**??**
CALVIN	Long-horizon	Language sequencing and continuous control from onboard sensors.	**??**
SimplerEnv	Visual Matching	Real-to-Sim evaluation via photorealistic matching of real setups.	**??**
ManiSkill	Dexterous Manip.	GPU-parallelized physics for generalizable skills.	**? ?**
Behavior-1K	Human-level Tasks	1000+ everyday activities in realistic scenes.	**?**
RoboCasa	Household Dynamics	Generative simulation of 100+ kitchen scenes.	**?**
TEACh	Dialogue	Conversational agents for embodied tasks.	**??**
ALFRED	Instruction Following	Long-horizon tasks with state changes.	**?**

Table 3: Comparison of Key Embodied AI Benchmarks (2025–2026).

7 Challenges and Future Directions

The rapid evolution of World Models for Embodied AI has fundamentally shifted the research landscape from proving algorithmic feasibility to addressing the rigorous demands of scalability, reliability, and safety in open-world deployment. While pioneering works demonstrated that generative models could simulate simplified game environments **?** or controlled robotic tasks, the transition to unstructured, real-world operation introduces a new class of challenges. As agents move from being "passive dataset consumers" to "active world modelers," the critical bottlenecks are no longer just representational capacity, but rather the fidelity of physical reasoning, the rigorous assurance of safety under uncertainty, and the ethical implications of persistent visual sensing **??**.

This section provides a strategic analysis of these hurdles, synthesizing technical limitations with broader societal impacts. We argue that the field is currently facing a "Physical Grounding Gap," where the semantic understanding of Foundation Models (FMs) vastly outpaces their kinematic and dynamic reliability. Bridging this gap requires not just larger models, but a fundamental rethinking of how we integrate perception, reasoning, and control.

7.1 Physical Consistency and the “Dream” Problem

The most pervasive technical challenge facing current generative world models is the tension between visual fidelity and physical consistency, a phenomenon often referred to as the **“Dream” Problem**. State-of-the-art video generation models, primarily driven by diffusion architectures, demonstrate exceptional photorealism, capable of rendering complex textures, lighting, and semantic details with high fidelity. However, they frequently fail to uphold fundamental physical laws over extended time horizons, leading to simulated futures that are visually plausible but physically impossible **??**.

7.1.1 Hallucination of Physics

Unlike Large Language Models (LLMs) where “hallucination” results in factual errors or semantic drift that may still be grammatically correct, hallucination in world models manifests as violations of conservation laws and geometric constraints. Common failure modes include:

- **Object Permanence Failures:** Objects spontaneously disappearing when occluded or reappearing in geometrically impossible locations. This is particularly problematic for manipulation tasks where an agent must track objects even when they are hidden by a robotic arm **?**.
- **Rigid Body Interpenetration:** Solid objects passing through one another or the robot end-effector clipping through table surfaces, violating collision constraints. This “ghosting” effect renders the world model useless for planning contact-rich interactions.

- **Inconsistent Contact Dynamics:** Deformable objects (e.g., cloth, fluids) exhibiting “floaty” or non-Newtonian behavior that defies gravity and friction. Liquids may flow upwards, and rigid blocks may deform like jelly under pressure.

This inconsistency arises fundamentally because most current architectures optimize for pixel-level reconstruction likelihood rather than state-space physical validity **?**. The standard Diffusion Loss $\mathcal{L}_{\text{diff}}$ focuses on denoising score matching:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right] \tag{32}$$

This objective ensures that the generated image x_0 looks statistically similar to the training distribution, but it imposes no explicit penalty for violating Newton’s laws. For example, a cup floating in mid-air is a perfectly valid "image" under the distribution of internet photos (which may contain magic tricks or special effects), but it is an invalid state for a robot planning a pick-and-place task.

7.1.2 Neural Physics vs. Symbolic Constraints

Overcoming this requires a paradigm shift toward **Physics-Informed Generative Learning**. Ray et al. **?** argue that achieving true embodied intelligence requires physical reasoning capabilities that transcend statistical correlations. They propose mechanisms to explicitly model causal physical properties, ensuring that generated predictions respect mass, momentum, and friction. Similarly, Zhang et al. **?** introduce VLAC, a framework that integrates rigorous constraint satisfaction into the generative process. By enforcing "hard" constraints on the latent space dynamics, VLAC mitigates the "dream" problem, ensuring that the model’s imagination remains grounded in feasible kinematic states. This represents a broader trend towards *Neural-Symbolic World Modeling*, where models must learn to respect "hard" constraints (gravity, solidity) while maintaining the "soft" flexibility of neural representations.

7.2 Safety Alignment and Constrained Control

As embodied agents are deployed in human-centric environments, ensuring safe interaction is paramount. Traditional safety constraints in Reinforcement Learning (RL) often fail to generalize to the open world. In robotics, safety is inherently *state-dependent* and dynamic; an action that is safe in an isolated workspace (e.g., high-speed swinging) becomes hazardous in the presence of humans or fragile objects.

7.2.1 Mathematical Formulation: CMDPs

We formally frame this as a **Constrained Markov Decision Process (CMDP)**, where the objective is to maximize the expected return $J(\pi)$ subject to a safety cost limit δ. The optimization problem is defined as:

$$\max_\pi J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T}\gamma^t r(s_t, a_t)\right] \tag{33}$$

$$\text{subject to} \quad J_C(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T}\gamma^t c(s_t, a_t)\right] \leq \delta \tag{34}$$

Here, $c(s_t, a_t)$ represents a safety cost function (e.g., distance to obstacle, force magnitude, or human proximity). Solving this often involves Lagrangian relaxation, converting the constrained problem into an unconstrained dual problem:

$$\min_{\lambda\geq 0}\max_\pi \left(J(\pi) - \lambda(J_C(\pi) - \delta)\right) \tag{35}$$

where λ is the Lagrange multiplier that penalizes safety violations. The multiplier is updated iteratively via gradient ascent:

$$\lambda_{k+1} \leftarrow \max(0, \lambda_k + \alpha(J_C(\pi_k) - \delta)) \tag{36}$$

This adaptive penalty ensures that the agent prioritizes safety only when necessary, avoiding overly conservative behaviors that stifle task performance.

7.2.2 Safety Critics and Verification

In the context of World Models, we require a **Safety Critic** or **Safety World Model** capable of predicting the probability of future constraint violations $\mathcal{P}(C > \delta | s_t, a_{t:t+H})$ *before* execution. Recent works like SafeVLA **?** and RECAP **?** incorporate these critics into the sampling process. By learning a safety value function $Q_{safe}(s, a)$ alongside the task value function, these models can effectively “veto” dangerous trajectories in the latent planning space.

However, a critical challenge remains in **Distributional Shift**: a safety critic trained on standard data may fail to recognize edge-case hazards. Addressing this, Zhang et al. **?** propose RobustVLA, which explicitly optimizes for worst-case performance under adversarial perturbations, ensuring that safety guarantees hold even when environmental conditions deviate from the training distribution. Complementing this, Liu et al. **?** introduce Eva-VLA, a rigorous evaluation framework designed to stress-test VLA models under extreme variations in lighting, texture, and physics parameters. These approaches underscore the necessity of **Conservative Uncertainty Estimation**, ensuring that the model defaults to caution when ensuring physical safety in novel states. Furthermore, techniques like Reinforcement Learning from Verifiable Rewards (RLVR) are emerging to ground safety not just in human preference, but in verifiable physical metrics **???**.

7.3 Interpretability and Mechanistic Transparency

As World Models grow in complexity, their internal decision-making processes become increasingly opaque, posing a significant barrier to trust and certification. Unlike traditional control theory, where stability margins are mathematically explicit, deep generative models often function as black boxes. Haon et al. **?** pioneeringly apply **Mechanistic Interpretability** to embodied agents, attempting to decompose high-dimensional neural activations into understandable physical concepts such as "mass," "velocity," or "obstacle." By identifying the specific circuits responsible for physical reasoning, they aim to verify whether the model is genuinely learning physics or merely overfitting to visual patterns.

Building on this, Wang et al. **?** introduce **Mechanistic Steering**, a technique that allows for active intervention in the model’s latent space. By locating the direction vectors corresponding to specific physical properties (e.g., "friction coefficient"), they demonstrate the ability to "steer" the model’s behavior, correcting physical misconceptions (such as underestimating an object’s weight) before they manifest as unsafe actions. This line of research is critical for moving from "black-box" neural networks to "glass-box" systems where safety properties can be audited and guaranteed.

7.4 Compute Constraints and Real-Time Inference

The computational cost of World Models presents a severe bottleneck for deployment. Real-time robotic control typically requires control loops running at 10Hz to 50Hz (20ms-100ms latency) to maintain stability during contact-rich tasks. However, high-fidelity autoregressive video prediction or diffusion denoising can take seconds per frame on consumer hardware, creating a massive "System 1 vs System 2" gap **?**.

7.4.1 The Autoregressive Bottleneck

The high dimensionality of visual observations means that standard "Next-Token Prediction" for video involves processing millions of pixels. Even with VQ-VAE compression, the sequence length for a few seconds of video can explode, making autoregressive generation prohibitively slow for closed-loop control. This latency introduces "blind" intervals where the robot acts on outdated information, leading to instability **?**.

7.4.2 Accelerating Inference

Several strategies are emerging to address this:

- **Frequency-Domain Tokenization:** Approaches like FAST **?** move away from raw pixel space, encoding actions and dynamics in frequency domains (e.g., Discrete Cosine Transform - DCT) to compress the sequence length. By predicting the top-k spectral coefficients rather than the full trajectory, FAST reduces the token count by orders of magnitude without losing high-frequency control information.
- **Speculative Decoding for Robotics:** Inspired by LLM acceleration, Spec-VLA **?** utilizes a small, fast "draft" model to propose trajectories which are then verified in parallel by a larger, more accurate world model. The draft model $\mathcal{M}_{draft}$ runs at high frequency (e.g., 50Hz), while the verifier $\mathcal{M}_{verify}$ corrects deviations at a lower frequency (e.g., 5Hz), effectively amortizing the cost of the large model.
- **Adaptive Computation:** Frameworks like AC^2-VLA selectively activate model components based on context, allocating more compute to novel or complex situations while using lightweight paths for routine actions **?**.

- **Efficient Pruning:** Post-training recovery methods like GLUESTICK **?** restore functionality in pruned VLA models, enabling deployment on edge devices without sacrificing safety.

Despite these advances, running a multi-billion parameter VLA model like π_0 **?** onboard a mobile robot remains a systems engineering challenge, pushing the field towards **Cloud-Edge Co-Design**, where heavy world modeling occurs in the cloud while safety-critical reflex loops run locally.

7.5 Data Quality and the Supervision Deficit

As the field moves toward "scaling laws" for embodied intelligence, data *quality* has emerged as a critical bottleneck. The reliance on internet-scale video data (e.g., YouTube, Ego4D) offers vast diversity but introduces severe noise **?**. We face a "Supervision Deficit" where we have abundant pixels but scarce actions.

7.5.1 The Action Label Deficiency

Most internet videos lack proprioceptive or action labels. Inferring the precise force or torque applied by a human hand from pixels alone is an ill-posed inverse problem. A model trained on video might learn *that* a cup moves, but not the subtle impedance control required to move it without spilling. This "Outcome vs. Action" gap limits the utility of passive video for learning contact-rich skills **?**.

7.5.2 World Models as Active Data Generators

To bridge this gap, World Models are increasingly used as **Active Data Engines**. Models like GigaBrain **?**, DriveVLA-W0 **?**, and World-Env **?** use the world model to generate synthetic training data, creating diverse, counterfactual scenarios ("what-if" reasoning) to bootstrap policy learning. This "Dreamer" paradigm allows agents to practice dangerous or rare skills (e.g., catching a falling vase) without real-world risk. By hallucinating physically consistent variations of tasks, these models effectively multiply the available training data, converting limited real-world demonstrations into massive synthetic datasets.

Recent datasets like Galaxea **?** and techniques like Interleave-VLA **?** are also enhancing the semantic richness of training data by combining cross-embodiment demonstrations with interleaved image-text instructions, further mitigating the supervision deficit.

7.6 Standardized Evaluation and Benchmarking

The evaluation of World Models for Embodied AI is fraught with inconsistency. Unlike Computer Vision, where static datasets (e.g., ImageNet) suffice, embodied agents require dynamic, interactive evaluation. Wang et al. **?** propose **VLATest**, a standardized benchmarking suite designed to assess long-horizon reasoning, physical consistency, and robustness across diverse manipulation tasks. VLATest moves beyond simple success rates, introducing fine-grained metrics for kinematic fidelity and error recovery.

However, Xie et al. **?** highlight **Emerging Challenges** in evaluation, arguing that current benchmarks are often too narrow to capture the open-ended nature of real-world interaction. They identify a need for benchmarks that can evaluate "physical common sense" and the ability to handle unforeseen physical phenomena (e.g., liquid dynamics, plastic deformation). This calls for a shift towards *Procedural Metrology*, where evaluation environments are generated procedurally to test specific capabilities, rather than relying on fixed task sets.

7.7 Ethical Implications of Egocentric Perception

The deployment of agents equipped with persistent, ego-centric World Models raises significant privacy and ethical concerns. Unlike fixed security cameras, embodied agents are mobile, interactive, and often operate in private spaces (homes, offices).

7.7.1 Privacy in the Loop

A robot that builds a "World Model" of a user's home effectively creates a searchable, 3D semantic database of their private life. It knows where valuables are stored, daily routines, and social interactions. If this model is uploaded to the cloud for inference (as per Section **??**), it creates vectors for surveillance capitalism. Techniques for **Privacy-Preserving World Modeling**, such as federated learning or running semantic abstraction on-device before transmission, are essential to ensure user trust **?**.

7.7.2 Bystander Consent and Gaze

Egocentric data collection inevitably captures non-consenting bystanders. While face blurring is a standard mitigation, it introduces a technical trade-off: gaze detection is crucial for human-robot interaction (HRI) and safety. Anonymizing faces destroys the very signal needed to predict human intent. Resolving this tension—perhaps through "feature-preserving anonymization" that retains gaze vectors but masks identity—is an open ethical-technical challenge.

7.8 Sovereign Physical Intelligence

As World Models become the "operating system" of physical reality, their ownership becomes a matter of geopolitical strategy. We term this **Sovereign Physical Intelligence**. Just as nations are concerned with sovereign capability in LLMs, the ability to automate physical labor, logistics, and manufacturing via Foundation World Models is a critical economic asset.

Dependency on a few proprietary "closed" World Models for national infrastructure poses risks of lock-in and reduced resilience. If the model controlling a factory's logistics requires an API call to a foreign provider, that physical infrastructure is no longer sovereign. This drives a need for **Open-Weight Physical Models** (like the π series **??**, OpenVLA, or VLA-Adapter **?**) that allow nations and industries to fine-tune and run controls entirely on-premise, ensuring security and stability of the physical supply chain.

7.9 Future Directions: Towards Persistent Intelligence

Looking toward 2026 and beyond, we anticipate several key trajectories that will define the next generation of Embodied AI.

World Models as Active Data Engines: We expect a shift from training *on* data to training *inside* data. High-fidelity World Models will serve as infinite generators of diverse, counterfactual scenarios ("what-if" reasoning) to bootstrap policy learning. This "Dreamer" paradigm allows agents to practice dangerous or rare skills without real-world risk.

Unified Neuro-Symbolic Architectures: Future architectures will likely dissolve the boundary between the "planner" (LLM) and the "controller" (Policy). Instead of chaining separate modules, we foresee Unified Neuro-Symbolic Transformers that process tokenized text, video, and action commands in a shared latent space. These models will possess both the semantic "common sense" of System 2 thinking ("Why should I clean this?") and the precise System 1 motor control ("How do I grasp this handle?") **??**.

Lifelong Learning Strategies: Finally, the static "train-then-deploy" paradigm must evolve into **Lifelong Learning**. Embodied agents will encounter novel objects and physics daily. They must be capable of **In-Context Physical Adaptation**—updating their internal physics model on-the-fly based on prediction errors (e.g., realizing a package is heavier than it looks) without catastrophic forgetting of previous skills. This moves us from specialized tools to true autonomous physical companions, capable of growing with their environment.

8 Conclusion and Future Outlook

This survey has systematically explored the transformative landscape of World Models for Embodied AI during the pivotal 2025–2026 period. We have traced the field's rapid evolution from specialized, model-free control policies toward generalist architectures grounded in high-fidelity predictive dynamics and fueled by massive, heterogeneous data engines. Our analysis reveals that the core innovation defining this era is **Physicalization**—the transition from "dreamy" video generation to precise, action-conditional world modeling. By internalizing the laws of physics and the causal structure of interaction, world models have graduated from auxiliary components to the foundational pillars of embodied intelligence, enabling agents to reason about temporal transitions, anticipate the consequences of their interventions, and generalize across the reality gap with unprecedented robustness.

The convergence of three critical trends has enabled this progress. First, the emergence of unified **Data Engines** has solved the historical data bottleneck by synthesizing internet-scale video priors with high-intensity robot interaction data, as exemplified by initiatives like Open X-Embodiment **?** and NVIDIA Cosmos **?**. Second, the shift toward **Flow Matching and Latent-Space Predictive Architectures** has provided the mathematical maturity required for stable, high-frequency control, overcoming the limitations of stochastic diffusion and discrete autoregression. Third, the transformation of **Simulation into an Active Data Generator**—through procedural "Infinite Worlds" and photorealistic digital twins—has allowed models to learn from a distribution of experiences far broader than what is physically attainable, effectively addressing the "long-tail" scenarios essential for safe real-world deployment.

Looking ahead, we anticipate the dissolution of traditional boundaries between perception, prediction, and control. The trajectory points toward a **Unified World Model**—a single, massive neural architecture that consumes multimodal history and outputs both future rollouts and control actions within a shared, actionable latent space. This next generation of "System-2" embodied agents will likely integrate the deliberative "Chain-of-Thought" reasoning of LLMs directly into the physics prediction loop, enabling strategic planning over long horizons. As we move closer to closing the Sim-to-Real loop through Real-to-Sim-to-Real experience replay, world models will serve as the "physical conscience" of autonomous systems, paving the way for the first generation of truly capable, generalist agents that can navigate, manipulate, and master the complexity of the physical world.

9 Comparative Summary Tables

A. Algorithm-Oriented Papers

Table 4: Algorithm-oriented literature summary.

Year	Model/Method	Technical Contribution	Reference
2022	Do As I Can, Not As I Say	We propose to provide real-world grounding by means of pretrained skills, which are used to constrain the model to propose natural language actions that are both feasible	**?**
2022	Continuous scene representations for embodied ai	Not explicitly specified.	**?**
2022	Dialfred	Not explicitly specified.	**?**
2022	Inner Monologue	We propose that by leveraging environment feedback, LLMs are able to form an inner monologue that allows them to more richly process and plan in robotic control scenarios	**?**
2022	Language models as zero-shot planners	Not explicitly specified.	**?**
2022	GenIE	Not explicitly specified.	**?**
2022	AACR Project GENIE	Not explicitly specified.	**?**
2022	A Generalist Agent	Inspired by progress in large-scale language modeling, we apply a similar approach towards building a single generalist agent beyond the realm of text outputs.	**?**
2023	RoboCat	Inspired by recent advances in foundation models for vision and language, we propose a multi-embodiment, multi-task generalist agent for robotic manipulation.	**?**
2023	Edgi	Not explicitly specified.	**?**
2023	Collaborating with language models for embodied reasoning	We present a set of tasks that require reasoning, test this system's ability to generalize zero-shot and investigate failure cases, and demonstrate how components of this	**?**
2023	Can an embodied agent find your "cat-shaped mug"$\pi llm - basedzero - shot$	Not explicitly specified.	**?**
2023	PaLM-E	We propose embodied language models to directly incorporate real-world continuous sensor modalities into language models and thereby establish the link between words and	**?**
2023	RT-Trajectory	We propose a policy conditioning method using such rough trajectory sketches, which we call RT-Trajectory, that is practical, easy to specify, and allows the policy to ef	**?**
2023	Grounded Decoding	Recent progress in large language models (LLMs) has demonstrated the ability to learn and leverage Internet-scale knowledge through pre-training with autoregressive model	**?**
2023	VoxPoser	We present a large-scale study of the proposed method in both simulated and real-robot environments, showcasing the ability to perform a large variety of everyday manipul	**?**
2023	Do embodied agents dream of pixelated sheep	Not explicitly specified.	**?**
2023	Genie in the Model	In this paper, we propose to push forward the frontiers of the DNN performance-resource trade-off by introducing human intelligence as a new design dimension.	**?**
2023	Voyager	We introduce Voyager, the first LLM-powered embodied lifelong learning agent in Minecraft that continuously explores the world, acquires diverse skills, and makes novel d	**?**
2023	Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware	Can learning enable low-cost and imprecise hardware to perform these fine manipulation tasks$\pi Wepresentalow - costsystemthatperformsend - to - endimitationlearningd$	**?**
2024	Genie	Not explicitly specified.	**?**

Year	Model/Method	Technical Contribution	Reference
2024	GR-2	We present GR-2, a state-of-the-art generalist robot agent for versatile and generalizable robot manipulation.	**?**
2024	Embodied LLM Agents Learn to Cooperate in Organized Teams	Inspired by human organizations, this paper introduces a framework that imposes prompt-based organization structures on LLM agents to mitigate these problems.	**?**
2024	The Essential Role of Causality in Foundation World Models for	Recent advances in foundation models, especially in large multi-modal models and conversational agents, have ignited interest in the potential of generally capable embodi	**?**
2024	Multiply	Not explicitly specified.	**?**
2024	Sara-rt	Not explicitly specified.	**?**
2024	Out of Many, One	In this work we introduce Genie 2, extending Genie to capture a larger and more diverse protein structure space through architectural innovations and massive data augment	**?**
2024	Embodied multi-modal agent trained by an llm from a parallel	Not explicitly specified.	**?**
2024	Physcene	Not explicitly specified.	**?**
2024	Achieving Faster and More Accurate Operation of Deep Predictive Learning	Achieving both high speed and precision in robot operations is a significant challenge for social implementation.	**?**
2024	Building Cooperative Embodied Agents Modularly with Large Language Models	In this work, we address challenging multi-agent cooperation problems with decentralized control, raw sensory observations, costly communication, and multi-objective task	**?**
2025	EVOLVE-VLA	We introduce EVOLVE-VLA, a test-time training framework enabling VLAs to continuously adapt through environment interaction with minimal or zero task-specific demonstrati	**?**
2025	Focusing on What Matters	Building on the insights of object-centric representation learning, our method introduces an inductive bias towards scene objects and the agent's own visual information.	**?**
2025	Semantic World Models	Planning with world models offers a powerful paradigm for robotic control.	**?**
2025	Motus	In this paper, we propose Motus, a unified latent action world model that leverages existing general pretrained models and rich, sharable motion information.	**?**
2025	WorldVLA	We present WorldVLA, an autoregressive action world model that unifies action and image understanding and generation.	**?**
2025	ConRFT	In this paper, we propose a reinforced fine-tuning approach for VLA models, named ConRFT, which consists of offline and online fine-tuning with a unified consistency-base	**?**
2025	InternVLA-M1	We introduce InternVLA-M1, a unified framework for spatial grounding and robot control that advances instruction-following robots toward scalable, general-purpose intelli	**?**
2025	villa-X	In this paper, we introduce villa-X, a novel Vision-Language-Latent-Action (ViLLA) framework that advances latent action modeling for learning generalizable robot manipul	**?**
2025	VLA+VLBA to ngVLA Transition Option Concepts	The next-generation Very Large Array (ngVLA) is intended to be the premier centimeter-wavelength facility for astronomy and astrophysics, building on the substantial scie	**?**
2025	Revla	Not explicitly specified.	**?**
2025	Understanding World or Predicting Future$\pi AComprehensiveSurveyofWorld$	The concept of world models has garnered significant attention due to advancements in multimodal large language models such as GPT-4 and video generation models such as S	**?**
2025	Bridging Perception, Language, and Action	Not explicitly specified.	**?**
2025	VITA-VLA	In this work, we propose a simple yet effective distillation-based framework that equips VLMs with action-execution capability by transferring knowledge from pretrained s	**?**
2025	Knowledge Insulating Vision-Language-Action Models	Vision-language-action (VLA) models provide a powerful approach to training control policies for physical systems, such as robots, by combining end-to-end learning with t	**?**

Year	Model/Method	Technical Contribution	Reference
2025	SQAP-VLA	We overcome the incompatibility by co-designing the quantization and token pruning pipeline, where we propose new quantization-aware token pruning criteria that work on a	**?**
2025	Embodied AI	Not explicitly specified.	**?**
2025	Spatial-Aware VLA Pretraining through Visual-Physical Alignment from Human Videos	To bridge this gap, we propose a Spatial-Aware VLA Pretraining paradigm that performs explicit alignment between visual space and physical space during pretraining, enabl	**?**
2025	MergeVLA	To address these challenges, we present MergeVLA, a merging-oriented VLA architecture that preserves mergeability by design.	**?**
2025	Embodied AI Agents	We propose that the development of world models is central to reasoning and planning of embodied AI agents, allowing these agents to understand and predict their environm	**?**
2025	VLA-OS	To systematically investigate the impacts of different planning paradigms and representations isolating from network architectures and training data, in this paper, we in	**?**
2025	Efficient Vision-Language-Action Models for Embodied Manipulation	Vision-Language-Action (VLA) models extend vision-language models to embodied control by mapping natural-language instructions and visual observations to robot actions.	**?**
2025	OmniVLA	We present OmniVLA, an omni-modality VLA model that integrates novel sensing modalities for physically-grounded spatial intelligence beyond RGB perception.	**?**
2025	Run-time observation interventions make vision-language-action models more visually robust	Not explicitly specified.	**?**
2025	Mechanistic interpretability for steering vision-language-action models	Motivated by advances in mechanistic interpretability for large language models, we introduce the first framework for interpreting and steering VLAs via their internal re	**?**
2025	Do what$\pi Teachingvision - language - actionmodelstorejecttheimpossible$	Not explicitly specified.	**?**
2025	Joint Optimization of Fine-grained Representation and Workflow Orchestration in Metaverse	Not explicitly specified.	**?**
2025	GraphCoT-VLA	Vision-language-action models have emerged as a crucial paradigm in robotic manipulation.	**?**
2025	Tactile-VLA	This paper introduces Tactile-VLA, a novel framework that deeply fuses vision, language, action, and tactile sensing.	**?**
2025	NORA	To address the limitations of existing VLA models, we propose NORA, a 3B-parameter model designed to reduce computational overhead while maintaining strong task performan	**?**
2025	$\pi_{0.5}$	In order for robots to be useful, they must perform practically relevant tasks in the real world, outside of the lab.	**?**
2025	Don't Run with Scissors	We introduce GLUESTICK, a post-pruning recovery method that restores much of the original model's functionality while retaining sparsity benefits.	**?**
2025	ContextVLA	In this paper, we introduce ContextVLA, a policy model that robustly improves robotic task performance by effectively leveraging multi-frame observations.	**?**
2025	Don't Blind Your VLA	The growing success of Vision-Language-Action (VLA) models stems from the promise that pretrained Vision-Language Models (VLMs) can endow agents with transferable world k	**?**
2025	RetoVLA	Recent Vision-Language-Action (VLA) models demonstrate remarkable generalization in robotics but are restricted by their substantial size and computational cost, limiting	**?**
2025	Embodied Understanding of Driving Scenarios	Not explicitly specified.	**?**
2025	3ds-vla	Not explicitly specified.	**?**
2025	Coa-vla	Not explicitly specified.	**?**

Year	Model/Method	Technical Contribution	Reference
2025	ControlVLA	To achieve this, we propose ControlVLA, a novel framework that bridges pre-trained VLA models with object-centric representations via a ControlNet-style architecture for	**?**
2025	Embodied Multi-Agent Systems	Not explicitly specified.	**?**
2025	MimicDreamer	To bridge this gap, we propose MimicDreamer, a framework that turns fast, low-cost human demonstrations into robot-usable supervision by jointly aligning vision, viewpoin	**?**
2025	Robonurse-vla	Not explicitly specified.	**?**
2025	SwitchVLA	We propose SwitchVLA, a unified, execution-aware framework that enables smooth and reactive task switching without external planners or additional switch-specific data.	**?**
2025	UrbanVLA	To this end, we propose UrbanVLA, a route-conditioned Vision-Language-Action (VLA) framework designed for scalable urban navigation.	**?**
2025	VLA Models Are More Generalizable Than You Think	To address this, we propose a one-shot adaptation framework that recalibrates visual representations through lightweight, learnable updates.	**?**
2025	Large Model Empowered Embodied AI	For embodied learning, we introduce mainstream learning methodologies, elaborating on how large models enhance imitation learning and reinforcement learning in-depth.	**?**
2025	Evo-0	We evaluate our method on a set of spatially challenging tasks in both simulation and the real world.	**?**
2025	VOTE	To address these issues, we develop a training framework to finetune VLA models for generating significantly fewer action tokens with high parallelism, effectively reduci	**?**
2025	HybridVLA	To address these limitations, we introduce HybridVLA, a unified framework that absorbs the continuous nature of diffusion-based actions and the contextual reasoning of au	**?**
2025	HybridVLA	Not explicitly specified.	**?**
2025	MLA	To this end, we introduce a multisensory language-action (MLA) model that collaboratively perceives heterogeneous sensory modalities and predicts future multisensory obje	**?**
2025	TTF-VLA	We propose Temporal Token Fusion (TTF), a training-free approach that intelligently integrates historical and current visual representations to enhance VLA inference qual	**?**
2025	VLA-Mark	Not explicitly specified.	**?**
2025	VLA-RL	We present VLA-RL, an algorithmic and systematic framework that leverages online reinforcement learning (RL) to improve pretrained auto-regressive VLAs in downstream task	**?**
2025	Improving Pre-Trained Vision-Language-Action Policies with Model-Based Search	We present Vision-Language-Action Planning	
	Search (VLAPS) – a novel framework and accompanying algorithms that embed model-based search into the inference procedure o	**?**	
2025	GraSP-VLA	In this paper we present a new neuro-symbolic approach, GraSP-VLA, a framework that uses a Continuous Scene Graph representation to generate a symbolic representation of	**?**
2025	Cosmos World Foundation Model Platform for Physical AI	In this paper, we present the Cosmos World Foundation Model Platform to help developers build customized world models for their Physical AI setups.	**?**
2025	ACG	In this paper, we present Action Coherence Guidance (ACG) for VLA models, a training-free test-time guidance algorithm that improves action coherence and thereby yields p	**?**
2025	Spatial Traces	Not explicitly specified.	**?**
2025	Bring the Apple, Not the Sofa	In this work, we present a novel systematic study of the robustness of state-of-the-art VLA models under linguistic perturbations.	**?**

Year	Model/Method	Technical Contribution	Reference
2025	SpatialVLA	Specifically, we introduce Ego3D Position Encoding to inject 3D information into the input observations of the visual-language-action model, and propose Adaptive Action G	**?**
2025	Exploration-Driven Generative Interactive Environments	Not explicitly specified.	**?**
2025	Leave No Observation Behind	We introduce Asynchronous Action Chunk Correction (A2C2), which is a lightweight real-time chunk correction head that runs every control step and adds a time-aware correc	**?**
2025	VLA-R	In this work, we present Vision-Language Action Retrieval (VLA-R), an open-world end-to-end autonomous driving (OW-E2EAD) framework that integrates open-world perception	**?**
2025	CEED-VLA	To address it, we introduce consistency distillation training to predict multiple correct action tokens in each iteration, thereby achieving acceleration.	**?**
2025	PD-VLA	To tackle this problem, we propose PD-VLA, the first parallel decoding framework for VLA models integrated with action chunking.	**?**
2025	CollabVLA	In this work, we present CollabVLA, a self-reflective vision-language-action framework that transforms a standard visuomotor policy into a collaborative assistant.	**?**
2025	From Perception to Action with Integrated VLA Systems	Not explicitly specified.	**?**
2025	Evaluating Uncertainty and Quality of Visual Language Action-enabled Robots	In this paper, we propose eight uncertainty metrics and five quality metrics specifically designed for VLA models for robotic manipulation tasks.	**?**
2025	WorldAgen	Not explicitly specified.	**?**
2025	Exploring the adversarial vulnerabilities of vision-language-action models in robotics	Not explicitly specified.	**?**
2025	Genie	Not explicitly specified.	**?**
2025	Spec-vla	Not explicitly specified.	**?**
2025	SpecPrune-VLA	We introduce SpecPrune-VLA, a training-free method with two-level pruning and heuristic control: (1) Static pruning at action level: uses global history and local context	**?**
2025	VLATest	To address this gap, we present VLATest, a fuzzing framework designed to generate robotic manipulation scenes for testing VLA models.	**?**
2025	Audio-VLA	Additionally, this paper introduces the Task Completion Rate (TCR) metric to systematically evaluate dynamic operational processes.	**?**
2025	DexVLA	This paper introduces DexVLA, a novel framework designed to enhance the efficiency and generalization capabilities of VLAs for complex, long-horizon tasks across diverse	**?**
2025	DiffusionVLA	Not explicitly specified.	**?**
2025	LLaDA-VLA	In this work, we present LLaDA-VLA, the first Vision-Language-Diffusion-Action model built upon pretrained d-VLMs for robotic manipulation.	**?**
2025	Tinyvla	Not explicitly specified.	**?**
2025	Momanipvla	Not explicitly specified.	**?**
2025	VLA Model-Expert Collaboration for Bi-directional Manipulation Learning	The emergence of vision-language-action (VLA) models has given rise to foundation models for robot manipulation.	**?**
2025	HyperVLA	In this paper, we propose HyperVLA to address this problem.	**?**
2025	STARE-VLA	Thereby, we present Stage-Aware Reinforcement (STARE), a module that decomposes a long-horizon action trajectory into semantically meaningful stages and provides dense, i	**?**

Year	Model/Method	Technical Contribution	Reference
2025	VLA-Cache	This paper introduces VLA-Cache, a training-free inference acceleration method that reduces computational overhead by adaptively caching and reusing static visual tokens	**?**
2025	WAM-Diff	End-to-end autonomous driving systems based on vision-language-action (VLA) models integrate multimodal sensor inputs and language instructions to generate planning and c	**?**
2025	FPC-VLA	To address these challenges, we propose FPC-VLA, a dual-model framework that integrates VLA with a supervisor for failure prediction and correction.	**?**
2025	Balancing Signal and Variance	Vision-Language-Action (VLA) models based on flow matching have shown excellent performance in general-purpose robotic manipulation tasks.	**?**
2025	InSpire	To tackle this challenge, we propose Intrinsic Spatial Reasoning (InSpire), a simple yet effective approach that mitigates the adverse effects of spurious correlations by	**?**
2025	MoLe-VLA	We introduce a Spatial-Temporal Aware Router (STAR) for MoLe to selectively activate only parts of the layers based on the robot's current state, mimicking the brain's di	**?**
2025	RobustVLA	In this work, we introduce RobustVLA, a lightweight online RL post-training method designed to explicitly enhance the resilience of VLA models.	**?**
2025	A Step Toward World Models	Autonomous agents are increasingly expected to operate in complex, dynamic, and uncertain environments, performing tasks such as manipulation, navigation, and decision-ma	**?**
2025	Cot-vla	Not explicitly specified.	**?**
2025	MoRE	This paper introduces a novel vision-language-action (VLA) model, mixture of robotic experts (MoRE), for quadruped robots that aim to introduce reinforcement learning (RL	**?**
2025	A Survey on Vision-Language-Action Models	The remarkable advancements of vision and language foundation models in multimodal understanding, reasoning, and generation has sparked growing efforts to extend such int	**?**
2025	ChatVLA-2	In this work, we introduce ChatVLA-2, a novel mixture-of-expert VLA model coupled with a specialized two-stage training pipeline designed to preserve the VLM's original s	**?**
2025	WMPO	We introduce World-Model-based Policy Optimization (WMPO), a principled framework for on-policy VLA RL without interacting with the real environment.	**?**
2026	Flowdreamer	Not explicitly specified.	**?**
2026	PI-VLA	Not explicitly specified.	**?**
2026	Pointvla	Not explicitly specified.	**?**
2026	Reflection-Based Task Adaptation for Self-Improving VLA	We introduce Reflective Self-Adaptation, a framework for rapid, autonomous task adaptation without human intervention.	**?**
2026	Aligning Agentic World Models via Knowledgeable Experience Learning	To bridge this gap, we introduce WorldMind, a framework that autonomously constructs a symbolic World Knowledge Repository by synthesizing environmental feedback.	**?**
2026	Vision-Language-Action (VLA) Models	Our methodology adopts a rigorous literature review framework, covering over 80 VLA models published in the past three years.	**?**
2026	Learning Action-Conditioned World Models for Cataract Surgery from Unlabeled Videos	Not explicitly specified.	**?**
2026	An Efficient and Multi-Modal Navigation System with One-Step World Model	To address this bottleneck, we propose a lightweight navigation world model that adopts a one-step generation paradigm and a 3D U-Net backbone equipped with efficient spa	**?**
2026	Visual Generation Unlocks Human-Like Reasoning through Multimodal World Models	Humans construct internal world models and reason by manipulating the concepts within these models.	**?**

Year	Model/Method	Technical Contribution	Reference
2026	Do What You Say	Reasoning Vision Language Action (VLA) models improve robotic instruction-following by generating step-by-step textual plans before low-level actions, an approach inspire	**?**
2026	ACoT-VLA	We introduce Action Chain-of-Thought (ACoT), a paradigm where the reasoning process itself is formulated as a structured sequence of coarse action intents that guide the	**?**
2026	Digital Twin AI	Digital twins, as precise digital representations of physical systems, have evolved from passive simulation tools into intelligent and autonomous entities through the int	**?**
–	Goal-VLA	Not explicitly specified.	**?**
–	Emerging Paradigms in Deep Learning	Not explicitly specified.	**?**
–	VLAC	Not explicitly specified.	**?**

B. Data-Oriented Papers

Table 5: Data-oriented literature summary.

Year	Dataset/Framework	Data Modality	Task Focus	Source Type	Scale	Reference
2020	Rearrangement	Vision + Language	General embodied tasks	Simulated/Generated	Not explicitly specified	**?**
2022	A survey of embodied ai	Not explicitly specified	General embodied tasks	Simulated/Generated	Not explicitly specified	**?**
2022	MineDojo	Vision + Language	Navigation/interaction	Simulated/Generated	Not explicitly specified	**?**
2023	RT-1	Vision + Language	General embodied tasks	Real-world	Not explicitly specified	**?**
2023	Q-Transformer	Vision + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Real-world	Not explicitly specified	**?**
2023	VIMA	Vision + Language + Action Trajectory	Robotic manipulation	Simulated/Generated	Not explicitly specified	**?**
2023	Code as Policies	Vision + Language + Action Trajectory	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**
2023	Open-Ended Instructable Embodied Agents with Memory-Augmented Large Language Models	Vision + Language + Action Trajectory	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2023	LLM-Planner	Language	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2023	Bridgedata v2	Not explicitly specified	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2023	Embodied Task Planning with Large Language Models	Vision + Language + Action Trajectory	General embodied tasks	Simulated/Generated	Not explicitly specified	**?**
2023	Plan, Eliminate, and Track – Language Models are Good Teachers	Language + Action Trajectory	Navigation/interaction	Not explicitly specified	Not explicitly specified	**?**
2023	Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	**?**
2023	Urban Generative Intelligence (UGI)	Language + Action Trajectory	Autonomous driving	Simulated/Generated	Not explicitly specified	**?**

Year	Dataset/Framework	Data Modality	Task Focus	Source Type	Scale	Reference
2024	OptiMUS	Language	General embodied tasks	Simulated/Generated	Not explicitly specified	**?**
2024	Diffusion Policy	Action Trajectory	Robotic manipulation	Not explicitly specified	Not explicitly specified	**?**
2024	CoPa	Vision + Language	Robotic manipulation	Real-world	Not explicitly specified	**?**
2024	An Embodied Generalist Agent in 3D World	Vision + Depth/3D + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	**?**
2024	Learning Generative Interactive Environments By Trained Agent Exploration	Action Trajectory + Force/Tactile/Audio	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2024	OpenVLA	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	29 tasks	**?**
2024	Behavior Generation with Latent Actions	Vision + Language + Action Trajectory	Autonomous driving	Not explicitly specified	Not explicitly specified	**?**
2024	Embodied agent interface	Not explicitly specified	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2024	Towards Generalist Robot Policies	Vision + Language + Action Trajectory	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**
2024	Vision-Language Foundation Models as Effective Robot Imitators	Vision + Language	Robotic manipulation	Not explicitly specified	Not explicitly specified	**?**
2024	Open x-embodiment	Not explicitly specified	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2024	Bringing the RT-1-X Foundation Model to a SCARA robot	Action Trajectory	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2024	Octo	Vision + Language + Action Trajectory	Robotic manipulation	Not explicitly specified	800k trajectories	**?**
2024	Genie	Action Trajectory	General embodied tasks	Simulated/Generated	Not explicitly specified	**?**
2024	Learning Manipulation by Predicting Interaction	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2024	3D-VLA	Vision + Depth/3D + Language + Action Trajectory	General embodied tasks	Real-world	Not explicitly specified	**?**
2025	cVLA	Vision + Depth/3D + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	3D CAVLA	Vision + Depth/3D + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	VLA-Touch	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	EdgeVLA	Vision + Language + Action Trajectory	Robotic manipulation	Not explicitly specified	Not explicitly specified	**?**
2025	RynnVLA-002	Vision + Language + Action Trajectory	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	Planning with Reasoning using Vision Language World Model	Vision + Language + Action Trajectory	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2025	Unified Diffusion VLA	Vision + Depth/3D + Language + Action Trajectory	General embodied tasks	Real-world	Not explicitly specified	**?**

Year	Dataset/Framework	Data Modality	Task Focus	Source Type	Scale	Reference
2025	Impromptu VLA	Vision + Language + Action Trajectory	Autonomous driving	Not explicitly specified	Not explicitly specified	**?**
2025	Open X-Embodiment	Vision + Language	Robotic manipulation	Not explicitly specified	160266 tasks	**?**
2025	End-to-End Dexterous Arm-Hand VLA Policies via Shared Autonomy	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	GraspVLA	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	Vision Language Action Models in Robotic Manipulation	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	Humanoid-VLA	Vision + Language + Action Trajectory	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2025	HiMoE-VLA	Vision + Language + Action Trajectory	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	Interleave-VLA	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	210k episodes	**?**
2025	Long-VLA	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	DualVLA	Vision + Language + Action Trajectory	Robotic manipulation	Not explicitly specified	Not explicitly specified	**?**
2025	LIBERO-Plus	Vision + Depth/3D + Language + Action Trajectory	Robotic manipulation	Not explicitly specified	Not explicitly specified	**?**
2025	VLA-0	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	Enhancing Generalization in Vision-Language-Action Models by Preserving Pretrained Representations	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	ManualVLA	Vision + Depth/3D + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	Improving Vision-Language-Action Model with Online Reinforcement Learning	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	VDRive	Vision + Language + Action Trajectory + Force/Tactile/Audio	Autonomous driving	Simulated/Generated	Not explicitly specified	**?**
2025	VLA-Reasoner	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	Percept-WAM	Vision + Depth/3D + Language + Action Trajectory	Autonomous driving	Not explicitly specified	Not explicitly specified	**?**
2025	Actions as Language	Vision + Language + Action Trajectory	Navigation/interaction	Real-world	Not explicitly specified	**?**

Year	Dataset/Framework	Data Modality	Task Focus	Source Type	Scale	Reference
2025	IA-VLA	Vision + Language + Action Trajectory	Robotic manipulation	Not explicitly specified	Not explicitly specified	**?**
2025	DriveAction	Vision + Language + Action Trajectory	Autonomous driving	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	OmniVLA	Vision + Language + Action Trajectory	Navigation/interaction	Real-world	Not explicitly specified	**?**
2025	Sample-Efficient Robot Skill Learning for Construction Tasks	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	AdaPower	Language	Robotic manipulation	Simulated/Generated	Not explicitly specified	**?**
2025	NORA-1.5	Vision + Language + Action Trajectory	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	$\pi^*_{0.6}$	Vision + Language + Action Trajectory + Force/Tactile/Audio	General embodied tasks	Real-world	Not explicitly specified	**?**
2025	The Better You Learn, The Smarter You Prune	Vision + Language + Action Trajectory	General embodied tasks	Real-world	Not explicitly specified	**?**
2025	Galaxea Open-World Dataset and G0 Dual-System VLA Model	Vision + Language + Action Trajectory	Robotic manipulation	Not explicitly specified	Not explicitly specified	**?**
2025	IRL-VLA	Vision + Language + Action Trajectory + Force/Tactile/Audio	Autonomous driving	Simulated/Generated	Not explicitly specified	**?**
2025	A Survey on Vision-Language-Action Models for Autonomous Driving	Vision + Language + Action Trajectory	Autonomous driving	Not explicitly specified	Not explicitly specified	**?**
2025	WholeBodyVLA	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	Dual-Actor Fine-Tuning of VLA Models	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	Refined Policy Distillation	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	DROID	Action Trajectory	Robotic manipulation	Real-world	350 hours; 84 tasks	**?**
2025	Fine-Tuning Vision-Language-Action Models	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	CogVLA	Vision + Language + Action Trajectory	General embodied tasks	Real-world	Not explicitly specified	**?**
2025	A Comprehensive Survey on World Models for Embodied AI	Vision + Action Trajectory	Autonomous driving	Simulated/Generated	Not explicitly specified	**?**
2025	DriveVLA-W0	Vision + Language + Action Trajectory	Autonomous driving	Not explicitly specified	Not explicitly specified	**?**

Year	Dataset/Framework	Data Modality	Task Focus	Source Type	Scale	Reference
2025	HAMSTER	Vision + Depth/3D + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	?
2025	JARVIS-VLA	Vision + Language + Action Trajectory	Navigation/interaction	Not explicitly specified	Not explicitly specified	?
2025	MAP-VLA	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	?
2025	QDepth-VLA	Vision + Depth/3D + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	?
2025	Scalable Vision-Language-Action Model Pretraining for Robotic Manipulation with Real-Life Human	Vision + Depth/3D + Language + Action Trajectory	Robotic manipulation	Real-world	1M episodes	?
2025	SimpleVLA-RL	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Real-world	Not explicitly specified	?
2025	Spatial Forcing	Vision + Depth/3D + Language + Action Trajectory + Force/Tactile/Audio	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	?
2025	Survey of Vision-Language-Action Models for Embodied Manipulation	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	?
2025	Towards Deploying VLA without Fine-Tuning	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	?
2025	VLA-RFT	Vision + Language + Action Trajectory + Force/Tactile/Audio	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	?
2025	Discrete Diffusion VLA	Vision + Language + Action Trajectory	General embodied tasks	Not explicitly specified	Not explicitly specified	?
2025	PixelVLA	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	?
2025	Genie Envisioner	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	?
2025	Evo-1	Vision + Language + Action Trajectory	General embodied tasks	Real-world	Not explicitly specified	?
2025	HiF-VLA	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	?
2025	Aligning Cyber Space with Physical World	Action Trajectory	General embodied tasks	Simulated/Generated	Not explicitly specified	?
2025	Eva-VLA	Vision + Depth/3D + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	?
2025	EvoVLA	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	?
2025	TrackVLA++	Vision + Language + Action Trajectory	General embodied tasks	Real-world	Not explicitly specified	?

Year	Dataset/Framework	Data Modality	Task Focus	Source Type	Scale	Reference
2025	Multimodal Data Storage and Retrieval for Embodied AI	Not explicitly specified	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2025	CognitiveDrone	Vision + Language + Action Trajectory	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2025	GR00T N1	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	Counterfactual VLA	Vision + Language + Action Trajectory	Autonomous driving	Not explicitly specified	Not explicitly specified	**?**
2025	FAST	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	10k hours	**?**
2025	WristWorld	Vision + Depth/3D + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	Physical AI	Vision + Action Trajectory	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	RaceVLA	Vision + Language + Action Trajectory	Navigation/interaction	Not explicitly specified	Not explicitly specified	**?**
2025	Large VLM-based Vision-Language-Action Models for Robotic Manipulation	Vision + Depth/3D + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Not explicitly specified	Not explicitly specified	**?**
2025	SmolVLA	Vision + Language + Action Trajectory	General embodied tasks	Real-world	Not explicitly specified	**?**
2025	OG-VLA	Vision + Depth/3D + Language + Action Trajectory	Robotic manipulation	Real-world	5 demonstrations	**?**
2025	RationalVLA	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	14,000 samples; 000 samples	**?**
2025	ReconVLA	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	100k trajectories	**?**
2025	GeoVLA	Vision + Depth/3D + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	ExpReS-VLA	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	12 demonstrations	**?**
2025	RealMirror	Vision + Depth/3D + Language + Action Trajectory	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	Interactive Post-Training for Vision-Language-Action Models	Vision + Language + Action Trajectory + Force/Tactile/Audio	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2025	Latent Chain-of-Thought World Modeling for End-to-End Driving	Vision + Language + Action Trajectory + Force/Tactile/Audio	Autonomous driving	Not explicitly specified	Not explicitly specified	**?**
2025	NinA	Vision + Language + Action Trajectory	General embodied tasks	Real-world	Not explicitly specified	**?**
2025	GigaBrain-0	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**

Year	Dataset/Framework	Data Modality	Task Focus	Source Type	Scale	Reference
2025	GigaWorld-0	Vision + Depth/3D + Language + Action Trajectory	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	Latent Action Pretraining Through World Modeling	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	Unified Vision-Language-Action Model	Vision + Language + Action Trajectory	Autonomous driving	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	VLA-Adapter	Vision + Language + Action Trajectory	General embodied tasks	Real-world	8 hours	**?**
2025	VLA Model Post-Training via Action-Chunked PPO and Self Behavior Cloning	Vision + Language + Action Trajectory + Force/Tactile/Audio	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2025	VQ-VLA	Vision + Language + Action Trajectory	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	dVLA	Vision + Language + Action Trajectory	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	Dual-Stream Diffusion for World-Model Augmented Vision-Language-Action Model	Vision + Language + Action Trajectory	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	World-Env	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	Model-agnostic Adversarial Attack and Defense for Vision-Language-Action Models	Vision + Language + Action Trajectory	General embodied tasks	Simulated/Generated	Not explicitly specified	**?**
2025	LeVERB	Vision + Language + Action Trajectory + Force/Tactile/Audio	Navigation/interaction	Simulated/Generated	150 tasks	**?**
2025	When Alignment Fails	Vision + Language + Action Trajectory	Robotic manipulation	Not explicitly specified	Not explicitly specified	**?**
2025	Beyond Human Demonstrations	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Simulated/Generated	Not explicitly specified	**?**
2025	EfficientVLA	Vision + Language + Action Trajectory	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2025	Learning to Feel the Future	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	VLA-R1	Vision + Language + Action Trajectory + Force/Tactile/Audio	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**

Year	Dataset/Framework	Data Modality	Task Focus	Source Type	Scale	Reference
2025	DeepThinkVLA	Vision + Language + Action Trajectory + Force/Tactile/Audio	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2025	ForceVLA	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Not explicitly specified	Not explicitly specified	**?**
2025	AutoDrive-R^2	Vision + Language + Action Trajectory + Force/Tactile/Audio	Autonomous driving	Not explicitly specified	Not explicitly specified	**?**
2025	DepthVLA	Vision + Depth/3D + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	RLinf-VLA	Vision + Language + Action Trajectory + Force/Tactile/Audio	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	A Vision-Language-Action-Critic Model for Robotic Real-World Reinforcement Learning	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	4D-VLA	Vision + Depth/3D + Language + Action Trajectory	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	Align-Then-stEer	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	DreamVLA	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	IRef-VLA	Vision + Depth/3D + Language + Action Trajectory	Navigation/interaction	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	Reasoning-VLA	Vision + Language + Action Trajectory + Force/Tactile/Audio	Autonomous driving	Not explicitly specified	Not explicitly specified	**?**
2025	SafeVLA	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	TA-VLA	Vision + Language + Action Trajectory + Force/Tactile/Audio	Autonomous driving	Not explicitly specified	Not explicitly specified	**?**
2025	UP-VLA	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	VTLA	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	VLA^2	Vision + Language + Action Trajectory	Robotic manipulation	Simulated/Generated	Not explicitly specified	**?**

Year	Dataset/Framework	Data Modality	Task Focus	Source Type	Scale	Reference
2025	JARVIS	Language + Action Trajectory	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2025	X-VLA	Vision + Language + Action Trajectory	General embodied tasks	Hybrid (real + simulated)	Not explicitly specified	**?**
2025	FlowVLA	Vision + Language + Action Trajectory	Robotic manipulation	Not explicitly specified	Not explicitly specified	**?**
2025	ChatVLA	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	**?**
2025	ObjectVLA	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	**?**
2026	$_0$	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	Not explicitly specified	**?**
2026	InternVLA-A1	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2026	BridgeV2W	Vision + Action Trajectory	General embodied tasks	Real-world	Not explicitly specified	**?**
2026	CombatVLA	Vision + Depth/3D + Language + Action Trajectory	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2026	Wow, wo, val! A Comprehensive Embodied World Model Evaluation Turing	Vision + Depth/3D + Language	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2026	Vision-Language-Action Models for Autonomous Driving	Vision + Language + Action Trajectory	Autonomous driving	Not explicitly specified	Not explicitly specified	**?**
2026	Flow Equivariant World Models	Vision + Depth/3D	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2026	On-the-Fly VLA Adaptation via Test-Time Reinforcement Learning	Vision + Language + Action Trajectory + Force/Tactile/Audio	General embodied tasks	Real-world	Not explicitly specified	**?**
2026	What Can RL Bring to VLA Generalization$\pi AnEmpiricalStudy$	Vision + Language + Action Trajectory + Force/Tactile/Audio	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2026	NitroGen	Vision + Depth/3D + Action Trajectory	General embodied tasks	Simulated/Generated	40,000 hours; 000 hours	**?**
2026	Video Generation Models in Robotics – Applications, Research Challenges, Future	Vision + Language + Action Trajectory + Force/Tactile/Audio	Autonomous driving	Simulated/Generated	Not explicitly specified	**?**
2026	ReWorld	Vision + Language + Force/Tactile/Audio	Robotic manipulation	Simulated/Generated	Not explicitly specified	**?**
2026	WorldBench	Vision + Action Trajectory	General embodied tasks	Real-world	Not explicitly specified	**?**
2026	A Mechanistic View on Video Generation as World Models	Vision + Language	General embodied tasks	Simulated/Generated	Not explicitly specified	**?**
2026	A Pragmatic VLA Foundation Model	Vision + Language + Action Trajectory	Robotic manipulation	Real-world	20,000 hours; 000 hours	**?**
2026	Parallels Between VLA Model Post-Training and Human Motor Learning	Vision + Language + Action Trajectory	Robotic manipulation	Not explicitly specified	Not explicitly specified	**?**

Year	Dataset/Framework	Data Modality	Task Focus	Source Type	Scale	Reference
2026	DynamicVLA	Vision + Language + Action Trajectory	Robotic manipulation	Hybrid (real + simulated)	Not explicitly specified	**?**
2026	LatentVLA	Vision + Language + Action Trajectory	Autonomous driving	Not explicitly specified	Not explicitly specified	**?**
2026	Vlaser	Vision + Language + Action Trajectory	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2026	Dream-VL					
	Dream-VLA	Vision + Language + Action Trajectory	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2026	Genie Sim 3.0	Vision + Language	Robotic manipulation	Hybrid (real + simulated)	10,000 hours; 000 hours	**?**
2026	AC^2-VLA	Vision + Depth/3D + Language + Action Trajectory	Robotic manipulation	Not explicitly specified	Not explicitly specified	**?**
2026	A Survey on Efficient Vision-Language-Action Models	Vision + Language + Action Trajectory	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**
2026	CompliantVLA-adaptor	Vision + Language + Action Trajectory + Force/Tactile/Audio	Robotic manipulation	Simulated/Generated	Not explicitly specified	**?**
2026	VLM4VLA	Vision + Depth/3D + Language + Action Trajectory	General embodied tasks	Not explicitly specified	Not explicitly specified	**?**

References

Open X.-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I. Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J. Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R. Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic Learning Datasets and RT-X Models, May 2025. URL `http://arxiv.org/abs/2310.08864`. arXiv:2310.08864 [cs].

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. $_0$: A Vision-Language-Action Flow Model for General Robot Control, January 2026. URL `http://arxiv.org/abs/2410.24164`. arXiv:2410.24164 [cs].

Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A Comprehensive Survey on World Models for Embodied AI. 2025a. doi:10.48550/ARXIV.2510.16732. URL `https://arxiv.org/abs/2510.16732`. Version Number: 2.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, Fengli Xu, and Yong Li. Understanding World or Predicting Futurepi A Comprehensive Survey of World Models. *ACM Comput. Surv.*, 58(3):57:1–57:38, September 2025a. ISSN 0360-0300. doi:10.1145/3746449. URL `https://dl.acm.org/doi/10.1145/3746449`.

Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, Arjun Majumdar, Andrea Madotto, Franziska Meier, Florian Metze, Louis-Philippe Morency, Théo Moutakanni, Juan Pino, Basile Terver, Joseph Tighe, Paden Tomasello, and Jitendra Malik. Embodied AI Agents: Modeling the World, July 2025. URL `http://arxiv.org/abs/2506.22355`. arXiv:2506.22355 [cs].

Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI, August 2025a. URL `http://arxiv.org/abs/2407.06886`. arXiv:2407.06886 [cs].

Tarun Gupta, Wenbo Gong, Chao Ma, Nick Pawlowski, Agrin Hilmkil, Meyer Scetbon, Marc Rigter, Ade Famoti, Ashley Juan Llorens, Jianfeng Gao, Stefan Bauer, Danica Kragic, Bernhard Schölkopf, and Cheng Zhang. The Essential Role of Causality in Foundation World Models for Embodied AI, April 2024. URL `http://arxiv.org/abs/2402.06665`. arXiv:2402.06665 [cs].

Kaizhi Zheng, Kaiwen Zhou, Jing Gu, Yue Fan, Jialu Wang, Zonglin Di, Xuehai He, and Xin Eric Wang. JARVIS: A Neuro-Symbolic Commonsense Reasoning Framework for Conversational Embodied Agents, September 2025a. URL `http://arxiv.org/abs/2208.13266`. arXiv:2208.13266 [cs].

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R. Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, Vitor Guizilini, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J. Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset, April 2025. URL `http://arxiv.org/abs/2403.12945`. arXiv:2403.12945 [cs].

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $_{0.5}$: a Vision-Language-Action Model with Open-World Generalization, April 2025a. URL `http://arxiv.org/abs/2504.16054`. arXiv:2504.16054 [cs].

Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachy Groom, Hunter Hancock, Karol Hausman, Gashon Hussein, Brian Ichter, Szymon Jakubczak, Rowan Jen, Tim Jones, Ben Katz, Liyiming Ke, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Yao Lu, Vishnu Mano, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Charvi Sharma, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Alex Swerdlow, James Tanner, Marcel Torne, Quan Vuong, Anna Walling, Haohuan Wang, Blake Williams, Sukwon Yoo, Lili Yu, Ury Zhilinsky, and Zhiyuan Zhou. $^{*}_{0.6}$: a VLA That Learns From Experience, November 2025b. URL `http://arxiv.org/abs/2511.14759`. arXiv:2511.14759 [cs].

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient Action Tokenization for Vision-Language-Action Models, January 2025. URL `http://arxiv.org/abs/2501.09747`. arXiv:2501.09747 [cs].

Johann Brehmer, Joey Bose, Pim De Haan, and Taco S. Cohen. Edgi: Equivariant diffusion for planning with embodied agents. *Advances in Neural Information Processing Systems*, 36: 63818–63834, 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/hash/c95c049637c5c549c2a08e8d6dcbca4b-Abstract-Conference.html`.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, and Ruohan Zhang. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:

100428–100534, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b631da756d1573c24c9ba9c702fde5a9-Abstract-Datasets_and_Benchmarks_Track.html.

Yixiang Chen, Peiyan Li, Jiabing Yang, Keji He, Xiangnan Wu, Yuan Xu, Kai Wang, Jing Liu, Nianfeng Liu, Yan Huang, and Liang Wang. BridgeV2W: Bridging Video Generation Models to Embodied World Models via Embodiment Masks, February 2026a. URL http://arxiv.org/abs/2602.03793. arXiv:2602.03793 [cs].

Zhuo Li, Weiran Wu, Yunlong Guo, Jian Sun, and Qing-Long Han. Embodied Multi-Agent Systems: A Review. *IEEE/CAA Journal of Automatica Sinica*, 12(6):1095–1116, 2025b. URL https://ieeexplore.ieee.org/abstract/document/11036708/.

Dhruv Batra, Angel X. Chang, Sonia Chernova, Andrew J. Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, Manolis Savva, and Hao Su. Rearrangement: A Challenge for Embodied AI, November 2020. URL http://arxiv.org/abs/2011.01975. arXiv:2011.01975 [cs].

Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. Physcene: Physically interactable 3d scene synthesis for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16262–16272, 2024a. URL http://openaccess.thecvf.com/content/CVPR2024/html/Yang_PhyScene_Physically_Interactable_3D_Scene_Synthesis_for_Embodied_AI_CVPR_2024_paper.html.

Jiaming Wang, Diwen Liu, Jizhuo Chen, Jiaxuan Da, Nuowen Qian, Minh Man Tram, and Harold Soh. Genie: A generalizable navigation system for in-the-wild environments. *IEEE Robotics and Automation Letters*, 2025a. URL https://ieeexplore.ieee.org/abstract/document/11206420/.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, and Chris Apps. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forumpiid=bJbSbJskOS.

Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, Yizhong Zhang, Xi Chen, Hao Chen, Lily Sun, Dong Chen, Jiaolong Yang, and Baining Guo. Scalable Vision-Language-Action Model Pretraining for Robotic Manipulation with Real-Life Human Activity Videos, October 2025c. URL http://arxiv.org/abs/2510.21571. arXiv:2510.21571 [cs].

Haoyun Li, Ivan Zhang, Runqi Ouyang, Xiaofeng Wang, Zheng Zhu, Zhiqin Yang, Zhentao Zhang, Boyuan Wang, Chaojun Ni, Wenkang Qin, Xinze Chen, Yun Ye, Guan Huang, Zhenbo Song, and Xingang Wang. MimicDreamer: Aligning Human and Robot Demonstrations for Scalable VLA Training, September 2025d. URL http://arxiv.org/abs/2509.22199. arXiv:2509.22199 [cs].

Cong Tai, Zhaoyu Zheng, Haixu Long, Hansheng Wu, Haodong Xiang, Zhengbin Long, Jun Xiong, Rong Shi, Shizhuang Zhang, Gang Qiu, He Wang, Ruifeng Li, Jun Huang, Bin Chang, Shuai Feng, and Tao Shen. RealMirror: A Comprehensive, Open-Source Vision-Language-Action Platform for Embodied AI, September 2025. URL http://arxiv.org/abs/2509.14687. arXiv:2509.14687 [cs].

GigaWorld Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jiagang Zhu, Kerui Li, Mengyuan Xu, Qiuping Deng, Siting Wang, Wenkang Qin, Xinze Chen, Xiaofeng Wang, Yankai Wang, Yu Cao, Yifan Chang, Yuan Xu, Yun Ye, Yang Wang, Yukun Zhou, Zhengyuan Zhang, Zhehao Dong, and Zheng Zhu. GigaWorld-0: World Models as Data Engine to Empower Embodied AI, November 2025a. URL http://arxiv.org/abs/2511.19861. arXiv:2511.19861 [cs].

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, Fan Lu, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. DreamVLA: A Vision-Language-Action Model Dreamed with Comprehensive World Knowledge, August 2025a. URL http://arxiv.org/abs/2507.04447. arXiv:2507.04447 [cs].

Anqi Li, Zhiyong Wang, Jiazhao Zhang, Minghan Li, Yunpeng Qi, Zhibo Chen, Zhizheng Zhang, and He Wang. UrbanVLA: A Vision-Language-Action Model for Urban Micromobility, October 2025e. URL http://arxiv.org/abs/2510.23576. arXiv:2510.23576 [cs].

Bahey Tharwat, Yara Nasser, Ali Abouzeid, and Ian Reid. Latent Action Pretraining Through World Modeling, September 2025. URL http://arxiv.org/abs/2509.18428. arXiv:2509.18428 [cs].

Minjie Zhu, Yichen Zhu, Jinming Li, Zhongyi Zhou, Junjie Wen, Xiaoyu Liu, Chaomin Shen, Yaxin Peng, and Feifei Feng. ObjectVLA: End-to-End Open-World Object Manipulation Without Demonstration, February 2025a. URL http://arxiv.org/abs/2502.19250. arXiv:2502.19250 [cs].

Angen Ye, Zeyu Zhang, Boyuan Wang, Xiaofeng Wang, Dapeng Zhang, and Zheng Zhu. VLA-R1: Enhancing Reasoning in Vision-Language-Action Models, October 2025a. URL http://arxiv.org/abs/2510.01623. arXiv:2510.01623 [cs].

Yihao Lu and Hao Tang. Multimodal Data Storage and Retrieval for Embodied AI: A Survey, August 2025. URL http://arxiv.org/abs/2508.13901. arXiv:2508.13901 [cs] version: 1.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR, 2022a. URL https://proceedings.mlr.press/v162/huang22a.html.

Yue Wu, So Yeon Min, Yonatan Bisk, Ruslan Salakhutdinov, Amos Azaria, Yuanzhi Li, Tom Mitchell, and Shrimai Prabhumoye. Plan, Eliminate, and Track – Language Models are Good Teachers for Embodied Agents, May 2023a. URL http://arxiv.org/abs/2305.02412. arXiv:2305.02412 [cs].

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models, March 2023. URL http://arxiv.org/abs/2212.04088. arXiv:2212.04088 [cs].

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An Open-Ended Embodied Agent with Large Language Models, October 2023a. URL http://arxiv.org/abs/2305.16291. arXiv:2305.16291 [cs].

Kolby Nottingham, Prithviraj Ammanabrolu, Alane Suhr, Yejin Choi, Hannaneh Hajishirzi, Sameer Singh, and Roy Fox. Do embodied agents dream of pixelated sheep: Embodied decision making using language guided world modelling. In *International Conference on Machine Learning*, pages 26311–26325. PMLR, 2023. URL https://proceedings.mlr.press/v202/nottingham23a.html.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. Building Cooperative Embodied Agents Modularly with Large Language Models, February 2024. URL http://arxiv.org/abs/2307.02485. arXiv:2307.02485 [cs].

Yifan Zhong, Fengshuo Bai, Shaofei Cai, Xuchuan Huang, Zhang Chen, Xiaowei Zhang, Yuanfei Wang, Shaoyang Guo, Tianrui Guan, Ka Nam Lui, Zhiquan Qi, Yitao Liang, Yuanpei Chen, and Yaodong Yang. A Survey on Vision-Language-Action Models: An Action Tokenization Perspective, July 2025a. URL http://arxiv.org/abs/2507.01925. arXiv:2507.01925 [cs].

Yu Cui, Yujian Zhang, Lina Tao, Yang Li, Xinyu Yi, and Zhibin Li. End-to-End Dexterous Arm-Hand VLA Policies via Shared Autonomy: VR Teleoperation Augmented by Autonomous Hand VLA Policy for Efficient Data Collection, December 2025. URL http://arxiv.org/abs/2511.00139. arXiv:2511.00139 [cs].

Tian-Yu Xiang, Ao-Qun Jin, Xiao-Hu Zhou, Mei-Jiang Gui, Xiao-Liang Xie, Shi-Qi Liu, Shuang-Yi Wang, Sheng-Bin Duang, Si-Cheng Wang, Zheng Lei, and Zeng-Guang Hou. VLA Model-Expert Collaboration for Bi-directional Manipulation Learning, March 2025. URL http://arxiv.org/abs/2503.04163. arXiv:2503.04163 [cs].

Weiqi Li, Quande Zhang, Ruifeng Zhai, Liang Lin, and Guangrun Wang. VLA Models Are More Generalizable Than You Think: Revisiting Physical and Spatial Modeling, December 2025f. URL http://arxiv.org/abs/2512.02902. arXiv:2512.02902 [cs].

Ranjan Sapkota, Yang Cao, Konstantinos I. Roumeliotis, and Manoj Karkee. Vision-Language-Action (VLA) Models: Concepts, Progress, Applications and Challenges, January 2026. URL http://arxiv.org/abs/2505.04769. arXiv:2505.04769 [cs].

Yihao Wang, Pengxiang Ding, Lingxiao Li, Can Cui, Zirui Ge, Xinyang Tong, Wenxuan Song, Han Zhao, Wei Zhao, Pengxu Hou, Siteng Huang, Yifan Tang, Wenhui Wang, Ru Zhang, Jianyi Liu, and Donglin Wang. VLA-Adapter: An Effective Paradigm for Tiny-Scale Vision-Language-Action Model, September 2025b. URL http://arxiv.org/abs/2509.09372. arXiv:2509.09372 [cs].

Shaoqi Dong, Chaoyou Fu, Haihan Gao, Yi-Fan Zhang, Chi Yan, Chu Wu, Xiaoyu Liu, Yunhang Shen, Jing Huo, Deqiang Jiang, Haoyu Cao, Yang Gao, Xing Sun, Ran He, and Caifeng Shan. VITA-VLA: Efficiently Teaching Vision-Language Models to Act via Action Expert Distillation, October 2025. URL http://arxiv.org/abs/2510.09607. arXiv:2510.09607 [cs].

Haoran Li, Yuhui Chen, Wenbo Cui, Weiheng Liu, Kai Liu, Mingcai Zhou, Zhengtao Zhang, and Dongbin Zhao. Survey of Vision-Language-Action Models for Embodied Manipulation, November 2025g. URL http://arxiv.org/abs/2508.15201. arXiv:2508.15201 [cs].

GigaBrain Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jie Li, Jiagang Zhu, Lv Feng, Peng Li, Qiuping Deng, Runqi Ouyang, Wenkang Qin, Xinze Chen, Xiaofeng Wang, Yang Wang, Yifan Li, Yilong Li, Yiran Ding, Yuan Xu, Yun Ye, Yukun Zhou, Zhehao Dong, Zhenan Wang, Zhichao Liu, and Zheng Zhu. GigaBrain-0: A World Model-Powered Vision-Language-Action Model, December 2025b. URL http://arxiv.org/abs/2510.19430. arXiv:2510.19430 [cs].

Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, Yiyu Ren, Kejia Zhang, Hui Yu, Jingmei Zhao, Shuai Zhou, Zhenqi Qiu, Houlong Xiong, Ziyu Wang, Zechen Wang, Ran Cheng, Yong-Lu Li, Yongtao Huang, Xing Zhu, Yujun Shen, and Kecheng Zheng. A Pragmatic VLA Foundation Model, January 2026a. URL `http://arxiv.org/abs/2601.18692`. arXiv:2601.18692 [cs].

Chia-Yu Hung, Qi Sun, Pengfei Hong, Amir Zadeh, Chuan Li, U.-Xuan Tan, Navonil Majumder, and Soujanya Poria. NORA: A Small Open-Sourced Generalist Vision Language Action Model for Embodied Tasks, April 2025a. URL `http://arxiv.org/abs/2504.19854`. arXiv:2504.19854 [cs].

Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, and Hongyang Li. WholeBodyVLA: Towards Unified Latent VLA for Whole-Body Loco-Manipulation Control, December 2025a. URL `http://arxiv.org/abs/2512.11047`. arXiv:2512.11047 [cs].

Rashid Turgunbaev. From Perception to Action with Integrated VLA Systems. *Technical Science Integrated Research*, 1(6):11–17, 2025. URL `https://altumnova.com/index.php/tsir/article/view/35`.

Muhayy Ud Din, Waseem Akram, Lyes Saad Saoud, Jan Rosell, and Irfan Hussain. Vision Language Action Models in Robotic Manipulation: A Systematic Review, July 2025. URL `http://arxiv.org/abs/2507.10672`. arXiv:2507.10672 [cs].

Lin Sun, Bin Xie, Yingfei Liu, Hao Shi, Tiancai Wang, and Jiale Cao. GeoVLA: Empowering 3D Representations in Vision-Language-Action Models, August 2025a. URL `http://arxiv.org/abs/2508.09071`. arXiv:2508.09071 [cs].

Chengmeng Li, Junjie Wen, Yaxin Peng, Yan Peng, and Yichen Zhu. Pointvla: Injecting the 3d world into vision-language-action models. *IEEE Robotics and Automation Letters*, 11(3):2506–2513, 2026a. URL `https://ieeexplore.ieee.org/abstract/document/11346992/`.

Xiaoqi Li, Liang Heng, Jiaming Liu, Yan Shen, Chenyang Gu, Zhuoyang Liu, Hao Chen, Nuowei Han, Renrui Zhang, and Hao Tang. 3ds-vla: A 3d spatial-aware vision language action model for robust multi-task manipulation. In *9th Annual Conference on Robot Learning*, 2025h. URL `https://openreview.net/forumpiid=dT45OMevL5`.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. 4D-VLA: Spatiotemporal Vision-Language-Action Pretraining with Cross-Scene Calibration, November 2025b. URL `http://arxiv.org/abs/2506.22242`. arXiv:2506.22242 [cs].

Ishika Singh, Ankit Goyal, Stan Birchfield, Dieter Fox, Animesh Garg, and Valts Blukis. OG-VLA: Orthographic Image Generation for 3D-Aware Vision-Language Action Model, November 2025. URL `http://arxiv.org/abs/2506.01196`. arXiv:2506.01196 [cs].

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Model, May 2025. URL `http://arxiv.org/abs/2501.15830`. arXiv:2501.15830 [cs].

Yixuan Li, Yuhui Chen, Mingcai Zhou, Haoran Li, Zhengtao Zhang, and Dongbin Zhao. QDepth-VLA: Quantized Depth Prediction as Auxiliary Supervision for Vision-Language-Action Models, December 2025i. URL `http://arxiv.org/abs/2510.14836`. arXiv:2510.14836 [cs].

M. A. Patratskiy, A. K. Kovalev, and A. I. Panov. Spatial Traces: Enhancing VLA Models with Spatial-Temporal Understanding. *Optical Memory and Neural Networks*, 34(S1):S72–S82, December 2025. ISSN 1060-992X, 1934-7898. doi:10.3103/S1060992X25601654. URL `https://link.springer.com/10.3103/S1060992X25601654`.

Jacob Berg, Chuning Zhu, Yanda Bao, Ishan Durugkar, and Abhishek Gupta. Semantic World Models, October 2025. URL `http://arxiv.org/abs/2510.19818`. arXiv:2510.19818 [cs].

Jiawen Yu, Hairuo Liu, Qiaojun Yu, Jieji Ren, Ce Hao, Haitong Ding, Guangyu Huang, Guofan Huang, Yan Song, Panpan Cai, Cewu Lu, and Wenqiang Zhang. ForceVLA: Enhancing VLA Models with a Force-aware MoE for Contact-rich Manipulation, September 2025. URL `http://arxiv.org/abs/2505.22159`. arXiv:2505.22159 [cs].

Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-VLA: Unlocking Vision-Language-Action Model's Physical Knowledge for Tactile Generalization, July 2025a. URL `http://arxiv.org/abs/2507.09160`. arXiv:2507.09160 [cs].

Jianxin Bi, Kevin Yuchen Ma, Ce Hao, Mike Zheng Shou, and Harold Soh. VLA-Touch: Enhancing Vision-Language-Action Models with Dual-Level Tactile Feedback, July 2025a. URL `http://arxiv.org/abs/2507.17294`. arXiv:2507.17294 [cs].

Roman Dolgopolyi and Anastasios Tsevas. Bridging Perception, Language, and Action: A Survey and Bibliometric Analysis of VLM & VLA Systems. 2025. URL `https://www.researchsquare.com/article/rs-7935378/latest`.

Heng Zhang, Wei-Hsing Huang, Qiyi Tong, Gokhan Solak, Puze Liu, Sheng Liu, Jan Peters, and Arash Ajoudani. CompliantVLA-adaptor: VLM-Guided Variable Impedance Action for Safe Contact-Rich Manipulation, January 2026a. URL http://arxiv.org/abs/2601.15541. arXiv:2601.15541 [cs].

Zongzheng Zhang, Haobo Xu, Zhuo Yang, Chenghao Yue, Zehao Lin, Huan-ang Gao, Ziwei Wang, and Hao Zhao. TA-VLA: Elucidating the Design Space of Torque-aware Vision-Language-Action Models, September 2025c. URL http://arxiv.org/abs/2509.07962. arXiv:2509.07962 [cs].

Linqing Zhong, Yi Liu, Yifei Wei, Ziyu Xiong, Maoqing Yao, Si Liu, and Guanghui Ren. ACoT-VLA: Action Chain-of-Thought for Vision-Language-Action Models, January 2026. URL http://arxiv.org/abs/2601.11404. arXiv:2601.11404 [cs].

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, and Chelsea Finn. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713, 2025a. URL http://openaccess.thecvf.com/content/CVPR2025/html/Zhao_CoT-VLA_Visual_Chain-of-Thought_Reasoning_for_Vision-Language-Action_Models_CVPR_2025_paper.html.

Helong Huang, Min Cen, Kai Tan, Xingyue Quan, Guowei Huang, and Hong Zhang. GraphCoT-VLA: A 3D Spatial-Aware Reasoning Vision-Language-Action Model for Robotic Manipulation with Ambiguous Instructions, August 2025b. URL http://arxiv.org/abs/2508.07650. arXiv:2508.07650 [cs].

Haoru Xue, Xiaoyu Huang, Dantong Niu, Qiayuan Liao, Thomas Kragerud, Jan Tommy Gravdahl, Xue Bin Peng, Guanya Shi, Trevor Darrell, Koushil Sreenath, and Shankar Sastry. LeVERB: Humanoid Whole-Body Control with Latent Vision-Language Instruction, September 2025. URL http://arxiv.org/abs/2506.13751. arXiv:2506.13751 [cs].

Yilin Wu, Anqi Li, Tucker Hermans, Fabio Ramos, Andrea Bajcsy, and Claudia Pérez-D'Arpino. Do What You Say: Steering Vision-Language-Action Models via Runtime Reasoning-Action Alignment Verification, January 2026b. URL http://arxiv.org/abs/2510.16281. arXiv:2510.16281 [cs].

Chenyang Gu, Jiaming Liu, Hao Chen, Runzhong Huang, Qingpo Wuwu, Zhuoyang Liu, Xiaoqi Li, Ying Li, Renrui Zhang, Peng Jia, Pheng-Ann Heng, and Shanghang Zhang. ManualVLA: A Unified VLA Model for Chain-of-Thought Manual Generation and Robotic Manipulation, December 2025. URL http://arxiv.org/abs/2512.02013. arXiv:2512.02013 [cs].

Dapeng Zhang, Zhenlong Yuan, Zhangquan Chen, Chih-Ting Liao, Yinda Chen, Fei Shen, Qingguo Zhou, and Tat-Seng Chua. Reasoning-VLA: A Fast and General Vision-Language-Action Reasoning Model for Autonomous Driving, November 2025d. URL http://arxiv.org/abs/2511.19912. arXiv:2511.19912 [cs].

Zhenghao "Mark" Peng, Wenhao Ding, Yurong You, Yuxiao Chen, Wenjie Luo, Thomas Tian, Yulong Cao, Apoorva Sharma, Danfei Xu, Boris Ivanovic, Boyi Li, Bolei Zhou, Yan Wang, and Marco Pavone. Counterfactual VLA: Self-Reflective Vision-Language-Action Model with Adaptive Reasoning, December 2025. URL http://arxiv.org/abs/2512.24426. arXiv:2512.24426 [cs].

Wen-Han Hsieh, Elvis Hsieh, Dantong Niu, Trevor Darrell, Roei Herzig, and David M. Chan. Do whatpi Teaching vision-language-action models to reject the impossible. *arXiv preprint arXiv:2508.16292*, 2, 2025. URL https://aclanthology.org/anthology-files/pdf/findings/2025.findings-emnlp.635.pdf.

Wenkai Guo, Guanxing Lu, Haoyuan Deng, Zhenyu Wu, Yansong Tang, and Ziwei Wang. VLA-Reasoner: Empowering Vision-Language-Action Models with Reasoning via Online Monte Carlo Tree Search, September 2025a. URL http://arxiv.org/abs/2509.22643. arXiv:2509.22643 [cs].

Shuhan Tan, Kashyap Chitta, Yuxiao Chen, Ran Tian, Yurong You, Yan Wang, Wenjie Luo, Yulong Cao, Philipp Krahenbuhl, Marco Pavone, and Boris Ivanovic. Latent Chain-of-Thought World Modeling for End-to-End Driving, December 2025a. URL http://arxiv.org/abs/2512.10226. arXiv:2512.10226 [cs].

Junjie Wen, Minjie Zhu, Jiaming Liu, Zhiyuan Liu, Yicun Yang, Linfeng Zhang, Shanghang Zhang, Yichen Zhu, and Yi Xu. dVLA: Diffusion Vision-Language-Action Model with Multimodal Chain-of-Thought, September 2025a. URL http://arxiv.org/abs/2509.25681. arXiv:2509.25681 [cs].

Cheng Yin, Yankai Lin, Wang Xu, Sikyuen Tam, Xiangrui Zeng, Zhiyuan Liu, and Zhouping Yin. DeepThinkVLA: Enhancing Reasoning Capability of Vision-Language-Action Models, October 2025. URL http://arxiv.org/abs/2511.15669. arXiv:2511.15669 [cs].

Zhide Zhong, Haodong Yan, Junfeng Li, Xiangchen Liu, Xin Gong, Tianran Zhang, Wenxuan Song, Jiayi Chen, Xinhu Zheng, Hesheng Wang, and Haoang Li. FlowVLA: Visual Chain of Thought-based Motion Reasoning for Vision-Language-Action Models, October 2025b. URL http://arxiv.org/abs/2508.18269. arXiv:2508.18269 [cs].

Wenxuan Song, Jiayi Chen, Wenxue Li, Xu He, Han Zhao, Can Cui, Pengxiang Ding Shiyan Su, Feilong Tang, Xuelian Cheng, Donglin Wang, Zongyuan Ge, Xinhu Zheng, Zhe Liu, Hesheng Wang, and Haoang Li. RationalVLA: A Rational Vision-Language-Action Model with Dual System, June 2025a. URL http://arxiv.org/abs/2506.10826. arXiv:2506.10826 [cs].

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, Lu Hou, Lue Fan, and Zhaoxiang Zhang. DriveVLA-W0: World Models Amplify Data Scaling Law in Autonomous Driving, December 2025j. URL http://arxiv.org/abs/2510.12796. arXiv:2510.12796 [cs].

Zezhong Qian, Xiaowei Chi, Yuming Li, Shizun Wang, Zhiyuan Qin, Xiaozhu Ju, Sirui Han, and Shanghang Zhang. WristWorld: Generating Wrist-Views via 4D World Models for Robotic Manipulation, October 2025. URL http://arxiv.org/abs/2510.07313. arXiv:2510.07313 [cs].

Jiacheng Ye, Shansan Gong, Jiahui Gao, Junming Fan, Shuang Wu, Wei Bi, Haoli Bai, Lifeng Shang, and Lingpeng Kong. Dream-VL & Dream-VLA: Open Vision-Language and Vision-Language-Action Models with Diffusion Language Model Backbone, January 2026. URL http://arxiv.org/abs/2512.22615. arXiv:2512.22615 [cs].

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. WorldVLA: Towards Autoregressive Action World Model, June 2025a. URL http://arxiv.org/abs/2506.21539. arXiv:2506.21539 [cs].

Jun Cen, Siteng Huang, Yuqian Yuan, Kehan Li, Hangjie Yuan, Chaohui Yu, Yuming Jiang, Jiayan Guo, Xin Li, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. RynnVLA-002: A Unified Vision-Language-Action and World Model, November 2025b. URL http://arxiv.org/abs/2511.17502. arXiv:2511.17502 [cs].

Fangqi Zhu, Zhengyang Yan, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. WMPO: World Model-based Policy Optimization for Vision-Language-Action Models, November 2025b. URL http://arxiv.org/abs/2511.09515. arXiv:2511.09515 [cs].

Hengtao Li, Pengxiang Ding, Runze Suo, Yihao Wang, Zirui Ge, Dongyuan Zang, Kexian Yu, Mingyang Sun, Hongyin Zhang, Donglin Wang, and Weihua Su. VLA-RFT: Vision-Language-Action Reinforcement Fine-tuning with Verified Rewards in World Simulators, October 2025k. URL http://arxiv.org/abs/2510.00406. arXiv:2510.00406 [cs].

Junjin Xiao, Yandan Yang, Xinyuan Chang, Ronghan Chen, Feng Xiong, Mu Xu, Wei-Shi Zheng, and Qing Zhang. World-Env: Leveraging World Model as a Virtual Environment for VLA Post-Training, November 2025. URL http://arxiv.org/abs/2509.24948. arXiv:2509.24948 [cs].

Anqing Jiang, Yu Gao, Yiru Wang, Zhigang Sun, Shuo Wang, Yuwen Heng, Hao Sun, Shichen Tang, Lijuan Zhu, Jinhao Chai, Jijun Wang, Zichong Gu, Hao Jiang, and Li Sun. IRL-VLA: Training an Vision-Language-Action Policy via Reward World Model, August 2025b. URL http://arxiv.org/abs/2508.06571. arXiv:2508.06571 [cs].

Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Zhaohui Yang, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, Dehui Wang, Dingxiang Luo, Yuchen Fan, Youbang Sun, Jia Zeng, Jiangmiao Pang, Shanghang Zhang, Yu Wang, Yao Mu, Bowen Zhou, and Ning Ding. SimpleVLA-RL: Scaling VLA Training via Reinforcement Learning, September 2025l. URL http://arxiv.org/abs/2509.09674. arXiv:2509.09674 [cs].

Hongzhi Zang, Mingjie Wei, Si Xu, Yongji Wu, Zhen Guo, Yuanqing Wang, Hao Lin, Liangzhi Shi, Yuqing Xie, Zhexuan Xu, Zhihao Liu, Kang Chen, Wenhao Tang, Quanlu Zhang, Weinan Zhang, Chao Yu, and Yu Wang. RLinf-VLA: A Unified and Efficient Framework for VLA+RL Training, October 2025. URL http://arxiv.org/abs/2510.06710. arXiv:2510.06710 [cs].

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A Unified Latent Action World Model, December 2025b. URL http://arxiv.org/abs/2512.13030. arXiv:2512.13030 [cs].

Hansen Jin Lillemark, Benhao Huang, Fangneng Zhan, Yilun Du, and Thomas Anderson Keller. Flow Equivariant World Models: Memory for Partially Observed Dynamic Environments, January 2026. URL http://arxiv.org/abs/2601.01075. arXiv:2601.01075 [cs].

Loïc Magne, Anas Awadalla, Guanzhi Wang, Yinzhen Xu, Joshua Belofsky, Fengyuan Hu, Joohwan Kim, Ludwig Schmidt, Georgia Gkioxari, Jan Kautz, Yisong Yue, Yejin Choi, Yuke Zhu, and Linxi "Jim" Fan. NitroGen: An Open Foundation Model for Generalist Gaming Agents, January 2026. URL http://arxiv.org/abs/2601.02427. arXiv:2601.02427 [cs].

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, Liliang Chen, Shuicheng Yan, Maoqing Yao, and Guanghui Ren. Genie Envisioner: A Unified World

Foundation Platform for Robotic Manipulation, November 2025. URL http://arxiv.org/abs/2508.05635. arXiv:2508.05635 [cs].

Chi Wan, Kangrui Wang, Yuan Si, Pingyue Zhang, Huang Huang, and Manling Li. WorldAgen: Unified State-Action Prediction with Test-Time World Model Training. In *NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning*, 2025. URL https://openreview.net/forumpiid=egbFo1gvYp.

Jun Guo, Xiaojian Ma, Yikai Wang, Min Yang, Huaping Liu, and Qing Li. Flowdreamer: A rgb-d world model with flow-based motion representations for robot manipulation. *IEEE Robotics and Automation Letters*, 11(3):2466–2473, 2026. URL https://ieeexplore.ieee.org/abstract/document/11345941/.

Yanfei Wang, Zhiwen Yu, Sicong Liu, Zimu Zhou, and Bin Guo. Genie in the Model: Automatic Generation of Human-in-the-Loop Deep Neural Networks for Mobile Applications. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(1):1–29, March 2023b. ISSN 2474-9567. doi:10.1145/3580815. URL https://dl.acm.org/doi/10.1145/3580815.

Jianhua Han, Meng Tian, Jiangtong Zhu, Fan He, Huixin Zhang, Sitong Guo, Dechang Zhu, Hao Tang, Pei Xu, Yuze Guo, Minzhe Niu, Haojie Zhu, Qichao Dong, Xuechao Yan, Siyuan Dong, Lu Hou, Qingqiu Huang, Xiaosong Jia, and Hang Xu. Percept-WAM: Perception-Enhanced World-Awareness-Action Model for Robust End-to-End Autonomous Driving, November 2025. URL http://arxiv.org/abs/2511.19221. arXiv:2511.19221 [cs].

Mingwang Xu, Jiahao Cui, Feipeng Cai, Hanlin Shang, Zhihao Zhu, Shan Luan, Yifang Xu, Neng Zhang, Yaoyi Li, Jia Cai, and Siyu Zhu. WAM-Diff: A Masked Diffusion VLA Framework with MoE and Online Reinforcement Learning for Autonomous Driving, December 2025a. URL http://arxiv.org/abs/2512.11872. arXiv:2512.11872 [cs].

Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiang Zhu, and Jianyu Chen. UP-VLA: A Unified Understanding and Prediction Model for Embodied Agent, June 2025e. URL http://arxiv.org/abs/2501.18867. arXiv:2501.18867 [cs].

Bear Häon, Kaylene Stocking, Ian Chuang, and Claire Tomlin. Mechanistic interpretability for steering vision-language-action models, August 2025. URL http://arxiv.org/abs/2509.00328. arXiv:2509.00328 [cs].

Luozhou Wang, Zhifei Chen, Yihua Du, Dongyu Yan, Wenhang Ge, Guibao Shen, Xinli Xu, Leyi Wu, Man Chen, Tianshuo Xu, Peiran Ren, Xin Tao, Pengfei Wan, and Ying-Cong Chen. A Mechanistic View on Video Generation as World Models: State and Dynamics, January 2026. URL http://arxiv.org/abs/2601.17067. arXiv:2601.17067 [cs].

Yantai Yang, Yuhao Wang, Zichen Wen, Luo Zhongwei, Chang Zou, Zhipeng Zhang, Chuan Wen, and Linfeng Zhang. EfficientVLA: Training-Free Acceleration and Compression for Vision-Language-Action Models, June 2025a. URL http://arxiv.org/abs/2506.10100. arXiv:2506.10100 [cs].

Zheng Xiong, Kang Li, Zilin Wang, Matthew Jackson, Jakob Foerster, and Shimon Whiteson. HyperVLA: Efficient Inference in Vision-Language-Action Models via Hypernetworks, October 2025. URL http://arxiv.org/abs/2510.04898. arXiv:2510.04898 [cs].

Hanzhen Wang, Jiaming Xu, Jiayi Pan, Yongkang Zhou, and Guohao Dai. SpecPrune-VLA: Accelerating Vision-Language-Action Models via Action-Aware Self-Speculative Pruning, September 2025c. URL http://arxiv.org/abs/2509.05614. arXiv:2509.05614 [cs].

Hengyu Fang, Yijiang Liu, Yuan Du, Li Du, and Huanrui Yang. SQAP-VLA: A Synergistic Quantization-Aware Pruning Framework for High-Performance Vision-Language-Action Models, September 2025a. URL http://arxiv.org/abs/2509.09090. arXiv:2509.09090 [cs].

Rongyu Zhang, Menghang Dong, Yuan Zhang, Liang Heng, Xiaowei Chi, Gaole Dai, Li Du, Yuan Du, and Shanghang Zhang. MoLe-VLA: Dynamic Layer-skipping Vision Language Action Model via Mixture-of-Layers for Efficient Robot Manipulation, April 2025f. URL http://arxiv.org/abs/2503.20384. arXiv:2503.20384 [cs].

Wenda Yu, Tianshi Wang, Fengling Li, Jingjing Li, and Lei Zhu. AC^2-VLA: Action-Context-Aware Adaptive Computation in Vision-Language-Action Models for Efficient Robotic Manipulation, January 2026a. URL http://arxiv.org/abs/2601.19634. arXiv:2601.19634 [cs].

Yanjiang Guo, Jianke Zhang, Xiaoyu Chen, Xiang Ji, Yen-Jen Wang, Yucheng Hu, and Jianyu Chen. Improving Vision-Language-Action Model with Online Reinforcement Learning, January 2025b. URL http://arxiv.org/abs/2501.16664. arXiv:2501.16664 [cs].

Han Zhao, Wenxuan Song, Donglin Wang, Xinyang Tong, Pengxiang Ding, Xuelian Cheng, and Zongyuan Ge. MoRE: Unlocking Scalability in Reinforcement Learning for Quadruped Vision-Language-Action Models, March 2025b. URL http://arxiv.org/abs/2503.08007. arXiv:2503.08007 [cs].

Yating Wang, Haoyi Zhu, Mingyu Liu, Jiange Yang, Hao-Shu Fang, and Tong He. VQ-VLA: Improving Vision-Language-Action Models via Scaling Vector-Quantized Action Tokenizers, July 2025d. URL `http://arxiv.org/abs/2507.01016`. arXiv:2507.01016 [cs].

Asher J. Hancock, Xindi Wu, Lihan Zha, Olga Russakovsky, and Anirudha Majumdar. Actions as Language: Fine-Tuning VLMs into VLAs Without Catastrophic Forgetting, September 2025a. URL `http://arxiv.org/abs/2509.22195`. arXiv:2509.22195 [cs].

Songsheng Wang, Rucheng Yu, Zhihang Yuan, Chao Yu, Feng Gao, Yu Wang, and Derek F. Wong. Spec-vla: speculative decoding for vision-language-action models with relaxed acceptance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26916–26928, 2025e. URL `https://aclanthology.org/2025.emnlp-main.1367/`.

Wenxuan Song, Jiayi Chen, Pengxiang Ding, Yuxin Huang, Han Zhao, Donglin Wang, and Haoang Li. CEED-VLA: Consistency Vision-Language-Action Model with Early-Exit Decoding, June 2025b. URL `http://arxiv.org/abs/2506.13725`. arXiv:2506.13725 [cs].

Wenxuan Song, Jiayi Chen, Pengxiang Ding, Han Zhao, Wei Zhao, Zhide Zhong, Zongyuan Ge, Zhijun Li, Donglin Wang, Lujia Wang, Jun Ma, and Haoang Li. PD-VLA: Accelerating Vision-Language-Action Model Integrated with Action Chunking via Parallel Decoding. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13162–13169, October 2025c. doi:10.1109/IROS60139.2025.11247519. URL `https://ieeexplore.ieee.org/document/11247519/`. ISSN: 2153-0866.

Denis Tarasov, Alexander Nikulin, Ilya Zisman, Albina Klepach, Nikita Lyubaykin, Andrei Polubarov, Alexander Derevyagin, and Vladislav Kurenkov. NinA: Normalizing Flows in Action. Training VLA Models with Normalizing Flows, October 2025. URL `http://arxiv.org/abs/2508.16845`. arXiv:2508.16845 [cs].

Paweł Budzianowski, Wesley Maa, Matthew Freed, Jingxiang Mo, Winston Hsiao, Aaron Xie, Tomasz Młoduchowski, Viraj Tipnis, and Benjamin Bolte. EdgeVLA: Efficient Vision-Language-Action Models, July 2025. URL `http://arxiv.org/abs/2507.14049`. arXiv:2507.14049 [cs].

Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Zhibin Tang, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, and Chaomin Shen. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*, 2025b. URL `https://ieeexplore.ieee.org/abstract/document/10900471/`.

Zhiying Du, Bei Liu, Yaobo Liang, Yichao Shen, Haidong Cao, Xiangyu Zheng, Zhiyuan Feng, Zuxuan Wu, Jiaolong Yang, and Yu-Gang Jiang. HiMoE-VLA: Hierarchical Mixture-of-Experts for Generalist Vision-Language-Action Policies, December 2025. URL `http://arxiv.org/abs/2512.05693`. arXiv:2512.05693 [cs].

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, Simon Alibert, Matthieu Cord, Thomas Wolf, and Remi Cadene. SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics, June 2025. URL `http://arxiv.org/abs/2506.01844`. arXiv:2506.01844 [cs].

Chenghao Liu, Jiachen Zhang, Chengxuan Li, Zhimu Zhou, Shixin Wu, Songfang Huang, and Huiling Duan. TTF-VLA: Temporal Token Fusion via Pixel-Attention Integration for Vision-Language-Action Models, November 2025b. URL `http://arxiv.org/abs/2508.19257`. arXiv:2508.19257 [cs].

Siyu Xu, Yunke Wang, Chenghao Xia, Dihao Zhu, Tao Huang, and Chang Xu. VLA-Cache: Efficient Vision-Language-Action Manipulation via Adaptive Token Caching, October 2025b. URL `http://arxiv.org/abs/2502.02175`. arXiv:2502.02175 [cs].

Ankit Goyal, Hugo Hadfield, Xuning Yang, Valts Blukis, and Fabio Ramos. VLA-0: Building State-of-the-Art VLAs with Zero Modification, October 2025. URL `http://arxiv.org/abs/2510.13054`. arXiv:2510.13054 [cs].

Puhao Li, Yingying Wu, Ziheng Xi, Wanlin Li, Yuzhe Huang, Zhiyuan Zhang, Yinghan Chen, Jianan Wang, Song-Chun Zhu, Tengyu Liu, and Siyuan Huang. ControlVLA: Few-shot Object-centric Adaptation for Pre-trained Vision-Language-Action Models, June 2025m. URL `http://arxiv.org/abs/2506.16211`. arXiv:2506.16211 [cs].

Yi Li, Yuquan Deng, Jesse Zhang, Joel Jang, Marius Memmel, Raymond Yu, Caelan Reed Garrett, Fabio Ramos, Dieter Fox, Anqi Li, Abhishek Gupta, and Ankit Goyal. HAMSTER: Hierarchical Action Models For Open-World Robot Manipulation, May 2025n. URL `http://arxiv.org/abs/2502.05485`. arXiv:2502.05485 [cs].

Zhixuan Liang, Yizhuo Li, Tianshuo Yang, Chengyue Wu, Sitong Mao, Tian Nian, Liuao Pei, Shunbo Zhou, Xiaokang Yang, Jiangmiao Pang, Yao Mu, and Ping Luo. Discrete Diffusion VLA: Bringing Discrete Diffusion to Action Decoding in Vision-Language-Action Policies, December 2025a. URL `http://arxiv.org/abs/2508.20072`. arXiv:2508.20072 [cs].

Wenqi Liang, Gan Sun, Yao He, Jiahua Dong, Suyan Dai, Ivan Laptev, Salman Khan, and Yang Cong. PixelVLA: Advancing Pixel-level Understanding in Vision-Language-Action Model, November 2025b. URL http://arxiv.org/abs/2511.01571. arXiv:2511.01571 [cs].

Xiangyi Wei, Haotian Zhang, Xinyi Cao, Siyu Xie, Weifeng Ge, Yang Li, and Changbo Wang. Audio-VLA: Adding Contact Audio Perception to Vision-Language-Action Model for Robotic Manipulation, November 2025. URL http://arxiv.org/abs/2511.09958. arXiv:2511.09958 [cs].

Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive Post-Training for Vision-Language-Action Models, May 2025b. URL http://arxiv.org/abs/2505.17016. arXiv:2505.17016 [cs].

Chaofan Zhang, Peng Hao, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. VTLA: Vision-Tactile-Language-Action Model with Preference Learning for Insertion Manipulation, May 2025g. URL http://arxiv.org/abs/2505.09577. arXiv:2505.09577 [cs].

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, Ya-Qin Zhang, Jiangmiao Pang, Jingjing Liu, Tai Wang, and Xianyuan Zhan. X-VLA: Soft-Prompted Transformer as Scalable Cross-Embodiment Vision-Language-Action Model, October 2025b. URL http://arxiv.org/abs/2510.10274. arXiv:2510.10274 [cs].

Yuxia Fu, Zhizhen Zhang, Yuqi Zhang, Zijian Wang, Zi Huang, and Yadan Luo. MergeVLA: Cross-Skill Model Merging Toward a Generalist Vision-Language-Action Agent, November 2025. URL http://arxiv.org/abs/2511.18810. arXiv:2511.18810 [cs].

Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea Open-World Dataset and G0 Dual-System VLA Model, August 2025c. URL http://arxiv.org/abs/2509.00576. arXiv:2509.00576 [cs].

Rushuai Yang, Hangxing Wei, Ran Zhang, Zhiyuan Feng, Xiaoyu Chen, Tong Li, Chuheng Zhang, Li Zhao, Jiang Bian, Xiu Su, and Yi Chen. Beyond Human Demonstrations: Diffusion-Based Reinforcement Learning to Generate Data for VLA Training, September 2025b. URL http://arxiv.org/abs/2509.19752. arXiv:2509.19752 [cs].

Guo Ye, Zexi Zhang, Xu Zhao, Shang Wu, Haoran Lu, Shihan Lu, and Han Liu. Learning to Feel the Future: DreamTacVLA for Contact-Rich Manipulation, December 2025b. URL http://arxiv.org/abs/2512.23864. arXiv:2512.23864 [cs].

Zhongyi Zhou, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. ChatVLA-2: Vision-Language-Action Model with Open-World Embodied Reasoning from Pretrained Knowledge, May 2025a. URL http://arxiv.org/abs/2505.21906. arXiv:2505.21906 [cs].

Yiguo Fan, Pengxiang Ding, Shuanghao Bai, Xinyang Tong, Yuyang Zhu, Hongchao Lu, Fengqi Dai, Wei Zhao, Yang Liu, Siteng Huang, Zhaoxin Fan, Badong Chen, and Donglin Wang. Long-VLA: Unleashing Long-Horizon Capability of Vision Language Action Model for Robot Manipulation, August 2025a. URL http://arxiv.org/abs/2508.19958. arXiv:2508.19958 [cs].

Cunxin Fan, Xiaosong Jia, Yihang Sun, Yixiao Wang, Jianglan Wei, Ziyang Gong, Xiangyu Zhao, Masayoshi Tomizuka, Xue Yang, Junchi Yan, and Mingyu Ding. Interleave-VLA: Enhancing Robot Manipulation with Interleaved Image-Text Instructions, October 2025b. URL http://arxiv.org/abs/2505.02152. arXiv:2505.02152 [cs].

Hanqing Liu, Jiahuan Long, Junqi Wu, Jiacheng Hou, Huili Tang, Tingsong Jiang, Weien Zhou, and Wen Yao. Eva-VLA: Evaluating Vision-Language-Action Models' Robustness Under Real-World Physical Variations, September 2025c. URL http://arxiv.org/abs/2509.18953. arXiv:2509.18953 [cs].

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: In-depth Robustness Analysis of Vision-Language-Action Models, December 2025. URL http://arxiv.org/abs/2510.13626. arXiv:2510.13626 [cs].

Zhuo Li, Junjia Liu, Zhipeng Dong, Tao Teng, Quentin Rouxel, Darwin Caldwell, and Fei Chen. Towards Deploying VLA without Fine-Tuning: Plug-and-Play Inference-Time VLA Policy Steering via Embodied Evolutionary Diffusion, November 2025o. URL http://arxiv.org/abs/2511.14178. arXiv:2511.14178 [cs].

Wenxuan Song, Ziyang Zhou, Han Zhao, Jiayi Chen, Pengxiang Ding, Haodong Yan, Yuxin Huang, Feilong Tang, Donglin Wang, and Haoang Li. ReconVLA: Reconstructive Vision-Language-Action Model as Effective Robot Perceiver, August 2025d. URL http://arxiv.org/abs/2508.10333. arXiv:2508.10333 [cs].

Rokas Bendikas, Daniel Dijkman, Markus Peschl, Sanjay Haresh, and Pietro Mazzaglia. Focusing on What Matters: Object-Agent-centric Tokenization for Vision Language Action models, September 2025. URL http://arxiv.org/abs/2509.23655. arXiv:2509.23655 [cs].

Asher J. Hancock, Allen Z. Ren, and Anirudha Majumdar. Run-time observation interventions make vision-language-action models more visually robust. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9499–9506. IEEE, 2025b. URL https://ieeexplore.ieee.org/abstract/document/11128017/.

Han Zhao, Jiaxuan Zhang, Wenxuan Song, Pengxiang Ding, and Donglin Wang. VLA^2: Empowering Vision-Language-Action Models with an Agentic Framework for Unseen Concept Manipulation, October 2025c. URL http://arxiv.org/abs/2510.14902. arXiv:2510.14902 [cs].

Daria Pugacheva, Andrey Moskalenko, Denis Shepelev, Andrey Kuznetsov, Vlad Shakhuro, and Elena Tutubalina. Bring the Apple, Not the Sofa: Impact of Irrelevant Context in Embodied AI Commands on VLA Models, October 2025. URL http://arxiv.org/abs/2510.07067. arXiv:2510.07067 [cs].

Yang Zhang, Chenwei Wang, Ouyang Lu, Yuan Zhao, Yunfei Ge, Zhenglong Sun, Xiu Li, Chi Zhang, Chenjia Bai, and Xuelong Li. Align-Then-stEer: Adapting the Vision-Language Action Models through Unified Latent Guidance, September 2025h. URL http://arxiv.org/abs/2509.02055. arXiv:2509.02055 [cs].

Yifan Yang, Zhixiang Duan, Tianshi Xie, Fuyu Cao, Pinxi Shen, Peili Song, Piaopiao Jin, Guokang Sun, Shaoqing Xu, Yangwei You, and Jingtai Liu. FPC-VLA: A Vision-Language-Action Framework with a Supervisor for Failure Prediction and Correction, December 2025c. URL http://arxiv.org/abs/2509.04018. arXiv:2509.04018 [cs].

Yuhan Hao, Zhengning Li, Lei Sun, Weilong Wang, Naixin Yi, Sheng Song, Caihong Qin, Mofan Zhou, Yifei Zhan, and Xianpeng Lang. DriveAction: A Benchmark for Exploring Human-like Driving Decisions in VLA Models, September 2025. URL http://arxiv.org/abs/2506.05667. arXiv:2506.05667 [cs].

Zhenlong Yuan, Chengxuan Qian, Jing Tang, Rui Chen, Zijian Song, Lei Sun, Xiangxiang Chu, Yujun Cai, Dapeng Zhang, and Shuo Li. AutoDrive-R^2: Incentivizing Reasoning and Self-Reflection Capacity for VLA Model in Autonomous Driving, December 2025a. URL http://arxiv.org/abs/2509.01944. arXiv:2509.01944 [cs].

Ziang Guo and Zufeng Zhang. VDRive: Leveraging Reinforced VLA and Diffusion Policy for End-to-end Autonomous Driving, October 2025. URL http://arxiv.org/abs/2510.15446. arXiv:2510.15446 [cs].

Taowen Wang, Cheng Han, James Liang, Wenhao Yang, Dongfang Liu, Luna Xinyu Zhang, Qifan Wang, Jiebo Luo, and Ruixiang Tang. Exploring the adversarial vulnerabilities of vision-language-action models in robotics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6948–6958, 2025f. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_Exploring_the_Adversarial_Vulnerabilities_of_Vision-Language-Action_Models_in_Robotics_ICCV_2025_paper.html.

Titong Jiang, Xuefeng Jiang, Yuan Ma, Xin Wen, Bailin Li, Kun Zhan, Peng Jia, Yahui Liu, Sheng Sun, and Xianpeng Lang. The Better You Learn, The Smarter You Prune: Towards Efficient Vision-language-action Models via Differentiable Token Pruning, September 2025d. URL http://arxiv.org/abs/2509.12594. arXiv:2509.12594 [cs].

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial Forcing: Implicit Spatial Representation Alignment for Vision-language-action Model, October 2025p. URL http://arxiv.org/abs/2510.12276. arXiv:2510.12276 [cs].

Vineet Bhat, Yu-Hsiang Lan, Prashanth Krishnamurthy, Ramesh Karri, and Farshad Khorrami. 3D CAVLA: Leveraging Depth and 3D Context to Generalize Vision Language Action Models for Unseen Tasks, May 2025. URL http://arxiv.org/abs/2505.05800. arXiv:2505.05800 [cs].

Yicheng Feng, Wanpeng Zhang, Ye Wang, Hao Luo, Haoqi Yuan, Sipeng Zheng, and Zongqing Lu. Spatial-Aware VLA Pretraining through Visual-Physical Alignment from Human Videos, December 2025a. URL http://arxiv.org/abs/2512.13080. arXiv:2512.13080 [cs].

Jinming Li, Yichen Zhu, Zhibin Tang, Junjie Wen, Minjie Zhu, Xiaoyu Liu, Chengmeng Li, Ran Cheng, Yaxin Peng, and Yan Peng. Coa-vla: Improving vision-language-action models via visual-text chain-of-affordance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9759–9769, 2025q. URL https://openaccess.thecvf.com/content/ICCV2025/html/Li_CoA-VLA_Improving_Vision-Language-Action_Models_via_Visual-Text_Chain-of-Affordance_ICCV_2025_paper.html.

Zhen Fang, Zhuoyang Liu, Jiaming Liu, Hao Chen, Yu Zeng, Shiting Huang, Zehui Chen, Lin Chen, Shanghang Zhang, and Feng Zhao. DualVLA: Building a Generalizable Embodied Agent via Partial Decoupling of Reasoning and Action, November 2025b. URL http://arxiv.org/abs/2511.22134. arXiv:2511.22134 [cs].

Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, Chengkai Hou, Mengdi Zhao, KC alex Zhou, Pheng-Ann Heng, and Shanghang Zhang. HybridVLA: Collaborative Diffusion and Autoregression in a Unified Vision-Language-Action Model, June 2025d. URL http://arxiv.org/abs/2503.10631. arXiv:2503.10631 [cs].

Piaopiao Jin, Qi Wang, Guokang Sun, Ziwen Cai, Pinjia He, and Yangwei You. Dual-Actor Fine-Tuning of VLA Models: A Talk-and-Tweak Human-in-the-Loop Approach, September 2025. URL `http://arxiv.org/abs/2509.13774`. arXiv:2509.13774 [cs].

Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Wenhao Zhang, Heming Cui, Zhizheng Zhang, and He Wang. GraspVLA: a Grasping Foundation Model Pre-trained on Billion-scale Synthetic Action Data, August 2025. URL `http://arxiv.org/abs/2505.03233`. arXiv:2505.03233 [cs].

Shunlei Li, Jin Wang, Rui Dai, Wanyu Ma, Wing Yin Ng, Yingbai Hu, and Zheng Li. Robonurse-vla: Robotic scrub nurse system based on vision-language-action model. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3986–3993. IEEE, 2025r. URL `https://ieeexplore.ieee.org/abstract/document/11246030/`.

Artem Lykov, Valerii Serpiva, Muhammad Haris Khan, Oleg Sautenkov, Artyom Myshlyaev, Grik Tadevosyan, Yasheerah Yaqoot, and Dzmitry Tsetserukou. CognitiveDrone: A VLA Model and Evaluation Benchmark for Real-Time Cognitive Task Solving and Reasoning in UAVs, March 2025. URL `http://arxiv.org/abs/2503.01378`. arXiv:2503.01378 [cs].

Valerii Serpiva, Artem Lykov, Artyom Myshlyaev, Muhammad Haris Khan, Ali Alridha Abdulkarim, Oleg Sautenkov, and Dzmitry Tsetserukou. RaceVLA: VLA-based Racing Drone Navigation with Human-like Behaviour, March 2025. URL `http://arxiv.org/abs/2503.02572`. arXiv:2503.02572 [cs].

Muyao Li, Zihao Wang, Kaichen He, Xiaojian Ma, and Yitao Liang. JARVIS-VLA: Post-Training Large-Scale Vision Language Models to Play Visual Games with Keyboards and Mouse, September 2025s. URL `http://arxiv.org/abs/2503.16365`. arXiv:2503.16365 [cs].

Pengxiang Ding, Jianfei Ma, Xinyang Tong, Binghong Zou, Xinxin Luo, Yiguo Fan, Ting Wang, Hongchao Lu, Panzhong Mo, Jinxin Liu, Yuefan Wang, Huaicheng Zhou, Wenshuo Feng, Jiacheng Liu, Siteng Huang, and Donglin Wang. Humanoid-VLA: Towards Universal Humanoid Control with Visual Integration, February 2025b. URL `http://arxiv.org/abs/2502.14795`. arXiv:2502.14795 [cs].

Peng Chen, Pi Bu, Yingyao Wang, Xinyi Wang, Ziming Wang, Jie Guo, Yingxiu Zhao, Qi Zhu, Jun Song, Siran Yang, Jiamang Wang, and Bo Zheng. CombatVLA: An Efficient Vision-Language-Action Model for Combat Tasks in 3D Action Role-Playing Games, January 2026b. URL `http://arxiv.org/abs/2503.09527`. arXiv:2503.09527 [cs].

Zhongyi Zhou, Yichen Zhu, Minjie Zhu, Junjie Wen, Ning Liu, Zhiyuan Xu, Weibin Meng, Yaxin Peng, Chaomin Shen, Feifei Feng, and Yi Xu. ChatVLA: Unified Multimodal Understanding and Robot Control with Vision-Language-Action Model. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5377–5395, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi:10.18653/v1/2025.emnlp-main.273. URL `https://aclanthology.org/2025.emnlp-main.273/`.

Meng Li, Zhen Zhao, Zhengping Che, Fei Liao, Kun Wu, Zhiyuan Xu, Pei Ren, Zhao Jin, Ning Liu, and Jian Tang. SwitchVLA: Execution-Aware Task Switching for Vision-Language-Action Models, June 2025t. URL `http://arxiv.org/abs/2506.03574`. arXiv:2506.03574 [cs].

Noriaki Hirose, Catherine Glossop, Dhruv Shah, and Sergey Levine. OmniVLA: An Omni-Modal Vision-Language-Action Model for Robot Navigation, September 2025. URL `http://arxiv.org/abs/2509.19480`. arXiv:2509.19480 [cs].

Heyu Guo, Shanmu Wang, Ruichun Ma, Shiqi Jiang, Yasaman Ghasempour, Omid Abari, Baining Guo, and Lili Qiu. OmniVLA: Physically-Grounded Multimodal VLA with Unified Multi-Sensor Perception for Robotic Manipulation, November 2025c. URL `http://arxiv.org/abs/2511.01210`. arXiv:2511.01210 [cs].

Huiwon Jang, Sihyun Yu, Heeseung Kwon, Hojin Jeon, Younggyo Seo, and Jinwoo Shin. ContextVLA: Vision-Language-Action Model with Amortized Multi-Frame Context, October 2025. URL `http://arxiv.org/abs/2510.04246`. arXiv:2510.04246 [cs].

Jiyeon Koo, Taewan Cho, Hyunjoon Kang, Eunseom Pyo, Tae Gyun Oh, Taeryang Kim, and Andrew Jaeyong Choi. RetoVLA: Reusing Register Tokens for Spatial Reasoning in Vision-Language-Action Models, September 2025. URL `http://arxiv.org/abs/2509.21243`. arXiv:2509.21243 [cs].

Sombit Dey, Jan-Nico Zaech, Nikolay Nikolov, Luc Van Gool, and Danda Pani Paudel. Revla: Reverting visual domain limitation of robotic foundation models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8679–8686. IEEE, 2025. URL `https://ieeexplore.ieee.org/abstract/document/11128823/`.

Sicong Jiang, Zilin Huang, Kangan Qian, Ziang Luo, Tianze Zhu, Yang Zhong, Yihong Tang, Menglin Kong, Yunlong Wang, Siwen Jiao, Hao Ye, Zihao Sheng, Xin Zhao, Tuopu Wen, Zheng Fu, Sikai Chen, Kun Jiang, Diange Yang, Seongjin Choi, and Lijun Sun. A Survey on Vision-Language-Action Models for Autonomous Driving, June 2025e. URL http://arxiv.org/abs/2506.24044. arXiv:2506.24044 [cs].

Wenlong Liang, Rui Zhou, Yang Ma, Bing Zhang, Songlin Li, Yijia Liao, and Ping Kuang. Large Model Empowered Embodied AI: A Survey on Decision-Making and Embodied Learning, August 2025c. URL http://arxiv.org/abs/2508.10399. arXiv:2508.10399 [cs] version: 1.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge, November 2022. URL http://arxiv.org/abs/2206.08853. arXiv:2206.08853 [cs].

Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S. Sukhatme. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4):10049–10056, 2022. URL https://ieeexplore.ieee.org/abstract/document/9837390/.

Xudong Guo, Kaixuan Huang, Jiale Liu, Wenhui Fan, Natalia Vélez, Qingyun Wu, Huazheng Wang, Thomas L. Griffiths, and Mengdi Wang. Embodied LLM Agents Learn to Cooperate in Organized Teams, May 2024. URL http://arxiv.org/abs/2403.12482. arXiv:2403.12482 [cs].

Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. Embodied multi-modal agent trained by an llm from a parallel textworld. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26275–26285, 2024b. URL http://openaccess.thecvf.com/content/CVPR2024/html/Yang_Embodied_Multi-Modal_Agent_trained_by_an_LLM_from_a_Parallel_CVPR_2024_paper.html.

Ishita Dasgupta, Christine Kaeser-Chen, Kenneth Marino, Arun Ahuja, Sheila Babayan, Felix Hill, and Rob Fergus. Collaborating with language models for embodied reasoning, February 2023. URL http://arxiv.org/abs/2302.00763. arXiv:2302.00763 [cs].

Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. Embodied Task Planning with Large Language Models, July 2023b. URL http://arxiv.org/abs/2307.01848. arXiv:2307.01848 [cs].

Haozhe Xie, Beichen Wen, Jiarui Zheng, Zhaoxi Chen, Fangzhou Hong, Haiwen Diao, and Ziwei Liu. DynamicVLA: A Vision-Language-Action Model for Dynamic Object Manipulation, January 2026a. URL http://arxiv.org/abs/2601.22153. arXiv:2601.22153 [cs].

Jianke Zhang, Xiaoyu Chen, Qiuyue Wang, Mingsheng Li, Yanjiang Guo, Yucheng Hu, Jiajun Zhang, Shuai Bai, Junyang Lin, and Jianyu Chen. VLM4VLA: Revisiting Vision-Language-Models in Vision-Language-Action Models, January 2026b. URL http://arxiv.org/abs/2601.03309. arXiv:2601.03309 [cs].

Zhaoshu Yu, Bo Wang, Pengpeng Zeng, Haonan Zhang, Ji Zhang, Zheng Wang, Lianli Gao, Jingkuan Song, Nicu Sebe, and Heng Tao Shen. A Survey on Efficient Vision-Language-Action Models, February 2026b. URL http://arxiv.org/abs/2510.24795. arXiv:2510.24795 [cs].

Trevor J. Pugh, Jonathan L. Bell, Jeff P. Bruce, Gary J. Doherty, Matthew Galvin, Michelle F. Green, Haley Hunter-Zinck, Priti Kumari, Michele L. Lenoue-Newton, and Marilyn M. Li. AACR Project GENIE: 100,000 cases and beyond. *Cancer Discovery*, 12(9):2044–2057, 2022. URL https://aacrjournals.org/cancerdiscovery/article-abstract/12/9/2044/708766.

Samir Yitzhak Gadre, Kiana Ehsani, Shuran Song, and Roozbeh Mottaghi. Continuous scene representations for embodied ai. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14849–14859, 2022. URL http://openaccess.thecvf.com/content/CVPR2022/html/Gadre_Continuous_Scene_Representations_for_Embodied_AI_CVPR_2022_paper.html.

Fengli Xu, Jun Zhang, Chen Gao, Jie Feng, and Yong Li. Urban Generative Intelligence (UGI): A Foundational Platform for Agents in Embodied City Environment, December 2023. URL http://arxiv.org/abs/2312.11813. arXiv:2312.11813 [cs].

Yining Hong, Zishuo Zheng, Peihao Chen, Yian Wang, Junyan Li, and Chuang Gan. Multiply: A multisensory object-centric embodied large language model in 3d world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26406–26416, 2024. URL http://openaccess.thecvf.com/content/CVPR2024/html/Hong_MultiPLY_A_Multisensory_Object-Centric_Embodied_Large_Language_Model_in_3D_CVPR_2024_paper.html.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An Embodied Generalist Agent in 3D World, May 2024a. URL http://arxiv.org/abs/2311.12871. arXiv:2311.12871 [cs].

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022. URL https://ieeexplore.ieee.org/abstract/document/9687596/.

Changyu Liu, Yiyang Liu, Taowen Wang, Qiao Zhuang, James Chenhao Liang, Wenhao Yang, Renjing Xu, Qifan Wang, Dongfang Liu, and Cheng Han. On-the-Fly VLA Adaptation via Test-Time Reinforcement Learning, January 2026a. URL http://arxiv.org/abs/2601.06748. arXiv:2601.06748 [cs].

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics Transformer for Real-World Control at Scale, August 2023. URL http://arxiv.org/abs/2212.06817. arXiv:2212.06817 [cs].

Guanxing Lu, Wenkai Guo, Chubin Zhang, Yuheng Zhou, Haonan Jiang, Zifeng Gao, Yansong Tang, and Ziwei Wang. VLA-RL: Towards Masterful and General Robotic Manipulation with Scalable Reinforcement Learning, May 2025. URL http://arxiv.org/abs/2505.18719. arXiv:2505.18719 [cs].

Baorui Peng, Wenyao Zhang, Liang Xu, Zekun Qi, Jiazhao Zhang, Hongsi Liu, Wenjun Zeng, and Xin Jin. ReWorld: Multi-Dimensional Reward Modeling for Embodied World Models, January 2026. URL http://arxiv.org/abs/2601.12428. arXiv:2601.12428 [cs].

Nedko Savov, Naser Kazemi, Mohammad Mahdi, Danda Pani Paudel, Xi Wang, and Luc Van Gool. Exploration-Driven Generative Interactive Environments. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27597–27607, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Savov_Exploration-Driven_Generative_Interactive_Environments_CVPR_2025_paper.html.

Zhaofeng Hu, Hongrui Yu, Vaidhyanathan Chandramouli, and Ci-Jyun Liang. Sample-Efficient Robot Skill Learning for Construction Tasks: Benchmarking Hierarchical Reinforcement Learning and Vision-Language-Action VLA Model, December 2025a. URL http://arxiv.org/abs/2512.14031. arXiv:2512.14031 [cs].

Feng Xu, Guangyao Zhai, Xin Kong, Tingzhong Fu, Daniel F. N. Gordon, Xueli An, and Benjamin Busam. STARE-VLA: Progressive Stage-Aware Reinforcement for Fine-Tuning Vision-Language-Action Models, December 2025c. URL http://arxiv.org/abs/2512.05107. arXiv:2512.05107 [cs].

Zechen Bai, Chen Gao, and Mike Zheng Shou. EVOLVE-VLA: Test-Time Training from Environment Feedback for Vision-Language-Action Models, December 2025. URL http://arxiv.org/abs/2512.14666. arXiv:2512.14666 [cs].

Cyrus Neary, Omar G. Younis, Artur Kuramshin, Ozgur Aslan, and Glen Berseth. Improving Pre-Trained Vision-Language-Action Policies with Model-Based Search, November 2025. URL http://arxiv.org/abs/2508.12211. arXiv:2508.12211 [cs].

Jiayi Chen, Wenxuan Song, Pengxiang Ding, Ziyang Zhou, Han Zhao, Feilong Tang, Donglin Wang, and Haoang Li. Unified Diffusion VLA: Vision-Language-Action Model via Joint Discrete Denoising Diffusion Process, November 2025a. URL http://arxiv.org/abs/2511.01718. arXiv:2511.01718 [cs].

Tobias Jülg, Wolfram Burgard, and Florian Walter. Refined Policy Distillation: From VLA Generalists to RL Experts, August 2025. URL http://arxiv.org/abs/2503.05833. arXiv:2503.05833 [cs].

Kohei Sendai, Maxime Alvarez, Tatsuya Matsushima, Yutaka Matsuo, and Yusuke Iwasawa. Leave No Observation Behind: Real-time Correction for VLA Action Chunks, September 2025. URL http://arxiv.org/abs/2509.23224. arXiv:2509.23224 [cs].

Si-Cheng Wang, Tian-Yu Xiang, Xiao-Hu Zhou, Mei-Jiang Gui, Xiao-Liang Xie, Shi-Qi Liu, Shuang-Yi Wang, Ao-Qun Jin, and Zeng-Guang Hou. VLA Model Post-Training via Action-Chunked PPO and Self Behavior Cloning, September 2025g. URL http://arxiv.org/abs/2509.25718. arXiv:2509.25718 [cs].

Shresth Grover, Akshay Gopalkrishnan, Bo Ai, Henrik I. Christensen, Hao Su, and Xuanlin Li. Enhancing Generalization in Vision-Language-Action Models by Preserving Pretrained Representations, September 2025. URL http://arxiv.org/abs/2509.11417. arXiv:2509.11417 [cs].

Hongyin Zhang, Shuo Zhang, Junxi Jin, Qixin Zeng, Runze Li, and Donglin Wang. RobustVLA: Robustness-Aware Reinforcement Post-Training for Vision-Language-Action Models, December 2025i. URL http://arxiv.org/abs/2511.01331. arXiv:2511.01331 [cs].

Hongyin Zhang, Shiyuan Zhang, Junxi Jin, Qixin Zeng, Yifan Qiao, Hongchao Lu, and Donglin Wang. Balancing Signal and Variance: Adaptive Offline RL Post-Training for VLA Flow Models, September 2025j. URL http://arxiv.org/abs/2509.04063. arXiv:2509.04063 [cs].

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success, April 2025. URL http://arxiv.org/abs/2502.19645. arXiv:2502.19645 [cs].

Naser Kazemi, Nedko Savov, Danda Paudel, and Luc Van Gool. Learning Generative Interactive Environments By Trained Agent Exploration, October 2024. URL http://arxiv.org/abs/2409.06445. arXiv:2409.06445 [cs].

Chia-Yu Hung, Navonil Majumder, Haoyuan Deng, Liu Renhang, Yankang Ang, Amir Zadeh, Chuan Li, Dorien Herremans, Ziwei Wang, and Soujanya Poria. NORA-1.5: A Vision-Language-Action Model Trained using World Model- and Action-based Preference Rewards, November 2025b. URL http://arxiv.org/abs/2511.14659. arXiv:2511.14659 [cs].

Tian-Yu Xiang, Ao-Qun Jin, Xiao-Hu Zhou, Mei-Jiang Gui, Xiao-Liang Xie, Shi-Qi Liu, Shuang-Yi Wang, Sheng-Bin Duan, Fu-Chao Xie, Wen-Kai Wang, Si-Cheng Wang, Ling-Yun Li, Tian Tu, and Zeng-Guang Hou. Parallels Between VLA Model Post-Training and Human Motor Learning: Progress, Challenges, and Trends, January 2026. URL http://arxiv.org/abs/2506.20966. arXiv:2506.20966 [cs].

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, and Ajinkya Jain. 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)*, pages 6892–6903. IEEE, 2024. URL https://ieeexplore.ieee.org/abstract/document/10611477/.

Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. Towards Generalist Robot Policies: What Matters in Building Vision-Language-Action Models, December 2024b. URL http://arxiv.org/abs/2412.14058. arXiv:2412.14058 [cs].

Gabriel Sarch, Yue Wu, Michael J. Tarr, and Katerina Fragkiadaki. Open-Ended Instructable Embodied Agents with Memory-Augmented Large Language Models, November 2023. URL http://arxiv.org/abs/2310.15127. arXiv:2310.15127 [cs].

Vishnu Sashank Dorbala, James F. Mullen, and Dinesh Manocha. Can an embodied agent find your "cat-shaped mugpi llm-based zero-shot object navigation. *IEEE Robotics and Automation Letters*, 9(5):4083–4090, 2023. URL https://ieeexplore.ieee.org/abstract/document/10373065/.

Wenlong Huang, Fei Xia, Dhruv Shah, Danny Driess, Andy Zeng, Yao Lu, Pete Florence, Igor Mordatch, Sergey Levine, Karol Hausman, and Brian Ichter. Grounded Decoding: Guiding Text Generation with Grounded Models for Embodied Agents, December 2023a. URL http://arxiv.org/abs/2303.00855. arXiv:2303.00855 [cs].

Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni, and Robert West. GenIE: Generative information extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4626–4643, 2022. URL https://aclanthology.org/2022.naacl-main.342/.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. Genie: Achieving Human Parity in Content-Grounded Datasets Generation, January 2024. URL http://arxiv.org/abs/2401.14367. arXiv:2401.14367 [cs].

Yunsong Zhou, Linyan Huang, Qingwen Bu, Jia Zeng, Tianyu Li, Hang Qiu, Hongzi Zhu, Minyi Guo, Yu Qiao, and Hongyang Li. Embodied Understanding of Driving Scenarios. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision piECCV 2024*, volume 15120, pages 129–148. Springer Nature Switzerland, Cham, 2025c. ISBN 978-3-031-73032-0 978-3-031-73033-7. doi:10.1007/978-3-031-73033-7_8. URL https://link.springer.com/10.1007/978-3-031-73033-7_8. Series Title: Lecture Notes in Computer Science.

Jonathan Salzer and Arnoud Visser. Bringing the RT-1-X Foundation Model to a SCARA robot, September 2024. URL http://arxiv.org/abs/2409.03299. arXiv:2409.03299 [cs].

Masaki Yoshikawa, Hiroshi Ito, and Tetsuya Ogata. Achieving Faster and More Accurate Operation of Deep Predictive Learning, August 2024. URL http://arxiv.org/abs/2408.10231. arXiv:2408.10231 [cs].

Delong Chen, Theo Moutakanni, Willy Chung, Yejin Bang, Ziwei Ji, Allen Bolourchi, and Pascale Fung. Planning with Reasoning using Vision Language World Model, September 2025b. URL http://arxiv.org/abs/2509.02722. arXiv:2509.02722 [cs].

Jiayuan Gu, Sean Kirmani, Paul Wohlhart, Yao Lu, Montserrat Gonzalez Arenas, Kanishka Rao, Wenhao Yu, Chuyuan Fu, Keerthana Gopalakrishnan, Zhuo Xu, Priya Sundaresan, Peng Xu, Hao Su, Karol Hausman, Chelsea Finn, Quan

Vuong, and Ted Xiao. RT-Trajectory: Robotic Task Generalization via Hindsight Trajectory Sketches, November 2023. URL http://arxiv.org/abs/2311.01977. arXiv:2311.01977 [cs].

Isabel Leal, Krzysztof Choromanski, Deepali Jain, Avinava Dubey, Jake Varley, Michael Ryoo, Yao Lu, Frederick Liu, Vikas Sindhwani, and Quan Vuong. Sara-rt: Scaling up robotics transformers with self-adaptive robust attention. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6920–6927. IEEE, 2024. URL https://ieeexplore.ieee.org/abstract/document/10611597/.

Danny Driess, Jost Tobias Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Z. Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, and Sergey Levine. Knowledge Insulating Vision-Language-Action Models: Train Fast, Run Fast, Generalize Better, May 2025. URL http://arxiv.org/abs/2505.23705. arXiv:2505.23705 [cs].

Minghui Lin, Pengxiang Ding, Shu Wang, Zifeng Zhuang, Yang Liu, Xinyang Tong, Wenxuan Song, Shangke Lyu, Siteng Huang, and Donglin Wang. HiF-VLA: Hindsight, Insight and Foresight through Motion Representation for Vision-Language-Action Models, December 2025a. URL http://arxiv.org/abs/2512.09928. arXiv:2512.09928 [cs].

Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, Jianyu Chen, and Jiang Bian. villa-X: Enhancing Latent Action Modeling in Vision-Language-Action Models, September 2025c. URL http://arxiv.org/abs/2507.23682. arXiv:2507.23682 [cs].

John Won, Kyungmin Lee, Huiwon Jang, Dongyoung Kim, and Jinwoo Shin. Dual-Stream Diffusion for World-Model Augmented Vision-Language-Action Model, November 2025. URL http://arxiv.org/abs/2510.27607. arXiv:2510.27607 [cs].

Partha Pratim Ray. Physical AI: Bridging the Sim-to-Real Divide Toward Embodied, Ethical, and Autonomous Intelligence, November 2025. URL https://www.techrxiv.org/users/913189/articles/1355704-physical-ai-bridging-the-sim-to-real-divide-toward-embodied-ethical-and-autonomous-intelligen
b33378a87ff47426d47b47d22f9ff745894b789d.

Qi Zhang, Shaopeng Zhai, Shengzhe Zhang, Litao Liu, Fuxian Huang, Zhang HaoranECNU, Ming Zhou, and Jiangmiao Pang. VLAC: A Generalist Action-Critic Model via Pair-wise Progress Understanding. URL https://openreview.net/forumpiid=PmYX0XiQQ0.

Borong Zhang, Yuhao Zhang, Jiaming Ji, Yingshan Lei, Josef Dai, Yuanpei Chen, and Yaodong Yang. SafeVLA: Towards Safety Alignment of Vision-Language-Action Model via Constrained Learning, November 2025k. URL http://arxiv.org/abs/2503.03480. arXiv:2503.03480 [cs].

Jason Jabbour, Dong-Ki Kim, Max Smith, Jay Patrikar, Radhika Ghosal, Youhui Wang, Ali Agha, Vijay Janapa Reddi, and Shayegan Omidshafiei. Don't Run with Scissors: Pruning Breaks VLA Models but They Can Be Recovered, October 2025. URL http://arxiv.org/abs/2510.08464. arXiv:2510.08464 [cs].

Zhijie Wang, Zhehua Zhou, Jiayang Song, Yuheng Huang, Zhan Shu, and Lei Ma. VLATest: Testing and Evaluating Vision-Language-Action Models for Robotic Manipulation. *Proceedings of the ACM on Software Engineering*, 2 (FSE):1615–1638, June 2025h. ISSN 2994-970X. doi:10.1145/3729343. URL https://dl.acm.org/doi/10.1145/3729343.

JunHao Xie. Emerging Paradigms in Deep Learning: Efficient Sequence Models, Visual Autoregressive Generation, World Models, and Diffusion Theory. URL https://www.researchgate.net/profile/Junhao-Xie-7/publication/399277627_Emerging_Paradigms_in_Deep_Learning_Efficient_Sequence_Models_Visual_Autoregressive_Generation_World_Models_and_Diffusion_Theory_-A_Continuously_Updated_Survey_for_UG_Academic_Training/links/6955f4f627359023a01273ee/Emerging-Paradigms-in-Deep-Learning-Efficient-Sequence-Models-Visual-Autoregressive-Generation-World-
pdf.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, August 2022. URL http://arxiv.org/abs/2204.01691. arXiv:2204.01691 [cs].

Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X. Lee, Maria Bauza, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, Antoine Laurens, Claudio Fantacci, Valentin Dalibard, Martina Zambelli, Murilo

Martins, Rugile Pevceviciute, Michiel Blokzijl, Misha Denil, Nathan Batchelor, Thomas Lampe, Emilio Parisotto, Konrad Żołna, Scott Reed, Sergio Gómez Colmenarejo, Jon Scholz, Abbas Abdolmaleki, Oliver Groth, Jean-Baptiste Regli, Oleg Sushkov, Tom Rothörl, José Enrique Chen, Yusuf Aytar, Dave Barker, Joy Ortiz, Martin Riedmiller, Jost Tobias Springenberg, Raia Hadsell, Francesco Nori, and Nicolas Heess. RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation, December 2023. URL http://arxiv.org/abs/2306.11706. arXiv:2306.11706 [cs].

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation, October 2024. URL http://arxiv.org/abs/2410.06158. arXiv:2410.06158 [cs].

Yuhui Chen, Shuai Tian, Shugao Liu, Yingting Zhou, Haoran Li, and Dongbin Zhao. ConRFT: A Reinforced Fine-tuning Method for VLA Models via Consistency Policy, April 2025d. URL http://arxiv.org/abs/2502.05450. arXiv:2502.05450 [cs].

Haonan Chen, Jingxiang Guo, Bangjun Wang, Tianrui Zhang, Xuchuan Huang, Boren Zheng, Yiwen Hou, Chenrui Tie, Jiajun Deng, and Lin Shao. Goal-VLA: Image-Generative VLMs as Object-Centric World Models Empowering Zero-shot Robot Manipulation. URL https://nus-lins-lab.github.io/goalvlaweb/static/data/paper.pdf.

Xinyi Chen, Yilun Chen, Yanwei Fu, Ning Gao, Jiaya Jia, Weiyang Jin, Hao Li, Yao Mu, Jiangmiao Pang, Yu Qiao, Yang Tian, Bin Wang, Bolun Wang, Fangjing Wang, Hanqing Wang, Tai Wang, Ziqin Wang, Xueyuan Wei, Chao Wu, Shuai Yang, Jinhui Ye, Junqiu Yu, Jia Zeng, Jingjing Zhang, Jinyu Zhang, Shi Zhang, Feng Zheng, Bowen Zhou, and Yangkun Zhu. InternVLA-M1: A Spatially Guided Vision-Language-Action Framework for Generalist Robot Policy, October 2025e. URL http://arxiv.org/abs/2510.13778. arXiv:2510.13778 [cs].

Alessandra Corsi, Joseph W. Lazio, Stefi Baum, Simona Giacintucci, George Heald, Patricia Henning, Ian Heywood, Daisuke Iono, Megan Johnson, Michael T. Lam, Adam Leroy, Laurent Loinard, Leslie Looney, Lynn Matthews, Ned Molter, Eric Murphy, Eva Schinnerer, Alex Tetarenko, Grazia Umana, and Alexander van der Horst. VLA+VLBA to ngVLA Transition Option Concepts, July 2025. URL http://arxiv.org/abs/2501.06333. arXiv:2501.06333 [astro-ph].

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An Embodied Multimodal Language Model, March 2023. URL http://arxiv.org/abs/2303.03378. arXiv:2303.03378 [cs].

Tongtong Feng, Xin Wang, Yu-Gang Jiang, and Wenwu Zhu. Embodied AI: From LLMs to World Models [Feature]. *IEEE Circuits and Systems Magazine*, 25(4):14–37, 2025b. URL https://ieeexplore.ieee.org/abstract/document/11317901/.

Chongkai Gao, Zixuan Liu, Zhenghao Chi, Junshan Huang, Xin Fei, Yiwen Hou, Yuxuan Zhang, Yudi Lin, Zhirui Fang, Zeyu Jiang, and Lin Shao. VLA-OS: Structuring and Dissecting Planning Representations and Paradigms in Vision-Language-Action Models, June 2025. URL http://arxiv.org/abs/2506.17561. arXiv:2506.17561 [cs].

Weifan Guan, Qinghao Hu, Aosheng Li, and Jian Cheng. Efficient Vision-Language-Action Models for Embodied Manipulation: A Systematic Survey, October 2025. URL http://arxiv.org/abs/2510.17111. arXiv:2510.17111 [cs].

Ruihan Hu, Xiangdong He, Feiyang Huang, Jiaxing Zhao, Xinrui Cheng, and Zhongjie Wang. Joint Optimization of Fine-grained Representation and Workflow Orchestration in Metaverse Articulated Manipulation Auto-generation by VLA Method. *IEEE Transactions on Services Computing*, 2025b. URL https://ieeexplore.ieee.org/abstract/document/11207517/.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner Monologue: Embodied Reasoning through Planning with Language Models, July 2022b. URL http://arxiv.org/abs/2207.05608. arXiv:2207.05608 [cs].

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models, November 2023b. URL http://arxiv.org/abs/2307.05973. arXiv:2307.05973 [cs].

Yina Jian, Tian Di, Zhen-Yuan Wei, Chen-Wei Liang, and Mu-Jiang-Shan Wang. PI-VLA: A Symmetry-Aware Predictive and Interactive Vision–Language–Action Framework for Robust Robotic Manipulation. 2026. URL https://www.preprints.org/manuscript/202601.0682.

Nikita Kachaev, Mikhail Kolosov, Daniil Zelezetsky, Alexey K. Kovalev, and Aleksandr I. Panov. Don't Blind Your VLA: Aligning Visual Representations for OOD Generalization, October 2025. URL `http://arxiv.org/abs/2510.25616`. arXiv:2510.25616 [cs].

Baicheng Li, Dong Wu, Zike Yan, Xinchen Liu, Zecui Zeng, Lusong Li, and Hongbin Zha. Reflection-Based Task Adaptation for Self-Improving VLA, January 2026b. URL `http://arxiv.org/abs/2510.12710`. arXiv:2510.12710 [cs].

Tao Lin, Gen Li, Yilei Zhong, Yanwen Zou, Yuxin Du, Jiting Liu, Encheng Gu, and Bo Zhao. Evo-0: Vision-Language-Action Model with Implicit Spatial Understanding, November 2025b. URL `http://arxiv.org/abs/2507.00416`. arXiv:2507.00416 [cs].

Yeqing Lin, Minji Lee, Zhao Zhang, and Mohammed AlQuraishi. Out of Many, One: Designing and Scaffolding Proteins at the Scale of the Structural Universe with Genie 2, May 2024. URL `http://arxiv.org/abs/2405.15489`. arXiv:2405.15489 [q-bio].

Juyi Lin, Amir Taherin, Arash Akbari, Arman Akbari, Lei Lu, Guangyu Chen, Taskin Padir, Xiaomeng Yang, Weiwei Chen, Yiqian Li, Xue Lin, David Kaeli, Pu Zhao, and Yanzhi Wang. VOTE: Vision-Language-Action Optimization with Trajectory Ensemble Voting, October 2025c. URL `http://arxiv.org/abs/2507.05116`. arXiv:2507.05116 [cs].

Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, and Mengzhen Liu. HybridVLA: Collaborative Autoregression and Diffusion in a Unified Vision-Language-Action Model. 2025e. URL `https://openreview.net/forumpiid=8VyjwyLuSl`.

Zhuoyang Liu, Jiaming Liu, Jiadong Xu, Nuowei Han, Chenyang Gu, Hao Chen, Kaichen Zhou, Renrui Zhang, Kai Chin Hsieh, Kun Wu, Zhengping Che, Jian Tang, and Shanghang Zhang. MLA: A Multisensory Language-Action Model for Multimodal Understanding and Forecasting in Robotic Manipulation, September 2025f. URL `http://arxiv.org/abs/2509.26642`. arXiv:2509.26642 [cs].

Shuliang Liu, Zheng Qi, Jesse Jiaxi Xu, Yibo Yan, Junyan Zhang, He Geng, Aiwei Liu, Peijie Jiang, Jia Liu, and Yik-Cheung Tam. VLA-Mark: A cross modal watermark for large vision-language alignment models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26420–26438, 2025g. URL `https://aclanthology.org/2025.emnlp-main.1342/`.

Maëlic Neau, Zoe Falomir, Paulo E. Santos, Anne-Gwenn Bosser, and Cédric Buche. GraSP-VLA: Graph-based Symbolic Action Representation for Long-Horizon Planning with VLA Policies, November 2025. URL `http://arxiv.org/abs/2511.04357`. arXiv:2511.04357 [cs].

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos World Foundation Model Platform for Physical AI, July 2025a. URL `http://arxiv.org/abs/2501.03575`. arXiv:2501.03575 [cs].

Minho Park, Kinam Kim, Junha Hyung, Hyojin Jang, Hoiyeong Jin, Jooyeol Yun, Hojoon Lee, and Jaegul Choo. ACG: Action Coherence Guidance for Flow-based VLA models, October 2025. URL `http://arxiv.org/abs/2510.22201`. arXiv:2510.22201 [cs].

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A Generalist Agent, November 2022. URL `http://arxiv.org/abs/2205.06175`. arXiv:2205.06175 [cs].

Baochang Ren, Yunzhi Yao, Rui Sun, Shuofei Qiao, Ningyu Zhang, and Huajun Chen. Aligning Agentic World Models via Knowledgeable Experience Learning, January 2026. URL `http://arxiv.org/abs/2601.13247`. arXiv:2601.13247 [cs].

Hyunki Seong, Seongwoo Moon, Hojin Ahn, Jehun Kang, and David Hyunchul Shim. VLA-R: Vision-Language Action Retrieval toward Open-World End-to-End Autonomous Driving, November 2025. URL `http://arxiv.org/abs/2511.12405`. arXiv:2511.12405 [cs].

Nisarg A. Shah, Mingze Xia, Shameema Sikder, S. Swaroop Vedula, and Vishal M. Patel. Learning Action-Conditioned World Models for Cataract Surgery from Unlabeled Videos. In *Medical Imaging with Deep Learning*, 2026. URL https://openreview.net/forumpiid=aYQYOVm2AB.

Wangtian Shen, Ziyang Meng, Jinming Ma, Mingliang Zhou, and Diyun Xiang. An Efficient and Multi-Modal Navigation System with One-Step World Model, January 2026. URL http://arxiv.org/abs/2601.12277. arXiv:2601.12277 [cs].

Nan Sun, Yongchang Li, Chenxu Wang, Huiying Li, and Huaping Liu. CollabVLA: Self-Reflective Vision-Language-Action Model Dreaming Together with Human, September 2025b. URL http://arxiv.org/abs/2509.14889. arXiv:2509.14889 [cs].

Pablo Valle, Chengjie Lu, Shaukat Ali, and Aitor Arrieta. Evaluating Uncertainty and Quality of Visual Language Action-enabled Robots, July 2025. URL http://arxiv.org/abs/2507.17049. arXiv:2507.17049 [cs].

Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. DexVLA: Vision-Language Model with Plug-In Diffusion Expert for General Robot Control, August 2025c. URL http://arxiv.org/abs/2502.05855. arXiv:2502.05855 [cs].

Junjie Wen, Yichen Zhu, Minjie Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Xiaoyu Liu, Chaomin Shen, Yaxin Peng, and Feifei Feng. DiffusionVLA: Scaling Robot Foundation Models via Unified Diffusion and Autoregression. In *Forty-second International Conference on Machine Learning*, 2025d. URL https://openreview.net/forumpiid=VdwdU81Uzy.

Yuqing Wen, Hebei Li, Kefan Gu, Yucheng Zhao, Tiancai Wang, and Xiaoyan Sun. LLaDA-VLA: Vision Language Diffusion Action Models, September 2025e. URL http://arxiv.org/abs/2509.06932. arXiv:2509.06932 [cs].

Zhenyu Wu, Yuheng Zhou, Xiuwei Xu, Ziwei Wang, and Haibin Yan. Momanipvla: Transferring vision-language-action models for general mobile manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1714–1723, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wu_MoManipVLA_Transferring_Vision-language-action_Models_for_General_Mobile_Manipulation_CVPR_2025_paper.html.

Jialong Wu, Xiaoying Zhang, Hongyi Yuan, Xiangcheng Zhang, Tianhao Huang, Changjing He, Chaoyi Deng, Renrui Zhang, Youbin Wu, and Mingsheng Long. Visual Generation Unlocks Human-Like Reasoning through Multimodal World Models, January 2026c. URL http://arxiv.org/abs/2601.19834. arXiv:2601.19834 [cs].

Ji Zhang, Shihan Wu, Xu Luo, Hao Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. InSpire: Vision-Language-Action Models with Intrinsic Spatial Reasoning, September 2025l. URL http://arxiv.org/abs/2505.13888. arXiv:2505.13888 [cs].

Peng-Fei Zhang, Ying Cheng, Xiaofan Sun, Shijie Wang, Fengling Li, Lei Zhu, and Heng Tao Shen. A Step Toward World Models: A Survey on Robotic Manipulation, November 2025m. URL http://arxiv.org/abs/2511.02097. arXiv:2511.02097 [cs].

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware, April 2023. URL http://arxiv.org/abs/2304.13705. arXiv:2304.13705 [cs].

Rong Zhou, Dongping Chen, Zihan Jia, Yao Su, Yixin Liu, Yiwen Lu, Dongwei Shi, Yue Huang, Tianyang Xu, Yi Pan, Xinliang Li, Yohannes Abate, Qingyu Chen, Zhengzhong Tu, Yu Yang, Yu Zhang, Qingsong Wen, Gengchen Mai, Sunyang Fu, Jiachen Li, Xuyu Wang, Ziran Wang, Jing Huang, Tianming Liu, Yong Chen, Lichao Sun, and Lifang He. Digital Twin AI: Opportunities and Challenges from Large Language Models to World Models, January 2026. URL http://arxiv.org/abs/2601.01321. arXiv:2601.01321 [cs].

Ali AhmadiTeshnizi, Wenzhi Gao, and Madeleine Udell. OptiMUS: Scalable Optimization Modeling with (MI)LP Solvers and Large Language Models, February 2024. URL http://arxiv.org/abs/2402.10172. arXiv:2402.10172 [cs].

Max Argus, Jelena Bratulic, Houman Masnavi, Maxim Velikanov, Nick Heppert, Abhinav Valada, and Thomas Brox. cVLA: Towards Efficient Camera-Space VLAs, December 2025. URL http://arxiv.org/abs/2507.02190. arXiv:2507.02190 [cs].

Junhao Cai, Zetao Cai, Jiafei Cao, Yilun Chen, Zeyu He, Lei Jiang, Hang Li, Hengjie Li, Yang Li, Yufei Liu, Yanan Lu, Qi Lv, Haoxiang Ma, Jiangmiao Pang, Yu Qiao, Zherui Qiu, Yanqing Shen, Xu Shi, Yang Tian, Bolun Wang, Hanqing Wang, Jiaheng Wang, Tai Wang, Xueyuan Wei, Chao Wu, Yiman Xie, Boyang Xing, Yuqiang Yang, Yuyin Yang, Qiaojun Yu, Feng Yuan, Jia Zeng, Jingjing Zhang, Shenghan Zhang, Shi Zhang, Zhuoma Zhaxi, Bowen Zhou, Yuanzhen Zhou, Yunsong Zhou, Hongrui Zhu, Yangkun Zhu, and Yuchen Zhu. InternVLA-A1: Unifying Understanding, Generation and Action for Robotic Manipulation, January 2026. URL http://arxiv.org/abs/2601.02456. arXiv:2601.02456 [cs].

Yevgen Chebotar, Quan Vuong, Alex Irpan, Karol Hausman, Fei Xia, Yao Lu, Aviral Kumar, Tianhe Yu, Alexander Herzog, Karl Pertsch, Keerthana Gopalakrishnan, Julian Ibarz, Ofir Nachum, Sumedh Sontakke, Grecia Salazar, Huong T. Tran, Jodilyn Peralta, Clayton Tan, Deeksha Manjunath, Jaspiar Singht, Brianna Zitkovich, Tomas Jackson, Kanishka Rao, Chelsea Finn, and Sergey Levine. Q-Transformer: Scalable Offline Reinforcement Learning via Autoregressive Q-Functions, October 2023. URL `http://arxiv.org/abs/2309.10150`. arXiv:2309.10150 [cs].

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion, March 2024. URL `http://arxiv.org/abs/2303.04137`. arXiv:2303.04137 [cs].

Haohan Chi, Huan-ang Gao, Ziming Liu, Jianing Liu, Chenyu Liu, Jinwei Li, Kaisen Yang, Yangcheng Yu, Zeda Wang, Wenyi Li, Leichen Wang, Xingtao Hu, Hao Sun, Hang Zhao, and Hao Zhao. Impromptu VLA: Open Weights and Open Data for Driving Vision-Language-Action Models, May 2025. URL `http://arxiv.org/abs/2505.23757`. arXiv:2505.23757 [cs].

Chun-Kai Fan, Xiaowei Chi, Xiaozhu Ju, Hao Li, Yong Bao, Yu-Kai Wang, Lizhang Chen, Zhiyuan Jiang, Kuangzhi Ge, Ying Li, Weishi Mi, Qingpo Wuwu, Peidong Jia, Yulin Luo, Kevin Zhang, Zhiyuan Qin, Yong Dai, Sirui Han, Yike Guo, Shanghang Zhang, and Jian Tang. Wow, wo, val! A Comprehensive Embodied World Model Evaluation Turing Test, January 2026. URL `http://arxiv.org/abs/2601.04137`. arXiv:2601.04137 [cs].

Eric Hannus, Miika Malin, Tran Nguyen Le, and Ville Kyrki. IA-VLA: Input Augmentation for Vision-Language-Action models in settings with semantically complex tasks, September 2025. URL `http://arxiv.org/abs/2509.24768`. arXiv:2509.24768 [cs].

Tianshuai Hu, Xiaolu Liu, Song Wang, Yiyao Zhu, Ao Liang, Lingdong Kong, Guoyang Zhao, Zeying Gong, Jun Cen, Zhiyu Huang, Xiaoshuai Hao, Linfeng Li, Hang Song, Xiangtai Li, Jun Ma, Shaojie Shen, Jianke Zhu, Dacheng Tao, Ziwei Liu, and Junwei Liang. Vision-Language-Action Models for Autonomous Driving: Past, Present, and Future, January 2026. URL `http://arxiv.org/abs/2512.16760`. arXiv:2512.16760 [cs].

Yuhang Huang, Shilong Zou, Jiazhao Zhang, Xinwang Liu, Ruizhen Hu, and Kai Xu. AdaPower: Specializing World Foundation Models for Predictive Manipulation, December 2025c. URL `http://arxiv.org/abs/2512.03538`. arXiv:2512.03538 [cs].

Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. CoPa: General Robotic Manipulation through Spatial Constraints of Parts with Foundation Models, March 2024b. URL `http://arxiv.org/abs/2403.08248`. arXiv:2403.08248 [cs].

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: General Robot Manipulation with Multimodal Prompts, May 2023. URL `http://arxiv.org/abs/2210.03094`. arXiv:2210.03094 [cs].

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An Open-Source Vision-Language-Action Model, September 2024. URL `http://arxiv.org/abs/2406.09246`. arXiv:2406.09246 [cs].

Seungjae Lee, Yibin Wang, Haritheja Etukuru, H. Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Behavior Generation with Latent Actions, June 2024. URL `http://arxiv.org/abs/2403.03181`. arXiv:2403.03181 [cs].

Wei Li, Renshan Zhang, Rui Shao, Jie He, and Liqiang Nie. CogVLA: Cognition-Aligned Vision-Language-Action Model via Instruction-Driven Routing & Sparsification, October 2025u. URL `http://arxiv.org/abs/2508.21046`. arXiv:2508.21046 [cs].

Runhao Li, Wenkai Guo, Zhenyu Wu, Changyuan Wang, Haoyuan Deng, Zhenyu Weng, Yap-Peng Tan, and Ziwei Wang. MAP-VLA: Memory-Augmented Prompting for Vision-Language-Action Model in Robotic Manipulation, November 2025v. URL `http://arxiv.org/abs/2511.09516`. arXiv:2511.09516 [cs].

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-Language Foundation Models as Effective Robot Imitators, February 2024c. URL `http://arxiv.org/abs/2311.01378`. arXiv:2311.01378 [cs].

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as Policies: Language Model Programs for Embodied Control, May 2023. URL `http://arxiv.org/abs/2209.07753`. arXiv:2209.07753 [cs].

Tao Lin, Yilei Zhong, Yuxin Du, Jingjing Zhang, Jiting Liu, Yinxinyu Chen, Encheng Gu, Ziyan Liu, Hongyi Cai, Yanwen Zou, Lixing Zou, Zhaoye Zhou, Gen Li, and Bo Zhao. Evo-1: Lightweight Vision-Language-Action Model with Preserved Semantic Alignment, December 2025d. URL `http://arxiv.org/abs/2511.04555`. arXiv:2511.04555 [cs].

Zeting Liu, Zida Yang, Zeyu Zhang, and Hao Tang. EvoVLA: Self-Evolving Vision-Language-Action Model, November 2025h. URL http://arxiv.org/abs/2511.16166. arXiv:2511.16166 [cs].

Jiahang Liu, Yunpeng Qi, Jiazhao Zhang, Minghan Li, Shaoan Wang, Kui Wu, Hanjing Ye, Hong Zhang, Zhibo Chen, Fangwei Zhong, Zhizheng Zhang, and He Wang. TrackVLA++: Unleashing Reasoning and Memory Capabilities in VLA Models for Embodied Visual Tracking, October 2025i. URL http://arxiv.org/abs/2510.07134. arXiv:2510.07134 [cs].

Jijia Liu, Feng Gao, Bingwen Wei, Xinlei Chen, Qingmin Liao, Yi Wu, Chao Yu, and Yu Wang. What Can RL Bring to VLA Generalizationpi An Empirical Study, January 2026b. URL http://arxiv.org/abs/2505.19789. arXiv:2505.19789 [cs].

Zhiting Mei, Tenny Yin, Ola Shorinwa, Apurva Badithela, Zhonghe Zheng, Joseph Bruno, Madison Bland, Lihan Zha, Asher Hancock, Jaime Fernández Fisac, Philip Dames, and Anirudha Majumdar. Video Generation Models in Robotics – Applications, Research Challenges, Future Directions, January 2026. URL http://arxiv.org/abs/2601.07823. arXiv:2601.07823 [eess].

NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots, March 2025b. URL http://arxiv.org/abs/2503.14734. arXiv:2503.14734 [cs].

Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. Large VLM-based Vision-Language-Action Models for Robotic Manipulation: A Survey, September 2025. URL http://arxiv.org/abs/2508.13073. arXiv:2508.13073 [cs].

Shahram Najam Syed, Yatharth Ahuja, Arthur Jakobsson, and Jeff Ichnowski. ExpReS-VLA: Specializing Vision-Language-Action Models Through Experience Replay and Retrieval, November 2025. URL http://arxiv.org/abs/2511.06202. arXiv:2511.06202 [cs].

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An Open-Source Generalist Robot Policy, May 2024. URL http://arxiv.org/abs/2405.12213. arXiv:2405.12213 [cs].

Rishi Upadhyay, Howard Zhang, Jim Solomon, Ayush Agrawal, Pranay Boreddy, Shruti Satya Narayana, Yunhao Ba, Alex Wong, Celso M. de Melo, and Achuta Kadambi. WorldBench: Disambiguating Physics for Diagnostic Evaluation of World Models, January 2026. URL http://arxiv.org/abs/2601.21282. arXiv:2601.21282 [cs].

Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, and Max Du. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023. URL https://proceedings.mlr.press/v229/walke23a.html.

Yuqi Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. Unified Vision-Language-Action Model, June 2025i. URL http://arxiv.org/abs/2506.19850. arXiv:2506.19850 [cs].

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation, December 2023c. URL http://arxiv.org/abs/2312.13139. arXiv:2312.13139 [cs].

Chengen Xie, Bin Sun, Tianyu Li, Junjie Wu, Zhihui Hao, XianPeng Lang, and Hongyang Li. LatentVLA: Efficient Vision-Language Models for Autonomous Driving via Latent Action Prediction, January 2026b. URL http://arxiv.org/abs/2601.05611. arXiv:2601.05611 [cs].

Haochuan Xu, Yun Sing Koh, Shuhuai Huang, Zirun Zhou, Di Wang, Jun Sakuma, and Jingfeng Zhang. Model-agnostic Adversarial Attack and Defense for Vision-Language-Action Models, October 2025d. URL http://arxiv.org/abs/2510.13237. arXiv:2510.13237 [cs].

Yuping Yan, Yuhan Xie, Yixin Zhang, Lingjuan Lyu, Handing Wang, and Yaochu Jin. When Alignment Fails: Multimodal Adversarial Attacks on Vision-Language-Action Models, December 2025. URL http://arxiv.org/abs/2511.16203. arXiv:2511.16203 [cs].

Ganlin Yang, Tianyi Zhang, Haoran Hao, Weiyun Wang, Yibin Liu, Dehui Wang, Guanzhou Chen, Zijian Cai, Junting Chen, Weijie Su, Wengang Zhou, Yu Qiao, Jifeng Dai, Jiangmiao Pang, Gen Luo, Wenhai Wang, Yao Mu, and

Zhi Hou. Vlaser: Vision-Language-Action Model with Synergistic Embodied Reasoning, January 2026. URL http://arxiv.org/abs/2510.11027. arXiv:2510.11027 [cs].

Chenghao Yin, Da Huang, Di Yang, Jichao Wang, Nanshu Zhao, Chen Xu, Wenjun Sun, Linjie Hou, Zhijun Li, Junhui Wu, Zhaobo Liu, Zhen Xiao, Sheng Zhang, Lei Bao, Rui Feng, Zhenquan Pang, Jiayu Li, Qian Wang, and Maoqing Yao. Genie Sim 3.0 : A High-Fidelity Comprehensive Simulation Platform for Humanoid Robot, January 2026. URL http://arxiv.org/abs/2601.02078. arXiv:2601.02078 [cs].

Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Zhuoguang Chen, Tao Jiang, and Hang Zhao. DepthVLA: Enhancing Vision-Language-Action Models with Depth-Aware Spatial Reasoning, October 2025b. URL http://arxiv.org/abs/2510.13375. arXiv:2510.13375 [cs].

Jia Zeng, Qingwen Bu, Bangjun Wang, Wenke Xia, Li Chen, Hao Dong, Haoming Song, Dong Wang, Di Hu, Ping Luo, Heming Cui, Bin Zhao, Xuelong Li, Yu Qiao, and Hongyang Li. Learning Manipulation by Predicting Interaction, June 2024. URL http://arxiv.org/abs/2406.00439. arXiv:2406.00439 [cs].

Shaopeng Zhai, Qi Zhang, Tianyi Zhang, Fuxian Huang, Haoran Zhang, Ming Zhou, Shengzhe Zhang, Litao Liu, Sixu Lin, and Jiangmiao Pang. A Vision-Language-Action-Critic Model for Robotic Real-World Reinforcement Learning, September 2025. URL http://arxiv.org/abs/2509.15937. arXiv:2509.15937 [cs].

Haochen Zhang, Nader Zantout, Pujith Kachana, Ji Zhang, and Wenshan Wang. IRef-VLA: A Benchmark for Interactive Referential Grounding with Imperfect Language in 3D Scenes, March 2025n. URL http://arxiv.org/abs/2503.17406. arXiv:2503.17406 [cs].

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3D-VLA: A 3D Vision-Language-Action Generative World Model, March 2024. URL http://arxiv.org/abs/2403.09631. arXiv:2403.09631 [cs].