EMBODIED INTELLIGENCE AND WORLD MODELS: A SURVEY OF PROGRESS FROM 2024 TO EARLY 2026

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 28, 2026

ABSTRACT

We present a unified, decision-oriented survey of **318 papers** (January 2024 – February 2026) on embodied intelligence and world models, revealing that explicit coupling between world-model prediction and downstream control optimization is the single strongest predictor of real-world deployment robustness across the systems studied. The field has undergone a decisive transition: modular perception-planning-control pipelines are giving way to large-scale vision-language-action (VLA) policies and world-model-guided control stacks that jointly optimize representation, prediction, and decision making. We formalize the embodied control problem as a partially observable Markov decision process and derive a shared learning objective coupling latent dynamics modeling with control optimization. We propose a **three-axis taxonomy**: (1) *Functionality*—decision-coupled models that directly optimize task-facing objectives versus general-purpose models trained for broad predictive transfer; (2) *Temporal Modeling*—sequential step-by-step rollouts versus global trajectory-segment predictors; (3) *Spatial Representation*—compact latent vectors, tokenized feature sequences, and geometry-aware rendering representations. We systematize data resources and evaluation metrics across five families—task success, control stability, prediction fidelity, generalization, and compute efficiency—and compare more than 30 representative systems across eight method families under a normalized decision utility framework. A recurrent two-stage recipe (large prior followed by decision-coupled adaptation) emerges as the dominant empirical pattern. The analysis distills seven open challenges: long-horizon physical consistency, embodiment-aware representation alignment, deployment-oriented evaluation, data governance and compute efficiency, safe continual adaptation, multi-agent coordination, and interpretability/trustworthiness. We maintain a curated bibliography at `https://github.com/rsea2z/review-embodied`.

Index Terms: embodied AI, world models, vision-language-action models, robotic foundation models, long-horizon planning, autonomous driving, embodied world modeling, multi-agent coordination, interpretability.

1 Introduction

Embodied AI studies agents that close the full interaction loop with the physical world: multimodal sensing, state estimation, goal-conditioned reasoning, action planning, and low-level motor execution under uncertainty, latency, and resource constraints. Unlike disembodied systems that operate purely on stored data, embodied agents must continuously reconcile internal representations with dynamic environmental feedback, handle partial observability and irreversibility, and recover from failures in real time. This fundamental challenge—bridging the semantic richness of large language and vision models with the physical precision of real-world control—defines the central research agenda of the current period.

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

Recent progress has been striking. As of early 2026, generalist robot policies can follow open-ended natural language instructions for dexterous household manipulation [Black et al., 2026, Intelligence et al., 2025a], humanoid systems demonstrate whole-body loco-manipulation in unstructured environments [NVIDIA et al., 2025a, Jiang et al., 2025a, Ding et al., 2025a], and world-model-guided stacks generate physically plausible synthetic data at scale to mitigate real-robot collection costs [Team et al., 2025a,b, NVIDIA et al., 2025b]. In parallel, VLA architectures have proliferated dramatically, with specialized variants addressing 3D spatial reasoning [Qu et al., 2025, Sun et al., 2025a, Zhang et al., 2025a, Li et al., 2025a], tactile and force feedback [Bi et al., 2025a, Huang et al., 2025a, Yu et al., 2025], chain-of-thought reasoning [Ye et al., 2025a, Zhao et al., 2025a, Zhang et al., 2025b, Huang et al., 2025b], reinforcement post-training [Li et al., 2025b, Lu et al., 2025, Chen et al., 2025a, Zang et al., 2025], efficiency and edge deployment [Pertsch et al., 2025, Yang et al., 2025a, Wen et al., 2025a, Shukor et al., 2025, Budzianowski et al., 2025], and autonomous driving [Guo and Zhang, 2025, Li et al., 2025c, Hao et al., 2025, Jiang et al., 2025b]. The breadth and pace of this progress demand a synthesis that organizes these methods by their decision-coupling strategy rather than their architectural lineage.

1.1 The Embodied Intelligence Imperative

The goal of building machines that act intelligently in the physical world has motivated AI research for decades. Early symbolic approaches modeled the physical world through explicit ontologies and geometric planning routines but struggled with perceptual grounding, uncertainty propagation, and the combinatorial complexity of real environments. Deep learning transformed this landscape by enabling end-to-end visuomotor mappings from raw pixels to actions, but at the cost of data hunger, poor sample efficiency, and weak generalization beyond training distributions.

The transformative insight of the current period is that *world models*—internal simulators of environment dynamics—can bridge this gap by providing:

1. **Predictive grounding**: anticipating the consequences of candidate actions before executing them, enabling safer and more deliberate behavior.
2. **Data amplification**: generating synthetic training data that would be prohibitively expensive or dangerous to collect physically.
3. **Counterfactual reasoning**: evaluating hypothetical action sequences under varying conditions to improve robustness to distribution shift.
4. **Representation learning**: structuring internal features around controllable scene dynamics rather than raw pixel correlation, enabling better transfer.

In parallel, large-scale vision-language pretraining has endowed policy networks with deep semantic understanding of instructions, objects, and goals. The synthesis of these two trends—language-grounded semantic reasoning and physically grounded world modeling—represents the defining architectural shift of 2024–2026.

1.2 Motivation and Historical Context

Cognitive foundations. The concept of an internal model of the environment was articulated by Batra et al. [2020] and formalized earlier by control theorists as the "internal model principle." From a cognitive science perspective, rich internal representations of world dynamics are considered foundational to intelligent planning and generalization in biological agents. This perspective motivated a long line of model-based reinforcement learning research and foreshadowed the current wave of neural world models for embodied control.

Phase 1: Task definition and environment grounding (2020–2022). The field began with clear delineation of canonical embodied tasks. Interactive instruction following, rearrangement-centered evaluation, and open-ended goal pursuit established the core challenge of bridging perception with executable skill libraries [Batra et al., 2020, Duan et al., 2022]. Early multimodal agents demonstrated that language could structure complex sequential behaviors, but relied heavily on symbolic planners that were brittle under perceptual ambiguity [Gao et al., 2022]. Open-ended survival and exploration settings underscored the importance of lifelong skill accumulation and self-guided curriculum construction [Fan et al., 2022]. These works collectively established the benchmark infrastructure that subsequent approaches would be measured against.

Phase 2: Language-grounded planning (2022–2023). The next phase recognized that large language models, trained on vast internet corpora, could serve as implicit world knowledge bases for embodied planning. SayCan demonstrated that LLM-generated action plans could be grounded by value functions that estimate physical feasibility [Ahn et al., 2022]. Inner Monologue showed that environmental language feedback—natural language descriptions of what happened after an action—dramatically improved task completion by closing the perception-action loop at the semantic level [Huang et al., 2022a]. Subsequent works refined this paradigm: LLM-Planner enabled open-vocabulary spatial navigation through in-context learning [Song et al., 2023]; Plan-and-Solve decomposed reasoning from execution [Wu et al., 2023a]; JARVIS combined neuro-symbolic reasoning with modular execution for dialogue-conditioned task completion [Zheng et al., 2025a]; VoxPoser synthesized 3D affordance and value maps from language descriptions to guide manipulation [Huang et al., 2023a]. The grounded decoding paradigm further showed that semantic constraints could directly modulate token generation for physical feasibility [Huang et al., 2023b]. Meanwhile, Voyager demonstrated that GPT-4 could iteratively design, refine, and accumulate executable skill libraries for open-ended agents [Wang et al., 2023].

Phase 3: Foundation policies and data scaling (2022–2023). A parallel line of work moved beyond language-only planners to build generalist visuomotor control policies at scale. Gato demonstrated that a single transformer could be trained on heterogeneous multi-domain data spanning games, robotic control, and language tasks [Reed et al., 2022]. RT-1 established transformer-based real-robot manipulation from diverse task demonstrations collected via large-scale data pipelines [Brohan et al., 2023]. Q-Transformer extended this to offline RL over large datasets while preserving the autoregressive token prediction interface [Chebotar et al., 2023]. RoboCat demonstrated few-shot adaptation across robot embodiments and tasks through self-improvement loops [Bousmalis et al., 2023]. PaLM-E showed that embodied policies could benefit from grounding large multimodal language models with physical sensorimotor data [Driess et al., 2023]. VIMA unified manipulation commands across heterogeneous task types through multimodal prompt engineering [Jiang et al., 2023]. RT-Trajectory and Code-as-Policies explored trajectory-sketch and code-mediated control interfaces [Gu et al., 2023, Liang et al., 2023]. BridgeData V2 and ACT-style teleoperation recipes accelerated data collection for low-cost manipulation systems [Walke et al., 2023, Zhao et al., 2023]. Diffusion Policy established denoising diffusion as a powerful generative framework for visuomotor control that naturally handles multimodal action distributions [Chi et al., 2024]. These works created the foundation on which the 2024–2026 generation would build.

Phase 4: VLA scaling and world-model integration (2024–2026). The current phase is characterized by three overlapping trends. First, open-source and closed VLA models scaled to 7B–70B parameters, with OpenVLA demonstrating state-of-the-art open-source manipulation across 29 tasks while outperforming closed models such as RT-2-X (55B) with 7× fewer parameters [Kim et al., 2024]. Second, specialized architectural innovations proliferated to address VLA limitations: flow-matching policies for dexterous control [Black et al., 2026], frequency-space action tokenization for high-frequency tasks [Pertsch et al., 2025], hybrid autoregressive-diffusion architectures [Liu et al., 2025a], and dual-system designs separating slow reasoning from fast motor execution [NVIDIA et al., 2025a]. Third, world models were tightly coupled with VLA pipelines: WorldVLA unified action prediction with future image synthesis in a single autoregressive stack [Cen et al., 2025a], BridgeV2W aligned coordinate-space actions with pixel-space predictions through URDF-rendered embodiment masks [Chen et al., 2026a], and GigaWorld-0/GigaBrain-0 established world models as data engines to generate training distributions at scale [Team et al., 2025a,b].

1.3 Why a New Survey Is Needed

Multiple surveys have examined embodied AI, VLA models, and world models from various angles in the 2024–2026 period. Liu et al. provide a broad panorama of multimodal large model alignment for embodied AI but give limited treatment to the algorithmic coupling between world modeling and control optimization [Liu et al., 2025b]. Liang et al. survey large model empowered embodied AI with strong coverage of hierarchical and end-to-end decision paradigms but a less systematic taxonomy of world-model design choices [Liang et al., 2025a]. Li et al. provide the most taxonomically complete world-model survey to date [Li et al., 2025d], while Ding et al. give a broader but less embodiment-focused treatment [Ding et al., 2025b]. Zhong et al. and Yu et al. survey the VLA methodology landscape [Zhong et al., 2025a, Yu et al., 2026a]. Jiang et al. cover VLA for autonomous driving specifically [Jiang et al., 2025b]. Fung et al. emphasize the world model as the core reasoning component for embodied agents [Fung et al., 2025]. More recent entries include Dolgopolyi et al.'s bibliometric analysis of VLM and VLA systems [Dolgopolyi and Tsevas, 2025], Li et al.'s comprehensive review of VLA models across five dimensions [Li et al., 2025e], Shao et al.'s first systematic taxonomy of large VLM-based VLA architectures [Shao et al., 2025], and Sapkota et al.'s review of VLA applications spanning autonomous vehicles, medical robotics, and humanoid systems [Sapkota et al., 2026].

Despite this growing body of work, three specific gaps remain unaddressed.

Gap 1: Decision-coupling perspective. Existing surveys organize methods either by application domain or by neural architecture class. Neither framing adequately captures the design choice that most predicts deployment behavior: whether the world model is directly coupled with the decision objective or decoupled for general-purpose pretraining. Decision-coupled models can achieve higher task-specific reliability but require careful data pipeline design; general-purpose models offer broader transfer but often need explicit post-training adaptation to reach target performance. This axis has not been systematically studied.

Gap 2: 2024–2026 VLA diversity. The VLA literature has expanded dramatically in 2025 to include models addressing spatial and geometric perception [Qu et al., 2025, Sun et al., 2025a, Li et al., 2025a, Bhat et al., 2025], multi-sensory fusion beyond vision [Bi et al., 2025a, Huang et al., 2025a, Yu et al., 2025, Wei et al., 2025], chain-of-thought and reinforcement reasoning [Ye et al., 2025a, Guo et al., 2025a, Yin et al., 2025], whole-body humanoid control [Jiang et al., 2025a, Ding et al., 2025a], multi-robot coordination [Sun et al., 2025b, Guo et al., 2024, Li et al., 2025f], and safety-aware deployment [Zhang et al., 2025c, Hancock et al., 2025a]. No existing survey provides systematic coverage of this breadth.

Gap 3: Deployment-centered evaluation. Recent work has begun exposing that standard benchmark scores are poor proxies for deployment reliability. WorldBench targets disentangled physical concept evaluation [Upadhyay et al., 2026]; Wu et al. analyze what video generation models understand about physics [Wu et al., 2026a]; Valle et al. argue for uncertainty and quality metrics beyond binary success [Valle et al., 2025]; Wu et al. systematically expose pragmatic failure modes [Wu et al., 2026b]. These diagnostic perspectives need integration into a unified framework.

1.4 Technical Lineage Before 2024

The current wave is built on three earlier lines of work. The first established canonical embodied tasks and open environments—rearrangement evaluation benchmarks and open-ended skill acquisition settings—motivating closed-loop success criteria and environment diversity requirements [Batra et al., 2020, Duan et al., 2022, Fan et al., 2022]. This line also developed multimodal dialogue-conditioned benchmarks such as DialFRED [Gao et al., 2022] and open-ended task completion in realistic household simulators.

The second line developed language-grounded planning with explicit feasibility checks and environment feedback [Ahn et al., 2022, Huang et al., 2022b,a, Song et al., 2023, Wu et al., 2023a,b, Huang et al., 2023b, Sarch et al., 2023, Zheng et al., 2025a]. In our notation, this work introduced the high-level/low-level policy factorization, where a language plan ξ_t conditions a motor-execution policy, foreshadowing modern VLA architectures that maintain semantic and motor heads simultaneously.

The third line established foundation-policy recipes for robot control at scale through heterogeneous imitation learning and transformer-based control [Reed et al., 2022, Brohan et al., 2023, Chebotar et al., 2023, Bousmalis et al., 2023, Driess et al., 2023, Jiang et al., 2023, Liang et al., 2023, Gu et al., 2023, Huang et al., 2023a, Walke et al., 2023, Zhao et al., 2023, Wang et al., 2023, Chi et al., 2024]. These developments yielded the action tokenization, data scaling, and VLM initialization insights that 2024–2026 systems inherit and extend. Complementary advances included equivariant diffusion for embodied planning [Brehmer et al., 2023], part-based spatial constraints for general robotic manipulation [Huang et al., 2024a], high-speed deep predictive motion generation [Yoshikawa et al., 2024], and generative interactive environments that demonstrated world models could be learned from unlabeled internet video [Bruce et al., 2024]. Urban embodied agents and navigation systems further extended the scope of embodied AI beyond tabletop manipulation [Xu et al., 2023].

1.5 Scope and Inclusion Criteria

This survey covers publications from **January 1, 2024 to February 27, 2026**. A work is in scope if it:

- proposes methods or benchmarks for embodied agents that interact with physical environments through robotic control, autonomous driving, or situated simulation;
- involves world modeling, VLA architectures, embodied data pipelines, or evaluation protocols for closed-loop physical tasks; or
- provides analysis directly relevant to coupling world models with embodied decision-making objectives.

Purely generic video generation models without embodiment-specific coupling, and general-purpose language models without grounding to physical action spaces, are excluded from the main technical analysis but discussed as precursors where historically relevant.

We adopt two synchronized analytical views:

- **Embodied pipeline view:** how perception, planning/reasoning, control, and adaptation components are composed and what interfaces connect them.
- **World-model design view:** functionality coupling, temporal modeling horizon, and spatial representation form.

1.6 Contributions

This survey makes five concrete contributions:

1. **Unified decision-oriented taxonomy**: we propose a three-axis taxonomy (functionality coupling, temporal modeling, spatial representation) that predicts deployment behavior more faithfully than architecture-centric or application-centric classifications.
2. **Mathematical formalization**: we derive a shared learning objective that links POMDP-based embodied control with latent world-model training, unifying formulations scattered across individual papers.
3. **Comprehensive 2024–2026 coverage**: we systematically analyze 318 papers across foundation VLA policies, world-model-guided control stacks, post-training reinforcement refinement, efficiency-oriented adaptation, humanoid and multi-modal systems, autonomous driving, and embodied benchmarks.
4. **Cross-family quantitative perspective**: we compare method families under a normalized decision utility framework across eight sub-families (foundation VLAs, world-model-guided control, post-training refinement, efficiency-oriented, 3D-aware, reasoning-augmented, multi-modal sensing, and domain-specific VLAs) and identify a recurrent two-stage recipe (large prior followed by decision-coupled adaptation) as the dominant empirical pattern.
5. **Deployment-oriented challenge synthesis**: we distill seven open challenges grounded in concrete empirical failures reported across the literature—including multi-agent coordination and interpretability/trustworthiness—each with specific research priority recommendations.

1.7 Paper Organization

Section 2 introduces mathematical foundations for embodied control and latent world-model training. Section 3 presents the three-axis taxonomy, interface contracts, and embodied pipeline mapping with comprehensive method coverage. Section 4 surveys data regimes, curation dimensions, benchmark categories, and evaluation metric families. Section 5 provides cross-family comparison including method-level tables, case studies, and the two-stage pattern analysis. Section 6 distills seven open challenges with near-term research priorities. Section 7 concludes with a synthesis of the current frontier and outlook.

2 Background and Mathematical Formulation

2.1 Embodied Interaction as a Partially Observable Control Process

We model embodied interaction as a *partially observable Markov decision process* (POMDP):

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, p, \Omega, r, \gamma), \tag{1}$$

where $s_t \in \mathcal{S}$ is the latent world state (geometry, object poses, joint configurations), $a_t \in \mathcal{A}$ is the control action (joint torques, end-effector deltas, waypoints, steering commands), and $o_t \in \mathcal{O}$ is the multimodal observation stream (RGB images, depth maps, force-torque readings, proprioception, language context). The world dynamics and observation model are given by:

$$s_{t+1} \sim p(s_{t+1} \mid s_t, a_t), \qquad o_t \sim \Omega(o_t \mid s_t). \tag{2}$$

The agent optimizes discounted cumulative reward:

$$J(\pi) = \mathbb{E}_{\pi,\, p}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right]. \tag{3}$$

The *embodied* qualifier on this standard RL problem imposes several additional requirements beyond simulator RL:

- **Real-time constraint**: the control policy π must produce actions within a hard latency budget, typically 10–100 ms depending on control frequency.
- **Safety constraint**: some state-action pairs carry irreversibility (dropped items, self-collision, road accidents), imposing a constrained optimization structure $\mathbb{E}[\sum_t c_k(s_t, a_t)] \leq d_k$ for each safety dimension k.
- **Partial observability**: s_t is never directly accessed; only o_t is available, making history-dependent policies $\pi(a_t \mid o_{\leq t}, g)$ necessary for tasks requiring memory of past interactions.
- **Distribution shift**: the test environment distribution p_{test} may differ substantially from training distribution p_{train}, requiring robust policy representations that generalize through physical geometry and semantic diversity.

2.2 Belief State Compression and History Encoding

Since s_t is unobservable, the agent must maintain a *belief state* $b_t = p(s_t \mid o_{\leq t}, a_{<t})$. Exact Bayesian belief updating is intractable for high-dimensional state spaces. In practice, embodied systems approximate b_t through four families of compact encoders:

1. **Recurrent state encoders**: $h_t = f_\psi(h_{t-1}, o_t, a_{t-1})$, where h_t is a hidden state compressing history. Widely used in model-based RL for robotics [Li et al., 2025d, Fung et al., 2025].
2. **Transformer attention windows**: attention over a fixed or growing context window $o_{t-W:t}$, providing temporal credit assignment without explicit recurrence. This is the dominant paradigm in current VLA architectures [Kim et al., 2024, Black et al., 2026, Intelligence et al., 2025a].
3. **3D spatial memory**: explicit voxel grids or point clouds that accumulate multi-view RGB-D observations into a persistent geometric representation [Bhat et al., 2025, Zhen et al., 2024, Qu et al., 2025, Sun et al., 2025a, Zhang et al., 2025a]. This design choice is especially valuable for manipulation tasks where object geometry and contact affordances critically determine feasibility.
4. **Language-conditioned belief**: belief compression guided by the goal instruction g, so that task-irrelevant perceptual details are suppressed and goal-relevant features are amplified [Huang et al., 2023a, Li et al., 2025g, Huang et al., 2025b].

Memory mechanisms for long-horizon control. A critical limitation of fixed-window attention is that task-relevant information may exceed the context window, particularly in long-horizon manipulation where the agent must remember previously visited locations, completed subgoals, or object states that have changed. Recent approaches address this through explicit memory augmentation. MAP-VLA introduces demonstration-derived memory prompts as a plug-and-play module for frozen VLAs, using a memory library of historical demonstrations with learnable soft prompts that yield up to 25% improvement on long-horizon real-robot tasks [Li et al., 2025h]. ContextVLA compresses past observations into a single context token that amortizes multi-frame observation benefits while reducing training and inference costs [Jang et al., 2025]. TrackVLA extends memory to embodied visual tracking with a Target Identification Memory module and gated update strategy for persistent target representation [Liu et al., 2025c]. Goal-VLA uses image-generative VLMs as object-centric world models to enable zero-shot manipulation by imagining goal states from current observations [Chen et al., 2025b]. Model-agnostic approaches to VLA improvement, including adversarial robustness through embedding disruption patches, highlight the need for memory and representation mechanisms that are resilient to distribution shift in the observation stream [Xu et al., 2025a].

2.3 Pre-2024 Design Motifs That Shaped Current Formulations

Several pre-2024 lines directly shaped current embodied modeling assumptions. Language-grounded planning works argued for explicit decomposition between high-level plan tokens and low-level motor execution, often with environment feedback and feasibility filters [Ahn et al., 2022, Huang et al., 2022b,a, Song et al., 2023, Wu et al., 2023a, Huang et al., 2023b]. In our notation, this motivates a latent plan variable ξ_t:

$$\pi(a_t \mid h_t, g) = \int \pi_{\text{low}}(a_t \mid h_t, \xi_t)\; \pi_{\text{high}}(\xi_t \mid h_t, g)\; d\xi_t, \tag{4}$$

where h_t is the observation-action history encoding and g is the natural language goal. This hierarchical decomposition appears in modern dual-system VLA architectures (e.g., GR00T N1's System 1/System 2 design [NVIDIA et al., 2025a]) and in chain-of-thought reasoning VLAs that generate explicit subgoal sequences before actions [Ye et al., 2025a, Zhao et al., 2025a, Zhang et al., 2025b, Yin et al., 2025].

Generalist transformer-control systems showed that heterogeneous action modalities can be cast as autoregressive token prediction over tokenized observation-action sequences [Reed et al., 2022, Brohan et al., 2023, Chebotar et al., 2023, Jiang et al., 2023]. This perspective strongly influenced VLA design choices on action tokenization, sequence conditioning, and VLM initialization. The insight that internet-scale pretraining provides strong semantic priors that transfer to physical manipulation was validated empirically by OpenVLA [Kim et al., 2024] and later systematically studied in VLM4VLA [Zhang et al., 2026a], which found that VLM quality and VLA quality correlate but not monotonically—embodied adaptation objectives remain essential.

Action representations evolved from simple per-dimension binning to more expressive forms: Diffusion Policy introduced denoising-based continuous action generation that handles multimodal action distributions and high-dimensional action spaces naturally [Chi et al., 2024], while FAST proposed DCT-based frequency-space tokenization to preserve dexterous high-frequency action structure [Pertsch et al., 2025]. More recently, VQ-VLA demonstrates that vector-quantized action tokenizers built on large-scale trajectory data can improve long-horizon real-world success rates by up to 30% compared to naive discretization [Wang et al., 2025a]. Discrete diffusion approaches unify discrete and diffusion-based action generation, achieving 96.3% on LIBERO through parallel decoding that breaks the autoregressive bottleneck [Liang et al., 2025b]. ACG introduces training-free test-time guidance that improves action coherence for flow-based VLA models by addressing jerks and pauses from imitation learning [Park et al., 2025]. Action tokenization is thus not merely a representation detail but a **systems lever** that influences training efficiency, dexterity, and deployment latency. The action tokenization reconstruction loss can be formalized as:

$$\mathcal{L}_{\text{act}} = \mathbb{E}\left[\sum_{t=1}^{T} \|a_t - \hat{a}_t\|^2 + \lambda_{\text{vq}} \|\text{sg}[z_e(a_t)] - e\|^2 + \beta_{\text{vq}} \|z_e(a_t) - \text{sg}[e]\|^2\right], \tag{5}$$

where z_e is the encoder, e is the nearest codebook entry, and $\text{sg}[\cdot]$ denotes the stop-gradient operator. The VQ commitment terms regularize the encoder-codebook alignment, and the balance between reconstruction fidelity and codebook utilization determines the downstream control quality.

2.4 Latent World Models for Embodied Control

A practical world model introduces a latent state z_t to represent controllable scene dynamics compactly:

$$z_t \sim q_\phi(z_t \mid o_{\leq t}, a_{<t}), \qquad \hat{z}_{t+1} \sim p_\theta(\hat{z}_{t+1} \mid z_t, a_t), \tag{6}$$

where q_ϕ is an encoder (posterior over latent states), and p_θ is a dynamics predictor. Decoder and task-prediction heads project latent trajectories back to observations and task-relevant signals:

$$\hat{o}_{t+1} \sim p_\theta(o_{t+1} \mid \hat{z}_{t+1}), \qquad \hat{y}_{t+1} = g_\psi(\hat{z}_{t+1}), \tag{7}$$

where $\hat{y}_{t+1}$ may denote future object states, occupancy, predicted rewards, contact events, or safety-constraint violations, depending on the downstream control stack [Li et al., 2025d, Fung et al., 2025, Team et al., 2025b, Berg et al., 2025].

The key design question is not whether this decomposition exists, but *where decision coupling is applied*. The options are:

- End-to-end VLA policy head: the latent z_t conditions an action head directly, combining perception, world modeling, and action generation in one network [Kim et al., 2024, Cen et al., 2025a].
- Model-predictive control over latent rollouts: the dynamics model p_θ is used to simulate futures, and a separate planner selects actions based on rollout reward [Wan et al., 2025, Guo et al., 2025a].
- World model as data engine: p_θ generates synthetic observations used to augment training data for a separately trained policy [Team et al., 2025a,b, Li et al., 2025b].
- Offline-to-online adaptation: the world model is used for simulator-free policy improvement during deployment [Li et al., 2025b, Lu et al., 2025, Chen et al., 2025a, Intelligence et al., 2025b].

2.5 Unified Training Objective

Most world-model+policy implementations optimize a joint objective that combines predictive, regularization, and task-facing terms:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{obs}}}_{\text{observation prediction}} + \beta \underbrace{\text{KL}[q_\phi(z_t \mid \cdot) \parallel p_\theta(z_t \mid z_{t-1}, a_{t-1})]}_{\text{dynamics consistency}} + \lambda \underbrace{\mathcal{L}_{\text{task}}}_{\text{control/planning utility}}, \tag{8}$$

where:

- $\mathcal{L}_{\text{obs}}$ may be a reconstruction loss (pixel MSE/LPIPS), a contrastive loss over future states, or a next-frame prediction cross-entropy depending on whether the representation is pixel-level, tokenized, or semantic.
- The KL term regularizes the posterior q_ϕ to remain close to the dynamics prior p_θ, preventing posterior collapse and improving rollout stability under long horizons.
- $\mathcal{L}_{\text{task}}$ includes action prediction loss (imitation learning), value estimates (offline RL), or contrastive task-conditioned objectives (instruction-conditioned control).

The hyperparameters (β, λ) implement a tradeoff between predictive fidelity and task specificity. Systems that pretrain with high β and low λ learn richer dynamics representations transferable across tasks; systems that deploy with high λ optimize directly for task success but may overfit the training distribution [NVIDIA et al., 2025b, Team et al., 2025a, Li et al., 2025d].

2.6 Decision Optimization with Learned Dynamics

Given a learned dynamics model, open-loop planning over a horizon H can be formulated as:

$$\mathbf{a}^*_{t:t+H-1} = \arg \max_{\mathbf{a}_{t:t+H-1}} \mathbb{E}_{p_\theta} \left[\sum_{k=0}^{H-1} \gamma^k \hat{r}_{t+k} \right], \tag{9}$$

where $\hat{r}_{t+k} = r(g_\psi(\hat{z}_{t+k}), a_{t+k})$ is the predicted reward at step $t + k$. In practice, pure open-loop planning (Eq. 9) is rarely sufficient. Embodied systems combine it with:

- **Receding horizon control**: re-optimize at each step to correct compounding errors [Li et al., 2025i, Lu et al., 2025].
- **Monte Carlo Tree Search**: guided tree expansion in latent action space [Guo et al., 2025a, Yin et al., 2025].
- **Intervention-guided online RL**: use human teleoperation corrections as additional reward signal during deployment [Intelligence et al., 2025b, Chen et al., 2025a].
- **Advantage-conditioned sampling**: RECAP conditions VLA token sampling on estimated advantage values for online RL without explicit value networks [Intelligence et al., 2025b].

2.7 The VLA Architecture Pattern

Modern VLA models implement a three-component stack: a vision encoder E_v, a language model E_l, and an action head π_a. The standard forward pass is:

$$v_t = E_v(\text{image}_t, \text{depth}_t, \ldots), \tag{10}$$

$$l_t = E_l(\text{instruction}_g, v_t), \tag{11}$$

$$a_t \sim \pi_a(a_t \mid l_t, h_{t-W:t}), \tag{12}$$

where $h_{t-W:t}$ is the observation-action history over a context window of width W. Variations include:

- **Autoregressive action head**: actions are discretized into tokens and predicted autoregressively, inheriting LLM's in-context learning capability [Kim et al., 2024, Li et al., 2024a, 2025j].
- **Diffusion action head** (flow matching or DDPM): continuous actions are denoised conditioned on l_t, preserving action continuity [Black et al., 2026, Chi et al., 2024, Wen et al., 2025b].
- **Hybrid head**: autoregressive token prediction augmented with diffusion denoising output, fusing reasoning and precision control [Liu et al., 2025a, Zhong et al., 2025b].
- **Dual-system architecture**: slow reasoning module (language-level planning, "System 2") coupled with fast diffusion motor module ("System 1"), enabling real-time dexterous execution with high-level task understanding [NVIDIA et al., 2025a, Won et al., 2025].
- **Multi-modal sensing action head**: extends v_t to include tactile readings, force-torque signals, or audio contact events, enabling fine-grained contact-rich manipulation [Bi et al., 2025a, Huang et al., 2025a, Yu et al., 2025, Wei et al., 2025].

The action representation also varies: per-dimension binning (simple but coarse), DCT-based frequency tokenization [Pertsch et al., 2025], continuous Gaussian [Chi et al., 2024], or hybrid discrete-continuous mixtures [Liu et al., 2025a]. This choice directly affects dexterity, training stability, and inference latency.

2.8 Action Chunking and Temporal Horizon

A key systemic design choice is *action chunking*: instead of producing a single action a_t, models predict a chunk $\mathbf{a}_{t:t+C-1}$ of C actions simultaneously. Chunking reduces the auto-correlation between high-frequency actions, enables diffusion-based smoothing over the chunk, and amortizes the VLM inference cost over multiple control timesteps. The tradeoff is that longer chunks reduce closed-loop correction frequency:

$$\text{correction bandwidth} \propto \frac{1}{C \cdot \Delta t_{\text{infer}}}, \tag{13}$$

where Δt_{infer} is model inference latency. Systems such as π_0 use $C \approx 50$ chunks at 50 Hz, giving a 1-second planning horizon; more reactive systems use $C = 4$–8 for contact-rich manipulation where rapid correction is essential [Black et al., 2026, Chen et al., 2025a, Li et al., 2025i].

2.9 Failure Modes in the 2024–2026 Regime

Contemporary embodied systems exhibit three characteristic failure patterns that motivate the taxonomy developed in Section 3:

Failure Mode 1: Long-horizon drift. Let ϵ_t denote the per-step world model prediction error in latent space. Under sequential rollout, errors accumulate approximately as:

$$\mathcal{E}_{t+H} \leq \sum_{k=0}^{H-1} L^k \epsilon_{t+k}, \tag{14}$$

where L is the Lipschitz constant of p_θ in latent space. For $L > 1$, errors grow exponentially toward the horizon, making long-horizon plans unreliable. Multi-stage household manipulation tasks, which may span dozens of component actions over minutes, are particularly vulnerable [Gupta et al., 2024, Team et al., 2025b, Wang et al., 2026]. WorldBench diagnoses this through isolated concept-level physical evaluation [Upadhyay et al., 2026], and recent work aims to reduce L via physics-informed regularization and contact-aware dynamics [Han et al., 2025].

Failure Mode 2: Representation mismatch. Coordinate-space action control and pixel-space visual prediction operate in fundamentally different reference frames. A wrist joint angle increment does not have a straightforward pixel-space interpretation without explicit embodiment geometry (URDF, camera extrinsics, contact normals). This mismatch causes two problems: (a) the world model generates visually plausible futures that are geometrically inconsistent with the robot's actual kinematic constraints; and (b) the policy cannot exploit geometric structure in visual predictions to improve action estimation. BridgeV2W [Chen et al., 2026a] addresses this through URDF-aligned embodiment masks rendered into the prediction pathway; FlowDreamer [Guo et al., 2026] uses optical flow as an intermediate representation that bridges pixel and action space. GeoVLA [Sun et al., 2025a] and 4D-VLA [Zhang et al., 2025a] incorporate depth and point cloud inputs to resolve this mismatch directly in the policy's observation space.

Failure Mode 3: Evaluation gaps. Standard success rate metrics aggregate task outcome into a binary signal, hiding causal and physical subtleties. A policy that achieves 80% success by exploiting benchmark-specific visual cues may have 30% success under natural scene variation or object substitution. This evaluation gap is especially dangerous because policy developers optimize for the visible metric, reinforcing it at the expense of genuine robustness [Valle et al., 2025, Wu et al., 2026b,a]. The remedy requires multi-dimensional evaluations that separately quantify task competence, intervention frequency, recovery capability, physical consistency, and out-of-distribution generalization—the metric families formalized in Section 4.

These three failure modes motivate the three axes of the taxonomy in Section 3: Functionality coupling (addressing Mode 1), Spatial Representation (addressing Mode 2), and Evaluation Protocol (addressing Mode 3).

3 Coupled Taxonomy of Embodied Intelligence and World Models

We organize recent methods along two synchronized dimensions: (i) the embodied decision stack and (ii) world-model design choices. This decomposition keeps algorithmic comparisons explicit while preserving system-level relevance.

3.1 Taxonomy Design Principles

We build the taxonomy around *decision coupling*, *temporal modeling*, and *spatial representation*, because these three choices consistently determine deployment behavior across manipulation, navigation, and driving settings. A purely architecture-centric taxonomy hides optimization targets and interface contracts; a purely task-centric taxonomy hides why similar tasks still diverge in stability, sample efficiency, and latency.

This design also aligns with historical development: early rearrangement and instruction-following studies separated task definition from policy mechanism [Batra et al., 2020, Gao et al., 2022]; language-grounded planners emphasized high-level symbolic decomposition with feasibility checks [Ahn et al., 2022, Huang et al., 2022a, Wu et al., 2023a]; and foundation-policy work emphasized unified token-based control with heterogeneous data [Reed et al., 2022, Brohan et al., 2023, Bousmalis et al., 2023]. The 2024–2026 systems can be interpreted as deeper integration of these once-separate axes.

3.2 Axis A: Functionality Coupling

Functionality coupling specifies *what the world model is for* inside the embodied decision stack: is it optimized as a reusable predictive prior, or is it optimized as a control-facing component whose value is measured by improved decisions? We separate recent systems into three regimes: **decision-coupled world models**, **general-purpose world models**, and **hybrid pretrain-then-couple** systems that explicitly switch objectives across phases.

3.2.1 Decision-Coupled World Models (Closed-Loop Optimization)

Decision-coupled world models are trained and evaluated for direct control impact: improved success rate under distribution shift, reduced interventions, safer recovery, or higher long-horizon utility. In this regime, the world model is not merely a predictor; it is an *optimization interface* for planning, policy gradients, and post-training refinement [Cen et al., 2025a, Chen et al., 2026a, Wan et al., 2025, Zhu et al., 2025a, Peng et al., 2026].

Online-refined VLA pipelines. A central 2025 trend is to treat VLA policies as strong priors and then close the deployment gap through decision-coupled refinement loops. $\pi^*_{0.6}$ (RECAP) emphasizes intervention- and correction-aware post-training that targets deployment reliability rather than only imitation fit [Intelligence et al., 2025b]. Related refinement styles include simulator- or reward-mediated fine-tuning for robustness and failure recovery [Li et al., 2025b, Lu et al., 2025, Zang et al., 2025, Zhang et al., 2025d, Guo et al., 2025b, Li et al., 2025k, Xu et al., 2025b]. ConRFT further emphasizes combining offline and online signals under a unified refinement objective, reflecting the broader shift from static imitation to deployment-aware learning [Chen et al., 2025a]. Complementary post-training strategies include action-chunked PPO with self behavior cloning for stable policy improvement [Wang et al., 2025b], adaptive offline RL that balances signal and variance for flow-based VLA models [Zhang et al., 2025e], interactive post-training that leverages iterative human feedback during adaptation [Tan et al., 2025a], fine-tuning protocols that jointly optimize speed and success rate under constrained compute [Kim et al., 2025], and refined policy distillation that transfers VLA generalist priors into RL-specialized experts for targeted deployment [Jülg et al., 2025].

Concretely, we observe several decision-coupling patterns in the 2024–2026 VLA refinement literature:

- **Intervention/correction coupling:** RECAP explicitly models corrections and uses post-training loops to reduce deployment failures [Intelligence et al., 2025b]; related studies emphasize the gap between offline imitation and online recovery [Li et al., 2025b, Lu et al., 2025, Chen et al., 2025a].
- **RL-from-policy-prior coupling:** RL-style updates are applied on top of pretrained VLA priors to improve robustness and reduce intervention needs under distribution shift [Zang et al., 2025, Li et al., 2025k, Xu et al., 2025b, Zhang et al., 2025d,f].
- **Safety and reliability coupling:** safety-oriented variants treat risk and intervention as first-class training/evaluation signals rather than as afterthoughts, motivating explicit reliability objectives and calibration [Zhang et al., 2025c, Liu et al., 2025d, Pugacheva et al., 2025, Wu et al., 2026b].
- **Critic/value coupling:** policy updates are shaped by critics or action-quality estimators that provide decision-relevant gradients beyond behavior cloning likelihood [Zhai et al., 2025, Park et al., 2025, Häon et al., 2025].
- **Action-interface coupling:** some works argue that improved action parameterizations (tokenization/chunking) are prerequisites for stable decision-coupled post-training [Pertsch et al., 2025, Wang et al., 2025a, Wen et al., 2025c].

The breadth of decision-coupled systems in 2024–2026 is substantial. Among refinement-oriented approaches, $\pi^*_{0.6}$/RECAP couples post-training directly to intervention and correction signals, reporting more than doubled throughput and roughly halved failure rates on difficult real-world tasks [Intelligence et al., 2025b]. VLA-RFT surpasses strong supervised baselines with fewer than 400 fine-tuning steps using simulator-driven RL [Li et al., 2025b], while VLA-RL demonstrates that on-policy RL from pretrained VLA priors improves robustness under distribution shift without catastrophic forgetting [Lu et al., 2025]. RLINF-VLA achieves 98.11% task success across 130 LIBERO tasks through hindsight relabeling and online RL on a 3B-parameter VLA prior [Zang et al., 2025], and STARE-VLA reaches 98.0% on SimplerEnv via stabilized reward-based training [Xu et al., 2025b]. RobustVLA and SimpleVLA-RL provide complementary evidence that lightweight RL coupling suffices to close substantial deployment gaps [Zhang et al., 2025d, Li et al., 2025k]. ReinboT integrates dense return prediction into VLA training to capture data-quality nuances, achieving state-of-the-art performance on the CALVIN mixed-quality dataset with improved out-of-distribution generalization [Zhang et al., 2025f]. ConRFT unifies offline and online refinement under a single consistency objective evaluated on eight real-world manipulation tasks [Chen et al., 2025a].

Among world-model-guided stacks, WorldVLA co-trains action generation and future prediction in a shared autoregressive backbone, demonstrating that predictive co-modeling improves manipulation success without separate planning modules [Cen et al., 2025a]. BridgeV2W converts coordinate actions into pixel-aligned embodiment masks rendered from URDF and camera parameters, achieving state-of-the-art cross-view transfer on BridgeData manipulation [Chen et al., 2026a]. FlowDreamer introduces flow-style intermediates to reduce pixel-action mismatch in contact-rich settings [Guo et al., 2026], while WMPO explicitly optimizes policy through learned world-model rollouts, highlighting model-bias management as a core challenge [Zhu et al., 2025a]. WorldAgen and ReWorld exemplify long-horizon planning via imagined rollouts and decision scoring [Wan et al., 2025, Peng et al., 2026]. Reward-conditioned and alignment-oriented models make coupling explicit by treating reward as an input/output channel for rollout ranking [Xiao et al., 2025, Ren et al., 2026]. Critically, decision coupling must be paired with monitoring and trust signals: SafeVLA and VLA-Mark introduce safety markers and policy diagnostics to prevent reward hacking or unsafe deployment under aggregate metric optimization [Zhang et al., 2025c, Liu et al., 2025d, Valle et al., 2025, Wu et al., 2026b].

World-model-guided decision stack patterns. We categorize world-model-guided decision stacks by *where* the world model enters the control loop. In **joint prediction-action stacks**, action generation and predictive rollout share a representation backbone, as in WorldVLA [Cen et al., 2025a]. **Interface-alignment stacks** use explicit embodiment alignment—URDF/camera-conditioned masks or flow intermediates—to connect coordinate actions to pixel-space predictions, addressing the representation mismatch that otherwise limits cross-view transfer [Chen et al., 2026a, Guo et al., 2026]. **Planner-over-rollouts stacks** employ a planner (search, reasoning, or constraint filtering) that scores imagined trajectories and selects actions based on long-horizon structure; VLA-Reasoner, WorldAgen, and AdaPower exemplify this pattern with different planning interfaces [Guo et al., 2025a, Chen et al., 2025c, Wan et al., 2025, Peng et al., 2026, Huang et al., 2025c, Liu et al., 2026a]. Finally, **benchmark-anchored rollout diagnosis** evaluates rollout quality at the concept level rather than the pixel level, making world-model failures auditable for downstream decision-making [Upadhyay et al., 2026, Wu et al., 2026a].

Reward-conditioned and alignment-oriented world models. Reward- or preference-conditioned predictive models make coupling explicit by treating reward as an input/output channel of the world model, supporting task-conditioned rollouts and plan ranking [Xiao et al., 2025, Ren et al., 2026]. The main tradeoff is that reward alignment errors can dominate if the representation does not encode embodiment constraints; this motivates evaluation that disentangles perceptual quality from physically correct decision utility [Valle et al., 2025, Wu et al., 2026b, Upadhyay et al., 2026].

A gradient-alignment view of "decision coupling." We can formalize decision coupling as a consistency condition between updates that optimize predictive fidelity and updates that optimize decision utility. Let θ parameterize a policy π_θ and let ϕ parameterize a world model p_ϕ. Consider a decision objective $J(\pi_\theta)$ (e.g., expected success or value) and a world-model loss $\mathcal{L}_{\mathrm{wm}}(\phi;\theta)$ that depends on the policy-induced data distribution. A minimal coupling requirement is that model-focused updates should not systematically oppose decision improvement:

$$\langle \nabla_\theta J(\pi_\theta),\ -\nabla_\theta \mathcal{L}_{\mathrm{wm}}(\phi;\theta) \rangle \geq 0. \tag{15}$$

While few papers report this dot-product explicitly, the design patterns above (intervention-aware objectives, embodiment-aware interfaces, and planning-time constraints) can be interpreted as choices that increase alignment between predictive training signals and control utility [Intelligence et al., 2025b, Chen et al., 2026a, Cen et al., 2025a, Zhu et al., 2025a].

3.2.2 General-Purpose World Models (Reusable Predictive Priors)

General-purpose world models prioritize broad predictive capability and transfer, and only later attach downstream controllers or use the model as a data engine. This family includes platform-level initiatives and scaling efforts that emphasize multimodal dynamics modeling, large heterogeneous datasets, and reusable priors [NVIDIA et al., 2025b, Team et al., 2025a,b, Fan et al., 2026, Yin et al., 2026].

Cosmos positions world models as infrastructure (data pipelines, tokenization, post-training) rather than as a single monolithic controller module [NVIDIA et al., 2025b]. GigaWorld-0 and GigaBrain-0 emphasize that predictive models can be used to generate or curate training data at scale, effectively shifting the bottleneck from robot collection to world-model quality and filtering [Team et al., 2025a,b]. Mechanistic and semantic-consistency efforts motivate structure beyond pixels, arguing that perceptual realism does not guarantee physically correct causal transitions [Wang et al., 2026, Berg et al., 2025, Upadhyay et al., 2026].

Within this general-purpose regime, the dominant tradeoff is **breadth vs. controllability**. Broad models aim to cover many domains and tasks, but control-specific deployment often reveals that the learned dynamics are too weakly grounded in embodiment constraints (kinematics, contacts, latency) to directly support decision making without adaptation [Valle et al., 2025, Wu et al., 2026b,a]. This motivates two complementary strategies: (i) augment general-purpose models with more structured latent interfaces that preserve action-relevant factors while remaining scalable [Bi et al., 2025b, Tharwat et al., 2025, Tan et al., 2025b, Lillemark et al., 2026], and (ii) use the general-purpose model as an upstream generator/curator of data, then learn a control-specialized policy in a decision-coupled stage [Team et al., 2025a,b, Black et al., 2026, Intelligence et al., 2025b].

Within the general-purpose regime, several recurring subtypes emerge. **Platform/infrastructure** approaches, exemplified by Cosmos, emphasize modular pipelines for world modeling and post-training rather than monolithic architectures [NVIDIA et al., 2025b]. **Data-engine world models** such as GigaWorld-0 and GigaBrain-0 position the world model as a scalable data generator and curator, shifting the training bottleneck from physical robot collection to model quality and output filtering [Team et al., 2025a,b]. **Foundation video/world models** in the GENIE lineage emphasize broad transfer across domains and tasks through massive-scale training [Yehudai et al., 2024, Liao et al., 2025, Yin et al., 2026], while WOW and related video forecasting systems push long-horizon coherence at scale [Fan et al., 2026, Mei et al., 2026, Ye et al., 2026]. **Flow- and latent-structured models** use structured latent representations and flow-style formulations to improve rollout efficiency and controllability [Bi et al., 2025b, Tharwat et al., 2025, Tan et al., 2025b, Lillemark et al., 2026]. **Mechanistic and semantic consistency models** move beyond purely perceptual metrics to enforce concept-level physical constraints, arguing that pixel realism does not guarantee causally correct state transitions [Berg et al., 2025, Wang et al., 2026, Upadhyay et al., 2026]. While these models often report impressive prediction quality, the survey evidence indicates that control relevance must be engineered through interfaces (action conditioning, embodiment masks, geometry, or reward conditioning) or through explicit post-training [Cen et al., 2025a, Chen et al., 2026a, Li et al., 2025b].

3.2.3 Hybrid Pretrain-Then-Couple (Objective Switching)

Many successful systems are **hybrids**: they pretrain a broad prior (policy or world model), then explicitly switch to decision-coupled objectives for target deployment. The π_0 lineage is an archetype: π_0 and $\pi_{0.5}$ emphasize large-scale pretraining, while $\pi^*_{0.6}$ emphasizes post-training and correction loops that improve robustness in real-world settings [Black et al., 2026, Intelligence et al., 2025a,b]. Similar pretrain-then-couple patterns also appear in foundation VLA series and multi-stage pipelines that separate representation learning from later decision calibration [Kim et al., 2024, Cen et al., 2025b, Cai et al., 2026, Li et al., 2025i]. UP-VLA exemplifies this pattern by unifying understanding and prediction in a single model that first learns broad visual-language priors and then couples them with embodied action prediction for downstream manipulation [Zhang et al., 2025g].

Operationally, hybrids specify *when* coupling occurs and *what signal* drives the switch. A common pattern is: (i) pretrain a foundation policy or predictive model on heterogeneous offline data, (ii) align the action interface and task conditioning for a target embodiment (often via supervised adaptation), and (iii) perform decision-coupled post-training using RL, interventions, or planning-time constraints [Intelligence et al., 2025a, Black et al., 2026, Li et al., 2025i, Lu et al., 2025, Li et al., 2025b]. This framing helps explain why many 2025 papers report large gains from relatively small amounts of closed-loop data: the pretraining phase provides broad competence, and the coupling phase calibrates failure recovery and domain shift [Intelligence et al., 2025b, Li et al., 2025b, Zang et al., 2025].

We emphasize that this hybrid switch is not merely a training schedule detail; it changes the *failure surface*. Pretraining-dominant models often fail via systematic long-horizon drift and brittle recovery, while decision-coupled post-training tends to reduce catastrophic failures at the cost of additional infrastructure (safe rollout collection, intervention design, reward specification) [Valle et al., 2025, Wu et al., 2026b,a].

Representative pretrain-then-couple systems span the full spectrum from industry-scale pipelines to academic adaptation studies. The $\pi_{0.5}$ and π_0 models serve as large-scale foundation priors trained on heterogeneous dexterous robot data with flow-matching action heads; $\pi^*_{0.6}$ (RECAP) then applies an explicit post-training phase using demonstrations, on-policy rollouts, and teleoperator corrections that substantially improve deployment reliability [Intelligence et al., 2025a, Black et al., 2026, Intelligence et al., 2025b]. Open-data foundation policies, including OpenVLA and Octo, couple broad pretraining on Open-X and BridgeData with downstream adaptation on target embodiments, demonstrating that even 7B-parameter open-source models can match or exceed larger closed models on standard manipulation benchmarks [Kim et al., 2024, Collaboration et al., 2025, Team et al., 2024]. Foundation VLA series such as RYNNVLA-002 and InternVLA-A1 separate representation learning from decision calibration through multi-stage training recipes [Cen et al., 2025b, Cai et al., 2026, Liu et al., 2025e]. Control-calibration variants like ControlVLA explicitly target action-interface alignment and closed-loop refinement as the final coupling stage [Li et al., 2025i, Lu et al., 2025, Li et al., 2025b].

3.3 Representative Method Evidence

Representative system reports indicate that recent gains are tied to explicit design decisions, not only scale.

- π_0 **lineage:** π_0 reports flow-matching policy design on top of pretrained VLM priors and heterogeneous dexterous robot data; $\pi_{0.5}$ emphasizes heterogeneous co-training with semantic subtask signals for open-world generalization; $\pi^*_{0.6}$ introduces RECAP with demonstrations, on-policy data, and teleoperated corrections for deployment improvement [Black et al., 2026, Intelligence et al., 2025a,b].

- **Tokenization as a systems lever:** FAST explicitly attributes failures of naive per-dimension binning in high-frequency dexterous control and proposes DCT-based tokenization, reporting up to $5\times$ training speedups [Pertsch et al., 2025].
- **Action-world co-modeling:** WorldVLA frames action generation and future image prediction as mutually beneficial in one autoregressive stack, while VLA-RFT and VLA-RL highlight RL-style fine-tuning for robustness under distribution shift [Cen et al., 2025a, Li et al., 2025b, Lu et al., 2025].
- **Embodiment-conditioned world modeling:** BridgeV2W converts coordinate actions into pixel-aligned embodiment masks (from URDF and camera parameters) to align action control with video prediction and cross-view consistency [Chen et al., 2026a].
- **Reasoning traces as control scaffolds:** reasoning-augmented VLAs generate explicit intermediate plans or CoT-style traces, aiming to improve multi-step execution and recovery under ambiguous instructions [Ye et al., 2025a, Zhao et al., 2025a, Huang et al., 2025b, Yin et al., 2025].
- **Efficiency as an interface problem:** speculative decoding, caching, and pruning approaches reduce latency and memory, making deployment feasible on constrained platforms [Wang et al., 2025c,d, Xu et al., 2025c, Fang et al., 2025a, Zhang et al., 2025h].
- **Multi-modal contact grounding:** tactile/force/audio augmentation makes contact and compliance observable, addressing failure modes that are invisible to RGB-only policies [Bi et al., 2025a, Huang et al., 2025a, Yu et al., 2025, Wei et al., 2025].
- **3D geometry for viewpoint and contact stability:** depth/point/3D-aware VLAs emphasize geometry as a robustness prior for manipulation and navigation under viewpoint shift [Qu et al., 2025, Sun et al., 2025a, Zhang et al., 2025a, Zhen et al., 2024, Bhat et al., 2025].

These results are consistent with the functionality axis: models that explicitly connect representation learning to downstream control objectives tend to report better real-world robustness than purely decoupled predictive modeling. Similar behavior was already visible in earlier grounding-focused methods that constrained language plans with executable skills or grounded objectives [Ahn et al., 2022, Huang et al., 2023b, Dasgupta et al., 2023].

3.4 Axis B: Temporal Modeling

Temporal modeling describes *how* the model commits to time: predicting step-by-step rollouts, forecasting long segments, or mixing both with corrective feedback. This axis is central because embodied tasks are long-horizon but must act at high frequency, which makes compounding error, latency, and correction bandwidth first-order deployment constraints [Gupta et al., 2024, Cen et al., 2025a, Wu et al., 2026b].

3.4.1 Sequential Rollout (Step-by-Step Simulation)

Sequential rollouts simulate future states step by step and align naturally with MPC-style control, but face compounding error over long horizons [Li et al., 2025d, Fung et al., 2025, Cen et al., 2025a, Shah et al., 2026]. They are attractive when the agent must react to contact events and micro-corrections (e.g., grasp slippage) and when planning requires action-conditioned branching [Qian et al., 2025, Cen et al., 2025a].

Let ϵ_t denote one-step model error in a representation space. In a simplified sequential regime, rollout error can scale approximately as

$$\mathcal{E}_{t+H} \propto \sum_{k=0}^{H-1} \|\epsilon_{t+k}\|, \tag{16}$$

which explains why long-horizon multi-stage tasks are sensitive to model drift even when one-step metrics are strong [Gupta et al., 2024, Upadhyay et al., 2026, Wang et al., 2026].

Sequential rollouts are most effective when paired with explicit **replanning and correction** loops: the world model is rolled out for a short horizon, an action chunk is selected, the system executes a few steps, and the model is reconditioned on new observations. This fits naturally with VLA chunking interfaces and with intervention-aware refinement, where the goal is not to perfectly predict far-future pixels, but to maintain decision-relevant accuracy over the next few correction windows [Cen et al., 2025a, Chen et al., 2026a, Intelligence et al., 2025b, Wu et al., 2026b].

Representative sequential-rollout systems span several design families. WorldVLA couples action generation with stepwise prediction within a single autoregressive backbone, keeping the rollout state control-relevant and avoiding the need for separate planning modules [Cen et al., 2025a]. WristWorld adopts an egocentric, wrist-centric conditioning strategy that stabilizes manipulation rollouts under occlusion and viewpoint motion inherent to first-person operation [Qian et al., 2025]. Control-focused world models such as WMPO and the learning-to-control framework of Shah et al. [2026] emphasize sequential prediction as a scaffold for long-horizon MPC-like optimization, where the world model provides gradients directly to the policy. Counterfactual evaluation designs exploit sequential rollouts to score alternative action sequences, enabling counterfactual learning objectives that improve robustness to spurious correlations [Peng et al., 2025]. Perception-conditioned action models pair sequential prediction with strong perception modules—Percept-WAM uses perceptual grounding to reduce rollout drift under contact and occlusion [Han et al., 2025].

3.4.2 Global/Segment Prediction (Parallel Long-Horizon Forecasting)

Global prediction methods forecast trajectory segments or long-horizon futures in parallel. This can improve training efficiency and reduce autoregressive accumulation steps, but typically requires stronger inductive biases (action conditioning, embodiment constraints, semantic consistency) to preserve causal coherence across the horizon [Wan et al., 2025, Mei et al., 2026, Wang et al., 2026, Xie et al., 2026a].

One useful way to interpret global predictors is as learning a segment function F_ϕ that outputs a chunk $\hat{\mathbf{s}}_{t:t+H}$ directly. The key issue becomes *temporal consistency*: overlapping segments predicted from adjacent contexts should agree on their shared portion. Let $\hat{\mathbf{s}}^{(t)}_{t:t+H}$ denote a segment predicted from context up to t, and $\hat{\mathbf{s}}^{(t+\Delta)}_{t+\Delta:t+\Delta+H}$ denote a later segment. A minimal consistency requirement is

$$\left\| \hat{\mathbf{s}}^{(t)}_{t+\Delta:t+H} - \hat{\mathbf{s}}^{(t+\Delta)}_{t+\Delta:t+H} \right\| \leq \kappa \cdot \left\| \mathbf{o}_{\leq t+\Delta} - \mathbf{o}_{\leq t} \right\|, \tag{17}$$

for some κ that captures the model's sensitivity to new evidence. Eq. (17) motivates designs that reduce sensitivity through action- and embodiment-aware constraints (so that segment updates reflect physical evidence rather than purely perceptual drift) [Chen et al., 2026a, Guo et al., 2026, Wang et al., 2026].

Global prediction becomes compelling when **parallelism** is a primary constraint: predicting long segments in one forward pass can reduce iterative decoding cost and can simplify training on long videos. However, the deployment risk is that segment predictors may produce futures that are visually coherent but physically inconsistent with actions (e.g., violating contact constraints). Recent work argues that this is best addressed by adding action- and embodiment-aware constraints or by attaching decision-coupled post-training objectives that expose physical inconsistencies through closed-loop rollouts [Wan et al., 2025, Wang et al., 2026, Chen et al., 2026a, Wu et al., 2026a].

Representative global/segment predictors include long-horizon agentic generation and general video forecast models [Wan et al., 2025, Mei et al., 2026, Fan et al., 2026], as well as latent-structured predictors that aim to preserve controllable factors while reducing bandwidth [Tan et al., 2025b, Xie et al., 2026a].

Additional representative global-prediction works (used either directly for forecasting or as priors for downstream coupling) include flow- and video-forecasting variants [Lillemark et al., 2026, Ye et al., 2026, Yang et al., 2026] and benchmark/diagnostic efforts that emphasize evaluating global coherence under controlled shifts [Xiang et al., 2026, Wu et al., 2026a, Upadhyay et al., 2026].

3.4.3 Hybrid Chunk-and-Correct (Coarse Prediction + Local Refinement)

Recent hybrids combine chunk-wise global prediction with local sequential correction: a coarse global proposal provides long-horizon structure, while a local controller performs high-frequency corrective updates [Team et al., 2025a, Shen et al., 2026, Sendai et al., 2025, Lin et al., 2025a]. This design is especially natural for VLA policies that already operate with action chunking and periodic replanning: chunk-level outputs amortize expensive VLM inference, while correction mechanisms restore closed-loop reactivity [Black et al., 2026, Chen et al., 2025a, Wen et al., 2025c].

Hybrid designs can be interpreted as implementing a **coarse-to-fine temporal contract**: the global model proposes a trajectory distribution over a longer horizon, and a local model or controller corrects deviations using fresh observations. This structure matches the practical constraints of robot control, where (i) the agent must act within a strict latency budget and (ii) long-horizon plans must be revisable when contact dynamics or other agents change the state unpredictably [Wu et al., 2026b,a].

Representative hybrid temporal designs include compute-aware hybridization [Shen et al., 2026], as well as correction-style training objectives (leave-one-out or chunk-and-correct losses) that directly optimize consistency under partial rollouts [Sendai et al., 2025, Lin et al., 2025a].

3.5 Axis C: Spatial Representation

Spatial representation specifies the state interface between perception, prediction, and action. We separate four common choices: **compact latents**, **tokenized representations**, **geometry-aware (3D) representations**, and **rendering-**

aware representations. This axis matters because embodiment breaks many purely-visual invariances: camera motion, occlusion, contact geometry, and kinematic constraints all create failure modes when representation choices hide the relevant structure [Gupta et al., 2024, Chen et al., 2026a, Upadhyay et al., 2026].

3.5.1 Compact Latent Representations

Compact latent representations compress observations into low-dimensional continuous states that support real-time rollout and control. They are attractive when compute budgets are tight and when the primary goal is closed-loop control rather than high-fidelity rendering [Lee et al., 2024, Tharwat et al., 2025, Tan et al., 2025b]. In VLA settings, compact latents are often combined with token-based language context to keep the control state small while retaining instruction conditioning [Chen et al., 2025d, Li et al., 2025g].

The main limitation is that compact latents can hide geometry- and contact-critical information unless explicitly structured or trained with control-facing objectives. This is one reason compact-latent designs frequently appear in decision-coupled systems (where closed-loop losses reveal what information the latent must preserve) or are combined with explicit 3D modules when viewpoint and occlusion dominate [Cen et al., 2025a, Chen et al., 2026a, Qu et al., 2025].

Representative compact-latent works include behavior- and control-oriented latent designs that compress state for efficient policy learning [Lee et al., 2024], latent-structured world models such as Motus that target efficient rollout by compressing dynamics factors into low-bandwidth representations [Tharwat et al., 2025, Tan et al., 2025b, Bi et al., 2025b], and latent interfaces embedded within broader VLA stacks where language context supplies semantic structure while the latent supplies fast control state [Chen et al., 2025d, Li et al., 2025g].

3.5.2 Tokenized Representations

Tokenized representations discretize visual or action spaces into sequences compatible with transformer decoding. This makes it easy to unify vision, language, and action under one autoregressive interface, and it enables shared attention over multimodal tokens [Pertsch et al., 2025, Wang et al., 2025a, Liang et al., 2025b, Zhang et al., 2025i]. A recurring design question is *what to tokenize*: raw pixels (high bandwidth), learned VQ codes (lower bandwidth), frequency-domain tokens (control efficient), or structured object tokens (more semantic) [Pertsch et al., 2025, Wang et al., 2025a, Liang et al., 2025b, Bendikas et al., 2025]. TTF-VLA introduces temporal token fusion via pixel-attention integration that enriches action tokens with temporally grounded visual context, improving manipulation accuracy under dynamic scenes [Liu et al., 2025f].

Tokenization is a **systems lever** because it influences both training efficiency and control fidelity. FAST argues that naive discretization can fail in high-frequency dexterous control and proposes frequency-domain tokenization to improve efficiency [Pertsch et al., 2025]. VQ-based action/state designs aim to compress action bandwidth while keeping a transformer-friendly interface [Wang et al., 2025a, Liang et al., 2025b]. Focusing/object-centric tokenization emphasizes allocating capacity to task-relevant objects and interactions [Bendikas et al., 2025, Patratskiy et al., 2025]. At inference time, caching/speculative mechanisms can exploit token structure to reduce latency [Xu et al., 2025c, Wang et al., 2025c,d].

Representative tokenized-interface designs include FAST's frequency-domain action tokenization, which uses DCT-based encoding to preserve dexterous high-frequency action structure while achieving up to $5\times$ training-time reduction [Pertsch et al., 2025]; VQ-style action/state tokenization that compresses continuous control into transformer-friendly discrete codes [Wang et al., 2025a, Liang et al., 2025b]; tokenized spatial attention mechanisms that focus transformer capacity on task-relevant objects and contact regions [Bendikas et al., 2025, Patratskiy et al., 2025]; discrete-token policy interfaces that standardize multi-embodiment control pipelines through shared vocabulary [Wang et al., 2025e]; and token-level runtime acceleration through caching and speculative decoding that exploit the sequential structure of tokenized actions to reduce inference latency to real-time budgets [Xu et al., 2025c, Wang et al., 2025c,d].

3.5.3 Geometry-Aware / 3D Representations

Geometry-aware representations incorporate depth, point clouds, multi-view fusion, or explicit 3D structure to stabilize viewpoint changes and preserve contact-relevant constraints. This is especially important for manipulation and navigation where pixel similarity is a weak proxy for physical feasibility [Qu et al., 2025, Sun et al., 2025a, Zhang et al., 2025a, Li et al., 2025a, Zhen et al., 2024]. Recent designs span depth-augmented policies and action heads [Yuan et al., 2025a, Li et al., 2025l], point- and object-centric state interfaces [Li et al., 2026a, Singh et al., 2025], and more explicit spatial reasoning stacks that bind language plans to 3D constraints [Feng et al., 2025a, Koo et al., 2025, Argus et al., 2025, Patratskiy et al., 2025].

The key tradeoff is **accuracy vs. complexity**. Geometry-aware pipelines often require calibrated sensors, multi-view fusion, and more complex preprocessing, but they can dramatically reduce viewpoint-induced brittleness and improve contact consistency. This is particularly visible in depth-centric action heads and 3D-aware VLA formulations [Bhat et al., 2025, Zhen et al., 2024, Yuan et al., 2025a, Li et al., 2025l]. In driving and navigation, geometry is also a stability prior for multi-agent interaction and long-horizon safety, motivating hybrid representations that combine semantic tokens with explicit spatial maps or 3D structure [Guo and Zhang, 2025, Hao et al., 2025].

Representative geometry-aware systems include explicit spatial reasoning VLAs that bind language-conditioned plans to 3D structure—SpatialVLA uses learned spatial priors that improve cross-view generalization by 15–30% compared to RGB-only baselines [Qu et al., 2025, Feng et al., 2025a, Koo et al., 2025, Argus et al., 2025, Patratskiy et al., 2025]; depth-centric VLA designs such as GeoVLA and DepthVLA that augment policy inputs or action heads with calibrated depth for contact stability [Sun et al., 2025a, Yuan et al., 2025a, Li et al., 2025l]; 3D and 4D VLA formulations that enforce temporal-spatial consistency for long-horizon manipulation and navigation, with 3D-VLA and 4D-VLA incorporating point cloud and 4D scene representations respectively [Zhang et al., 2025a, Li et al., 2025a, Zhen et al., 2024, Bhat et al., 2025]; point- and object-centric interfaces such as PointVLA and OG-VLA that treat objects and points as primary state variables for compositional reasoning [Li et al., 2026a, Singh et al., 2025]; and spatially grounded policy learning studies that motivate geometry as a robustness prior across viewpoint and scene variations [Li et al., 2025m].

3.5.4 Rendering-Aware Representations

Rendering-aware representations introduce intermediate spatial fields or warping variables (e.g., optical flow) that better preserve pixel-space consistency under motion and viewpoint change. FlowDreamer uses flow-style intermediates to reduce pixel/action mismatch, while WristWorld highlights egocentric/wrist-centric modeling as a practical representation strategy for manipulation [Guo et al., 2026, Qian et al., 2025]. Simulation suites such as PhyScene motivate rendering-aware evaluation by controlling lighting, camera, and interaction conditions to probe physical generalization [Yang et al., 2024a].

Rendering-aware state interfaces are often used as **bridges** between pixel prediction and coordinate action control: optical flow, warping fields, and egocentric views make it easier to connect action-conditioned motion to visual change. This is complementary to geometry-aware design: geometry emphasizes physical structure, while rendering-aware design emphasizes view consistency and motion cues that improve prediction stability under camera motion [Chen et al., 2026a, Guo et al., 2026].

Representative rendering-aware systems include flow/warping intermediates [Guo et al., 2026], egocentric/wrist-centric interfaces [Qian et al., 2025], embodiment-mask rendering for pixel/action alignment [Chen et al., 2026a], and simulation suites that stress-test rendering variation and physical interaction [Yang et al., 2024a, Tai et al., 2025].

From a deployment perspective:

- **Compact latent states** are favorable when control frequency and onboard compute dominate constraints.
- **Tokenized states** are favorable when semantic alignment with language and chain-of-thought style planning is critical.
- **Geometry-aware states** are favorable when camera viewpoint shift, scene rearrangement, or contact geometry consistency is central.

No single representation is dominant across all tasks. Systems that report robust real-world transfer commonly use representation mixtures (e.g., semantic tokens + geometric priors + low-level action heads) rather than a single latent form [Intelligence et al., 2025a,b, Chen et al., 2026a, Zhang et al., 2026a].

An additional observation is that VLM quality alone is an imperfect predictor of downstream VLA behavior: VLM4VLA reports consistent benefits from VLM initialization but weak monotonicity between generic VLM capability and embodied-policy quality, reinforcing the need for embodied adaptation objectives [Zhang et al., 2026a].

3.6 Embodied Pipeline Mapping

Across 2024–2026 papers, we observe a recurrent template:

1. foundation pretraining over heterogeneous robot or video data,
2. adaptation via task conditioning and action-space alignment,
3. post-training or online correction for deployment robustness.

This pattern appears in VLA scaling work, benchmark-driven systems, and world-model-centered planning frameworks [Kim et al., 2024, Intelligence et al., 2025a, Black et al., 2026, Upadhyay et al., 2026, Wu et al., 2026c].

To make this mapping operational, we define three interface contracts:

- **Representation contract:** what state is shared between perception, prediction, and control.
- **Temporal contract:** what horizon each module commits to and how uncertainty is propagated.
- **Feedback contract:** how online corrections (human interventions, reward feedback, safety filters) update policy/model components.

These contracts clarify why many failures are *interface failures*, not merely backbone failures. Two systems with similar backbone scale can show different field behavior because they differ in interface consistency across planning, control, and adaptation loops [Li et al., 2025b, Wang et al., 2025f, Wu et al., 2026a].

3.7 VLA Architecture Sub-Taxonomy

Beyond the three world-model axes, we add a **VLA architecture sub-taxonomy** that captures how embodied policies instantiate perception-to-action mappings in practice. This sub-taxonomy is not a replacement for the three axes; rather, it provides an *implementation lens* that explains why two papers placed in the same coupling/temporal/spatial class can still behave differently in deployment.

3.7.1 Foundation VLAs

Foundation VLAs aim to cover broad instruction distributions and diverse embodiments with one model family. Open-data pipelines such as OpenVLA and Octo emphasize scalable data aggregation and consistent interfaces for pretraining, with OpenVLA demonstrating state-of-the-art open-source manipulation across 29 tasks using a 7B-parameter model [Kim et al., 2024, Collaboration et al., 2025, Team et al., 2024, O'Neill et al., 2024]. The π_0 lineage represents a foundation-policy direction that emphasizes large-scale heterogeneous pretraining with flow-matching action heads and then post-training for deployment robustness; $\pi_{0.5}$ extends this with semantic subtask signals for open-world generalization across multiple dexterous robots [Black et al., 2026, Intelligence et al., 2025a,b]. GR00T N1 introduces an explicit System 1/System 2 architecture, separating fast diffusion-based motor control from slow language-level reasoning to meet real-time constraints while preserving high-level task understanding [NVIDIA et al., 2025a, Cheang et al., 2024]. Additional foundation-style VLA series include VLA-0, which provides a standardized training and evaluation framework [Goyal et al., 2025], and the NORA series, which emphasizes modular and scalable foundation training [Hung et al., 2025a,b]. Broader consolidation efforts and training-recipe studies further codify the foundation VLA pattern [Li et al., 2024a, 2025j, 2024b]. InternVLA-M1 introduces a spatially guided framework that leverages explicit spatial reasoning within a generalist robot policy, demonstrating improved cross-scene manipulation through structured spatial conditioning [Chen et al., 2025e]. Domain-specialized foundation models extend the paradigm to specific manipulation competences while retaining broad conditioning, including MoManipVLA for mobile manipulation and GraspVLA for grasping [Wu et al., 2025, Deng et al., 2025, Neau et al., 2025, Du et al., 2025]. Early-2024 system studies and definition works established robust embodied training and evaluation protocols that informed subsequent developments [AhmadiTeshnizi et al., 2024, Huang et al., 2024b, Kazemi et al., 2024, Li et al., 2024c, Lin et al., 2024, Salzer and Visser, 2024, Zeng et al., 2024, Yang et al., 2024b]. Continual improvement directions include EvoVLA and evolve-style frameworks that emphasize iterative capability expansion rather than one-shot training [Liu et al., 2025g, Bai et al., 2025, Lin et al., 2025b,c], and pi-style policy variants beyond the main π_0 line [Jian et al., 2026, Xiang et al., 2025]. Platform-level and system reports further define data and training recipes for embodied foundations [Zhou et al., 2025a, Li et al., 2025f].

3.7.2 Action Head Variants

The **action head** determines how actions are represented and generated, and is often decisive for dexterity and latency. Recent variants include:

- **Autoregressive heads** (token-by-token actions), which integrate naturally with unified multimodal token streams but can be slow at high control frequency [Pertsch et al., 2025, Wang et al., 2025a, Liang et al., 2025b].
- **Diffusion/flow heads**, which generate action chunks with smoother trajectories and can improve robustness for contact-rich behaviors [Wen et al., 2025b,d,c,e, Tarasov et al., 2025, Zhong et al., 2025b].
- **Hybrid discrete-continuous heads**, which combine token interfaces with continuous residuals to balance controllability and expressiveness [Liu et al., 2025a,h, Chen et al., 2025f, Wang et al., 2025e].

- **Dual-system heads**, which separate slow reasoning/planning from fast motor execution to meet real-time constraints [Won et al., 2025, Fang et al., 2025b, Song et al., 2025a].

The choice of action head is often decisive for both dexterity and inference latency. Diffusion-based action generation is explored both as a modeling choice for smoother chunk generation and as a robustness lever under contact uncertainty: DexVLA applies diffusion heads specifically for dexterous bimanual manipulation, reporting improved grasp stability on contact-rich tasks [Chi et al., 2024, Wen et al., 2025b,d]. Flow-style action heads based on continuous-time or flow-matching formulations reduce sampling steps at inference while maintaining trajectory smoothness [Zhong et al., 2025b, Tarasov et al., 2025]. Hybrid heads and action-critic variants combine discrete token control with continuous residuals or quality estimators for precision—ACG uses action-conditioned critics to reshape policy updates [Liu et al., 2025a, Park et al., 2025, Zhai et al., 2025]. Expression- and reconstruction-focused variants further demonstrate that action-head design interacts with representation learning and training stability in ways that pure backbone scaling cannot address [Syed et al., 2025, Song et al., 2025b].

3.7.3 Multi-Modal Sensing VLAs

Multi-modal sensing VLAs extend beyond RGB to capture contact-relevant signals. Tactile and force augmentation improves manipulation reliability by making slip/contact observable [Bi et al., 2025a, Huang et al., 2025a, Zhang et al., 2025i, Yu et al., 2025, Zhang et al., 2025j, 2026b]. Audio augmentation is a complementary channel for contact events and tool interactions [Wei et al., 2025]. Multi-sensor policy reports emphasize that the main challenge is not only fusion architecture, but also data synchronization and interface contracts between sensors and action heads [Guo et al., 2025c, Hong et al., 2024, Hirose et al., 2025]. MLA further extends multi-modal integration by combining visual, tactile, and proprioceptive streams in a unified language-action model for multimodal understanding and forecasting in robotic manipulation [Liu et al., 2025i].

Tactile-augmented VLAs treat touch as a primary signal for slip and contact state estimation: VLA-Touch integrates GelSight-style tactile readings alongside RGB to improve grasp reliability under occlusion [Bi et al., 2025a, Huang et al., 2025a, Zhang et al., 2025i]. Force and compliance augmentation enables safer grasping under uncertain dynamics, with ForceVLA reporting reduced object damage rates through explicit force-feedback conditioning [Yu et al., 2025, Zhang et al., 2025j, 2026b]. Audio augmentation provides a complementary channel for contact events and tool interactions, particularly useful for tasks where visual feedback alone cannot disambiguate success (e.g., insertion clicks, material friction) [Wei et al., 2025]. Multi-sensor integration systems such as OmniVLA emphasize that the main challenge is not only fusion architecture but also data synchronization and interface contracts between heterogeneous sensors and action heads, with OmniVLA achieving 84% task success on diverse multi-sensor benchmarks [Guo et al., 2025c, Hirose et al., 2025, Hong et al., 2024].

3.7.4 Reasoning-Augmented VLAs

Reasoning-augmented VLAs explicitly generate intermediate reasoning traces (e.g., CoT-style plans) before or alongside actions. These approaches aim to improve multi-step task decomposition, robustness to ambiguous instructions, and error recovery through explicit intermediate representations [Ye et al., 2025a, Zhao et al., 2025a, Zhang et al., 2025b, Huang et al., 2025b, Yin et al., 2025, Gu et al., 2025, Li et al., 2025n, Zhong et al., 2026]. System-level reasoning layers (e.g., OS-style abstractions, reflection, or knowledge augmentation) can be interpreted as adding a slow planner that queries tools, memory, or world-model rollouts [Gao et al., 2025, Li et al., 2026b, Driess et al., 2025, Liu et al., 2026a].

The CoT-style VLA series exposes explicit intermediate plans before action generation: VLA-R1 adapts reinforcement learning on reasoning traces to improve multi-step execution, while CoT-VLA generates symbolic subgoal sequences that are then grounded in action space [Ye et al., 2025a, Zhao et al., 2025a, Zhang et al., 2025b]. Graph-structured and deep-thinking variants such as GraphCoT-VLA and DeepThinkVLA extend this with structured reasoning traces that capture spatial relationships and task hierarchies for long-horizon planning [Huang et al., 2025b, Yin et al., 2025]. ManualVLA and VLA-OS provide controlled reasoning interfaces that emphasize reliable tool and memory use through explicit OS-style abstractions [Gu et al., 2025, Gao et al., 2025]. Reflection and knowledge augmentation serve as slow-loop planning layers over fast action execution, with ChatVLA demonstrating that conversational feedback loops improve task completion through iterative self-correction [Li et al., 2026b, Driess et al., 2025, Liu et al., 2026a, Zhou et al., 2025b,c]. Action-oriented reasoning augmentation, including CoA-VLA and ACoT-VLA, ties plans explicitly to action affordances and control constraints rather than treating reasoning as a purely semantic process [Li et al., 2025n, Zhong et al., 2026]. VOTE optimizes VLA inference through trajectory ensemble voting, aggregating multiple candidate action sequences to improve robustness without additional training [Lin et al., 2025d].

3.7.5 Efficiency-Oriented VLAs

Efficiency-oriented VLAs target latency, memory, and training cost constraints. Recent methods include pruning/compression and structural sparsity [Fang et al., 2025a, Jabbour et al., 2025, Zhang et al., 2025h, Xiong et al., 2025], small/edge deployment models [Wen et al., 2025a, Shukor et al., 2025, Budzianowski et al., 2025, Wang et al., 2025e], and speculative/caching mechanisms that reduce inference overhead [Wang et al., 2025c,d, Xu et al., 2025c, Yu et al., 2026b]. FAST is a representative systems lever on the action interface side, showing that tokenization choice can be as impactful as backbone scale for training throughput [Pertsch et al., 2025]. PD-VLA accelerates VLA inference by integrating action chunking with parallel decoding, reducing latency without sacrificing action quality [Song et al., 2025c]. MergeVLA introduces cross-skill model merging that combines independently trained skill-specific VLAs into a unified generalist agent without retraining, enabling compositional skill transfer [Fu et al., 2025].

Efficiency-oriented methods span several complementary strategies. Structured pruning and compression reduce memory and latency without full retraining: SQAP-VLA applies structured quantization-aware pruning that preserves 95% of manipulation accuracy at 2× compression [Fang et al., 2025a, Jabbour et al., 2025, Zhang et al., 2025h, Xiong et al., 2025]. Edge and small-policy designs target resource-constrained platforms, with TinyVLA demonstrating effective manipulation from models under 1B parameters and SmolVLA providing an efficient open-source alternative [Wen et al., 2025a, Shukor et al., 2025, Budzianowski et al., 2025]. Speculative decoding and caching mechanisms exploit the sequential structure of token-based action interfaces for runtime reduction, with Spec-VLA reporting up to 3× inference speedup through speculative action token generation [Wang et al., 2025c,d, Xu et al., 2025c, Yu et al., 2026b]. Runtime analyses treat inference cost as a first-class systems constraint, providing actionable guidance on latency-accuracy tradeoffs for specific robot platforms [Hancock et al., 2025b,a]. Action tokenization choices that trade fidelity for compute enable both faster training and lower-latency inference [Pertsch et al., 2025, Xue et al., 2025].

3.7.6 Domain-Specific VLAs

Domain-specific VLAs specialize architecture and data to a domain with distinctive dynamics and safety constraints. Driving VLAs emphasize long-horizon prediction under multi-agent interaction, rare-event robustness, and stringent safety requirements [Guo and Zhang, 2025, Li et al., 2025c, Chi et al., 2025, Seong et al., 2025, Yuan et al., 2025b, Xu et al., 2025d, Hu et al., 2026, Hao et al., 2025]. Humanoid and whole-body policies emphasize high-dimensional control and stability constraints [Jiang et al., 2025a, Ding et al., 2025a]. Drone policies emphasize fast perception-control loops under viewpoint shift [Lykov et al., 2025, Serpiva et al., 2025]. Medical and assistive robotics emphasize reliability and compliance [Li et al., 2025o, Zhang et al., 2026b]. Game/task automation emphasizes tool use and long-horizon reasoning in interactive environments [Li et al., 2025p, Chen et al., 2026b, Wang et al., 2023].

Driving-specific VLAs address safety-critical interactive forecasting with long-horizon prediction under multi-agent interaction: VDrive emphasizes safety-aware interaction modeling, while DriveVLA-W0 integrates world-model-guided prediction with driving-specific action spaces [Guo and Zhang, 2025, Li et al., 2025c, Chi et al., 2025, Seong et al., 2025, Yuan et al., 2025b, Hao et al., 2025]. Diffusion-style driving policies emphasize smooth long-horizon control under uncertainty, with WAM-Diff applying diffusion denoising to generate safe trajectory distributions [Xu et al., 2025d]. Humanoid and whole-body control VLAs such as WholeBodyVLA address high-dimensional stability constraints across dozens of joints, requiring specialized action representations that preserve balance during loco-manipulation [Jiang et al., 2025a, Ding et al., 2025a]. Drone and racing VLAs emphasize fast control loops under aggressive viewpoint shift, with CognitiveDrone demonstrating outdoor navigation and RaceVLA targeting high-speed racing scenarios [Lykov et al., 2025, Serpiva et al., 2025]. Medical and assistive VLAs such as RoboNurse-VLA emphasize compliance, trust, and safety as primary design constraints rather than task throughput [Li et al., 2025o, Zhang et al., 2026b]. Interactive and tool-augmented VLAs support long-horizon task execution through explicit tool use and memory, with JARVIS-VLA demonstrating complex tool-mediated manipulation chains [Li et al., 2025p, Chen et al., 2026b, Zheng et al., 2025a,b].

3.8 Comprehensive Tables

We provide landscape tables to make the taxonomy auditable and to support quick cross-paper lookup. Fields marked "–" indicate values not explicitly reported in the cited papers.

Table 2: Foundation VLA models (Table 2): representative model families and implementation choices.

Model	Year	Backbone VLM	Action Head	Taxonomy Class	Params	Key Benchmarks	Real-World
OpenVLA [Kim et al., 2024]	2024	Llama-2 7B	token/AR	Hybrid pretrain-then-couple	7B	29 tasks; outperforms RT-2-X (55B)	Yes
SARA-RT [Leal et al., 2024]	2024	RT-2 backbone	token/AR	Foundation VLA policy	55B	RT-2 tasks	Yes
GR-2 [Cheang et al., 2024]	2024	custom VLM	diffusion	Foundation VLA policy	2.6B	GR-series tasks	Yes
Octo [Team et al., 2024]	2024	custom transformer	diffusion	Foundation VLA policy	93M	9 robot platforms; 800k trajs	Yes
$\pi_{0.5}$ [Intelligence et al., 2025a]	2025	PaliGemma 2 3B	flow	Hybrid pretrain-then-couple	3B	multi-robot dexterous	Yes
$\pi^*_{0.6}$ (RECAP) [Intelligence et al., 2025b]	2025	PaliGemma 2 3B	flow	Decision-coupled refinement	3B	2× throughput, 0.5× failure	Yes
π_0 [Black et al., 2026]	2026	PaliGemma 3B	flow	Hybrid pretrain-then-couple	3B	dexterous manipulation	Yes
GR00T N1 [NVIDIA et al., 2025a]	2025	Eagle 2 VLM	dual-system	Foundation VLA policy	–	humanoid loco-manipulation	Yes
InternVLA-A1 [Cai et al., 2026]	2026	InternVL-2.5	hybrid	Hybrid pretrain-then-couple	2B	LIBERO, real-world	Yes
FAST [Pertsch et al., 2025]	2025	any VLM	token/DCT	Efficiency-oriented (tokenization)	–	5× training speedup	Yes
VQ-VLA [Wang et al., 2025a]	2025	Qwen-2.5-VL	token/VQ	Tokenized action interface	3B	+30% long-horizon real-world	Yes
Discrete Diffusion [Liang et al., 2025b]	2025	Qwen-2.5-VL	discrete diff.	Tokenized action interface	3B	LIBERO 96.3%	–
DexVLA [Wen et al., 2025d]	2025	Qwen-2-VL 2B	diffusion	Dexterous manipulation VLA	2B	bimanual dexterous	Yes
FlowVLA [Zhong et al., 2025b]	2025	Qwen-2-VL	flow	Flow action head	2B	LIBERO, SimplerEnv	Yes
HybridVLA [Liu et al., 2025a]	2025	Qwen-2-VL	hybrid AR+diff	Hybrid action interface	2B	LIBERO, real-world	Yes
EdgeVLA [Budzianowski et al., 2025]	2025	Qwen-2.5-VL 3B	token/AR	Small/edge VLA	3B	real-time on edge	Yes
TinyVLA [Wen et al., 2025a]	2025	SigLIP+Phi-2	diffusion	Small/edge VLA	1B	LIBERO, real-world	Yes
SmolVLA [Shukor et al., 2025]	2025	SmolVLM	token/AR	Small/edge VLA	0.5B	LIBERO	Yes
VITA-VLA [Dong et al., 2025]	2025	Qwen-2-VL	distilled	Distillation-based VLA	2B	LIBERO 97.3%, real 82%	Yes
RLINF-VLA [Zang et al., 2025]	2025	Qwen-2-VL	token/AR	Decision-coupled RL	3B	LIBERO 98.11% (130 tasks)	Yes
STARE-VLA [Xu et al., 2025b]	2025	Qwen-2-VL	token/AR	Decision-coupled RL	3B	SimplerEnv 98.0%	Yes

Model	Year	Backbone VLM	Action Head	Taxonomy Class	Params	Key Benchmarks	Real-World
ReinboT [Zhang et al., 2025f]	2025	–	token/AR	RL return prediction	–	CALVIN SOTA	Yes
HAMSTER [Li et al., 2025q]	2025	Qwen-2-VL	hierarchical	Hierarchical VLA	7B	+20% over OpenVLA	Yes

Table 3: World model systems (Table 3): coupling, temporal, and spatial design choices.

System	Year	Coupling	Temporal	Spatial	Architecture	Domain	Key Result (as reported)
Cosmos [NVIDIA et al., 2025b]	2025	general-purpose	hybrid	token/latent	platform WM	general	infrastructure view
GigaWorld-0 [Team et al., 2025a]	2025	general-purpose	global	–	generation engine	general	scalable data engine
GigaBrain-0 [Team et al., 2025b]	2025	general-purpose	global	–	generation engine	general	scalable data engine
WorldVLA [Cen et al., 2025a]	2025	decision-coupled	sequential	token/latent	joint action+future	manipulation	co-modeling benefit
BridgeV2W [Chen et al., 2026a]	2026	decision-coupled	hybrid	rendering-aware	embodiment masks	manipulation	better alignment
WorldAgen [Wan et al., 2025]	2025	decision-coupled	global	–	agentic generation	general	long-horizon rollouts
ReWorld [Peng et al., 2026]	2026	decision-coupled	hybrid	–	WM-guided control	general	decision improvements
FlowDreamer [Guo et al., 2026]	2026	decision-coupled	hybrid	flow/rendering	flow intermediate	manipulation	reduced mismatch
WristWorld [Qian et al., 2025]	2025	decision-coupled	sequential	egocentric	video-conditioned	manipulation	stable egocentric prediction
WMPO [Zhu et al., 2025a]	2025	decision-coupled	sequential	–	WM policy opt.	robotics	policy improvement
VLA-Reasoner [Guo et al., 2025a]	2025	decision-coupled	hybrid	–	WM + reasoning	general	planning quality
Planning (WM-guided) [Chen et al., 2025c]	2025	decision-coupled	hybrid	–	planning augmentation	robotics	improved reliability
World-Env [Xiao et al., 2025]	2025	decision-coupled	global	–	reward-conditioned	general	task-conditioned rollouts
Aligning WM [Ren et al., 2026]	2026	decision-coupled	global	–	preference/reward	general	alignment focus
Mechanistic [Wang et al., 2026]	2026	general-purpose	global	structured	mechanistic prior	general	stronger physical consistency
Semantic WM [Berg et al., 2025]	2025	general-purpose	–	semantic	structured latent	general	semantic consistency
Motus [Bi et al., 2025b]	2025	general-purpose	sequential	latent	latent dynamics	general	compact rollout
Latent structured [Tharwat et al., 2025]	2025	general-purpose	–	latent	latent-space WM	general	efficient dynamics
Latent temporal [Tan et al., 2025b]	2025	general-purpose	–	latent	latent predictor	general	temporal modeling
Flow WM [Lillemark et al., 2026]	2026	general-purpose	global	rendering-aware	flow WM	general	scalable generation
GENIE (2025) [Liao et al., 2025]	2025	general-purpose	global	–	foundation WM	general	broad transfer
GENIE (2026) [Yin et al., 2026]	2026	general-purpose	global	–	foundation WM	general	broad transfer
WOW [Fan et al., 2026]	2026	general-purpose	global	–	foundation WM	general	large-scale modeling

System	Year	Coupling	Temporal	Spatial	Architecture	Domain	Key Result (as reported)
Learning WM [Shah et al., 2026]	2026	decision-coupled	sequential	–	WM for control	robotics	long-horizon control
Digital WM [Zhou et al., 2026]	2026	decision-coupled	hybrid	–	simulation-aligned	general	deployment relevance

Table 4: Domain-specific and multi-modal VLAs (Table 4): specialization targets and sensing interfaces.

Model	Year	Domain	Sensors	Key Architecture	Key Result
VDrive [Guo and Zhang, 2025]	2025	driving	RGB(+map)	VLA for driving	safety-aware interaction modeling
DriveVLA-W0 [Li et al., 2025c]	2025	driving	RGB	WM-guided driving	world-model-integrated driving
Impromptu [Chi et al., 2025]	2025	driving	RGB	long-horizon VLA	long-horizon planning
IRL-VLA [Jiang et al., 2025c]	2025	driving	RGB	inverse RL reward WM	1st runner-up CVPR2025 Grand Challenge
DriveAction [Hao et al., 2025]	2025	driving	RGB	action-driven benchmark	16k QA pairs from 2.6k scenarios
WholeBodyVLA [Jiang et al., 2025a]	2025	humanoid	RGB+proprio	whole-body control	high-DoF loco-manipulation
Humanoid-VLA [Ding et al., 2025a]	2025	humanoid	RGB+proprio	humanoid VLA	stability-constrained control
CognitiveDrone [Lykov et al., 2025]	2025	drone	RGB	fast perception loop	outdoor navigation
RoboNurse [Li et al., 2025o]	2025	medical	RGB	assistive VLA	safety and compliance focus
UrbanVLA [Li et al., 2025r]	2025	urban nav.	RGB	route-conditioned VLA	+55% over baselines on SocialNav
NitroGen [Magne et al., 2026]	2026	gaming	RGB	vision-action foundation	+52% on unseen games (1000+ games)
VLA-Touch [Bi et al., 2025a]	2025	manipulation	tactile+RGB	multimodal fusion	improved grasp under occlusion
Tactile-VLA [Huang et al., 2025a]	2025	manipulation	tactile+RGB	multimodal fusion	slip/contact detection
ForceVLA [Yu et al., 2025]	2025	manipulation	force+RGB	multimodal fusion	reduced object damage
Audio-VLA [Wei et al., 2025]	2025	manipulation	audio+RGB	multimodal fusion	audio-disambiguated contact
OmniVLA [Guo et al., 2025c]	2025	multi-domain	multi-sensor	multimodal VLA	84% task success, sensor diversity
DreamTacVLA [Ye et al., 2025b]	2025	manipulation	tactile+RGB	tactile world model	up to 95% on contact-rich tasks
JARVIS-VLA [Li et al., 2025p]	2025	interactive	RGB+tools	tool-augmented	long-horizon tool-mediated chains

Having established the taxonomy axes, VLA architecture sub-families, and their representative implementations, we turn next to the data regimes and evaluation metrics that ground these systems in empirical evidence. The coupling, temporal, and spatial design choices catalogued here only become meaningful when measured against standardized benchmarks and deployment-relevant metrics—the subject of Section 4.

4 Data Resources and Evaluation Metrics

The taxonomy in Section 3 characterizes systems by their design choices; this section complements that analysis by surveying the data regimes, benchmark suites, and evaluation metrics that determine how those design choices are empirically validated.

4.1 Data Regimes

Recent embodied research relies on four complementary data regimes, each with distinct strengths, biases, and scaling properties. We discuss each regime with concrete datasets and citation coverage to make the landscape auditable.

Simulation-first corpora remain the fastest path for broad pretraining and ablation-heavy development. PhyScene provides physically interactable 3D scene synthesis for embodied AI, enabling large-scale training on contact-rich manipulation with controlled physical properties [Yang et al., 2024a]. The GENIE lineage—spanning foundation world modeling and comprehensive simulation platforms—provides unified environments that support policy learning, evaluation, and data generation at scale; Genie Sim 3.0 introduces LLM-powered scene generation with over 10,000 hours of synthetic data across 200+ tasks and validates zero-shot sim-to-real transfer [Liao et al., 2025, Yin et al., 2026]. Exploration-driven generative interactive environments extend this paradigm by generating novel training scenarios through agent-driven exploration rather than hand-designed curricula [Savov et al., 2025]. MimicDreamer addresses the simulation-to-reality gap by aligning human demonstrations with robot-usable supervision through video diffusion and inverse kinematics, reporting a 14.7% improvement across six manipulation tasks [Li et al., 2025s].

Interactive benchmark suites improve comparability across methods but can overfit to narrow task interfaces when evaluation protocols are static. LIBERO and its extensions have emerged as central benchmarks: the original suite provides standardized manipulation tasks, while LIBERO-Plus introduces controlled perturbations across seven dimensions to expose VLA vulnerabilities—performance drops from 95% to below 30% under modest perturbations reveal fragility that headline numbers conceal [Fei et al., 2025]. VLATest provides a fuzzing framework that generates robotic manipulation scenes for systematic testing, revealing that all evaluated VLA models lack robustness under confounding objects, lighting variations, and instruction mutations [Wang et al., 2025g]. Eva-VLA offers the first unified framework for evaluating VLA robustness via continuous optimization across three variation domains (object 3D transformations, illumination, adversarial patches), with all variations triggering over 60% failure rates and up to 97.8% in long-horizon tasks [Liu et al., 2025e]. IRef-VLA provides a benchmark for interactive referential grounding with imperfect language in 3D scenes, evaluating how VLAs handle ambiguous and incomplete instructions in spatially complex environments [Zhang et al., 2025k]. For autonomous driving, DriveAction provides the first action-driven benchmark with 16,185 QA pairs from 2,610 driving scenarios organized in an action-rooted tree-structured evaluation framework [Hao et al., 2025].

Large offline robot datasets enable large-scale behavioral priors but inherit teleoperation and sensor bias. Open X-Embodiment aggregates data from 22 robots across 21 institutions with 527 skills and 160,266 tasks in a standardized format, demonstrating positive transfer across platforms in the resulting RT-X model [Collaboration et al., 2025]. DROID provides 76,000 trajectories and 350 hours of diverse manipulation data collected by 50 operators across 564 scenes on three continents [Khazatsky et al., 2025]. Octo leverages 800,000 trajectories from Open X-Embodiment for foundation policy training, demonstrating effective fine-tuning to new sensory and action spaces within hours on consumer GPUs [Team et al., 2024]. Scalable pretraining from human activity videos extends data availability further: Li et al. [2025t] pretrain on 1 million episodes and 26 million frames from unscripted egocentric human videos, using automated holistic analysis to generate atomic-level segments with 3D hand and camera motion, yielding strong zero-shot capabilities and favorable scaling behavior.

Real-world deployment logs are the only reliable source for intervention dynamics, recovery behavior, and edge-case calibration. The $\pi^*_{0.6}$ (RECAP) deployment pipeline explicitly collects and uses teleoperator corrections as training signal, demonstrating that intervention-aware post-training more than doubles throughput while halving failure rates [Intelligence et al., 2025b]. Galaxea provides a large-scale diverse robot behavior dataset collected in authentic environments, paired with a dual-system framework coupling VLM planning with VLA execution through a three-stage curriculum [Jiang et al., 2025d]. These deployment-oriented collections complement simulation data by providing the failure modes, edge cases, and intervention patterns that simulation environments systematically underrepresent [Wu et al., 2026b].

4.2 Data Curation Dimensions

Beyond raw scale, four curation dimensions strongly affect downstream behavior:

1. **Embodiment diversity** (single-arm, dual-arm, mobile manipulation, driving stacks, humanoid whole-body).
2. **Task horizon composition** (short atomic skills vs. multi-stage household workflows vs. long-horizon navigation).
3. **Interaction richness** (contact-heavy manipulation, tool use, intervention events, deformable objects).
4. **Annotation granularity** (language instructions, subgoal labels, proprioceptive traces, safety annotations, pixel-level segmentations).

Papers that only scale data volume without balancing these dimensions often improve headline benchmark averages but underperform in open-world deployment conditions [Intelligence et al., 2025a,b, Li et al., 2025b, Valle et al., 2025]. Interleave-VLA demonstrates the value of annotation richness by introducing an automated pipeline that converts Open X-Embodiment text instructions into interleaved image-text format (210,000 episodes), achieving 2× improvement in out-of-domain generalization [Fan et al., 2025a]. Similarly, PixelVLA shows that pixel-level annotations enable finer-grained reasoning: the Pixel-160K dataset with pixel-level labels yields 10.1–17.8% improvement over OpenVLA at only 1.5% of pretraining cost [Liang et al., 2025c].

Data sourcing strategies increasingly leverage non-robotic sources. Beyond human demonstration data, Yang et al. [2025b] propose using diffusion-based RL to generate high-quality training trajectories, achieving 81.9% success on LIBERO—5.3% above human-collected data and 12.6% above Gaussian RL data. This suggests that data quality, not merely data source, is the binding constraint for foundation policy performance.

4.3 Metric Families

We group evaluation metrics into five families that collectively capture the diagnostic dimensions needed for deployment-relevant assessment.

Task success and completion quality remains the most commonly reported metric family, but recent work reveals its limitations when used in isolation. Success rate (SR) aggregates task outcome into a binary signal, hiding execution quality, near-misses, and path efficiency. Long-horizon completion metrics decompose multi-stage tasks into subtask chains, revealing where policies fail—Long-VLA introduces the L-CALVIN benchmark specifically to evaluate long-horizon task chains rather than atomic skills [Fan et al., 2025b]. Throughput metrics (tasks completed per unit time) are increasingly reported alongside SR to capture operational efficiency [Intelligence et al., 2025b].

Control stability and safety metrics quantify the agent's reliability under sustained operation. Intervention rate (IR) measures human corrections per unit time and is central to deployment viability assessment. Collision rate, recovery latency, and safety constraint violation costs are increasingly reported: SafeVLA reduces cumulative safety violation cost by 83.58% while maintaining task success through constrained optimization [Zhang et al., 2025c]. Run-time analyses emphasize that inference latency itself is a safety-relevant metric, as policies that exceed real-time budgets create hazardous gaps in closed-loop control [Hancock et al., 2025a].

Prediction fidelity metrics assess the quality of world-model outputs. For video-based prediction, Fréchet Video Distance (FVD) and Fréchet Inception Distance (FID) quantify distributional similarity between generated and ground-truth sequences:

$$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\Big(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\Big), \tag{18}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of inception features from real and generated distributions, respectively. FVD extends this to temporal sequences using video features. However, perceptual quality alone is insufficient: Gupta et al. [2024] demonstrate that high FID/FVD scores can coexist with physically incorrect causal transitions, motivating concept-level evaluation that disentangles visual fidelity from dynamics correctness [Upadhyay et al., 2026, Wu et al., 2026a].

Generalization metrics measure transfer across distribution shifts. We distinguish four generalization axes: new scenes (novel spatial configurations), new objects (unseen instances or categories), new instructions (paraphrased or compositionally novel commands), and cross-embodiment transfer (different robot morphologies or sensor configurations). VLA^2 reports 44.2% improvement over OpenVLA on hard-level generalization by leveraging external knowledge modules for unseen concept manipulation [Zhao et al., 2025b]. ObjectVLA achieves 64% success on 100 novel objects not seen during training [Zhu et al., 2025b]. The generalization gap—the difference between in-distribution and shifted-distribution performance—can be formalized as:

$$\Delta_{\text{gen}} = \text{SR}_{\text{ID}} - \text{SR}_{\text{OOD}}, \tag{19}$$

where large Δ_{gen} indicates that reported benchmark performance overstates real-world capability. Recent evidence suggests that Δ_{gen} often exceeds 40 percentage points for state-of-the-art VLAs under modest distribution shift [Fei et al., 2025, Liu et al., 2025e, Pugacheva et al., 2025].

Efficiency metrics capture computational cost as a first-class constraint. Token and action efficiency measure the ratio of useful action tokens to total inference computation. Runtime latency must be compared against control frequency requirements: real-time factor (RTF) captures whether inference meets deployment timing budgets. Memory and compute cost determine hardware feasibility for edge deployment. CEED-VLA achieves $4\times$ inference acceleration through consistency distillation and early-exit decoding while maintaining task success [Song et al., 2025d]. LightVLA reduces FLOPs and latency by 59.1% and 38.2% respectively through differentiable token pruning with 2.6% success rate improvement [Jiang et al., 2025e].

4.4 Representative Quantitative Signals

Although protocols differ across papers, concrete reported numbers illustrate why evaluation must go beyond a single success metric. FAST reports up to $5\times$ training-time reduction under high-frequency dexterous settings through DCT-based action tokenization [Pertsch et al., 2025]. RECAP ($\pi^*_{0.6}$) more than doubles throughput and roughly halves failure rate on difficult tasks through intervention-aware post-training [Intelligence et al., 2025b]. VLA-RFT surpasses strong supervised baselines with fewer than 400 fine-tuning steps in simulator-driven RL [Li et al., 2025b]. RLINF-VLA achieves 98.11% task success across 130 LIBERO tasks [Zang et al., 2025]. VITA-VLA reaches 97.3% on LIBERO through efficient distillation from pretrained action experts [Dong et al., 2025]. ConRFT demonstrates evaluation on eight real-world manipulation tasks with a unified offline+online consistency objective [Chen et al., 2025a]. Valle et al. [2025] highlight that pure task success masks uncertainty and execution quality, motivating dedicated uncertainty and quality metrics.

These signals jointly support the same conclusion: **evaluation must be multi-objective**, combining competence, reliability, and efficiency.

A minimal closed-loop metric set can be formalized as:

$$\text{SR} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\text{task } i \text{ succeeds}], \tag{20}$$

$$\text{IR} = \frac{1}{N}\sum_{i=1}^{N} \frac{n_i^{\text{intervention}}}{T_i}, \tag{21}$$

$$\text{RTF} = \frac{\text{inference + planning time}}{\text{control horizon time}}, \tag{22}$$

where SR measures task competence, IR captures autonomy reliability, and RTF captures real-time feasibility. This triad is often more diagnostic than isolated success-rate reporting.

4.5 Dataset and Benchmark Catalog

Table 5 provides a structured overview of major datasets and benchmarks used across the 2024–2026 embodied AI literature.

4.6 Evaluation Protocol Recommendations

For reproducible and decision-relevant reporting, we recommend that embodied AI papers adopt a minimal reporting standard. First, at least one metric from each family (task success, control stability, prediction fidelity, generalization, efficiency) should be reported to prevent single-metric optimization from hiding deployment-critical weaknesses. Second, in-distribution and shifted-distribution performance should be separated, since the generalization gap Δ_{gen} (Eq. 19) is often the most informative signal for deployment readiness. Third, intervention-aware curves—plotting success rate against the number of allowed human interventions—should replace binary success reporting, as they

Table 5: Dataset and benchmark catalog (Table 5): representative data resources for embodied AI (2024–2026).

Resource	Year	Type	Domain	Scale	Key Feature
Open X-Embodiment [Collaboration et al., 2025]	2024	offline dataset	multi-robot	160k tasks	22 robots, 21 institutions, standardized format
DROID [Khazatsky et al., 2025]	2025	offline dataset	manipulation	76k trajs	564 scenes, 50 operators, 3 continents
LIBERO [Fei et al., 2025]	2024	benchmark	manipulation	130 tasks	standardized evaluation suite
LIBERO-Plus [Fei et al., 2025]	2025	robustness bench.	manipulation	7 perturbation axes	systematic vulnerability analysis
VLATest [Wang et al., 2025g]	2025	testing framework	manipulation	7 VLA models	fuzzing-based robustness evaluation
Eva-VLA [Liu et al., 2025e]	2025	robustness bench.	manipulation	3 variation domains	continuous optimization for physical variations
WorldBench [Upadhyay et al., 2026]	2026	diagnostic bench.	world models	concept-level	disentangled physical understanding
PhyScene [Yang et al., 2024a]	2024	simulation	embodied AI	large-scale	physically interactable 3D synthesis
Genie Sim 3.0 [Yin et al., 2026]	2026	simulation	humanoid	10k+ hours	LLM-powered scene generation, sim-to-real
Genie Envisioner [Liao et al., 2025]	2025	platform	manipulation	EWMBench	unified WM platform: policy, eval, sim
DriveAction [Hao et al., 2025]	2025	benchmark	driving	16k QA pairs	action-driven tree-structured evaluation
Galaxea [Jiang et al., 2025d]	2025	offline dataset	multi-task	large-scale	authentic environments, G0 dual-system
DOM [Xie et al., 2026b]	2026	benchmark	dynamic manip.	200k syn.+2k real	dynamic object manipulation
L-CALVIN [Fan et al., 2025b]	2025	benchmark	long-horizon	multi-stage	long-horizon task chain evaluation
Pixel-160K [Liang et al., 2025c]	2025	annotation set	manipulation	160k	pixel-level annotations for VLA
NitroGen dataset [Magne et al., 2026]	2026	game dataset	gaming	40k hours	1000+ games, auto action extraction

reveal the autonomy frontier more clearly than aggregate numbers. Fourth, compute budget, control frequency, and model update policy should be documented to enable fair cross-method comparison under resource constraints.

These elements are increasingly present in recent benchmark-oriented work and should become default for embodied world-model evaluation [Upadhyay et al., 2026, Wu et al., 2026a, Wang et al., 2025f]. WorldBench exemplifies this direction by emphasizing concept-level disambiguation rather than entangled aggregate physics tests, making failure attribution more actionable for model iteration [Upadhyay et al., 2026]. The VLATest fuzzing framework complements static benchmarks by generating adversarial evaluation scenarios that probe specific failure modes [Wang et al., 2025g].

4.7 Data Storage and Retrieval Infrastructure

As embodied AI datasets grow in scale and modality diversity, storage and retrieval infrastructure becomes a bottleneck. Lu and Tang [2025] survey five storage architectures and five retrieval paradigms for embodied AI data, reviewing over 180 studies. Key challenges include the physical grounding gap (aligning stored representations with real-world dynamics), cross-modal integration (maintaining temporal synchronization across RGB, depth, force, and proprioceptive streams), and open-world generalization (retrieving relevant experience from large heterogeneous corpora for novel deployment scenarios). IA-VLA demonstrates that augmenting VLA inputs with large VLM pre-processing can address semantic complexity in retrieval, particularly for tasks with visually indistinguishable objects that require external knowledge [Hannus et al., 2025].

Table 6: Qualitative comparison of representative embodied AI method families (2024–2026).

Family	Typical Strength	Typical Limitation	Representative Works	Deployment Fit
Foundation VLA policies	Strong instruction following, broad skill prior	Data and compute intensive; brittle OOD recovery	[Kim et al., 2024, Intelligence et al., 2025a, Black et al., 2026]	General-purpose manipulation
World-model-guided control	Better planning signal, sample efficiency, counterfactual reasoning	Model bias and rollout drift at long horizon	[Cen et al., 2025a, Wan et al., 2025, Chen et al., 2026a]	Long-horizon decision tasks
Post-training RL/refinement	Improves task throughput and robustness in deployment	Requires safe data collection and intervention design	[Intelligence et al., 2025b, Li et al., 2025b, Lu et al., 2025]	Continuous improvement loops
Efficiency-oriented methods	Lower latency and memory cost; easier edge use	Potential capability drop if over-compressed	[Pertsch et al., 2025, Yang et al., 2025a, Shen et al., 2026]	Resource-constrained systems
3D-aware VLAs	Viewpoint stability, contact consistency	Calibrated sensors, complex preprocessing	[Qu et al., 2025, Sun et al., 2025a, Zhang et al., 2025a]	Contact-rich manipulation
Reasoning-augmented VLAs	Multi-step decomposition, error recovery	Inference latency, reasoning hallucination	[Ye et al., 2025a, Zhao et al., 2025a, Guo et al., 2025a]	Ambiguous, long-horizon tasks
Multi-modal sensing VLAs	Contact/slip awareness, compliance	Sensor cost, data synchronization	[Bi et al., 2025a, Yu et al., 2025, Guo et al., 2025c]	Contact-rich, delicate manipulation
Domain-specific VLAs	Optimized for domain constraints	Limited cross-domain transfer	[Guo and Zhang, 2025, Jiang et al., 2025a, Li et al., 2025r]	Driving, humanoid, urban nav.

4.8 Current Gaps

Despite progress, several metric and data gaps persist. First, metric mismatch remains common: image-level prediction quality may not imply physically correct interaction outcomes, and short-horizon gains may not transfer to multi-stage tasks [Gupta et al., 2024, Valle et al., 2025, Wang et al., 2025f]. Second, adversarial robustness is largely untested: VLA-Fool demonstrates that multimodal adversarial attacks (textual, visual, and cross-modal misalignment) can severely degrade VLA outputs, while model-agnostic adversarial patches placed in camera view disrupt semantic alignment between visual and textual representations [Yan et al., 2025, Xu et al., 2025a]. Third, data licensing, robot-operator privacy, and intervention traceability increasingly constrain which datasets can be reused for large-scale pretraining, creating governance challenges that technical solutions alone cannot address. Fourth, sample efficiency varies dramatically: while some approaches require hundreds of thousands of trajectories, Hu et al. [2025a] demonstrate that VLAs can achieve 60–100% success on construction tasks with few-shot data when hierarchical task decomposition is properly designed.

These gaps collectively motivate evaluation protocols that jointly report dynamics realism, decision quality, deployment behavior, and adversarial robustness—moving beyond the current single-benchmark paradigm toward comprehensive deployment readiness assessment.

5 Cross-Family Comparison and Practical Tradeoffs

The data regimes and evaluation metrics surveyed in Section 4 provide the empirical substrate for comparing method families. In this section, we apply a normalized decision utility framework to compare the eight VLA sub-families identified in Section 3, using concrete reported numbers to anchor each family's strengths and deployment-relevant tradeoffs.

Table 6 summarizes high-level differences among major method families. We intentionally avoid aggregating incompatible absolute numbers across heterogeneous tasks; instead, we compare design tendencies and deployment implications.

5.1 Comparison Protocol

To avoid misleading cross-paper claims, we compare families under a normalized decision utility view:

$$\mathcal{U} = \alpha \cdot \mathrm{SR} - \beta \cdot \mathrm{IR} - \gamma \cdot \mathrm{RTF}, \tag{23}$$

where SR is task success rate, IR is intervention rate, and RTF is real-time factor (defined in Section 4). Coefficients (α, β, γ) are application-specific (e.g., higher β for safety-critical manipulation).

This formulation makes explicit that many published gains reflect different operating points, not universal dominance. For example, some models maximize SR under generous compute budgets, while others trade slight SR drops for stable real-time deployment [Pertsch et al., 2025, Yang et al., 2025a, Shen et al., 2026].

5.2 Where Each Family Wins

Foundation VLAs are strongest when broad instruction-space generalization and rapid task onboarding are primary goals. OpenVLA demonstrates that a 7B-parameter open-source model can match or exceed much larger closed models on standard manipulation benchmarks [Kim et al., 2024]. The $\pi_{0.5}$ lineage extends this to heterogeneous multi-robot co-training with semantic subtask signals for open-world generalization across diverse dexterous platforms [Intelligence et al., 2025a, Black et al., 2026]. However, foundation VLAs exhibit intervention-heavy recovery under compounding distribution shift, and Li et al. [2025u] show that much of the observed brittleness arises from spatial modeling misalignment rather than fundamental physical understanding deficits—Feature Token Modulation improves viewpoint accuracy from 48.5% to 87.1% with only 4,000 additional parameters.

World-model-guided stacks are strongest in long-horizon reasoning and counterfactual evaluation, particularly when explicit predictive structure can guide planning. WorldVLA co-trains action generation with future prediction in a shared autoregressive backbone [Cen et al., 2025a]. BridgeV2W achieves state-of-the-art cross-view transfer through embodiment-aligned interfaces [Chen et al., 2026a]. The weakness is representation mismatch and rollout bias when embodiment-specific constraints are weakly encoded—FlowDreamer addresses this through flow-style intermediates [Guo et al., 2026]. IRL-VLA demonstrates the strength of this family in driving, achieving first runner-up in the CVPR 2025 Autonomous Grand Challenge through inverse RL reward world models [Jiang et al., 2025c].

Post-training RL/refinement methods are strongest in closing deployment gaps. RECAP demonstrates more than doubled throughput and halved failure rates through intervention-aware post-training [Intelligence et al., 2025b]. RLINF-VLA achieves 98.11% across 130 LIBERO tasks through hindsight relabeling and online RL [Zang et al., 2025]. STARE-VLA reaches 98.0% on SimplerEnv via stabilized reward-based training [Xu et al., 2025b]. ConRFT unifies offline and online refinement under a single consistency objective evaluated on eight real-world tasks [Chen et al., 2025a]. These results indicate that static imitation pretraining is no longer sufficient for robust field behavior.

Efficiency-focused methods are strongest for latency-constrained and edge scenarios. FAST achieves up to 5× training speedup through DCT-based tokenization [Pertsch et al., 2025]. CEED-VLA achieves 4× inference acceleration through consistency distillation and early-exit decoding [Song et al., 2025d]. LightVLA reduces FLOPs by 59.1% and latency by 38.2% through differentiable token pruning [Jiang et al., 2025e]. The main risk is capacity loss if compression is applied without task-specific calibration [Guan et al., 2025, Yang et al., 2025a, Shen et al., 2026].

3D-aware VLAs are strongest when camera viewpoint shift, scene rearrangement, or contact geometry consistency is central. SpatialVLA uses learned spatial priors that improve cross-view generalization by 15–30% compared to RGB-only baselines [Qu et al., 2025]. GeoVLA and 4D-VLA incorporate depth and point cloud inputs to resolve pixel-action mismatch directly in the observation space [Sun et al., 2025a, Zhang et al., 2025a]. InSpire demonstrates that intrinsic spatial reasoning—prepending spatial relation queries to VLA inputs—mitigates spurious correlations without additional training data [Zhang et al., 2025l].

Reasoning-augmented VLAs are strongest for ambiguous instructions and multi-step task decomposition. VLA-R1 adapts RL on reasoning traces to improve multi-step execution [Ye et al., 2025a]. CoT-VLA generates symbolic subgoal sequences grounded in action space [Zhao et al., 2025a]. VLAPS embeds modified MCTS into VLA inference, achieving up to 67 percentage point improvements in success rates by controlling test-time compute [Neary et al., 2025]. The tradeoff is inference latency: explicit reasoning adds computation that may violate real-time constraints for high-frequency control.

Multi-modal sensing VLAs are strongest for contact-rich manipulation where visual feedback alone cannot disambiguate success. DreamTacVLA grounds VLA in contact physics through hierarchical tactile-visual perception and a tactile world model, achieving up to 95% success on contact-rich tasks [Ye et al., 2025b]. OmniVLA achieves 84% task success across diverse multi-sensor benchmarks [Guo et al., 2025c]. The primary challenge is data synchronization across heterogeneous sensor modalities.

Domain-specific VLAs excel when domain constraints dominate. UrbanVLA achieves 55% improvement over baselines on urban micromobility navigation through route-conditioned two-stage training [Li et al., 2025r]. Generalizable navigation systems extend embodied AI to in-the-wild environments through robust visual-language grounding [Wang et al., 2025h]. NitroGen demonstrates 52% relative improvement on unseen games through a vision-action foundation model trained on 40,000 hours of gameplay across 1,000+ games [Magne et al., 2026]. Joint optimization of fine-grained representation and workflow orchestration in metaverse articulated manipulation extends VLA methods to virtual environments with complex kinematics [Hu et al., 2025b]. Dexterous manipulation benefits from specialized architectures: end-to-end arm-hand VLA policies with shared autonomy achieve 90% success across diverse objects including unseen instances [Cui et al., 2025].

5.3 Quantitative Method Comparison

Table 7 provides concrete reported numbers for representative systems across method families. All numbers are reproduced from published results and should be interpreted within the context of each paper's specific evaluation protocol.

Table 7: Quantitative method comparison (Table 6): representative reported results across method families.

Method	Family	Benchmark	Params	Key Metric	Value	Key Design Choice
RLINF-VLA [Zang et al., 2025]	post-training RL	LIBERO (130 tasks)	3B	success rate	98.11%	hindsight relabeling + online RL
STARE-VLA [Xu et al., 2025b]	post-training RL	SimplerEnv	3B	success rate	98.0%	stabilized reward training
VITA-VLA [Dong et al., 2025]	distillation	LIBERO	2B	success rate	97.3%	action expert distillation
Discrete Diff. VLA [Liang et al., 2025b]	tokenization	LIBERO	3B	success rate	96.3%	discrete diffusion, parallel decoding
VQ-VLA [Wang et al., 2025a]	tokenization	real-world long-horizon	3B	success rate	+30%	VQ action tokenizer
RECAP [Intelligence et al., 2025b]	post-training	deployment	3B	throughput	2×	intervention-aware post-training
FAST [Pertsch et al., 2025]	efficiency	dexterous control	–	training time	5× reduction	DCT-based tokenization
CEED-VLA [Song et al., 2025d]	efficiency	inference	–	inference speed	4×	consistency distillation + early exit
LightVLA [Jiang et al., 2025e]	efficiency	LIBERO	–	FLOPs reduction	59.1%	differentiable token pruning
SafeVLA [Zhang et al., 2025c]	safety	safety violations	–	violation cost	−83.58%	constrained CMDP optimization
SpatialVLA [Qu et al., 2025]	3D-aware	cross-view	–	generalization	+15–30%	spatial priors
HAMSTER [Li et al., 2025q]	hierarchical	7 gen. axes	7B	success rate	+20% vs OpenVLA	hierarchical 2D path + 3D control
DreamTacVLA [Ye et al., 2025b]	multi-modal	contact-rich	–	success rate	95%	tactile world model
OmniVLA [Guo et al., 2025c]	multi-modal	multi-sensor	–	success rate	84%	multi-sensor fusion
UrbanVLA [Li et al., 2025r]	domain-specific	SocialNav/MetaUrban	–	success rate	+55%	route-conditioned VLA
VLA^2 [Zhao et al., 2025b]	generalization	hard-level gen.	–	success rate	+44.2%	agentic external knowledge
ObjectVLA [Zhu et al., 2025b]	generalization	novel objects	–	success rate	64% (100 novel)	VL pair implicit knowledge
CollabVLA [Sun et al., 2025b]	multi-agent	collaboration	–	time reduction	2×	self-reflective MoE
VLAPS [Neary et al., 2025]	reasoning	task planning	–	success rate	+67pp	MCTS at inference
Align-Then-Steer [Zhang et al., 2025m]	adaptation	cross-embodiment	–	real-world	+32%	VAE latent guidance
DynamicVLA [Xie et al., 2026b]	dynamic manip.	DOM benchmark	0.4B	continuous inference	real-time	convolutional encoder + streaming

Method	Family	Benchmark	Params	Key Metric	Value	Key Design Choice
MoRE [Zhao et al., 2025c]	quadruped	6 skills	–	OOD generalization	SOTA	RL Q-function MoE
SwitchVLA [Li et al., 2025v]	task switching	reactive switching	–	switch quality	smooth	execution-aware conditioning
Dual-Actor [Jin et al., 2025]	human-in-loop	3 tasks	–	success rate	100%	talk-and-tweak, 101 min online

5.4 Representative Case Studies

To anchor the comparison in concrete method behavior, we examine four system-level case studies that illustrate different points on the design space.

Scale-first foundation policy. The $\pi_0/\pi_{0.5}$ lineage emphasizes heterogeneous multi-robot and multimodal co-training to improve open-world manipulation coverage. π_0 introduces flow-matching policy design on top of pretrained VLM priors with heterogeneous dexterous robot data; $\pi_{0.5}$ extends this with semantic subtask signals for open-world generalization; and $\pi^*_{0.6}$ introduces RECAP with demonstrations, on-policy data, and teleoperated corrections [Black et al., 2026, Intelligence et al., 2025a,b]. This lineage demonstrates the pretrain-then-couple pattern at industry scale.

Deployment-first refinement. VLA-RFT, VLA-RL, and SimpleVLA-RL emphasize that distribution-shift robustness requires explicit post-deployment adaptation beyond static imitation. VLA-RFT surpasses supervised baselines with fewer than 400 online steps [Li et al., 2025b]. FPC-VLA takes this further with a dual-model framework integrating a VLA with a supervisor for failure prediction and correction, where the supervisor evaluates action viability via vision-language queries and generates corrective strategies [Yang et al., 2025c]. DreamVLA integrates dynamic-region-guided world knowledge prediction with spatial and semantic cues, achieving 76.7% real robot success and 4.44 average task length on CALVIN ABC-D [Zhang et al., 2025n].

Dexterous and contact-rich manipulation. DexVLA applies diffusion heads specifically for bimanual dexterous manipulation [Wen et al., 2025d]. End-to-end arm-hand VLA policies divide control between human VR teleoperation for arm macro-motions and autonomous DexGrasp-VLA for hand micro-motions, achieving 90% success across diverse objects including unseen instances [Cui et al., 2025]. Grover et al. demonstrate that preserving pretrained representations through dual-encoder design and string-based action tokenizers improves robustness to visual perturbations and novel instructions in dexterous settings [Grover et al., 2025].

Continual and adaptive VLAs. SwitchVLA models reactive task switching as behavior modulation conditioned on execution state, enabling smooth mid-execution task changes without external planners [Li et al., 2025v]. Dual-Actor fine-tuning achieves 100% success across three tasks within 101 minutes of online adaptation through a talk-and-tweak human-in-the-loop scheme that converts corrections into language commands [Jin et al., 2025]. On-the-fly VLA adaptation via test-time RL (TT-VLA) enables policy adaptation during inference using dense reward signals from step-by-step task-progress monitoring [Liu et al., 2026b].

5.5 Pareto Optimality Analysis

A formal Pareto analysis reveals that no single method family dominates across all dimensions simultaneously. We define the Pareto frontier over the multi-objective space $(\mathrm{SR}, -\mathrm{IR}, -\mathrm{RTF}, -\mathrm{Cost})$:

$$\mathcal{P} = \{m \in \mathcal{M} \mid \nexists\, m' \in \mathcal{M} : m' \succeq m \text{ on all objectives and } m' \succ m \text{ on at least one}\}, \tag{24}$$

where $\mathcal{M}$ is the set of evaluated methods. In practice, the Pareto frontier shows a clear structure: foundation VLAs and post-training methods occupy the high-SR, high-cost region; efficiency-oriented methods occupy the low-cost, moderate-SR region; and 3D-aware and multi-modal methods occupy specialized high-SR niches for specific task types at moderate cost.

This structure implies that **method selection is deployment-context-dependent**: there is no universally best approach, and the optimal choice depends on the specific tradeoff between task competence, autonomy, latency, and hardware budget.

5.6 Historical Continuity Across Families

Current families are not isolated inventions. Foundation VLAs inherit multi-embodiment token-policy ideas from Gato, RT-1, and RoboCat [Reed et al., 2022, Brohan et al., 2023, Bousmalis et al., 2023]. World-model-guided and planner-policy hybrids extend earlier language-grounding and feasibility-constrained planning lines [Ahn et al., 2022, Huang et al., 2023b, Wu et al., 2023a]. Data-scaling and adaptation loops connect to BridgeData-style collection, trajectory- and code-mediated control interfaces, and lifelong skill-library designs [Walke et al., 2023, Gu et al., 2023, Liang et al., 2023, Wang et al., 2023]. Pre-2024 generalist models for robot manipulation, including GR-1 which achieved 94.9% on CALVIN with zero-shot 85.4% on unseen scenes, established the viability of video-pretrained manipulation policies [Wu et al., 2023c]. Continuous scene representations and embodied agents with language-guided world modelling provided foundational representational ideas [Gadre et al., 2022, Nottingham et al., 2023, Dorbala et al., 2023]. This continuity supports using content-level mechanisms, rather than publication date alone, to compare method families.

5.7 Observed System-Level Pattern

Across recent systems, we observe a stable two-stage recipe:

1. build a large prior (foundation VLA or general world model),
2. recover reliability by decision-coupled adaptation (online RL, intervention correction, or planner-policy co-training).

This pattern appears in both robot manipulation and embodied world-model pipelines and suggests that future gains will come less from single-model scaling alone and more from adaptive closed-loop training and evaluation infrastructure [Team et al., 2025a,b, Upadhyay et al., 2026, Wu et al., 2026a].

Three practical tradeoffs dominate implementation decisions. **Breadth vs. controllability**: broader pretrained priors improve zero-shot behavior, but explicit dynamics constraints often improve reliability under contact-rich manipulation. **Long-horizon quality vs. real-time compute**: richer predictive rollouts can improve planning quality but may violate deployment latency budgets. **Offline scale vs. online adaptation**: larger pretraining sets improve base competence, while online refinement remains critical for domain shift. Recent systems such as DynamicVLA address the latency tradeoff directly with a compact 0.4B-parameter VLA using continuous inference that overlaps reasoning and execution, coupled with latent-aware action streaming for temporal alignment [Xie et al., 2026b].

These tradeoffs and system-level patterns set the stage for the open challenges discussed next: each challenge arises from a specific tension within the design space—between prediction fidelity and control relevance, between breadth and reliability, and between benchmark success and deployment readiness.

6 Open Challenges and Outlook

The cross-family analysis in Section 5 reveals that no single method family dominates across all deployment dimensions. The following seven challenges distill the most persistent gaps—each grounded in concrete empirical failures documented in the literature—and identify near-term research priorities that would most effectively advance the field.

6.1 Challenge 1: Long-Horizon Physical Consistency

Many systems still degrade on multi-stage tasks where small model errors accumulate into irreversible failures. As formalized in Section 2 (Eq. 14), rollout error can grow exponentially with the Lipschitz constant of the dynamics model, making long-horizon plans unreliable even when single-step prediction is accurate. WorldBench diagnoses this through concept-level physical evaluation that isolates specific failure modes rather than reporting entangled aggregate metrics [Upadhyay et al., 2026].

Two technical gaps are central. First, weak causal invariants under contact and object rearrangement allow models to exploit visual shortcuts that fail under scene perturbation—Eva-VLA demonstrates that all tested VLA models exhibit over 60% failure rates under modest physical variations, with up to 97.8% failure in long-horizon settings [Liu et al., 2025e]. Second, limited uncertainty calibration in long-horizon rollouts causes planners to over-trust model predictions, producing brittle control sequences under distribution shift. Mechanistic and semantic-consistency approaches address this by enforcing physically grounded prediction structure [Wang et al., 2026, Berg et al., 2025]. Future work should prioritize physically grounded temporal constraints and intervention-aware planning objectives, not only visual realism metrics [Gupta et al., 2024, Team et al., 2025b].

Long-VLA provides a concrete architectural response: a phase-aware input masking strategy segments subtasks into moving and interaction phases, enabling the first end-to-end VLA model for long-horizon tasks and introducing the L-CALVIN benchmark for systematic long-horizon evaluation [Fan et al., 2025b]. DreamVLA addresses long-horizon consistency through dynamic-region-guided world knowledge prediction that integrates spatial and semantic cues, achieving 4.44 average task length on CALVIN ABC-D [Zhang et al., 2025n].

6.2 Challenge 2: Embodiment-Aware Representation Alignment

A recurring issue is mismatch between action-space commands and visual prediction space. Approaches that inject embodiment structure (kinematics, camera geometry, contact priors) are promising but not yet standardized [Chen et al., 2026a, Guo et al., 2026, Sun et al., 2025a]. Li et al. [2025u] provide a precise diagnosis: VLA brittleness arises primarily from spatial modeling misalignment rather than physical modeling deficits. Their Feature Token Modulation improves viewpoint accuracy from 48.5% to 87.1% with only 4,000 parameters, while Feature Linear Adaptation achieves 90.8%—demonstrating that alignment, not capacity, is the binding constraint. ReVLA addresses a complementary visual domain limitation by reverting distribution shift in the visual encoder, restoring manipulation accuracy under novel backgrounds and lighting [Dey et al., 2025]. Similarly, Kachaev et al. [2025] demonstrate that aligning visual representations for out-of-distribution generalization—rather than scaling backbone capacity—yields large robustness gains under domain shift.

This mismatch is no longer a niche issue; it affects generalization across robot morphologies, viewpoint changes, and tool-based manipulation. Align-Then-Steer addresses cross-embodiment transfer through a VAE-based plug-and-play adaptation framework that embeds target actions into the pre-training latent distribution. This approach achieves +9.8% in simulation and +32% in real-world cross-embodiment settings [Zhang et al., 2025m]. InSpire demonstrates that intrinsic spatial reasoning—prepending "In which direction is [object] relative to robot?" queries to VLA inputs—mitigates spurious correlations without additional training data [Zhang et al., 2025l]. A key direction is to define representation interfaces that are simultaneously planner-friendly, control-grounded, and computationally efficient.

6.3 Challenge 3: Evaluation for Deployment, Not Only Benchmarks

Benchmark success remains an incomplete proxy for field reliability. LIBERO-Plus reveals that performance drops from 95% to below 30% under modest perturbations across seven dimensions. Moreover, models are largely insensitive to language variations, suggesting they tend to ignore instructions entirely rather than failing to parse them [Fei et al., 2025]. VLATest's fuzzing framework systematically exposes lack of robustness under confounding objects, lighting, camera poses, unseen objects, and instruction mutations across all seven evaluated VLA models [Wang et al., 2025g]. The community needs shared protocols that jointly evaluate safety, recovery, intervention rate, and sustained task throughput under shift [Upadhyay et al., 2026, Wu et al., 2026a, Valle et al., 2025].

In particular, evaluation suites should move from single-episode success to *session-level reliability*, including repeated-task stability, failure recovery quality, and operator load over extended runtime. The systematic review by Din et al. [2025], analyzing 102 VLA models, 26 datasets, and 12 simulation platforms, provides a comprehensive framework for understanding the current evaluation landscape, while Zhang et al. [2025o] examine approaches exhibiting core world model capabilities to identify what a fully realized evaluation framework should assess.

6.4 Challenge 4: Data Governance and Compute Efficiency

Scaling trends improve capability but increase data, compute, and reproducibility burdens. Efficient adaptation, model compression, and transparent data curation are central for practical adoption [Guan et al., 2025, Yang et al., 2025a, Shen et al., 2026].

Data governance is equally important: licensing, robot-operator privacy, and intervention traceability will increasingly influence which datasets can be reused for large-scale embodied pretraining. The survey by Lu and Tang [2025] identifies five storage architectures and five retrieval paradigms for embodied AI data, highlighting that the physical grounding gap and cross-modal integration challenges grow with dataset heterogeneity. As embodied datasets approach internet scale—Open X-Embodiment already aggregates data from 22 robots across 21 institutions [Collaboration et al., 2025]—governance frameworks must evolve beyond current ad-hoc practices.

On the efficiency front, recent results demonstrate that substantial capability can be preserved under aggressive compression. DynamicVLA achieves real-time dynamic object manipulation with a 0.4B-parameter model through continuous inference and latent-aware action streaming [Xie et al., 2026b]. CEED-VLA achieves 4× inference acceleration through consistency distillation [Song et al., 2025d]. These results suggest that the efficiency-capability frontier can be pushed further with careful architectural choices.

6.5 Challenge 5: Continual Adaptation Under Safety Constraints

Recent post-training results indicate that online adaptation is a major performance driver, but safe adaptation protocols are still immature [Intelligence et al., 2025b, Li et al., 2025b, Lu et al., 2025]. Open questions include how to schedule exploration under hard safety budgets, how to integrate teleoperator corrections without destabilizing pretrained priors, and how to prevent catastrophic forgetting during continual specialization.

SafeVLA provides a concrete starting point by framing VLA safety as a constrained MDP problem, reducing cumulative safety violation cost by 83.58% while maintaining task success through min-max optimization [Zhang et al., 2025c].

FPC-VLA introduces a dual-model framework where a supervisor evaluates action viability via vision-language queries and generates corrective strategies, outperforming baselines on both SIMPLER and LIBERO [Yang et al., 2025c]. On-the-fly VLA adaptation via test-time RL (TT-VLA) enables policy adaptation during inference using dense reward signals from step-by-step task-progress monitoring while preserving SFT and RL-trained priors [Liu et al., 2026b]. Dual-Actor fine-tuning demonstrates that human corrections can be converted into language commands for RL-based adaptation, achieving 100% success within 101 minutes of online training [Jin et al., 2025]. However, scaling these approaches to diverse deployment scenarios while maintaining safety guarantees remains an open challenge.

6.6 Challenge 6: Multi-Agent and Social Coordination

Beyond single-agent adaptation, embodied systems increasingly operate alongside other agents—both cooperative and adversarial. In driving and cooperative manipulation, multi-agent dynamics create failure modes that are invisible in single-agent benchmarks. The transition from single-agent to multi-agent embodied systems introduces coordination, communication, and social compliance as first-class design requirements.

CollabVLA introduces a self-reflective framework that transforms standard visuomotor policies into collaborative assistants through an MoE design integrating VLM-based reflective reasoning with diffusion-based action generation. This design cuts normalized task time by approximately $2\times$ and dream counts by $4\times$ compared to generative agent baselines [Sun et al., 2025b]. CoELA demonstrates that modular frameworks integrating LLMs with perception, memory, and execution modules enable effective multi-agent cooperation, with GPT-4-driven agents surpassing planning-based methods on cooperative household tasks [Zhang et al., 2024]. Organized team structures imposed through prompt-based designs improve cooperation quality through criticize-reflect processes for enhanced coordination [Guo et al., 2024]. The review by Li et al. [2025f] provides a comprehensive treatment of embodied multi-agent systems.

Key open problems include: (i) establishing shared world models that support coordinated planning across agents with different sensing and actuation capabilities; (ii) defining coordination metrics that capture both task efficiency and social compliance; (iii) scaling multi-agent embodied learning beyond tabletop cooperative tasks to household-scale and urban-scale environments; and (iv) handling adversarial or uncooperative agents in shared spaces.

6.7 Challenge 7: Interpretability and Trustworthiness

As VLA systems move toward deployment in safety-critical settings, interpretability and trustworthiness become necessary rather than optional. The opacity of end-to-end VLA models creates certification challenges and limits operator trust.

Häon et al. [2025] provide the first framework for mechanistic interpretation and steering of VLAs. By projecting feedforward activations onto token embedding bases, they identify sparse semantic directions (speed, direction) that can steer VLA behavior at inference time without fine-tuning—demonstrating that VLA internals are more structured and interpretable than previously assumed. Pugacheva et al. [2025] expose a complementary vulnerability: semantically similar irrelevant context in embodied AI commands can trigger up to 50% quality decline in VLA outputs, while LLM-based filtering recovers up to 98.5% of original performance. VLA-Mark introduces cross-modal watermarking for policy provenance and integrity verification, addressing the growing need for model authentication in deployment [Liu et al., 2025d]. VLA-OS provides OS-style abstractions for structured planning and monitoring, with controlled experiments revealing that visually grounded planning paradigms generally outperform language-only planning [Gao et al., 2025]. SafeVLA frames the trustworthiness problem formally through constrained optimization, reducing safety violations by 83.58% while preserving task performance [Zhang et al., 2025c]. Beyond safety, Hsieh et al. [2025] demonstrate that VLAs can be trained to reject physically impossible instructions, introducing a capability-awareness dimension that prevents execution of infeasible commands. Wang et al. [2025i] systematically explore adversarial vulnerabilities of VLA models in robotic settings, revealing that targeted perturbations to visual inputs can cause catastrophic policy failures even under benign task conditions.

Key open problems include: (i) developing causal rather than correlative explanations of VLA behavior under novel scenarios; (ii) creating online monitoring systems that can detect impending failures before they occur; (iii) establishing certification frameworks for embodied AI systems in regulated domains (medical, automotive, aviation); and (iv) designing human-VLA interfaces that convey appropriate levels of uncertainty and decision rationale.

6.8 Scope Boundaries

This survey focuses on embodied and decision-relevant world modeling within 2024–2026. Broader non-embodied world-model literature is therefore not covered in depth in the core analysis. The feature article by Feng et al. [2025b] provides a high-level overview of the progression from LLMs to world models in embodied AI. Existing surveys provide complementary perspectives: Dolgopolyi and Tsevas [2025] offer bibliometric analysis of VLM and VLA systems, Li et al. [2025e] provide a comprehensive review across five dimensions of VLA models, Shao et al. [2025] give the first systematic taxonomy of large VLM-based VLA architectures, and Sapkota et al. [2026] review VLA applications across autonomous vehicles, medical robotics, agriculture, and humanoid systems. Turgunbaev [2025] provide a concise overview of integrated VLA systems from perception to action. In fast-moving boundary areas, such as generic video world models later adapted to robotics, category boundaries will likely continue to shift as deployment evidence accumulates.

6.9 Near-Term Research Priorities

We identify five near-term priorities that would most effectively advance embodied intelligence:

Disentangled diagnostic evaluation should shift from monolithic benchmark scores to concept-isolated physical diagnostics and reasoning-action faithfulness checks. WorldBench and the physics-understanding analysis by Wu et al. [2026a] provide templates, but the community needs standardized toolkits that are as easy to deploy as existing benchmark suites [Upadhyay et al., 2026].

Action-world alignment under embodiment constraints should improve coordinate-to-pixel and language-to-control alignment using geometry-aware conditioning and consistency objectives. The success of BridgeV2W's embodiment masks and InSpire's spatial reasoning queries demonstrates that explicit alignment mechanisms can yield large improvements with minimal architectural overhead [Chen et al., 2026a, Wang et al., 2025f, Chen et al., 2025a, Zhang et al., 2025l].

Scalable but safe adaptation loops should combine synthetic or world-model-generated data with intervention-aware online refinement to improve robustness without uncontrolled exploration cost. The π_0 lineage and VLA-RFT demonstrate the recipe; extending it with formal safety guarantees (as in SafeVLA) is the key remaining challenge [Team et al., 2025a,b, Intelligence et al., 2025b, Li et al., 2025b, Zhang et al., 2025c].

Multi-agent coordination benchmarks are needed to move beyond single-agent evaluation. Current multi-agent embodied work is largely limited to tabletop cooperation and dialogue-conditioned tasks; scaling to household-scale and urban-scale coordination with heterogeneous agents remains underexplored [Sun et al., 2025b, Guo et al., 2024, Li et al., 2025f].

Interpretability-guided deployment should make mechanistic understanding actionable for deployment decisions. The sparse semantic directions identified by Häon et al. [2025] suggest that VLA internals can be monitored and steered in principled ways; integrating such tools into deployment pipelines would improve both safety and operator trust.

6.10 Outlook

We expect the next phase of embodied intelligence to converge on hybrid systems that combine reusable foundation priors, decision-coupled world models, online adaptation under safety constraints, and standardized evaluation pipelines tied to real deployment targets.

The strongest near-term gains will likely come from better coupling between predictive modeling and actionable control feedback. The evidence from 2024–2026 consistently shows that even modest amounts of decision-coupled post-training (hundreds of online RL steps, teleoperator corrections over minutes rather than hours) can close substantial deployment gaps that pretraining alone cannot address [Intelligence et al., 2025b, Li et al., 2025b, Zang et al., 2025].

Mid-term progress will depend on three developments. First, standardized deployment-centric evaluation must replace benchmark-centric evaluation as the primary measure of progress—the current gap between benchmark scores and real-world reliability is unsustainable. Second, safer continual learning protocols that maintain formal guarantees during online adaptation will enable broader deployment in regulated domains. Third, multi-agent and social coordination capabilities will extend embodied AI from single-robot manipulation to household-scale, urban-scale, and eventually societal-scale collaboration.

The convergence of foundation-scale pretraining, efficient adaptation, world-model-guided reasoning, and deployment-oriented evaluation creates a favorable moment for embodied AI. The systems surveyed here demonstrate meaningful progress toward physically grounded, decision-aware intelligent agents, though the gap between benchmark performance and sustained real-world reliability remains substantial and defines the next phase of the field.

7 Conclusion

7.1 Recap

This survey synthesized embodied AI and world-model research from 2024 to early 2026 under a coupled framework that links system-level embodied decision stacks with model-level dynamics design choices. We covered 318 papers spanning foundation VLA policies, world-model-guided control, post-training reinforcement refinement, efficiency-oriented adaptation, 3D-aware and reasoning-augmented VLAs, multi-modal sensing systems, and domain-specific applications across manipulation, driving, humanoid control, urban navigation, and gaming.

The three-axis taxonomy—functionality coupling, temporal modeling, and spatial representation—provides a principled decomposition that predicts deployment behavior more faithfully than architecture-centric or application-centric classifications. Functionality coupling, in particular, emerged as the most predictive axis: systems that explicitly connect representation learning to downstream control objectives consistently report better real-world robustness than purely decoupled predictive modeling, regardless of backbone scale.

7.2 Central Finding

The central conclusion is that strong embodied performance now depends on explicit coordination among representation, prediction, and control, rather than progress in any single module. The technical trajectory is cumulative: pre-2024 advances in task definition, language grounding, and early generalist robot policies established the interfaces that 2024–2026 systems now optimize at scale. Recent progress reflects integration of planning, world modeling, and policy adaptation into a single closed-loop training and deployment stack, extending rather than replacing the interfaces established before 2024.

The normalized decision utility framework (Section 5) and the Pareto analysis make explicit what the raw benchmark numbers often obscure: method selection is deployment-context-dependent, and there is no universally best approach. The two-stage recipe (large prior + decision-coupled adaptation) appears consistently across the strongest reported systems, suggesting that this pattern will remain dominant in the near term.

7.3 Future Outlook

Our overall reading of the current frontier is pragmatic: foundation-scale pretraining has become necessary but not sufficient. Reliable embodied intelligence increasingly requires decision-coupled post-training, representation interfaces aligned with embodiment constraints, and deployment-oriented evaluation protocols that quantify not just task completion, but sustained autonomy quality.

Seven open challenges structure the path forward. Long-horizon physical consistency and embodiment-aware alignment are primarily technical challenges amenable to architectural innovation. Deployment-oriented evaluation and data governance require community coordination and infrastructure investment. Safe continual adaptation straddles the boundary between technical methods and regulatory frameworks. Multi-agent coordination and interpretability/trustworthiness represent emerging frontiers that will become increasingly central as embodied AI systems move from laboratory demonstrations to real-world deployment.

7.4 Closing Statement

The convergence of foundation-scale pretraining, efficient adaptation, world-model-guided reasoning, and deployment-oriented evaluation creates a uniquely favorable moment for embodied AI. The systems surveyed here—achieving 98%+ success on standardized manipulation benchmarks while operating at 3B–7B parameters—demonstrate meaningful progress toward physically grounded, decision-aware intelligent agents, though substantial gaps remain before full deployment readiness. The pace of progress—from the first open-source 7B VLA in 2024 to 98%+ success rates on standardized manipulation benchmarks in 2025—suggests that the next phase will be defined not by whether these systems work, but by how reliably, safely, and broadly they are deployed.

A Full-Coverage Citation Map

A.1 In-Scope Citation Coverage (2024–2026)

This appendix groups all in-scope references by survey bucket and publication year to make coverage auditable.

A.2 agent-architecture

2026 Li et al. [2026a].

2025 Bai et al. [2025], Bendikas et al. [2025], Dey et al. [2025], Fang et al. [2025a], Guo et al. [2025c], Hancock et al. [2025a], Hsieh et al. [2025], Intelligence et al. [2025a], Jabbour et al. [2025], Jang et al. [2025], Li et al. [2025a,n,o], Lin et al. [2025d], Liu et al. [2025h,f], Patratskiy et al. [2025], Pertsch et al. [2025]. Pugacheva et al. [2025], Sendai et al. [2025], Song et al. [2025c], Tan et al. [2025a], Turgunbaev [2025], Wang et al. [2025i,c,d], Wen et al. [2025d,b,e,a], Wu et al. [2025], Xiang et al. [2025], Xiong et al. [2025].

2024 Leal et al. [2024], Yang et al. [2024b].

A.3 data-benchmark-eval

2026 Black et al. [2026], Cai et al. [2026], Chen et al. [2026a,b], Hu et al. [2026], Jian et al. [2026], Lillemark et al. [2026], Liu et al. [2026a], Magne et al. [2026], Mei et al. [2026], Peng et al. [2026], Ren et al. [2026], Upadhyay et al. [2026], Wang et al. [2026], Wu et al. [2026b,c,a], Xiang et al. [2026]. Xie et al. [2026b,a], Yang et al. [2026], Ye et al. [2026], Yu et al. [2026b], Zhang et al. [2026b,a].

2025 Argus et al. [2025], Bhat et al. [2025], Bi et al. [2025a], Cen et al. [2025b], Chen et al. [2025c,f], Chi et al. [2025], Collaboration et al. [2025], Cui et al. [2025], Deng et al. [2025], Din et al. [2025], Ding et al. [2025a], Du et al. [2025], Fan et al. [2025a,b], Fang et al. [2025b], Fei et al. [2025], Goyal et al. [2025]. Grover et al. [2025], Guo et al. [2025b], Guo and Zhang [2025], Han et al. [2025], Hancock et al. [2025b], Hannus et al. [2025], Hao et al. [2025], Hirose et al. [2025], Hu et al. [2025a], Huang et al. [2025c], Hung et al. [2025b], Jiang et al. [2025e,d,c,a], Jin et al. [2025], Jülg et al. [2025], Khazatsky et al. [2025]. Kim et al. [2025], Li et al. [2025g,c,q,p,h,l,t,m], Liang et al. [2025b,c], Liao et al. [2025], Lin et al. [2025b,c,a], Liu et al. [2025e,g,i]. Liu et al. [2025c], Lykov et al. [2025], NVIDIA et al. [2025a], Peng et al. [2025], Qian et al. [2025], Qu et al. [2025], Serpiva et al. [2025], Shukor et al. [2025], Singh et al. [2025], Song et al. [2025a,b], Sun et al. [2025a], Syed et al. [2025], Tan et al. [2025b], Tarasov et al. [2025], Team et al. [2025a], Tharwat et al. [2025], Valle et al. [2025]. Wang et al. [2025f,e,b,a], Wei et al. [2025], Wen et al. [2025c], Won et al. [2025], Xiao et al. [2025], Xu et al. [2025a], Xue et al. [2025], Yan et al. [2025], Yang et al. [2025b,a,c], Ye et al. [2025b,a], Yin et al. [2025], Yu et al. [2025]. Yuan et al. [2025b,a], Zhai et al. [2025], Zhang et al. [2025a,m,k,b,f,c,j,g,i], Zhao et al. [2025b], Zheng et al. [2025a,b], Zhong et al. [2025b], Zhou et al. [2025b].

2024 AhmadiTeshnizi et al. [2024], Chi et al. [2024], Huang et al. [2024b], Kazemi et al. [2024], Kim et al. [2024], Lee et al. [2024], Li et al. [2024c,a,b], Lin et al. [2024], O'Neill et al. [2024], Salzer and Visser [2024], Team et al. [2024], Yehudai et al. [2024], Zeng et al. [2024], Zhen et al. [2024].

A.4 foundation-definition

2025 Zhou et al. [2025a], Li et al. [2025f], Wang et al. [2025h].

2024 Cheang et al. [2024], Hong et al. [2024], Yang et al. [2024a].

A.5 planning-reasoning

2026 Li et al. [2026b], Sapkota et al. [2026], Zhong et al. [2026].

2025 Budzianowski et al. [2025], Chen et al. [2025e,d], Driess et al. [2025], Feng et al. [2025a], Gao et al. [2025], Gu et al. [2025], Hu et al. [2025b], Huang et al. [2025b,a], Hung et al. [2025a], Koo et al. [2025], Li et al. [2025k,v], Liu et al. [2025a], Neary et al. [2025], Neau et al. [2025], Seong et al. [2025]. Song et al. [2025d], Sun et al. [2025b], Wang et al. [2025g], Xu et al. [2025c,d], Zang et al. [2025], Zhang et al. [2025l,h], Zhao et al. [2025a,c], Zhou et al. [2025c].

2024 Guo et al. [2024], Huang et al. [2024a], Yoshikawa et al. [2024], Zhang et al. [2024].

A.6 policy-learning

2026 Liu et al. [2026b].

2025 Chen et al. [2025a], Dong et al. [2025], Fu et al. [2025], Häon et al. [2025], Intelligence et al. [2025b], Kachaev et al. [2025], Li et al. [2025i,s,j,r,u], Liu et al. [2025d], Lu et al. [2025], Park et al. [2025], Xu et al. [2025b], Zhang et al. [2025e,d], Zhu et al. [2025b].

A.7 survey-meta

2026 Fan et al. [2026], Yin et al. [2026], Yu et al. [2026a].

2025 Ding et al. [2025b], Dolgopolyi and Tsevas [2025], Guan et al. [2025], Jiang et al. [2025b], Li et al. [2025d,e], Liang et al. [2025a], Liu et al. [2025b], Lu and Tang [2025], Shao et al. [2025], Tai et al. [2025], Zhang et al. [2025n,o], Zhong et al. [2025a].

A.8 world-model-core

2026 Guo et al. [2026], Shah et al. [2026], Shen et al. [2026], Zhou et al. [2026].

2025 Berg et al. [2025], Bi et al. [2025b], Cen et al. [2025a], Chen et al. [2025b], Feng et al. [2025b], Fung et al. [2025], Guo et al. [2025a], Li et al. [2025b], NVIDIA et al. [2025b], Team et al. [2025b], Wan et al. [2025], Zhu et al. [2025a].

2024 Gupta et al. [2024].

B Exclusion Audit Log

B.1 Exclusion Audit

Entries excluded from the main in-scope set are listed by reason. Full row-level details are in `ref/paper_audit.csv`.

B.2 not_embodied_related

Bruce et al. [2024], Savov et al. [2025].

B.3 out_of_window

Ahn et al. [2022], Batra et al. [2020], Bousmalis et al. [2023], Brehmer et al. [2023], Brohan et al. [2023], Chebotar et al. [2023], Dasgupta et al. [2023], Dorbala et al. [2023], Driess et al. [2023], Duan et al. [2022], Fan et al. [2022], Gadre et al. [2022], Gao et al. [2022], Gu et al. [2023], Huang et al. [2023b, 2022a,b, 2023a]. Jiang et al. [2023], Liang et al. [2023], Nottingham et al. [2023], Reed et al. [2022], Sarch et al. [2023], Song et al. [2023], Walke et al. [2023], Wang et al. [2023], Wu et al. [2023b,a,c], Xu et al. [2023], Zhao et al. [2023].

References

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. π_0: A Vision-Language-Action Flow Model for General Robot Control, January 2026. URL `http://arxiv.org/abs/2410.24164`. arXiv:2410.24164 [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. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization, April 2025a. URL `http://arxiv.org/abs/2504.16054`. arXiv:2504.16054 [cs].

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 2025a. URL http://arxiv.org/abs/2503.14734. arXiv:2503.14734 [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].

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 2025a. URL http://arxiv.org/abs/2502.14795. arXiv:2502.14795 [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].

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].

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 2025b. URL http://arxiv.org/abs/2501.03575. arXiv:2501.03575 [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].

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].

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 2025a. URL http://arxiv.org/abs/2506.22242. arXiv:2506.22242 [cs].

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*, 2025a. URL https://openreview.net/forum?id=dT45OMevL5.

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].

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].

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].

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].

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.

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 2025b. URL http://arxiv.org/abs/2511.19912. arXiv:2511.19912 [cs].

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].

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 2025b. URL http://arxiv.org/abs/2510.00406. arXiv:2510.00406 [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].

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 2025a. URL http://arxiv.org/abs/2502.05450. arXiv:2502.05450 [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].

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].

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].

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*, 2025a. URL https://ieeexplore.ieee.org/abstract/document/10900471/.

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].

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].

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].

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 2025c. URL http://arxiv.org/abs/2510.12796. arXiv:2510.12796 [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].

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 2025b. URL http://arxiv.org/abs/2506.24044. arXiv:2506.24044 [cs].

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].

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/.

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/.

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].

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].

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 2022a. URL http://arxiv.org/abs/2207.05608. arXiv:2207.05608 [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].

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].

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].

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 2023a. URL http://arxiv.org/abs/2307.05973. arXiv:2307.05973 [cs].

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 2023b. URL http://arxiv.org/abs/2303.00855. arXiv:2303.00855 [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 2023. URL http://arxiv.org/abs/2305.16291. arXiv:2305.16291 [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].

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].

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].

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].

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].

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].

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].

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].

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`.

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].

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].

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].

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 2025a. URL `http://arxiv.org/abs/2503.10631`. arXiv:2503.10631 [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].

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].

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 2025b. URL http://arxiv.org/abs/2407.06886. arXiv:2407.06886 [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 2025a. URL http://arxiv.org/abs/2508.10399. arXiv:2508.10399 [cs] version: 1.

Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A Comprehensive Survey on World Models for Embodied AI. 2025d. 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 Future? A Comprehensive Survey of World Models. *ACM Comput. Surv.*, 58(3):57:1–57:38, September 2025b. ISSN 0360-0300. doi:10.1145/3746449. URL https://dl.acm.org/doi/10.1145/3746449.

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].

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 2026a. URL http://arxiv.org/abs/2510.24795. arXiv:2510.24795 [cs].

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].

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.

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 2025e. URL http://arxiv.org/abs/2508.15201. arXiv:2508.15201 [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].

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].

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].

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].

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].

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].

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].

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].

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, 2025f. URL https://ieeexplore.ieee.org/abstract/document/11036708/.

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 2025c. URL http://arxiv.org/abs/2503.03480. arXiv:2503.03480 [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, 2025a. URL https://ieeexplore.ieee.org/abstract/document/11128017/.

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].

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 2026a. URL http://arxiv.org/abs/2510.16281. arXiv:2510.16281 [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].

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 2026b. URL http://arxiv.org/abs/2601.18692. arXiv:2601.18692 [cs].

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, 2022b. URL https://proceedings.mlr.press/v162/huang22a.html.

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].

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].

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.

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

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/forum?id=bJbSbJskOS.

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].

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].

Wei Li, Renshan Zhang, Rui Shao, Jie He, and Liqiang Nie. CogVLA: Cognition-Aligned Vision-Language-Action Model via Instruction-Driven Routing & Sparsification, October 2025g. 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 2025h. URL http://arxiv.org/abs/2511.09516. arXiv:2511.09516 [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].

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 2025c. URL http://arxiv.org/abs/2510.07134. arXiv:2510.07134 [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, September 2025b. URL http://arxiv.org/abs/2506.23919. arXiv:2506.23919 [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 2025a. URL http://arxiv.org/abs/2510.13237. arXiv:2510.13237 [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 2026a. URL http://arxiv.org/abs/2601.03309. arXiv:2601.03309 [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 2025a. URL http://arxiv.org/abs/2507.01016. arXiv:2507.01016 [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 2025b. URL http://arxiv.org/abs/2508.20072. arXiv:2508.20072 [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].

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].

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/forum?id=egbFo1gvYp.

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. $\pi^*_{0.6}$: a VLA That Learns From Experience, November 2025b. URL http://arxiv.org/abs/2511.14759. arXiv:2511.14759 [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 2025i. URL http://arxiv.org/abs/2506.16211. arXiv:2506.16211 [cs].

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 2024a. URL http://arxiv.org/abs/2412.14058. arXiv:2412.14058 [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 2025j. URL http://arxiv.org/abs/2511.14178. arXiv:2511.14178 [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*, 2025b. URL https://openreview.net/forum?id=VdwdU81Uzy.

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].

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].

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].

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].

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].

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/.

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 2025a. URL http://arxiv.org/abs/2511.09515. arXiv:2511.09515 [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].

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 2025d. URL http://arxiv.org/abs/2511.01331. arXiv:2511.01331 [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].

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 2025k. URL http://arxiv.org/abs/2509.09674. arXiv:2509.09674 [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 2025b. URL http://arxiv.org/abs/2512.05107. arXiv:2512.05107 [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 2025b. URL http://arxiv.org/abs/2509.25718. arXiv:2509.25718 [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 2025e. URL http://arxiv.org/abs/2509.04063. arXiv:2509.04063 [cs].

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

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].

Hongyin Zhang, Zifeng Zhuang, Han Zhao, Pengxiang Ding, Hongchao Lu, and Donglin Wang. ReinboT: Amplifying Robot Visual-Language Manipulation with Reinforcement Learning, May 2025f. URL http://arxiv.org/abs/2505.07395. arXiv:2505.07395 [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, 2025d. URL https://aclanthology.org/2025.emnlp-main.1342/.

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].

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].

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].

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 2025c. URL http://arxiv.org/abs/2509.25681. arXiv:2509.25681 [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].

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].

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

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

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].

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].

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].

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 2025b. URL http://arxiv.org/abs/2512.10226. arXiv:2512.10226 [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].

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].

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].

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].

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, 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].

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].

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 2025g. URL http://arxiv.org/abs/2501.18867. arXiv:2501.18867 [cs].

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].

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].

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 2025e. URL http://arxiv.org/abs/2509.18953. arXiv:2509.18953 [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, 2025c. URL https://aclanthology.org/2025.emnlp-main.1367/.

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 2025d. URL http://arxiv.org/abs/2509.05614. arXiv:2509.05614 [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 2025c. URL http://arxiv.org/abs/2502.02175. arXiv:2502.02175 [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 2025h. URL http://arxiv.org/abs/2503.20384. arXiv:2503.20384 [cs].

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].

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/forum?id=aYQYOVm2AB.

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].

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].

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 2026a. URL http://arxiv.org/abs/2601.05611. arXiv:2601.05611 [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].

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].

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].

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].

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].

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].

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 2025d. URL http://arxiv.org/abs/2507.23682. arXiv:2507.23682 [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 2025i. URL http://arxiv.org/abs/2505.09577. arXiv:2505.09577 [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].

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 2025f. URL http://arxiv.org/abs/2508.19257. arXiv:2508.19257 [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.

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 2025e. URL http://arxiv.org/abs/2509.09372. arXiv:2509.09372 [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 2025a. URL http://arxiv.org/abs/2510.13375. arXiv:2510.13375 [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 2025l. URL http://arxiv.org/abs/2510.14836. arXiv:2510.14836 [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/.

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].

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].

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].

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].

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 2025m. URL http://arxiv.org/abs/2510.12276. arXiv:2510.12276 [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.

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].

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].

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

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].

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].

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].

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].

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 2024b. URL http://arxiv.org/abs/2311.01378. arXiv:2311.01378 [cs].

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].

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.

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].

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].

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].

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].

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 2024b. URL http://arxiv.org/abs/2311.12871. arXiv:2311.12871 [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].

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, 2024c. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b631da756d1573c24c9ba9c702fde5a9-Abstract-Datasets_and_Benchmarks_Track.html.

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].

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].

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].

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.

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

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].

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 2025c. URL http://arxiv.org/abs/2511.04555. arXiv:2511.04555 [cs].

Yina Jian, Di Tian, Xuan-Jing Chen, Zhen-Yuan Wei, Chen-Wei Liang, and Mu-Jiang-Shan Wang. PI-VLA: Adaptive Symmetry-Aware Decision-Making for Long-Horizon Vision-Language-Action Manipulation. *Symmetry*, 18(3):394, March 2026. ISSN 2073-8994. doi:10.3390/sym18030394. URL https://www.mdpi.com/2073-8994/18/3/394.

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].

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 – ECCV 2024*, volume 15120, pages 129–148. Springer Nature Switzerland, Cham, 2025a. 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.

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 2025d. URL http://arxiv.org/abs/2502.05855. arXiv:2502.05855 [cs].

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].

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].

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. 2025h. URL https://openreview.net/forum?id=8VyjwyLuSl.

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 2025f. URL http://arxiv.org/abs/2511.01718. arXiv:2511.01718 [cs].

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].

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].

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].

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 2025b. URL http://arxiv.org/abs/2508.10333. arXiv:2508.10333 [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 2025j. URL http://arxiv.org/abs/2509.07962. arXiv:2509.07962 [cs].

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 2026b. URL http://arxiv.org/abs/2601.15541. arXiv:2601.15541 [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].

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.

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].

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 2025i. URL http://arxiv.org/abs/2509.26642. arXiv:2509.26642 [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].

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, 2025n. 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.

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].

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].

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].

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].

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/.

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 2025c. URL http://arxiv.org/abs/2505.21906. arXiv:2505.21906 [cs].

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 2025d. URL http://arxiv.org/abs/2507.05116. arXiv:2507.05116 [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].

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].

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 2026b. URL http://arxiv.org/abs/2601.19634. arXiv:2601.19634 [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.

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].

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

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].

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].

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 2025b. URL `http://arxiv.org/abs/2509.01944`. arXiv:2509.01944 [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 2025d. URL `http://arxiv.org/abs/2512.11872`. arXiv:2512.11872 [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].

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].

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, 2025o. URL `https://ieeexplore.ieee.org/abstract/document/11246030/`.

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 2025p. URL `http://arxiv.org/abs/2503.16365`. arXiv:2503.16365 [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].

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].

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/`.

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].

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 2025q. URL `http://arxiv.org/abs/2502.05485`. arXiv:2502.05485 [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].

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 2025c. URL `http://arxiv.org/abs/2508.06571`. arXiv:2508.06571 [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 2025r. URL `http://arxiv.org/abs/2510.23576`. arXiv:2510.23576 [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].

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].

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`.

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 2025s. URL `http://arxiv.org/abs/2509.22199`. arXiv:2509.22199 [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].

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 2025g. ISSN 2994-970X. doi:10.1145/3729343. URL `https://dl.acm.org/doi/10.1145/3729343`.

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 2025k. URL `http://arxiv.org/abs/2503.17406`. arXiv:2503.17406 [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].

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 2025t. URL http://arxiv.org/abs/2510.21571. arXiv:2510.21571 [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 2025d. URL http://arxiv.org/abs/2509.00576. arXiv:2509.00576 [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 2025a. URL http://arxiv.org/abs/2505.02152. arXiv:2505.02152 [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 2025c. URL http://arxiv.org/abs/2511.01571. arXiv:2511.01571 [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].

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 2025b. URL http://arxiv.org/abs/2508.19958. arXiv:2508.19958 [cs].

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 2025b. URL http://arxiv.org/abs/2510.14902. arXiv:2510.14902 [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 2025b. URL http://arxiv.org/abs/2502.19250. arXiv:2502.19250 [cs].

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 2025d. URL http://arxiv.org/abs/2506.13725. arXiv:2506.13725 [cs].

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 2025e. URL http://arxiv.org/abs/2509.12594. arXiv:2509.12594 [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 2026b. URL http://arxiv.org/abs/2601.22153. arXiv:2601.22153 [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.

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].

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].

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].

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 2025u. URL http://arxiv.org/abs/2512.02902. arXiv:2512.02902 [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].

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].

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].

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*, 2025h. URL https://ieeexplore.ieee.org/abstract/document/11206420/.

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/.

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].

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 2025m. URL http://arxiv.org/abs/2509.02055. arXiv:2509.02055 [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 2025c. URL http://arxiv.org/abs/2503.08007. arXiv:2503.08007 [cs].

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 2025v. URL http://arxiv.org/abs/2506.03574. arXiv:2506.03574 [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].

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].

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 2025n. URL http://arxiv.org/abs/2507.04447. arXiv:2507.04447 [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].

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 2026b. URL http://arxiv.org/abs/2601.06748. arXiv:2601.06748 [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].

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.

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.

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

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/.

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].

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].

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 2025o. URL http://arxiv.org/abs/2511.02097. arXiv:2511.02097 [cs].

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].

Wen-Han Hsieh, Elvis Hsieh, Dantong Niu, Trevor Darrell, Roei Herzig, and David M. Chan. Do what? 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.

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, 2025i. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_Exploring_the_Adversarial_Vulnerabilities_of_Vision-Language-Action_Models_in_Robotics_ICCV_2025_paper.html.

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/.

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.