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 27, 2026

ABSTRACT

Embodied intelligence requires agents to perceive multimodal environments, execute goal-directed actions under physical constraints, and anticipate how those actions reshape future world states—capabilities central to both robotic manipulation and autonomous driving. Over the past two years, 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-based control stacks that jointly optimize representation, prediction, and decision making (**???????**). This survey presents a unified framework covering **318 papers** from January 2024 to February 2026. We formalize the embodied control problem as a partially observable Markov decision process and derive a shared learning objective that couples latent dynamics modeling with downstream 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 robotics, autonomous driving, and embodied simulation, covering five metric families—task success, control stability, prediction fidelity, generalization, and compute efficiency—and compare more than 30 representative systems across four method families: foundation VLA policies (**????**), world-model-guided control (**??**), post-training reinforcement refinement (**????**), and efficiency-oriented adaptation (**???**). The analysis distills six open challenges: long-horizon physical consistency, embodiment-aware representation alignment, deployment-oriented evaluation, compute governance, safe continual adaptation, and multi-agent coordination. 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.

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

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

large language and vision models with the physical precision of real-world control—defines the central research agenda of the current period.

Recent progress has been striking. As of early 2026, generalist robot policies can follow open-ended natural language instructions for dexterous household manipulation (**??**), humanoid systems demonstrate whole-body loco-manipulation in unstructured environments (**???**), and world-model-guided stacks generate physically plausible synthetic data at scale to mitigate real-robot collection costs (**???**). In parallel, VLA architectures have proliferated dramatically, with specialized variants addressing 3D spatial reasoning (**????**), tactile and force feedback (**???**), chain-of-thought reasoning (**????**), reinforcement post-training (**????**), efficiency and edge deployment (**?????**), and autonomous driving (**????**). The breadth and pace of this progress make a unified, decision-oriented synthesis both timely and necessary.

1.1 Hook: 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 **?** 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 (**??**). Early multimodal agents demonstrated that language could structure complex sequential behaviors, but relied heavily on symbolic planners that were brittle under perceptual ambiguity (**?**). Open-ended survival and exploration settings underscored the importance of lifelong skill accumulation and self-guided curriculum construction (**?**). 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 (**?**). 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 (**?**). Subsequent works refined this paradigm: LLM-Planner enabled open-vocabulary spatial navigation through in-context learning (**?**); Plan-and-Solve decomposed reasoning from execution (**?**); JARVIS combined neuro-symbolic reasoning with modular execution for dialogue-conditioned task completion (**?**); VoxPoser synthesized 3D affordance and value maps from language descriptions to guide manipulation (**?**). The grounded decoding paradigm further showed that semantic constraints could directly modulate token generation for physical feasibility (?). Meanwhile, Voyager demonstrated that GPT-4 could iteratively design, refine, and accumulate executable skill libraries for open-ended agents (?).

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 (?). RT-1 established transformer-based real-robot manipulation from diverse task demonstrations collected via large-scale data pipelines (?). Q-Transformer extended this to offline RL over large datasets while preserving the autoregressive token prediction interface (?). RoboCat demonstrated few-shot adaptation across robot embodiments and tasks through self-improvement loops (?). PaLM-E showed that embodied policies could benefit from grounding large multimodal language models with physical sensorimotor data (?). VIMA unified manipulation commands across heterogeneous task types through multimodal prompt engineering (?). RT-Trajectory and Code-as-Policies explored trajectory-sketch and code-mediated control interfaces (??). BridgeData V2 and ACT-style teleoperation recipes accelerated data collection for low-cost manipulation systems (??). Diffusion Policy established denoising diffusion as a powerful generative framework for visuomotor control that naturally handles multimodal action distributions (?). 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 (?). Second, specialized architectural innovations proliferated to address VLA limitations: flow-matching policies for dexterous control (?), frequency-space action tokenization for high-frequency tasks (?), hybrid autoregressive-diffusion architectures (?), and dual-system designs separating slow reasoning from fast motor execution (?). Third, world models were tightly coupled with VLA pipelines: WorldVLA unified action prediction with future image synthesis in a single autoregressive stack (?), BridgeV2W aligned coordinate-space actions with pixel-space predictions through URDF-rendered embodiment masks (?), and GigaWorld-0/GigaBrain-0 established world models as data engines to generate training distributions at scale (??).

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 (?). 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 (?). Li et al. provide the most taxonomically complete world-model survey to date (?), while Ding et al. give a broader but less embodiment-focused treatment (?). Zhong et al. and Yu et al. survey the VLA methodology landscape (??). Jiang et al. cover VLA for autonomous driving specifically (?). Fung et al. emphasize the world model as the core reasoning component for embodied agents (?).

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 (????), multi-sensory fusion beyond vision (????), chain-of-thought and reinforcement reasoning (???), whole-body humanoid control (??), multi-robot coordination (???), and safety-aware deployment (??). 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 (?); Wu et al. analyze what video generation models understand about physics (?); Valle et al. argue for uncertainty and quality metrics beyond binary success (?); Wu et al. systematically expose pragmatic failure modes (?). 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 (**???**). This line also developed multimodal dialogue-conditioned benchmarks such as DialFRED (**?**) and open-ended task completion in realistic household simulators.

The second line developed language-grounded planning with explicit feasibility checks and environment feedback (**?????????**). 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 (**?????????????**). These developments yielded the action tokenization, data scaling, and VLM initialization insights that 2024–2026 systems inherit and extend.

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 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 six open challenges grounded in concrete empirical failures reported across the literature, each with specific research priority recommendations.

1.7 Paper Organization

Section **??** introduces mathematical foundations for embodied control and latent world-model training. Section **??** presents the three-axis taxonomy, interface contracts, and embodied pipeline mapping with comprehensive method coverage. Section **??** surveys data regimes, curation dimensions, benchmark categories, and evaluation metric families. Section **??** provides cross-family comparison including method-level tables, case studies, and the two-stage pattern analysis. Section **??** distills six open challenges with near-term research priorities. Section **??** 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 (**??**).
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 (**???**).
3. **3D spatial memory**: explicit voxel grids or point clouds that accumulate multi-view RGB-D observations into a persistent geometric representation (**?????**). 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 (**???**).

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 (**??????**). 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 (**?**)) and in chain-of-thought reasoning VLAs that generate explicit subgoal sequences before actions (**????**).

Generalist transformer-control systems showed that heterogeneous action modalities can be cast as autoregressive token prediction over tokenized observation-action sequences (**????**). 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 (**?**) and later systematically studied in VLM4VLA (**?**), 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 (**?**), while FAST proposed DCT-based frequency-space tokenization to preserve dexterous high-frequency action structure (**?**).

2.4 Latent World Models for Embodied Control

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

$$z_t \sim q_\phi(z_t \mid o_{\le t}, a_{<t}), \qquad \hat{z}_{t+1} \sim p_\theta(\hat{z}_{t+1} \mid z_t, a_t), \tag{5}$$

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:

$$\hat{o}_{t+1} \sim p_\theta(o_{t+1} \mid \hat{z}_{t+1}), \qquad \hat{y}_{t+1} = g_\psi(\hat{z}_{t+1}), \tag{6}$$

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 (**????**).

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 (**??**).
- 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 (**??**).
- World model as data engine: p_θ generates synthetic observations used to augment training data for a separately trained policy (**???**).
- Offline-to-online adaptation: the world model is used for simulator-free policy improvement during deployment (**????**).

2.5 Unified Training Objective

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

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{obs}}}_{\text{observation prediction}} + \beta \underbrace{\mathrm{KL}[q_\phi(z_t \mid \cdot) \parallel p_\theta(z_t \mid z_{t-1}, a_{t-1})]}_{\text{dynamics consistency}} + \lambda \underbrace{\mathcal{L}_{\text{task}}}_{\text{control/planning utility}}, \tag{7}$$

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 pretraining 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 (**???**).

2.6 Decision Optimization with Learned Dynamics

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

$$\mathbf{a}^*_{t:t+H-1} = \arg \max_{\mathbf{a}_{t:t+H-1}} \mathbb{E}_{p_\theta} \left[\sum_{k=0}^{H-1} \gamma^k \hat{r}_{t+k} \right], \tag{8}$$

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. **??**) is rarely sufficient. Embodied systems combine it with:

- **Receding horizon control**: re-optimize at each step to correct compounding errors (**??**).
- **Monte Carlo Tree Search**: guided tree expansion in latent action space (**??**).
- **Intervention-guided online RL**: use human teleoperation corrections as additional reward signal during deployment (**??**).
- **Advantage-conditioned sampling**: RECAP conditions VLA token sampling on estimated advantage values for online RL without explicit value networks (**?**).

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:

$$v_t = E_v(\text{image}_t, \text{depth}_t, \ldots), \tag{9}$$

$$l_t = E_l(\text{instruction}_g, v_t), \tag{10}$$

$$a_t \sim \pi_a(a_t \mid l_t, h_{t-W:t}), \tag{11}$$

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 (**???**).
- **Diffusion action head** (flow matching or DDPM): continuous actions are denoised conditioned on l_t, preserving action continuity (**???**).
- **Hybrid head**: autoregressive token prediction augmented with diffusion denoising output, fusing reasoning and precision control (**??**).
- **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 (**??**).
- **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 (**????**).

The action representation also varies: per-dimension binning (simple but coarse), DCT-based frequency tokenization (**?**), continuous Gaussian (**?**), or hybrid discrete-continuous mixtures (**?**). 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:

$$\text{correction bandwidth} \propto \frac{1}{C \cdot \Delta t_{\text{infer}}}, \tag{12}$$

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 (**???**).

2.9 Failure Modes in the 2024–2026 Regime

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

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:

$$\mathcal{E}_{t+H} \leq \sum_{k=0}^{H-1} L^k \epsilon_{t+k}, \tag{13}$$

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 (**???**). WorldBench diagnoses this through isolated concept-level physical evaluation (**?**), and recent work aims to reduce L via physics-informed regularization and contact-aware dynamics (**??**).

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 (**?**) addresses this through URDF-aligned embodiment masks rendered into the prediction pathway; FlowDreamer (**?**) uses optical flow as an intermediate representation that bridges pixel and action space. GeoVLA (**?**) and 4D-VLA (**?**) 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 (**???**). 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 **??**.

These three failure modes motivate the three axes of the taxonomy in Section **??**: 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 (**??**); language-grounded planners emphasized high-level symbolic decomposition with feasibility checks (**???**); and foundation-policy work emphasized unified token-based control with heterogeneous data (**???**). The 2024–2026 systems can be interpreted as deeper integration of these once-separate axes.

3.2 Axis A: Functionality Coupling

Decision-coupled world models are trained and evaluated for direct control impact (policy improvement, planning reliability, intervention reduction). Representative examples include online-refined VLA pipelines and world-model-guided policy optimization (**????**).

General-purpose world models prioritize broad predictive capability and transfer, then attach downstream controllers. This line includes large pretraining efforts and multimodal dynamics models used as reusable priors (**????**).

In practice, the key separator is *optimization target*. Decision-coupled models directly optimize task-facing losses (success, intervention, or value improvement) under closed-loop rollout constraints. General-purpose models prioritize reusable predictive competence, often scaling with heterogeneous data and delaying control coupling to post-training.

The two settings are complementary rather than contradictory: many successful systems pretrain in a general-purpose regime and then switch to decision-coupled adaptation for target deployment (**?????**).

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 (**???**).
- **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 (**?**).
- **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 (**???**).
- **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 (**?**).

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 (**???**).

3.4 Axis B: Temporal Modeling

Sequential rollouts simulate future states step by step and align naturally with MPC-style control, but face compounding error over long horizons (**???**).

Global prediction methods forecast larger trajectory segments or future differences in parallel and can improve efficiency, but require stronger structural priors to preserve causal consistency (**???**).

This temporal choice can be viewed as a bias-variance-compute tradeoff. Let ϵ_t denote one-step model error in latent space. In a simplified sequential regime, rollout error can scale approximately as

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

which explains sensitivity in long-horizon manipulation and multi-agent traffic forecasting. Global predictors reduce iterative accumulation steps but can underfit local control-relevant transitions unless they include action- and embodiment-aware constraints (**???**).

Recent hybrids combine chunk-wise global prediction with local sequential correction, effectively using coarse global proposals and fine-grained control-time refinement (**???**).

3.5 Axis C: Spatial Representation

Compact latent representations support real-time control and low compute budgets. **Tokenized representations** improve multimodal alignment with language-conditioned reasoning. **Geometry-aware or rendering-aware representations** better preserve view consistency and object-level structure for manipulation and driving scenarios (**????**).

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 (**????**).

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 (**?**).

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 (**?????**).

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 (**???**).

4 Data Resources and Evaluation Metrics

4.1 Data Regimes

Recent embodied research uses four complementary data regimes:

- **Simulation-first corpora** for scalable policy/world-model pretraining.
- **Interactive benchmark suites** for closed-loop reproducibility.
- **Large offline robot datasets** for foundation model initialization.
- **Real-world deployment logs** for post-training and robustness analysis.

Representative resources include OpenVLA/Open-X style pipelines, DROID-scale data, and newer embodied world-model benchmarks focused on rollout quality and control relevance (**?????**).

These resources extend earlier scaling efforts in diverse ways: web-scale open-world embodied environments, large real-robot manipulation corpora, and multimodal prompt-driven simulation suites (**???**).

Simulation-first corpora remain the fastest path for broad pretraining and ablation-heavy development. **Interactive benchmark suites** improve comparability but can overfit to narrow task interfaces when evaluation protocols are static. **Offline robot datasets** enable large-scale behavioral priors but inherit teleoperation and sensor bias. **Real-world logs** are the only reliable source for intervention dynamics, recovery behavior, and edge-case calibration (**????**).

4.2 Data Curation Dimensions

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

1. **Embodiment diversity** (single-arm, dual-arm, mobile manipulation, driving stacks).
2. **Task horizon composition** (short atomic skills vs. multi-stage household workflows).
3. **Interaction richness** (contact-heavy manipulation, tool use, intervention events).
4. **Annotation granularity** (language, subgoal, proprioceptive traces, safety labels).

Papers that only scale data volume without balancing these dimensions often improve headline benchmark averages but underperform in open-world deployment conditions (**????**).

4.3 Metric Families

We group evaluation metrics into five families:

1. **Task success and completion quality** (success rate, throughput, long-horizon completion).
2. **Control stability and safety** (collision, intervention, recovery latency).
3. **Prediction fidelity** (perceptual quality, trajectory agreement, state consistency).
4. **Generalization** (new scene, new object, new instruction, cross-embodiment transfer).
5. **Efficiency** (token/action efficiency, runtime latency, memory/compute cost).

Several recent papers explicitly report tradeoffs between closed-loop gains and compute/latency costs, making efficiency metrics first-class rather than optional (**????**).

4.4 Representative Quantitative Signals

Although protocols differ across papers, several 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 (**?**).
- RECAP ($\pi^*_{0.6}$) reports more than doubling throughput and roughly halving failure rate on difficult tasks (**?**).
- VLA-RFT reports surpassing strong supervised baselines with fewer than 400 fine-tuning steps in simulator-driven RL fine-tuning (**?**).
- ConRFT reports evaluation on eight real-world manipulation tasks with a unified offline+online consistency objective (**?**).
- Valle et al. 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:

$$\mathrm{SR} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\text{task } i \text{ succeeds}], \tag{15}$$

$$\mathrm{IR} = \frac{1}{N}\sum_{i=1}^{N} \frac{n_i^{\text{intervention}}}{T_i}, \tag{16}$$

$$\mathrm{RTF} = \frac{\text{inference + planning time}}{\text{control horizon time}}, \tag{17}$$

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.

Table 1: 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	(**???**)	General-purpose manipulation
World-model-guided control	Better planning signal, sample efficiency, counterfactual reasoning	Model bias and rollout drift at long horizon	(**???**)	Long-horizon decision tasks
Post-training RL/refinement for VLAs	Improves task throughput and robustness in deployment	Requires safe data collection and intervention design	(**???**)	Continuous improvement loops
Efficiency-oriented compression/adaptation	Lower latency and memory cost; easier edge use	Potential capability drop if over-compressed	(**???**)	Resource-constrained systems

4.5 Evaluation Protocol Recommendations

For reproducible and decision-relevant reporting, we recommend:

- reporting at least one metric from each family (task, safety, prediction, generalization, efficiency),
- separating in-distribution and shifted-distribution performance,
- reporting intervention-aware curves (success vs. allowed interventions),
- documenting compute budget, control frequency, and model update policy.

These elements are increasingly present in recent benchmark-oriented work and should become default for embodied world-model evaluation (**???**).

In addition, recent diagnostic benchmarks explicitly target disentangled physical understanding. WorldBench emphasizes concept-level disambiguation rather than entangled aggregate physics tests, making failure attribution more actionable for model iteration (**?**).

4.6 Current Gaps

Despite progress, 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 (**???**). This gap motivates evaluation protocols that jointly report dynamics realism, decision quality, and deployment behavior.

5 Cross-Family Comparison and Practical Tradeoffs

Table **??** 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:

$$\mathcal{U} = \alpha \cdot \mathrm{SR} - \beta \cdot \mathrm{IR} - \gamma \cdot \mathrm{RTF}, \tag{18}$$

where SR is task success rate, IR is intervention rate, and RTF is real-time factor (defined in Section **??**). 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 (**???**).

5.2 Where Each Family Wins

Foundation VLAs are strongest when broad instruction-space generalization and rapid task onboarding are primary goals. Their weakness is often intervention-heavy recovery under compounding distribution shift (**???**).

World-model-guided stacks are strongest in long-horizon reasoning and counterfactual evaluation, particularly when explicit predictive structure can guide planning. Their weakness is representation mismatch and rollout bias when embodiment-specific constraints are weakly encoded (**????**).

Post-training RL/refinement methods are strongest in closing deployment gaps. Notably, several reports show substantial throughput and failure-rate improvements after online or intervention-aware refinement, indicating that static imitation pretraining is no longer sufficient for robust field behavior (**????**).

Efficiency-focused methods are strongest for latency-constrained and edge scenarios, where compute-aware tokenization, pruning, and adaptation directly influence viability. Their main risk is capacity loss if compression is applied without task-specific calibration (**????**).

5.3 Representative Case Studies

To anchor the comparison in concrete method behavior:

- **Scale-first foundation policy:** $\pi_0/\pi_{0.5}$ emphasizes heterogeneous multi-robot and multimodal co-training to improve open-world manipulation coverage (**??**).
- **Deployment-first refinement:** $\pi^*_{0.6}$ (RECAP), VLA-RFT, and VLA-RL emphasize online or simulator-mediated reinforcement fine-tuning, arguing that distribution-shift robustness requires explicit post-deployment adaptation (**???**).
- **World-model-as-data-engine:** GigaWorld-0 and GigaBrain-0 present a synthesis view where world models are used to generate scalable embodied training data and reduce dependence on expensive physical collection (**??**).
- **Platform-level foundation world models:** Cosmos positions world foundation models as customizable infrastructure (data curation, tokenization, post-training), rather than a single monolithic policy component (**?**).

5.4 Historical Continuity Across Families

Current families are not isolated inventions. Foundation VLAs inherit multi-embodiment token-policy ideas from Gato, RT-1, and RoboCat (**???**). World-model-guided and planner-policy hybrids extend earlier language-grounding and feasibility-constrained planning lines (**???**). Data-scaling and adaptation loops connect to BridgeData-style collection, trajectory- and code-mediated control interfaces, and lifelong skill-library designs (**????**). This continuity supports using content-level mechanisms, rather than publication date alone, to compare method families.

5.5 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 (**????**).

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.

6 Open Challenges and Outlook

6.1 Challenge 1: Long-Horizon Physical Consistency

Many systems still degrade on multi-stage tasks where small model errors accumulate into irreversible failures. Future work should prioritize physically grounded temporal constraints and intervention-aware planning objectives, not only visual realism metrics (**???**).

Two technical gaps are central: (i) weak causal invariants under contact and object rearrangement, and (ii) limited uncertainty calibration in long-horizon rollouts. Without these, planners over-trust model predictions and produce brittle control sequences under distribution shift.

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 (**???**).

This mismatch is no longer a niche issue; it affects generalization across robot morphologies, viewpoint changes, and tool-based manipulation. 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. The community needs shared protocols that jointly evaluate safety, recovery, intervention rate, and sustained task throughput under shift (**???**).

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.

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 (**???**).

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.

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 (**???**). Open questions include:

- how to schedule exploration under hard safety budgets,
- how to integrate teleoperator corrections without destabilizing pretrained priors,
- how to prevent catastrophic forgetting during continual specialization.

6.6 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. 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.

6.7 Near-Term Research Priorities

We identify three near-term priorities:

- **Disentangled diagnostic evaluation:** shift from monolithic benchmark scores to concept-isolated physical diagnostics and reasoning-action faithfulness checks (**??**).
- **Action-world alignment under embodiment constraints:** improve coordinate-to-pixel and language-to-control alignment using geometry-aware conditioning and consistency objectives (**???**).

- **Scalable but safe adaptation loops:** combine synthetic or world-model-generated data with intervention-aware online refinement to improve robustness without uncontrolled exploration cost (**????**).

6.8 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,
- 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, while mid-term progress will depend on standardized deployment-centric evaluation and safer continual learning protocols.

7 Conclusion

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. 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 therefore looks less like a paradigm replacement and more like integration of planning, world modeling, and policy adaptation into a single closed-loop training and deployment stack.

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.

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 **??**.

2025 **??????????????????**. **????????????????**.

2024 **??**.

A.3 data-benchmark-eval

2026 **??????????????????**. **??????**.

2025 **??????????????????**. **??????????????????**. **??????????????????**. **??????????????????**. **??????????????????**. **????????????????**.

2024 **????????????????**.

A.4 foundation-definition

2025 ???.

2024 ???.

A.5 planning-reasoning

2026 ???.

2025 ??????????????????. ???????????.

2024 ????.

A.6 policy-learning

2026 ?.

2025 ??????????????????.

A.7 survey-meta

2026 ???.

2025 ??????????????.

A.8 world-model-core

2026 ????.

2025 ???????????.

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 missing_year

???.

B.3 not_embodied_related

??.

B.4 out_of_window

??????????????????. ????????????????.

References

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.

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

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

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].

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].

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

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

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].

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].

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].

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].

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.

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].

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/.

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].

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].

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 2025a. ISSN 0360-0300. doi:10.1145/3746449. URL https://dl.acm.org/doi/10.1145/3746449.

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

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].

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].

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].

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.

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].

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

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].

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].

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

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

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].

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].

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

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].

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/.

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].

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].

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].

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].

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].

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

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].

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

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.

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.

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.

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].

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].

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].

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].

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].

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].

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].

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 2025c. URL http://arxiv.org/abs/2506.16211. arXiv:2506.16211 [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].

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/.

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].

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].

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].

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].

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].

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].

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].

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].

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].

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 2025d. URL http://arxiv.org/abs/2510.12276. arXiv:2510.12276 [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 2025b. URL http://arxiv.org/abs/2506.22242. arXiv:2506.22242 [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].

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

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].

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].

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].

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].

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].

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

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/.

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].

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].

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

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/.

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].

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 2025a. URL http://arxiv.org/abs/2511.01210. arXiv:2511.01210 [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/.

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.

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].

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].

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*, 2025e. URL https://openreview.net/forum-id=dT45OMevL5.

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, 2025f. 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.

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

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

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

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 2025c. 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.

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].

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].

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 2025a. doi:10.1109/IROS60139.2025.11247519. URL https://ieeexplore.ieee.org/document/11247519/. ISSN: 2153-0866.

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].

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.

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

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].

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

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

Yuqing Wen, Hebei Li, Kefan Gu, Yucheng Zhao, Tiancai Wang, and Xiaoyan Sun. LLaDA-VLA: Vision Language Diffusion Action Models, September 2025c. URL http://arxiv.org/abs/2509.06932. arXiv:2509.06932 [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*, 2025d. URL https://ieeexplore.ieee.org/abstract/document/10900471/.

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.

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].

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].

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/.

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, 2024a. URL http://openaccess.thecvf.com/content/CVPR2024/html/Yang_Embodied_Multi-Modal_Agent_trained_by_an_LLM_from_a_Parallel_CVPR_2024_paper.html.

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].

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].

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].

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].

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].

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].

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].

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].

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].

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

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

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].

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].

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].

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].

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].

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].

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].

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].

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

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

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].

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].

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].

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

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].

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].

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].

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].

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].

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].

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].

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].

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].

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].

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].

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].

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].

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].

Yuhang Huang, Shilong Zou, Jiazhao Zhang, Xinwang Liu, Ruizhen Hu, and Kai Xu. AdaPower: Specializing World Foundation Models for Predictive Manipulation, December 2025a. URL `http://arxiv.org/abs/2512.03538`. arXiv:2512.03538 [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 2025a. URL `http://arxiv.org/abs/2511.14659`. arXiv:2511.14659 [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 2025a. URL `http://arxiv.org/abs/2509.12594`. arXiv:2509.12594 [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 2025b. URL `http://arxiv.org/abs/2509.00576`. arXiv:2509.00576 [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].

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 2025d. URL `http://arxiv.org/abs/2512.11047`. arXiv:2512.11047 [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].

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].

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].

Wei Li, Renshan Zhang, Rui Shao, Jie He, and Liqiang Nie. CogVLA: Cognition-Aligned Vision-Language-Action Model via Instruction-Driven Routing & Sparsification, October 2025h. URL `http://arxiv.org/abs/2508.21046`. arXiv:2508.21046 [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 2025i. URL `http://arxiv.org/abs/2510.12796`. arXiv:2510.12796 [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 2025j. URL http://arxiv.org/abs/2502.05485. arXiv:2502.05485 [cs].

Muyao Li, Zihao Wang, Kaichen He, Xiaojian Ma, and Yitao Liang. JARVIS-VLA: Post-Training Large-Scale Vision Language Models to Play Visual Games with Keyboards and Mouse, September 2025k. URL http://arxiv.org/abs/2503.16365. arXiv:2503.16365 [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 2025l. URL http://arxiv.org/abs/2511.09516. arXiv:2511.09516 [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 2025m. URL http://arxiv.org/abs/2510.14836. arXiv:2510.14836 [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 2025n. URL http://arxiv.org/abs/2510.21571. arXiv:2510.21571 [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].

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].

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].

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].

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 2025d. URL http://arxiv.org/abs/2512.09928. arXiv:2512.09928 [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 2025d. URL http://arxiv.org/abs/2509.18953. arXiv:2509.18953 [cs].

Zeting Liu, Zida Yang, Zeyu Zhang, and Hao Tang. EvoVLA: Self-Evolving Vision-Language-Action Model, November 2025e. URL http://arxiv.org/abs/2511.16166. arXiv:2511.16166 [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 2025f. URL http://arxiv.org/abs/2509.26642. arXiv:2509.26642 [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 2025g. URL http://arxiv.org/abs/2510.07134. arXiv:2510.07134 [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].

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].

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].

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].

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

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].

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].

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].

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 2025b. URL http://arxiv.org/abs/2506.10826. arXiv:2506.10826 [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 2025c. URL http://arxiv.org/abs/2508.10333. arXiv:2508.10333 [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].

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].

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].

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].

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].

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 2025f. URL http://arxiv.org/abs/2509.25718. arXiv:2509.25718 [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 2025g. URL http://arxiv.org/abs/2507.01016. arXiv:2507.01016 [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].

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

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].

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].

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].

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].

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].

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].

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 2025a. URL http://arxiv.org/abs/2512.23864. arXiv:2512.23864 [cs].

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

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].

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

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

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].

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

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

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

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 2025f. URL http://arxiv.org/abs/2503.03480. arXiv:2503.03480 [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 2025g. URL http://arxiv.org/abs/2509.07962. arXiv:2509.07962 [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 2025h. URL http://arxiv.org/abs/2501.18867. arXiv:2501.18867 [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].

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 2025a. URL http://arxiv.org/abs/2510.14902. arXiv:2510.14902 [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].

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 2025a. URL http://arxiv.org/abs/2508.18269. arXiv:2508.18269 [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 2025a. 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/.

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].

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].

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

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].

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].

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, and Ruohan Zhang. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37: 100428–100534, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b631da756d1573c24c9ba9c702fde5a9-Abstract-Datasets_and_Benchmarks_Track.html.

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

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

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].

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/.

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].

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].

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].

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].

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, 2025b. 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.

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

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/.

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].

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.

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, 2024b. URL http://openaccess.thecvf.com/content/CVPR2024/html/Yang_PhyScene_Physically_Interactable_3D_Scene_Synthesis_for_Embodied_AI_CVPR_2024_paper.html.

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].

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].

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].

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].

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 2025d. URL http://arxiv.org/abs/2510.13778. arXiv:2510.13778 [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 2025e. URL http://arxiv.org/abs/2507.23682. arXiv:2507.23682 [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].

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].

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].

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].

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/.

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].

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 2025c. URL http://arxiv.org/abs/2507.09160. arXiv:2507.09160 [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 2025b. URL http://arxiv.org/abs/2504.19854. arXiv:2504.19854 [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].

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 2025p. URL http://arxiv.org/abs/2509.09674. arXiv:2509.09674 [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 2025q. URL http://arxiv.org/abs/2506.03574. arXiv:2506.03574 [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 2025h. URL http://arxiv.org/abs/2503.10631. arXiv:2503.10631 [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].

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].

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].

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].

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].

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

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

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 2025c. URL http://arxiv.org/abs/2512.11872. arXiv:2512.11872 [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 2025j. URL http://arxiv.org/abs/2505.13888. arXiv:2505.13888 [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 2025k. URL http://arxiv.org/abs/2503.20384. arXiv:2503.20384 [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, 2025b. 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.

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].

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].

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].

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

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].

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].

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].

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].

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].

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].

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].

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 2025r. URL http://arxiv.org/abs/2509.22199. arXiv:2509.22199 [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 2025s. URL http://arxiv.org/abs/2511.14178. arXiv:2511.14178 [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 2025t. URL http://arxiv.org/abs/2510.23576. arXiv:2510.23576 [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].

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, 2025i. URL https://aclanthology.org/2025.emnlp-main.1342/.

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].

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 2025d. URL http://arxiv.org/abs/2512.05107. arXiv:2512.05107 [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 2025l. URL http://arxiv.org/abs/2509.04063. arXiv:2509.04063 [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 2025m. URL http://arxiv.org/abs/2511.01331. arXiv:2511.01331 [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].

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.

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

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 2025v. URL `http://arxiv.org/abs/2508.15201`. arXiv:2508.15201 [cs].

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

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].

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 2025b. URL `http://arxiv.org/abs/2507.01925`. arXiv:2507.01925 [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`.

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].

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/`.

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 2025c. URL `http://arxiv.org/abs/2509.22643`. arXiv:2509.22643 [cs].

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

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

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

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`.

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`.

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/`.

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

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

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

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].