Review

What Matters: Datasets or Robust Frameworks in Modern Robot Learning?

Md Selim Sarowar *

Posted Date: 17 June 2026

doi: 10.20944/preprints202606.1149.v1

Keywords: vision-language-action models; robot learning; datasets; policy evaluation; robot manipulation

 doi:10.20944/preprints202606.1149.v1

Review

What Matters: Datasets or Robust Frameworks in Modern Robot Learning?

Md Selim Sarowar

Independent Researcher, Bangladesh; selim.sarowar12@gmail.com

Abstract

Recent progress in robot learning has relied on two investments: larger datasets and more capable models. Vision-language-action (VLA) policies now report success rates above 90% on standard benchmarks, yet perturbation studies show the same policies collapsing to near 0% when object positions, instructions, or scene layouts shift, exposing memorization where competence was claimed. This survey asks whether progress comes mainly from data, from models, or from an interaction that current evaluations often obscure. We review more than 200 papers spanning VLA architectures, world models, reinforcement-learning post-training, robot manipulation datasets, data generation pipelines, scaling studies, and perturbation benchmarks, including a structured analysis of a 100-paper survey set centered on the ICLR 2026 world-model literature. We catalogue every major public manipulation dataset with size, embodiment coverage, collection method, and known weaknesses; we reconstruct the evidence on data scaling laws and data quality; and we trace the evaluation crisis from benchmark inflation through memorization diagnoses to factor-level robustness decompositions. Our synthesis is that the question is ill-posed as a dichotomy: data diversity dominates in-distribution gains, model class and training objective dominate out-of-distribution retention, and current benchmarks confound the two because train and test conditions coincide. We state the conditions under which each answer holds, identify bottlenecks per subfield, and propose falsifiable research directions, including counterfactually structured datasets, world-model-regularized policies, and factor-controlled evaluation protocols.

Keywords: vision-language-action models; robot learning; datasets; policy evaluation; robot manipulation

1. Introduction

Two explanations are common in modern robot manipulation. The first is a data narrative: many leading systems were built from large, diverse demonstration datasets, from the 130,000 teleoperated episodes behind RT-1 [1] through the million-trajectory cross-embodiment pooling of Open X-Embodiment [2] to vertically integrated collection factories such as AgiBot World [3]. In this view, manipulation follows the language-modeling recipe: scale the data and keep the training recipe simple. The second is a model narrative: the main gains came from architecture and training objectives, including the transfer of web-scale vision-language backbones into action prediction [4,5], expressive action heads based on diffusion and flow matching [6,7], world models that learn predictive structure without action labels [8,9], and reinforcement-learning post-training that repairs the known pathologies of behavior cloning [10,11]. On this view, data is necessary, but the policy class and objective determine what the system learns from it.

Recent perturbation results make this disagreement hard to ignore. Perturbation studies of the strongest available policies report that headline benchmark numbers measure memorization to a degree the field did not intend. On LIBERO [12], the de facto standard simulation benchmark for VLA evaluation, leading models exceed 90% average success; under the perturbed evaluation of LIBERO-PRO, in which object positions, instructions, objects, and scene layouts change while the task semantics remain reasonable, the same models drop to 0.0% [13]. Models continue to execute grasp

 doi:10.20944/preprints202606.1149.v1

sequences when the target object has been replaced by an irrelevant item, and their outputs remain unchanged under corrupted instruction tokens [13]. The factor-level decomposition of LIBERO-Plus shows that the damage concentrates in camera viewpoint and robot initial state, and that several models barely attend to language at all [14]. THE COLOSSEUM reports qualitatively identical degradation across 14 perturbation factors in a different simulator and correlates the degradation with real-robot behavior [15]. If benchmark success can be achieved by layout-keyed trajectory replay, then benchmark success alone cannot tell us whether data or modeling choices deserve credit.

This survey is organized around the question: *what matters more for robust robot manipulation, datasets or models, and under what conditions does each answer hold?* We do not assume that either side is sufficient on its own. The split is imperfect, but it mirrors many practical choices about data collection, model design, and compute. Our method is to separate three bodies of evidence that are usually argued together: evidence about models (Section 4), evidence about data (Section 5), and evidence about evaluation (Section 6), and only then to synthesize (Section 7).

1.1. Scope of Evidence

We analyze more than 200 papers. These include a structured 100-paper document set drawn primarily from ICLR 2026 and its World Models workshop, which provides a focused sample of recent world-model work; the major VLA systems and their ablation literature; major public robot manipulation datasets (Table 3); the data generation and augmentation literature; the data scaling studies; and the perturbation benchmark literature that constitutes the evaluation crisis. Where evidence is incomplete, we say so rather than treating reported success as attribution.

1.2. Contributions

This survey makes five contributions. First, a taxonomy of the field organized by mechanism rather than chronology, separating what a method changes (data distribution, representation, objective, policy class, inference procedure) from what it claims (Figure 1). Second, a reference comparison of robot manipulation datasets covering scale, embodiments, modalities, collection method, license posture, and documented weaknesses (Table 3). Third, a reconstruction of the evaluation crisis with a factor-level account of what breaks under which perturbation and why standard protocols could not detect it (Section 6, Figure 4). Fourth, an evidence matrix that assigns published findings to the data side, the model side, or the confounded middle, with explicit identification of the confounds (Table 5). Fifth, a set of bottlenecks and falsifiable research directions stated concretely enough to act on (Sections 8 and 9).

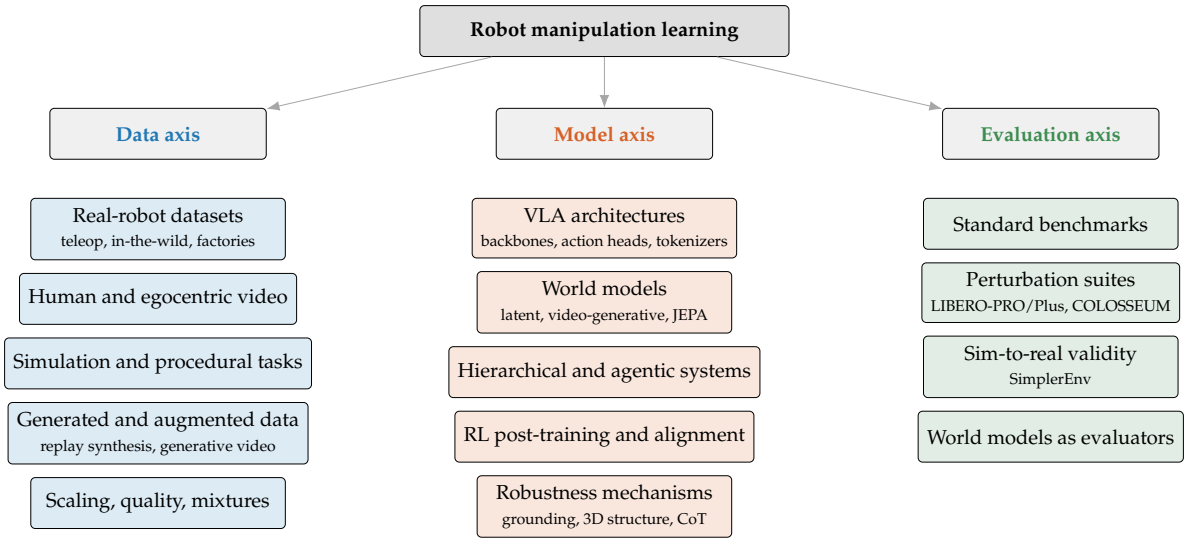

Figure 1. Taxonomy used in this survey. Methods are organized by the mechanism they modify. The data axis changes the training distribution; the model axis changes the hypothesis class, objective, or inference procedure; the evaluation axis determines what either is credited with. Sections 4, 5, and 6 follow this structure.

 doi:10.20944/preprints202606.1149.v1

1.3. Organization

Section 2 fixes terminology, inclusion criteria, and the survey's own limitations. Section 3 reviews problem formulations, policy classes, and evaluation protocols, and introduces the taxonomy (Figure 1) and timeline (Figure 2) used throughout. Section 4 reviews the model axis: VLA architectures and action interfaces, world models inside and beyond robotics, hierarchical and agentic decompositions, reinforcement-learning post-training, and dedicated robustness mechanisms, closing with a statement of exactly what the model axis has and has not demonstrated. Section 5 reviews the data axis symmetrically: real-robot datasets, human video, simulation, generated data, collection economics, scaling evidence, and quality effects. Section 6 reconstructs the evaluation crisis and derives requirements for valid measurement. Section 7 assembles the evidence matrix, names the confounds, states the conditional answer with its falsifiers, and adds a practitioner decision framework and a failure taxonomy. Sections 8 and 9 close with bottlenecks and an executable research agenda (Table 6).

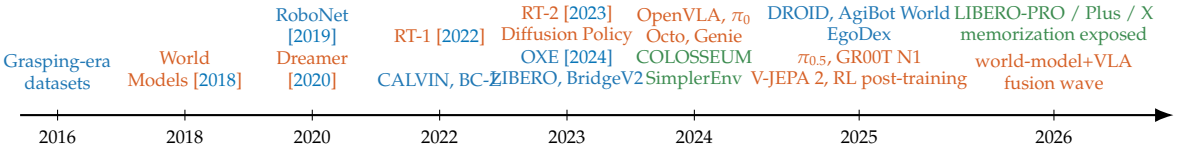

Figure 2. Milestones, colored by axis (data, model, evaluation). The evaluation axis matured years after the other two, which is why the memorization pathology of 2025–2026 went undetected through the scaling years.

1.4. Answer in Brief

Data diversity is the clearest source of in-distribution competence and of interpolation within the training support; none of the model changes surveyed here replaces it. Architecture, objective, and post-training matter most when conditions shift, especially when behavior cloning can solve the training tasks through shortcuts. These effects interact, and many current benchmarks blur them because training and evaluation conditions overlap. Section 7 gives the detailed argument and states tests that would falsify it.

2. Scope and Method

2.1. Inclusion Criteria and Search Protocol

We include a paper if it satisfies at least one of the following: (i) it introduces or substantially modifies a policy learning method evaluated on robot manipulation; (ii) it introduces a dataset, simulator, or benchmark used for manipulation policy learning or evaluation; (iii) it provides controlled evidence about data scale, data quality, data composition, or evaluation validity; (iv) it develops world models with stated relevance to embodied control. We exclude purely navigation, locomotion-only, and autonomous driving work except where it supplies evidence that transfers directly (e.g.,, world-model architectures [16]).

We assembled the document set in two passes. The first pass processed a fixed local collection of 100 papers, dominated by ICLR 2026 main-conference and World Models workshop publications, and extracted problem, method, data, evaluation, and limitation fields into a structured table. The second pass expanded coverage through open sources (arXiv, Semantic Scholar, OpenReview, project repositories), targeting the areas listed in Section 1.

2.2. Terminology and the Shape of the Question

We use *dataset* to mean the empirical distribution a policy is trained on, including demonstrations, instructions, and any auxiliary supervision; *model* to mean the policy class, architecture, training objective, and post-training procedure. For this survey, the title question means: holding compute fixed, which marginal investment most improves task success under the deployment distribution, and which most improves retention of success under distribution shift? These are different quantities, and the literature is easier to interpret when they are kept separate. A policy can be data-rich and

 doi:10.20944/preprints202606.1149.v1

shift-brittle [13]; it can be data-poor and shift-robust within a narrow envelope [17]. The survey keeps the two quantities distinct throughout.

2.3. Limitations of This Survey

Three limitations qualify our claims. First, the local survey set oversamples world-model research relative to the field as a whole; we use it as a depth probe for that axis and rely on the expanded literature for balance. Second, several frontier systems (e.g.,, Gemini Robotics [18]) report results on proprietary platforms that cannot be independently verified; we report their claims as claims. Third, the evaluation literature is young and concentrated on one benchmark family; where we generalize from LIBERO-derived evidence we say so explicitly.

2.4. Relation to Prior Surveys

Earlier surveys map foundation models onto robotics broadly [19,20], catalogue VLA architectures [21,22], or target efficiency [23]. This survey differs in organizing principle: we treat the data-versus-model question, and the evaluation evidence that arbitrates it, as the primary structure, and we incorporate the 2025–2026 perturbation literature that postdates prior surveys and changes their conclusions. The position paper closest in spirit argues that VLAs and world models are jointly insufficient for robust autonomy [24]; we treat its thesis as one hypothesis among several and test it against the assembled evidence in Section 7.

3. Background

3.1. Problem Formulations

Manipulation policy learning is dominated by three formulations. *Behavior cloning* (BC) fits a policy $\pi_\theta(a \mid o, \ell)$ to demonstration tuples of observation o, language instruction ℓ, and action a, by maximum likelihood or a regression surrogate; essentially every system in Section 4 trains this way at some stage [1,5,6]. *Reinforcement learning* optimizes expected return through environment interaction, used in manipulation mainly as post-training on top of a BC initialization because exploration from scratch is uneconomical on hardware [10,11]. *Model-based control* learns a dynamics model and derives behavior by planning or by policy optimization inside the model; this is the world-model program, with lineage from Ha and Schmidhuber [25] through the Dreamer family [8,26,27] and TD-MPC2 [28], and with renewed momentum from video-generative and latent-predictive instantiations [9,29].

The action interface differentiates systems more than is commonly acknowledged. Discrete-token action heads treat control as next-token prediction over binned actions [1,4,5]; compressed tokenizations reduce sequence length and improve throughput [30]. Continuous heads predict action chunks by regression [31,32], by denoising diffusion [6,33], or by flow matching [7]. Chunked prediction, in which the policy emits tens of future actions per inference, is now near universal because it suppresses compounding error and reduces inference frequency [6,31,32].

3.2. Policy Classes in Brief

Figure 1 organizes the field by mechanism. Vision-language-action models initialize from a pretrained vision-language backbone and fine-tune for action prediction, importing web-scale semantics into control [4,5,7,34], with lineage through language-conditioned imitation [35,36] and the generalist-agent program [37,38]. Visuomotor policies trained from scratch occupy the small-data regime [6,31]. World-model approaches learn predictive structure first and extract behavior second, either by planning in the model [28,39], by policy optimization in imagination [8,40], or by using the model as an evaluator and data generator [41–43]. Hierarchical and agentic systems decompose control into a semantic planner and a motor executor [44–46]. These classes are not exclusive; $\pi_{0.5}$ is simultaneously a VLA and a hierarchy [47], and several 2025–2026 systems fuse world models with VLA policies [48,49].

 doi:10.20944/preprints202606.1149.v1

3.3. Evaluation Protocols

Manipulation has no held-out test set in the supervised-learning sense; evaluation means rolling out the policy. Standard practice evaluates on the training tasks with modest initial-state variation drawn from the same distribution as training [12,51], which is precisely the protocol the perturbation literature later showed to be insufficient (Section 6). Real-robot evaluation suffers from low statistical power and irreproducibility across labs; simulation evaluation suffers from the sim-to-real gap, partially quantified by SimplerEnv's correlation analyses [52]. A third option emerged in 2025: evaluating policies inside learned world models, which trades physical fidelity for throughput and control [41,42, 53]. All three options are in active use, and their disagreements are themselves informative (Section 6).

3.4. The Compounding-Error Debate and the Case for Interaction

A background dispute affects both sides of the comparison: whether the dominant failure of behavior cloning is statistical (covariate shift compounding over a rollout) or representational (shortcut features that never bind to the causal task variables). The classical analysis attributes BC failure to compounding error, motivating chunked action prediction [6,31] and interactive correction. The perturbation findings of Section 6 weigh in for the representational account: policies fail at $t{=}0$ under static scene perturbations, before any error can compound [13,14]. This matters for resource allocation because the two diagnoses imply different interventions. Compounding error is treatable with more data along visited states; shortcut reliance is treatable only by changing what the objective rewards, through interaction [11], predictive auxiliary structure [9,54], or grounded intermediates [55]. The local document set provides related evidence from outside manipulation: probing studies find learned environment simulators encoding policy-relevant but causally shallow state [56], goal-directedness evaluations find agent behavior driven by surface regularities [57], and compositional-generalization analyses tie generalization to objective continuity rather than data volume [58].

4. The Model Axis

This section reviews what changes when researchers change the model: backbone and action interface (Section 4.1), predictive world models (Section 4.2), hierarchy and agency (Section 4.5), reinforcement-learning post-training (Section 4.7), and explicit robustness mechanisms (Section 4.9). The organizing question for each family is the same: which failure mode of plain behavior cloning does this mechanism remove, and what evidence shows that it removes it rather than relocating it?

4.1. Vision-language-action Architectures

The VLA recipe has three load-bearing decisions: which pretrained backbone supplies the representation, how actions are emitted, and what mixture of data flows through fine-tuning. RT-2 established the value of the first decision by co-fine-tuning a web-scale vision-language model on robot trajectories and reporting emergent semantic generalization, including manipulation of categories never demonstrated on the robot [4]. OpenVLA replicated the finding in the open, fine-tuning a 7B Prismatic-style backbone on 970k Open X-Embodiment trajectories and outperforming the 55B RT-2-X while releasing weights and code [5]. The backbone's contribution is semantic, not motoric: it supplies category-level visual grounding and instruction parsing, which is exactly the component that perturbation studies later found underused at the action head (Section 6).

The action interface is the second decision. Discrete binning inherited from RT-1 [1] creates long token sequences and quantization artifacts; FAST's frequency-domain tokenization compresses chunks and accelerates training [30]; π_0 replaced tokens with a flow-matching head over continuous action chunks and demonstrated dexterous, long-horizon behavior [7]; OpenVLA-OFT's ablation isolated the interface's contribution, showing that parallel decoding, continuous actions, and an L1 chunk regression objective lift LIBERO average success from 76.5% to 97.1% while accelerating inference [32]. The systematic study of Li and et al. [59] reaches a consistent conclusion across backbones and formulations: architecture choices interact strongly with data mixture, and several

 doi:10.20944/preprints202606.1149.v1

celebrated components matter less than tuning details. The honest reading of this literature is that action-interface engineering produced some of the largest single-paper gains of the period, which complicates any claim that data alone drives progress, but those gains are measured on standard protocols and therefore inherit the standard protocols' blindness to memorization [13].

The third decision, fine-tuning mixture and co-training, blurs into the data axis. $\pi_{0.5}$ co-trains on robot data, web vision-language data, subtask prediction, and verbal plans, and attributes its open-world generalization, including cleaning unseen kitchens, to the mixture rather than to capacity [47]. GR00T N1 formalizes a dual-system design with a vision-language module and a diffusion-transformer motor module, trained on a pyramid of real robot data, human video, and synthetic trajectories [34]. Gemini Robotics transfers a frontier multimodal backbone into manipulation with claims of strong instruction following [18]. Cross-embodiment training, pioneered at scale by RT-X and Octo and extended by HPT and CrossFormer, consistently buys transfer at the cost of per-embodiment optimality [2,60–62]. RDT-1B and CogACT represent the diffusion-backbone branch at the billion-parameter scale [33,63]. Specialized variants inject structure: SpatialVLA's 3D position encodings [64], TraceVLA's visual trace prompting [65], and embodied chain-of-thought supervision [55] each report generalization gains attributable to representation rather than data volume.

Table 1. manipulation policy systems. Parameters and training data are as reported in the cited papers; dashes mark undisclosed values. Success figures are deliberately omitted because cross-paper success rates are not comparable (Section 6).

System	Year	Params	Action head	Primary training data
RT-1 [1]	2022	35M	discrete tokens	130k teleop episodes
RT-2 [4]	2023	12B/55B	discrete tokens	web VQA + RT-1 data
Diffusion Policy [6]	2023	∼100M	diffusion chunks	per-task demos
ACT [31]	2023	∼80M	chunked regression	50 demos/task
Octo [60]	2024	93M	diffusion chunks	800k OXE trajectories
OpenVLA [5]	2024	7B	discrete tokens	970k OXE trajectories
π_0 [7]	2024	3.3B	flow matching	cross-embodiment collection + OXE
RDT-1B [33]	2024	1.2B	diffusion	1M+ multi-robot episodes
CogACT [63]	2024	7B+	diffusion	OXE subset
OpenVLA-OFT [32]	2025	7B	parallel continuous	LIBERO / ALOHA fine-tunes
GR00T N1 [34]	2025	2B	diffusion transformer	robot + human video + synthetic
$\pi_{0.5}$ [47]	2025	–	flow + subtask tokens	heterogeneous co-training
Gemini Robotics [18]	2025	–	–	proprietary

4.2. World Models

World models are the model axis's strongest claim to a qualitative, rather than incremental, answer to brittleness: a policy that predicts consequences cannot ignore state, whereas a behavior-cloned policy can and demonstrably does [13,14]. The classical line established feasibility: latent dynamics models support policy learning in imagination across Atari, continuous control, and physical robots [25–27,66], with transformer instantiations establishing sample efficiency [67], culminating in DreamerV3's single-configuration mastery of diverse domains [8] and TD-MPC2's scalable model-predictive control [28]. MuZero demonstrated that the model need only predict decision-relevant quantities [68], a lesson the field periodically relearns.

The 2024–2026 wave splits into three programs with different bets. The *video-generative* program scales action-conditioned video prediction into general-purpose simulators: Genie learns latent-action interactive environments from unlabeled video [69,70], UniSim and GAIA-1 build interactive simulators for manipulation and driving [16,71], Cosmos provides pretrained world-foundation backbones for physical AI [29], diffusion world models reach real-time game engines [72–74], text-guided video prediction acts directly as a universal policy or subgoal generator [75,76], and video-generative pretraining transfers into manipulation policies at scale in GR-1 and GR-2 [77,78]. Its core liability is physical fidelity: controlled studies find video generators violating rigid-body constraints and failing to internalize physical law from scaling alone [79–81]. The *latent-predictive* program bets that

 doi:10.20944/preprints202606.1149.v1

pixels are the wrong target: JEPA-style architectures predict in representation space [82–84], V-JEPA 2 scales this to internet video with an action-conditioned variant that supports zero-shot manipulation planning [9,85], and DINO-WM shows that planning over frozen pretrained features is competitive without reconstruction [39]. The *robotics-native* program couples world models to manipulation policies directly: Ctrl-World rolls out policy-conditioned futures to evaluate and improve instruction following without robot time [41], WorldGym and World Action Verifier push world models as evaluation environments [42,53], World4RL refines policies inside diffusion world models [40], and DreamGen uses video world models as data generators for policies that generalize to new verbs and environments [43].

The local ICLR 2026 document set sharpens this picture with unusual resolution. On the architecture front, scaling 3D-structured world models for in-the-wild manipulation [86], knowledge-encoded trajectory world models across robots [87], latent motion world models for long-horizon prediction [88], hierarchical latent action models [89], and next-embedding prediction objectives [54] all push the latent-predictive bet; geometry-aware positional encodings stabilize video world models [90]; cross-view consistency emerges as its own subproblem [91]. On the control front, world-action models act as zero-shot policies [48], world models guide robotic task and motion planning [92], RL inside world models trains deployable robot policies [49,93], uncertainty-aware world models make offline model-based RL work on physical hardware [94,95], differentiable world models support offline MPC [96], and simulation-pretrained world models adapt rapidly to reality [97]. On the skeptical front, probing studies ask what world models actually learn [56,98], physics-grounded metrics such as temporal reversal asymmetry expose dynamical implausibility [99], diagnostic environments test language-grounded world modeling [100], spatial world-model probes question whether LLMs maintain coherent spatial state [101], and reproducibility infrastructure arrives in the form of standardized world-model evaluation stacks [102]. The range is broad, but there is still little, within this document set on evaluation on perturbation-controlled manipulation benchmarks, which is precisely where world-model claims of robustness would be tested (Section 6).

What world models have not yet shown is the decisive manipulation result: a policy whose LIBERO-PRO-style retention is high *because* of a predictive objective, with an ablation isolating that objective. The closest published evidence is indirect: V-JEPA 2-AC's zero-shot planning transfers across labs [9], DreamGen's generated data lifts behavior beyond the demonstrated distribution [43], and Ctrl-World's model-based fine-tuning improves instruction following on unseen objects [41]. Each is consistent with the hypothesis that predictive structure buys shift-robustness; none isolates it. We return to this as a priority experiment in Section 9.

4.3. World Models beyond Manipulation: Evidence of a General Program

The ICLR 2026 survey set also shows how widely world models are being studied outside robotics; those domains often provide cleaner tests of predictive objectives than manipulation currently does. Driving world models encode physical priors for closed-loop control [16,103]; medical instantiations model disease progression and treatment timing as latent dynamics [104,105]; epidemiology is argued as a natural world-model domain [106]; program repair reframes execution semantics as a world model [107]; computer using agents learn models of GUI dynamics [108]; cognitive digital twins target real-time operational decisions [109]. Methodological infrastructure is consolidating in parallel: reproducible world-model research stacks [102], lightweight JEPA libraries [110], diagnostic grid-world judges [111], and physics-grounded evaluation metrics [80,99]. Reasoning-oriented variants connect world models to deliberation: latent imagination as a reasoning substrate [112], energy-based world models bridged to language generation [113], visual generation as a medium for multimodal reasoning [114,115], and synergized reasoning-acting-simulating agents [116]. Two further threads bear directly on manipulation despite non-robotic framing: quasimetric structure emerging in intrinsic-energy JEPAs constrains what plannable latent spaces look like [117], and navigable latent-energy world models couple representation shape to controllability [118]. The takeaway for the title question is comparative: in domains where the predictive objective is the only supervision available, world

 doi:10.20944/preprints202606.1149.v1

models demonstrably carry the load; in manipulation, where demonstrations are available, the field has not yet forced the comparison.

4.4. Generative Video and 3D Scene World Models

The video-generative program's manipulation relevance runs through three capabilities. Controllability: action-conditioned generation must respond to fine-grained control rather than prompt-level conditioning, addressed by frame-level action conditioning with memory retrieval [41], reward-guided autoregressive generation [119], and identity-preserving multi-actor synthesis [120,121]. Consistency: long-horizon spatial memory remains fragile, motivating geometry-aware positional encodings [90], explicit 3D memory in panoramic generation [122], dexterous interaction simulation with spatial caches [123], and camera-pose recovery from dynamic videos [124]. Physical fidelity: the audits cited in Section 4.2 [79–81] plus fluid-dynamics stress tests [125] and rocket-landing control benchmarks [126] consistently find that visual quality outruns dynamical correctness. The 3D-native alternative bypasses pixels: point-cloud world models pretrained at scale for in-the-wild manipulation [86], Gaussian-splatting compression for spatial substrates [127], 4D-informed retrieval for active exploration [128], and spatial-intelligence benchmarks for the foundation models these systems consume [129–131]. For manipulation, the 3D-native line connects directly to the viewpoint-robustness evidence of Section 4.9: the perturbation axis that most damages 2D policies [14] is the one 3D structure is built to remove [64,132].

4.5. Hierarchical and Agentic Systems

Hierarchy converts one hard problem into two easier ones: a semantic layer that decides what to do and a motor layer that does it. The language-model planning line, from SayCan's affordance-weighted plan scoring [44] through Inner Monologue's closed-loop replanning [133], Code as Policies' program synthesis [134], VoxPoser's language-shaped 3D value maps [135], and MOKA's mark-based visual prompting [136], demonstrated that frozen foundation models can supply the semantic layer zero-shot. PaLM-E showed the two layers can share one backbone [137]. The current generation internalizes the hierarchy: RT-H inserts a language-motion layer between instruction and action [45], Hi Robot couples an open-ended deliberative layer to a VLA executor [46], and $\pi_{0.5}$ trains subtask prediction and motor control jointly [47]. Agentic framings extend this with memory, tool use, and explicit world-model consultation [57,116,138,139].

The evidence pattern for hierarchy is consistent and underappreciated in the data-versus-model debate: hierarchical systems degrade gracefully on *semantic* novelty (new instructions, recomposed goals, unseen object categories) because the planner generalizes compositionally, while remaining exactly as brittle as their motor layer on *physical* novelty (shifted positions, new viewpoints). LIBERO-Plus finds $\pi_{0.5}$-style hierarchical models retaining the most performance under language and layout perturbations while still failing under camera shifts [13,14]. Hierarchy is therefore not an alternative to robust motor learning; it is a multiplier on whatever motor robustness exists.

4.6. Instruction Following and Language Grounding

The language channel deserves separate treatment because it is where the model axis fails most quietly. Nominally, every VLA is language-conditioned; empirically, LIBERO-Plus finds several leading models nearly invariant to instruction perturbation, and LIBERO-PRO finds outputs unchanged under meaningless token sequences [13,14]. The mechanism is a training-distribution property: when each scene admits exactly one demonstrated behavior, the instruction is redundant given the image, and gradient descent prunes the redundant channel. Mechanisms that keep the channel alive all share one property, making language non-redundant: embodied chain-of-thought routes action prediction through language-dependent intermediates [55]; subtask-prediction co-training forces verbalization of intent [47]; hierarchical interfaces make the motor layer consume language-shaped commands [45,46]; counterfactual data pairs scenes with multiple instructions (Section 9); and instruction-sensitivity regularizers penalize invariance directly. Evaluation-side instruments exist in embodied-cognition and

 doi:10.20944/preprints202606.1149.v1

progress-reasoning probes for VLMs [140,141] and language-grounded world-model diagnostics [100]. The state of the evidence is unambiguous about the symptom and thin on cures: no published VLA demonstrates both high standard success and high instruction sensitivity under the LIBERO-Plus protocol.

4.7. Reinforcement-Learning Post-Training

RL post-training addresses the optimization-level pathology that no dataset fixes: behavior cloning matches action distributions without ever being penalized for relying on spurious cues. The 2025 literature converged quickly. Controlled studies find that PPO-style on-policy optimization improves semantic and execution robustness over supervised fine-tuning, while preference and group-relative methods imported from language modeling (DPO, GRPO) struggle in the partially observed, sparse-reward robotic setting [11]. VLA-RL builds the systems scaffolding for trajectory-level RL on OpenVLA and reports consistent gains over the SFT baseline on LIBERO [142]. SimpleVLA-RL pushes the data-efficiency claim furthest: starting from a single demonstration per task, RL lifts OpenVLA-OFT from 17.3% to 91.7% on LIBERO long-horizon tasks and improves sim-to-real transfer [10]. RIPT-VLA stabilizes interactive post-training with dynamic sampling [17]; π_{RL} makes online RL tractable for flow-based policies whose log-likelihoods are not directly available [143]; RobustVLA targets robustness explicitly with noise-injected reinforcement objectives [144]; GRAPE aligns policies at trajectory level from preferences [145]; V-GPS shows that even pure inference-time value re-ranking improves generalist policies without touching their weights [146]. Infrastructure has kept pace, with open frameworks supporting parallel rollout RL on the LIBERO family directly and open policy implementations lowering the entry cost [147,148], and world models increasingly supplying the rollout environment instead of the simulator [40,42,49,93].

Table 2. Reinforcement-learning post-training for VLA policies. All methods initialize from a behavior-cloned base; "substrate" is where rollouts occur.

Method	Algorithm	Substrate	Base policy	Headline evidence
VLA-RL [142]	PPO + process reward	simulator	OpenVLA	gains over SFT on LIBERO
SimpleVLA-RL [10]	GRPO-style	simulator	OpenVLA-OFT	17.3→91.7% from 1 demo/task
RIPT-VLA [17]	interactive, dyn. sampling	simulator	OpenVLA-OFT	low-data stabilization
GRAPE [145]	trajectory preference	offline	OpenVLA	success + safety objectives
π_{RL} [143]	on-policy for flow heads	simulator	π_0-class	RL for flow policies
RobustVLA [144]	robustness-aware RL	simulator	OpenVLA-class	perturbation-targeted reward
V-GPS [146]	value re-ranking	inference only	Octo/OpenVLA	gains without weight updates
World4RL [40]	RL in diffusion WM	learned model	pretrained policies	refinement without simulator
World-Gymnast [93]	RL in video WM	learned model	BC base	physical-interaction-free RL

Three caveats discipline the enthusiasm. First, nearly all of this evidence lives on LIBERO-family simulators, so RL is being credited partly for adapting to the evaluation distribution; held-out-axis protocols (Section 6) are rarely used. Second, sparse-reward RL from a collapsed initialization yields no gradient, so every successful recipe depends on a competent BC base, making RL a complement to, not a substitute for, data [10,17]. Third, reward specification beyond binary task success remains unsolved at scale, and reward-learning identifiability has known theoretical obstructions [149,150].

4.8. Cross-Embodiment Transfer and Action-Space Unification

Cross-embodiment learning is where the data and model axes meet most directly: the data axis supplies heterogeneous datasets [2,151,152], and the model axis must decide how one policy consumes incompatible action spaces, camera rigs, and control frequencies. Solutions stratify by where unification happens. Input-level unification standardizes observations and tokenizes whatever remains, the RT-X and Octo approach, which works but pushes embodiment differences into the data mixture [2,60]. Architecture-level unification gives each embodiment its own stem and head around a shared trunk, the HPT design, which scales to 52 datasets and shows positive trunk transfer [61]; CrossFormer extends one policy across manipulation, navigation, and locomotion without aligned action spaces [62].

 doi:10.20944/preprints202606.1149.v1

Representation-level unification seeks embodiment-invariant intermediates: latent actions extracted from video [69,89], robot-object interaction fields for grasping that transfer across hands [153], and trajectory world models pretrained across robot morphologies [87]. The empirical scoreboard favors pragmatism: every demonstrated transfer gain to date comes from input- or architecture-level unification trained on pooled real data; representation-level approaches carry the stronger invariance argument but remain at proof-of-concept scale. For the title question, cross-embodiment results are the data axis's best exhibit that composition can substitute for per-platform collection, and the model axis's best exhibit that the substitution rate is set by architectural choices [61,154].

4.9. Robustness Mechanisms

A final family modifies representations or objectives specifically to remove brittleness. Grounding mechanisms force the policy to locate what it is told to manipulate: affordance-internalized VLAs [155], embodied chain-of-thought traces that route prediction through plans and bounding boxes [55], visual trace prompting [65], and mask-based pretraining objectives that make object permanence explicit [156]. Structural 3D inductive biases reduce viewpoint sensitivity: point-cloud policies [132], 3D scene-token diffusion [157], view transformers [158,159], spatial position encodings [64], and unified robot-object interaction representations for cross-embodiment grasping [153]. Size- and shape-aware contrastive objectives target the object-replacement axis directly [160]. Pretrained visual representations for control [161–165] promise invariance inherited from diverse video, with the sobering caveat from the cortex study that no existing representation dominates across embodied tasks [164]. Test-time mechanisms close the loop without retraining: value-guided action re-ranking [146], world-model verified action selection [53,166], and execution-acceleration wrappers [167].

The pattern across this family mirrors the hierarchy result: each mechanism buys robustness on the axis it structurally encodes (3D structure buys viewpoint, grounding buys object identity, language-routed prediction buys instruction sensitivity) and approximately nothing elsewhere. No single mechanism, and no published combination, has yet been shown to close the LIBERO-PRO gap. That is an open experiment, not a theorem.

4.10. Efficiency as a Model-Axis Variable

Execution efficiency is usually filed under engineering, but it couples to the scientific question through evaluation throughput and deployment realism. The efficient-VLA literature catalogues compression, quantization, and architectural acceleration [23]; FAST-style tokenization [30] and OFT-style parallel decoding [32] fold efficiency into the action interface itself; plug-and-play acceleration policies speed execution without retraining the base model [167]; and structured sparsity results target the backbone [168]. Slow policies are evaluated less, on fewer seeds, with fewer perturbations; efficiency is therefore an input to evaluation quality, not only to deployment cost.

4.11. What the Model Axis Buys

The model-side evidence is narrower: model interventions are the only demonstrated remedies for failure modes that are *invariant to data quantity*, namely shortcut reliance under the BC objective [11,13], instruction insensitivity [14,55], and viewpoint fragility [14,132]. Architecture and objective changes also delivered the largest measured same-data gains of the period [32]. What the model axis has not demonstrated is competence creation: no architectural or algorithmic intervention has produced broad manipulation skill without a large demonstration collection somewhere in its lineage. The strongest model-side results are all dependent on the data investments of Section 5, which is why the title question cannot be answered by choosing only one side.

5. The Data Axis

5.1. Real-robot Demonstration Datasets

Table 3 consolidates the public datasets that define the field's empirical base. Three collection regimes dominate. *Single-platform teleoperation at scale* began with RT-1's 130k episodes over 17 months

 doi:10.20944/preprints202606.1149.v1

of office-kitchen collection [1] and BridgeData V2's 60k trajectories on a low-cost WidowX arm across 24 environments [169]; its current extremes are AgiBot World's million-plus trajectories from a fleet of over one hundred robots in standardized facilities [3] and RoboMIND's roughly 107k trajectories spanning four embodiments and 479 tasks with annotated failures [151]. *Federated in-the-wild collection* is represented by DROID, with 76k trajectories and 350 hours across 564 scenes collected by 50 operators on a standardized Franka cell in 13 institutions [170], and by RH20T's contact-rich, force-annotated episodes paired with human videos [171]. *Aggregation* pools existing datasets: Open X-Embodiment unified over one million trajectories from 22 embodiments and demonstrated positive cross-embodiment transfer with RT-X [2], the substrate for Octo, OpenVLA, and most subsequent generalists [5,60]. Specialized datasets extend coverage to bimanual platforms [152,172], multimodal contact-rich sensing [173], and even surgical robotics [174]. Historical context matters: RoboNet's 15 million video frames across seven platforms anticipated the aggregation thesis years earlier [50], and its limited impact relative to OXE illustrates that aggregation pays only once policies exist that can absorb heterogeneity.

Table 3. Public robot manipulation datasets and simulation task suites. Sizes are as reported in the cited papers; h = hours, traj = trajectories, emb = embodiments. "Weakness" summarizes limitations documented in the source or in subsequent evaluation literature.

Dataset	Year	Scale	Embodiments	Collection	Documented weakness
Real-robot datasets					
RoboNet [50]	2019	15M frames	7 arms	scripted	weak action semantics
RT-1 [1]	2022	130k traj	1 (Everyday Robot)	teleop	single site, single emb
BridgeData V2 [169]	2023	60k traj	1 (WidowX)	teleop	toy-scale objects, 1 emb
RH20T [171]	2023	110k+ traj, 40+h	4 arms	teleop + force	short horizons
OXE [2]	2023	1M+ traj	22	aggregation	heterogeneous quality, schema drift
DROID [170]	2024	76k traj, 350h, 564 scenes	1 (Franka)	federated teleop	policy results initially weak in-distribution
RoboMIND [151]	2024	∼107k traj	4	teleop	lab scenes
AgiBot World [3]	2025	1M+ traj, 217 tasks	1 fleet	factory teleop	single platform family
RoboCOIN [152]	2025	bimanual, multi-emb	many	consortium teleop	recency, uneven density
Kaiwu [173]	2025	multimodal episodes	1 cell	teleop + tactile/audio	scale
Open-H [174]	2026	large, medical	surgical	consortium	domain-specific
Human and egocentric video					
Ego4D [175]	2022	3,670h	human	worn cameras	no actions, no robot morphology
Ego-Exo4D [176]	2024	1,286h skilled	human	multi-view	same
EgoDex [177]	2025	∼800h + 3D hands	human	AVP capture	retargeting gap
EgoLive [178]	2026	large, task-oriented	human	head-mounted	recency
Simulation suites and generated datasets					
Meta-World [179]	2019	50 tasks	1 (Sawyer)	scripted	state obs, no language
RLBench [180]	2019	100 tasks	1 (Franka)	planner	rendering realism
CALVIN [51]	2022	34 tasks, play data	1 (Franka)	teleop play	4 fixed scenes
ManiSkill2/3 [181,182]	2023–24	20+ families, 2k+ objects	several	planner/RL	object-centric, short tasks
LIBERO [12]	2023	130 tasks × 50 demos	1 (Franka)	teleop	one scene/instruction per task; memorization-prone [13]
MimicGen [183]	2023	50k demos from ∼200	several	auto-synthesis	inherits seed-demo biases
BEHAVIOR-1K [184]	2024	1,000 activities	several	sampled	evaluation cost
RoboCasa [185]	2024	100 tasks, 2.5k+ assets	several	MimicGen-expanded	kitchen-domain bound
DexMimicGen [186]	2025	bimanual dexterous	humanoid hands	auto-synthesis	same

5.2. Human and Egocentric Video

Human video is the cheapest motion data in existence and the hardest to use. Ego4D and Ego-Exo4D supply thousands of hours of egocentric activity without action labels or robot morphology [175,176]; EgoDex adds calibrated 3D hand tracking at the hundreds-of-hours scale, making retargeting to dexterous hands tractable [177]; EgoLive pushes ecological validity with task-oriented daily routines [178]. Exploitation strategies span representation pretraining [161,165], latent-action extraction that treats video as actionless trajectories [69,88,89], co-training pyramids that mix human video below robot data [34], and direct retargeting through portable capture hardware: UMI's hand-held grippers [187], DexCap's portable mocap [188], immersive teleoperation rigs [189], and humanoid shadowing [190]. The unresolved question is conversion efficiency: no controlled study yet quantifies

 doi:10.20944/preprints202606.1149.v1

how many hours of egocentric video substitute for one hour of teleoperation at matched downstream success, a gap we flag in Section 9.

5.3. Simulation

Simulation supplies the only data whose generating process the researcher fully controls, which makes it simultaneously the best instrument for controlled science and the easiest place to overfit. The suite inventory (Table 3) ranges from procedurally narrow (Meta-World's 50 single-scene tasks [179]) through language-conditioned long-horizon play (CALVIN [51]) to ecosystem-scale activity simulation (BEHAVIOR-1K [184]) and GPU-parallel stacks built for RL throughput (ManiSkill3 [182]). LIBERO deserves specific attention because it became the field's default VLA benchmark while having exactly one scene, one instruction phrasing, and one goal per task, a design choice that makes layout a sufficient statistic for action prediction and thereby invites the memorization that Section 6 documents [12–14]. Domain randomization remains the standard bridge across the reality gap [191–193], with simulation-pretrained world models as a newer variant [97], and SimplerEnv supplying the calibration methodology that makes simulated evaluation predictive of real success [52].

5.4. Lessons from the Flagship Collections

Three large datasets are especially informative beyond their headline statistics. *DROID* demonstrated that in-the-wild diversity is collectible at consortium scale, and simultaneously that diversity alone does not guarantee superior policies: early evaluations found DROID-trained policies underperforming narrower-dataset baselines in-distribution, with the benefits appearing in robustness and scene transfer rather than raw success [170]. The collection has since become useful for world-model training because of its scene diversity [41]. *Open X-Embodiment* demonstrated positive cross-embodiment transfer and also exposed the costs of aggregation: heterogeneous action spaces, camera conventions, and annotation schemas forced every consumer into bespoke normalization pipelines, and mixture weights became a hidden hyperparameter that Re-Mix later showed to be decision-relevant [2,60,154]. *AgiBot World* demonstrated that treating collection as manufacturing, with standardized cells and human-in-the-loop verification, changes the quality-quantity tradeoff rather than just the quantity; its accompanying policy results attribute gains to this verification layer as much as to scale [3]. The pattern across all three: each dataset's lasting contribution reflected its *distributional* innovation (scenes, embodiments, verification), not its trajectory count, consistent with the scaling evidence of Section 5.8.

5.5. Documentation, Licensing, and the Missing Metadata

Dataset papers in manipulation rarely report the statistics that the scaling laws identify as decision-relevant: number of distinct environments, object instances per category, instruction phrasings per task, operator counts, and failure-trajectory fractions. Table 3 could be assembled only because a minority of papers report a minority of these fields. Licensing is similarly uneven: dataset licenses range from permissive to research-only, and aggregated datasets inherit the most restrictive license of any constituent, a constraint that OXE consumers navigate case by case [2]. The community has standards to import: datasheets and model cards from the broader ML community, and the annotated-failure precedent of RoboMIND [151]. We treat reporting standards as a bottleneck rather than a courtesy because Section 7's confounds are, in part, missing-metadata problems: train-test coincidence cannot even be detected when scene and instruction inventories go unreported.

5.6. Generated and Augmented Data

Between collection and simulation sits a third source: data synthesized from existing data. Replay-based synthesis exploits the structure of demonstrations themselves. MimicGen segments demonstrations into object-relative phases and re-targets them across poses, objects, and scene variants, multiplying roughly 200 human demonstrations into 50,000 [183]; DexMimicGen extends the recipe to bimanual dexterous platforms [186]; DemoGen performs the equivalent operation directly in point-cloud space, achieving spatial generalization from one demonstration per configuration [194];

 doi:10.20944/preprints202606.1149.v1

RoboCasa wraps the pipeline in a large procedural kitchen world [185]. Generative-model synthesis replaces geometry with learned priors: semantic image editing rewrites scenes while preserving action validity [195–197], language models author tasks and reward code wholesale [198–201], world models hallucinate entire trajectories that train policies for unseen verbs and environments [41,43,48], and image-to-3D pipelines lift open-world photographs into manipulable scenes [202].

Two properties make generated data analytically interesting for the title question. First, it is the one lever that changes the training distribution without new physical interaction, so its successes are evidence that the binding constraint is distributional coverage rather than physical experience per se [43,194]. Second, every generator inherits the biases of its seed data and its model class, so generated datasets can turn model assumptions into apparent data effects; physics violations documented in video generators [79,80] become silent label noise downstream. The field currently lacks any standard for auditing generated datasets, and data attribution tools for world-model training remain embryonic [203,204].

5.7. Collection Economics

The data narrative is ultimately an economic argument, and its numbers deserve scrutiny. Teleoperation throughput is bounded by one operator producing roughly one demonstration per minute on tabletop tasks; RT-1's collection consumed 17 months across 13 robots [1], and DROID's 350 hours required a 13-institution consortium and a year of coordination [170]. Vertical integration changes the slope: AgiBot World reports a standardized facility with over one hundred robots and human-in-the-loop quality verification producing a million-trajectory collection within a year [3]. Portable capture amortizes hardware away entirely [187,188], egocentric capture rides consumer devices [177,178], and replay synthesis multiplies whatever exists by two to three orders of magnitude at pure compute cost [183,194]. The unresolved economic question is marginal value, not marginal cost: the scaling studies below indicate that an additional identically distributed demonstration is worth far less than an additional environment or object, which implies that most of the money spent on same-distribution volume buys little generalization [205]. No published cost model combines collection prices with diversity-adjusted value; Section 9 proposes one.

5.8. Data Scaling Evidence

The controlled evidence on scale is thinner than the rhetoric around it, but it is consistent. Lin et al. [205] provide the cleanest manipulation result: across UMI-collected tasks, generalization to unseen environments and objects follows an approximate power law in the number of training environments and objects, while the number of demonstrations per environment saturates around tens; a policy trained in 32 diverse environments with roughly 50 demonstrations each approaches 90% success in entirely new settings (Figure 3). Pearce and et al. [206] establish that pre-training loss for behavior cloning and world modeling follows compute power laws with architecture-dependent coefficients, transferring the language-model scaling toolkit to embodied objectives. Cross-embodiment aggregation shows positive but uneven transfer: RT-X models outperform their single-domain ancestors by roughly 50% on average in underrepresented-lab evaluations [2], while Octo and OpenVLA document both the gains and the interference that mixture design must manage [5,60]. Mixture weights are themselves a first-class variable: Re-Mix shows distributionally robust reweighting of OXE domains materially changes downstream success [154], and the co-training mixtures of $\pi_{0.5}$ and GR00T N1 attribute generalization to composition rather than raw volume [34,47]. The summary with the strongest support is: *diversity scales, volume saturates, composition mediates*.

 doi:10.20944/preprints202606.1149.v1

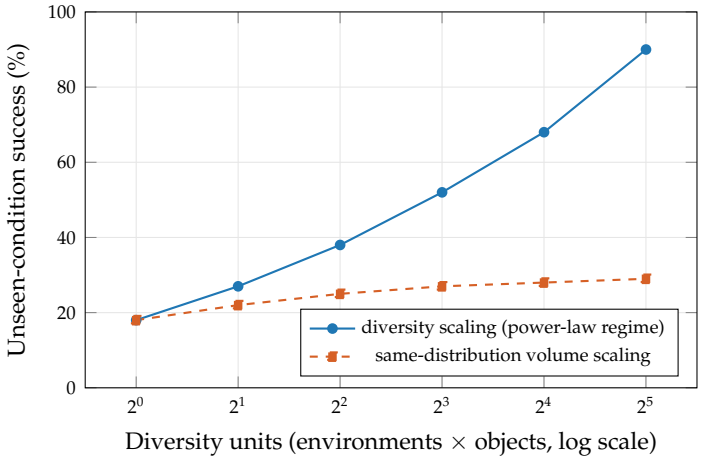

Figure 3. Qualitative redrawing of the central data-scaling finding [205]: unseen-environment success follows an approximate power law in environment and object *diversity* (solid), while adding demonstrations from already-covered conditions saturates quickly (dashed). Curves are schematic; quantitative values appear in the cited study, which reports roughly 90% success in novel environments after collection in 32 diverse environments with tens of demonstrations each.

5.9. Quality Versus Quantity

Quality effects are large enough to invert quantity conclusions. The robomimic study found that on identical tasks, demonstration quality (operator skill, consistency) changes achievable success more than dataset size, and that naive BC on mixed-quality data underperforms BC on a curated subset [207]. Belkhale et al. [208] formalize quality as action consistency and coverage, properties invisible to volume metrics. RoboMIND's annotated failure trajectories make error modes a usable signal rather than noise [151], and AgiBot World's human-in-the-loop verification treats quality as a manufacturing process [3]. At the collection level, weighting beats collecting: Re-Mix's reweighting gains came without a single new demonstration [154]. The data side is therefore not a single variable: volume, diversity, quality, and composition scale differently, are priced differently, and are conflated in nearly every public claim that "more data" produced a given system.

5.10. What the Data Axis Buys

The data evidence supports its own precise claim, symmetric to Section 4: distributional coverage is the only demonstrated source of broad competence, and diversity is its active ingredient [2,205]. Nothing on the model axis manufactures skills whose preconditions the training distribution lacks. What the data axis has not demonstrated is robustness to the perturbations that matter: every dataset in Table 3 was consumed by policies that subsequently failed factor-controlled evaluation [13–15], and the one benchmark designed around 130 tasks of curated diversity (LIBERO) became the canonical demonstration of memorization. More identically structured data does not fix an objective that permits shortcuts; that repair lives on the model axis. The two axes are complements with different failure coverage, which Section 7 makes exact.

6. The Evaluation Crisis

6.1. Benchmark Inflation and the Memorization Diagnosis

The crisis has a precise empirical core. Under LIBERO's standard protocol, in which evaluation episodes reuse the training tasks, scenes, and instruction strings with initial states drawn from the same distribution as training, OpenVLA-class and π_0-class models exceed 90% average success [13, 32]. LIBERO-PRO then perturbs along four reasonable dimensions (manipulated objects, initial states, instructions, environments) while keeping task semantics intact, and measures 0.0% in the fully generalized setting [13]. The qualitative failures identify the mechanism: policies execute the memorized grasp sequence when the target object is replaced by an irrelevant item, and produce

 doi:10.20944/preprints202606.1149.v1

unchanged outputs under corrupted or meaningless instruction tokens [13]. This is layout-keyed trajectory replay, not task competence. LIBERO-Plus decomposes the brittleness across seven factors and finds the damage concentrated in camera viewpoint and robot initial state, with several models effectively insensitive to language, confirming that the instruction channel is undertrained relative to its nominal role [14]. LIBERO-X extends the diagnosis with hierarchical perturbation protocols and cumulative-perturbation stress tests [209]. None of this is unique to one benchmark family: THE COLOSSEUM measured substantial degradation across 14 perturbation factors in RLBench-derived tasks two years earlier, established perturbation-ranking correlation between simulation and real robots [15], and the multi-benchmark VLA evaluations of Guruprasad et al. [210] found strong sensitivity to environmental factors across model families.

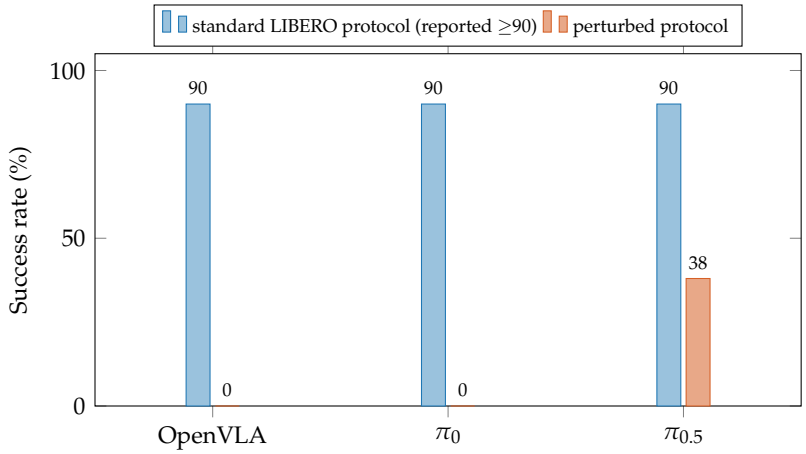

Figure 4. The generalization gap as reported by LIBERO-PRO [13]. Models exceeding 90% under the standard protocol drop to 0.0% in the fully generalized setting; the only non-zero retention reported among these models is $\pi_{0.5}$ at 38% on libero-goal under position perturbation alone, shown here as the perturbed bar. Standard-protocol bars are drawn at the conservative 90% bound stated in the source.

6.2. Why Standard Protocols Could Not Detect It

The blindness is structural, not negligent. First, train and test conditions coincide: LIBERO's design gives each task one scene, one instruction string, and one goal, so scene layout is a sufficient statistic for the correct action sequence, and a policy minimizing the BC objective has no gradient pressure to consult language or object identity [12,13]. Second, success-rate aggregation hides mechanism: a 97% average [32] is consistent both with robust competence and with high-fidelity replay, and only counterfactual probes (garbage instructions, swapped objects) separate the two [13,14]. Third, real-robot evaluation, the nominal gold standard, is statistically underpowered and irreproducible across labs, so the community substituted simulation benchmarks precisely because they are repeatable, then inherited their confounds; SimplerEnv exists because even measuring the substitution error required new methodology [52]. Fourth, evaluation matured late (Figure 2): COLOSSEUM's warning predated the VLA benchmark monoculture's peak, and was largely not adopted by the papers reporting LIBERO state of the art [15].

6.3. Sim-to-Real Validity and Learned Evaluators

Two further validity questions bound what any simulation number means. SimplerEnv shows that with careful system identification and visual matching, simulated success *rankings* correlate with real-robot rankings for generalist policies, making calibrated simulation a legitimate screening instrument while leaving absolute numbers uninterpretable [52]. GenManip extends controlled evaluation to LLM-driven scene generation and reports that modular foundation-model systems generalize better than end-to-end policies under its protocol, a finding that directly feeds the synthesis in Section 7 [201]. The newest instrument is the learned evaluator: world models that roll out policy-conditioned futures [41,42], verify action feasibility [53,166], or judge progress from observation

 doi:10.20944/preprints202606.1149.v1

streams [111,141]. Learned evaluators inherit their own validity problem (an evaluator trained on the same distribution as the policy can share its blind spots), and physics-faithfulness audits of the underlying video models remain discouraging [79,80,99]; we treat them as promising instruments requiring calibration studies of the SimplerEnv kind, not as solutions.

6.4. Statistical Power and the Reproducibility Layer

Beneath the validity questions sits a power question that the field systematically underweights. A manipulation evaluation with n rollouts per task bounds the confidence interval of a success estimate at roughly $\pm 1/\sqrt{n}$; at the common $n{=}50$ this is ± 14 percentage points at 95% confidence for mid-range success rates, wider than many claimed improvements. Real-robot evaluations frequently use $n \leq 20$ per condition, and cross-paper comparisons inherit additional variance from camera placement, controller tuning, and reset procedures that no protocol standardizes. Simulation removes the variance but, as Section 6 establishes, substitutes a validity problem. The practical consequences are visible in the literature: rank reversals between simulated and real evaluations [52], and benchmark deltas within the noise floor presented as method contributions. Reporting standards exist in adjacent fields and adapt directly: per-condition rollout counts, confidence intervals, seed counts, and pre-registered perturbation protocols. The evaluation instruments of Table 4 make this cheap in simulation; there is no equivalent fix for hardware beyond multi-lab replication, which the DROID consortium model shows to be organizationally feasible [170].

Table 4. Evaluation instruments for manipulation policies. "Perturbation axes" counts deliberately controlled factors of variation at evaluation time.

Instrument	Substrate	Perturbation axes	Validity evidence
LIBERO [12]	sim (MuJoCo)	none (standard protocol)	exposed by [13]
CALVIN [51]	sim	env split A–D	long-horizon chains
RoboCasa [185]	sim	scene/object sampling	–
THE COLOSSEUM [15]	sim (RLBench)	14 factors	sim-real ranking correlation
LIBERO-PRO [13]	sim	4–5 axes	memorization probes
LIBERO-Plus [14]	sim	7 factors	factor decomposition
LIBERO-X [209]	sim	hierarchical, cumulative	–
GenManip [201]	sim (Isaac)	LLM-generated scenes	human-in-loop audit
SimplerEnv [52]	calibrated sim	visual matching	explicit real correlation
World-model evaluators [41,42]	learned	arbitrary in principle	uncalibrated

6.5. What Good Evaluation Requires

The perturbation literature converges on four requirements. (i) *Factor control*: perturbation axes varied independently with per-axis reporting, not aggregate scores [14,15]. (ii) *Held-out axes*: when training-time augmentation targets known axes, honest evaluation requires axes the training pipeline never saw, otherwise robustness training reduces to benchmark adaptation one level up. (iii) *Mechanism probes*: counterfactual instructions and object swaps that distinguish competence from replay [13]. (iv) *Calibration*: any simulated or learned evaluator must carry quantified correlation with physical outcomes [52]. No published evaluation of a frontier VLA satisfies all four simultaneously as of this writing; the nearest miss is the LIBERO-Plus factor analysis, which satisfies (i) and (iii).

7. Evidence Synthesis: Data versus Models

7.1. The Evidence Matrix

Table 5 assigns the principal findings surveyed above to the data side, the model side, or the confounded middle. The assignment criterion is strict: a finding counts as evidence for an axis only if the study varied that axis while controlling the other.

 doi:10.20944/preprints202606.1149.v1

Table 5. Evidence matrix. Each row is a finding with controlled support; the final column names the confound that limits its scope.

Finding	Supports	Sources	Scope limitation
Generalization scales as a power law in environment/object diversity; per-environment volume saturates	data (diversity)	[205]	UMI tasks; single embodiment
Cross-embodiment aggregation improves underrepresented domains ∼50%	data (pooling)	[2]	evaluation on contributing labs
Mixture reweighting changes success without new data	data (composition)	[154]	OXE-scale only
Demonstration quality outweighs quantity on fixed tasks	data (quality)	[207,208]	small-scale, pre-VLA
Action-interface redesign lifts LIBERO 76.5→97.1% on identical data	model (architecture)	[32]	standard protocol; gap unprobed
Web-pretrained backbones transfer semantics to control	model (representation)	[4,5]	semantic, not motoric, transfer
RL post-training improves robustness over SFT on matched data; PPO > DPO/GRPO	model (objective)	[10,11]	LIBERO-family sims
One demo + RL reaches 91.7% where SFT yields 17.3%	model (objective)	[10]	needs competent base; sim
3D structure buys viewpoint robustness	model (inductive bias)	[14,132,158]	axis-specific
Hierarchy buys semantic-perturbation retention only	model (decomposition)	[14,201]	motor brittleness persists
≥90% standard success collapses to 0% under perturbation	evaluation	[13,15]	train/test coincidence is the cause
Co-training mixtures yield open-world generalization	confounded	[34,47]	data and model changed together
Generated data lifts unseen-verb/environment success	confounded	[43,194]	generator is itself a model

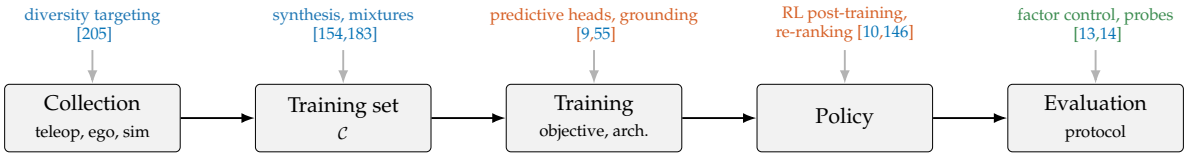

Figure 5. Where the surveyed interventions act on the learning pipeline. Data-axis levers (blue) change the training set; model-axis levers (orange) change what is learned from it or how it is used at inference; evaluation levers (green) change what either is credited with. The synthesis of Section 7 is that the blue levers set the ceiling, the orange levers set retention under shift, and the green levers determine which is visible.

7.2. Identified Confounds

Three confounds make the comparison difficult. *Train-test coincidence*: nearly every headline success number was measured where training and evaluation distributions coincide, which structurally favors the data narrative because matching the training distribution is exactly what data buys [12,13]. *Bundled interventions*: flagship systems change backbone, action head, data mixture, and post-training simultaneously [18,34,47]; their results are existence proofs, not attributions. *Generated-data attribution*: synthetic-data successes are often credited to the data axis even though the generator is itself a model, so the same result can support both narratives [43,183]. Future studies that claim to answer the title question need to avoid all three confounds; few surveyed here do, with [11,32,205] as the nearest exceptions on their respective axes.

7.3. The Conditional Answer

The evidence points to three claims. First, broad in-support competence comes from distributional coverage, with diversity as the active ingredient and quality and composition as multipliers; no model-side intervention creates skills absent from the training support [2,154,205,207]. Second, retention

 doi:10.20944/preprints202606.1149.v1

under distribution shift depends on objective, inductive bias, and post-training; more demonstrations with the same layout do not repair shortcut learning when layout already predicts action [11,13,55,132]. Third, the protocol changes the apparent winner: coincident train-test protocols favor the data story, whereas factor-controlled protocols expose failures that require model-side changes [14,15,52]. For system builders, the implication is simple: add diversity until coverage saturates, then invest in objectives and structure, and treat coincident train-test numbers with caution.

7.4. A Failure Taxonomy and Likely Interventions

The perturbation literature suggests a practical taxonomy of manipulation failures. *Perceptual failures*: the policy misreads scene content under viewpoint, lighting, background, or sensor-noise shift; concentrated damage under camera and visual factors [14,15]; addressed jointly by data (visual diversity is collectible and synthesizable at near-zero cost [195,196]) and by model structure (3D representations remove the viewpoint factor specifically [132,158]). *Semantic failures*: the policy misidentifies what to do; object replacement and instruction perturbation [13]; addressed mainly by the model side, since the training collection typically contains the needed semantics and the objective fails to bind them (Section 4.6), with hierarchy and grounding as demonstrated mitigations [14,55,201]. *Motoric failures*: the policy knows the task but cannot produce a valid trajectory from a shifted initial state; the largest single factor in LIBERO-Plus [14]; addressed jointly by data (spatial coverage via replay synthesis [183,194]) and by objective (interactive post-training [10,11]). *Systemic failures*: compounding drift, lack of recovery behavior, and unsafe trajectories; essentially unmeasured by current benchmarks, with closed-loop verification and uncertainty-aware models as candidate interventions [53,95,145]. The point of the taxonomy is diagnostic: a factor-controlled evaluation can identify the failure type and narrow the set of useful interventions. Its limitation is interaction effects, which remain unquantified (Section 7.5).

7.5. A Decision Framework for Practitioners

This suggests a practical rule for allocating effort. If the deployment conditions are already covered by the training collection, improve data quality and add per-condition examples until returns saturate [205,207]. If deployment varies a factor held fixed in training, collect that variation or use a model-side mechanism aimed at it: 3D structure for viewpoint, grounding for object identity, counterfactual pairing or hierarchy for instructions, and interactive post-training for initial-state robustness [10,14,55,132]. If the failure factor is unknown, run a factor-controlled audit first [14,15]. What remains largely untested is how these interventions interact when applied together.

Falsifiers. The first claim fails if a model-side method demonstrates broad novel-skill competence from a deliberately impoverished collection under factor-controlled evaluation; current world-model results do not yet meet this bar (Section 4.2). The second fails if scaling identically structured demonstrations closes a LIBERO-PRO-style gap without objective changes; LIBERO-X's diversified-training-data protocol is the natural venue for this test [209]. We consider both experiments executable today.

7.6. Threats to the Validity of This Synthesis

Four threats qualify the conclusions of this section. *Benchmark concentration*: the memorization evidence derives overwhelmingly from one simulator family; if LIBERO's single-scene task design is unusually pathological, the gap measured there overstates the field-wide problem, although COLOSSEUM's independent substrate argues otherwise [13,15]. *Survivor bias in the model evidence*: published model-side gains are selected for success, and the ablation studies we lean on [11,32,59] were run by groups invested in their methods. *Temporal asymmetry*: the data-axis evidence accumulated over five years while the factor-controlled evaluation evidence is months old; later perturbation studies may revise factor rankings, as LIBERO-Plus already partially revises LIBERO-PRO's aggregate framing [14]. *Proprietary opacity*: the strongest claimed open-world results sit on closed data and platforms [18,47], and nothing in the public record lets us decompose their gains across the axes. These threats do not overturn the argument, but they limit how strongly it should be read.

 doi:10.20944/preprints202606.1149.v1

8. Bottlenecks

8.1. Data

Collection remains supply-limited at the diversity margin: the scaling evidence prices environments and objects above demonstrations [205], yet most pipelines, including factory-scale ones, optimize demonstration throughput within few facilities [3]. Human-video conversion lacks a measured exchange rate [175,177]. Generated data lacks auditing standards, and attribution tooling is embryonic [79,203]. Dataset documentation rarely reports the diversity statistics that the scaling laws identify as decision-relevant.

8.2. Models

VLAs underuse their language channel and overuse layout [13,14]; no published mechanism or combination closes the perturbation gap. World models still lack the decisive isolation experiment connecting predictive objectives to manipulation robustness (Section 4.2), are unreliable on basic physics [79–81], and exhibit training instabilities that are only beginning to be characterized [98]. RL post-training depends on competent initializations and binary rewards, with reward learning facing identifiability obstructions [10,149]. Flow and diffusion heads complicate the very RL methods that fix BC's pathologies [143].

8.3. Evaluation

No current benchmark satisfies all four requirements of Section 6; real-robot evaluation remains statistically underpowered; learned evaluators are uncalibrated [42,52]. Most urgently, the perturbation suites that exposed memorization publish their full perturbation pools, so the next generation of models can train on the test distribution and the same overfitting problem can return [13,14].

8.4. Process

Two community practices make these technical problems harder to fix. Benchmark monoculture concentrated the field's evidence base on one simulator family until 2025, so conclusions and confounds were correlated across hundreds of papers; the perturbation suites inherit the same substrate and the same risk. Publication incentives also favor novelty over attribution: the bundled-intervention confound of Section 7 persists because unbundling is publishable only as an ablation table, while bundling is publishable as a system. Venues that solicit controlled negative results and factorial studies, as the reproducibility tracks of adjacent fields do, would repair the incentive at low cost.

9. Future Directions

We close with directions that can be tested directly.

1. The isolation experiment for world models. Train matched-capacity policies on identical LIBERO data, differing only in an auxiliary action-conditioned predictive head; evaluate on LIBERO-PRO/Plus with held-out axes. This is a direct low-cost test of whether predictive objectives improve robustness (Sections 4.2, 7), and infrastructure for it exists [102,147].

2. Counterfactually structured datasets. The memorization pathology traces to layout-instruction confounding in dataset design (one scene, one instruction, one goal). Datasets should include counterfactual pairs: identical scenes with differing instructions and differing correct behaviors, so that language is informative by design. Same-scene multi-task groups in existing datasets (e.g., LIBERO-goal's shared scene) already permit a retrofit, and replay synthesis machinery can manufacture the pairs at scale [13,183,194].

3. A diversity-adjusted cost model for collection. Combine per-modality collection prices (teleoperation, portable capture, egocentric video, synthesis) with the diversity exponents of Lin et al. [205] to produce dollars-per-unit generalization estimates, replacing trajectory counts as the headline statistic of dataset papers. The human-video exchange rate (hours of video per hour of teleoperation at matched success) is the key missing measurement [177,187].

 doi:10.20944/preprints202606.1149.v1

4. Held-out-axis evaluation as standard practice. Benchmarks should partition perturbation axes into public (trainable) and sealed (evaluation-only) sets, rotating the seal periodically, with mechanism probes (counterfactual instructions, object swaps) reported alongside success [13,209]. Without sealed axes, robustness leaderboards may overfit to public perturbations.

5. Calibrated learned evaluators. Replicate the SimplerEnv methodology for world-model evaluators: measure ranking correlation between world-model rollout success and physical success across a policy population, per perturbation factor [41,42,52]. A world model that ranks policies faithfully under perturbation would reduce the cost of the factor-controlled protocols this survey advocates.

6. Attribution for generated data. Extend data-attribution methods to world-model-generated datasets so that physics violations and distributional artifacts can be traced and filtered before they become policy failures [79,203,204].

7. Unified world-action training. Recent work across DreamZero-style world-action models [48], V-JEPA 2-AC [9], and WM-in-the-loop RL for VLAs [40,49] suggests a concrete experiment: a single model trained jointly as policy and predictor, evaluated under requirements (i)–(iv) protocols. The claim that robotics needs more than the VLA-plus-world-model stack [24] can be treated as a falsifiable null hypothesis.

8. Counterfactual probes as training-time monitors. The garbage-instruction and object-swap probes of Zhou and et al. [13] are cheap enough to run every epoch. Instruction-sensitivity divergence (the change in the action distribution under instruction substitution within a shared scene) is a direct, differentiable proxy for the language-channel health that LIBERO-Plus measures post hoc [14], and tracking it during training would show when the model stops using language.

9. Multimodal contact data at scale. Vision-only datasets dominate Table 3, yet contact-rich manipulation is where visual prediction is least sufficient. Force-annotated and tactile-instrumented datasets exist only at small scale [171,173]; no scaling study prices tactile diversity; and no perturbation benchmark varies contact dynamics. A comparable scaling study for force data would fill a major gap.

10. Safety-aware evaluation. Current protocols score success and ignore collision, force violations, and unsafe trajectories; GRAPE's multi-objective alignment is a rare exception [145], and evidential world models supply calibrated uncertainty that evaluation could consume [94,95]. A perturbation benchmark that reports success jointly with safety violations under shift would expose whether robustness interventions trade safety for retention, which is currently unknown.

Table 6. The research agenda of Section 9 in summary form. Cost is relative engineering plus compute effort for a single research group.

#	Direction	Hypothesis under test	Instrument	Cost
1	World-model isolation experiment	predictive objectives buy shift retention	LIBERO-PRO/Plus, held-out axes	low
2	Counterfactual dataset structure	layout-instruction confounds cause memorization	retrofitted LIBERO; new datasets	low–mid
3	Diversity-adjusted cost model	diversity, not volume, prices generalization	collection logs + [205]	low
4	Sealed-axis benchmarks	public perturbations can be overfit	rotating perturbation pools	mid
5	Calibrated learned evaluators	world models can rank policies faithfully	SimplerEnv-style correlation	mid
6	Generated-data attribution	generator artifacts propagate to policies	attribution tooling	mid
7	Unified world-action training	joint predictor-policy beats both alone	factor-controlled manipulation	high
8	Counterfactual training monitors	loss of language sensitivity is observable online	instruction-divergence probes	low
9	Contact-data scaling study	tactile diversity scales like visual diversity	force-instrumented collection	high
10	Safety-aware perturbation evaluation	robustness gains may trade against safety	joint success-safety reporting	mid

 doi:10.20944/preprints202606.1149.v1

10. Conclusions

What matters, datasets or robust models? The answer depends on the setting: datasets determine what a manipulation policy can do; models determine what it still does when the world stops resembling the training set; and evaluation design determines which contribution is visible. The research agenda follows from that split. On the data side: measure diversity, not only volume; include counterfactual structure; document distributions. On the model side: keep the objectives that punish shortcuts (predictive heads, interactive post-training, grounded intermediates) and submit them to the isolation experiments they have so far avoided. On the evaluation side: factor control, sealed axes, mechanism probes, calibration. The needed tools now exist, but experiments must separate data, model, and evaluation effects more carefully. The title question is useful because it forces that separation. Data and models are complements; their contributions become separable only under the newer factor-controlled evaluations. Future claims of progress should therefore be tested under better-controlled evaluations.

References

1. Brohan, A.; et al.. RT-1: Robotics Transformer for Real-World Control at Scale. *arXiv preprint arXiv:2212.06817* **2022**.
2. O'Neill, A.; et al.. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2024.
3. Bu, Q.; et al.. AgiBot World Colosseo: A Large-Scale Manipulation Platform for Scalable and Intelligent Embodied Systems. *arXiv preprint arXiv:2503.06669* **2025**.
4. Brohan, A.; et al.. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In Proceedings of the Conference on Robot Learning (CoRL), 2023.
5. Kim, M.J.; et al.. OpenVLA: An Open-Source Vision-Language-Action Model. In Proceedings of the Conference on Robot Learning (CoRL), 2024.
6. Chi, C.; et al.. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In Proceedings of the Robotics: Science and Systems (RSS), 2023.
7. Black, K.; et al.. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164* **2024**.
8. Hafner, D.; Pasukonis, J.; Ba, J.; Lillicrap, T. Mastering Diverse Control Tasks through World Models. *Nature* **2025**, *640*, 647–653.
9. Assran, M.; et al.. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. *arXiv preprint arXiv:2506.09985* **2025**.
10. Li, H.; et al.. SimpleVLA-RL: Scaling VLA Training via Reinforcement Learning. *arXiv preprint arXiv:2509.09674* **2025**.
11. Liu, J.; et al.. What Can RL Bring to VLA Generalization? An Empirical Study. *arXiv preprint arXiv:2505.19789* **2025**.
12. Liu, B.; et al.. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks, 2023.
13. Zhou, X.; et al.. LIBERO-PRO: Towards Robust and Fair Evaluation of Vision-Language-Action Models Beyond Memorization. *arXiv preprint arXiv:2510.03827* **2025**.
14. Fei, S.; et al.. LIBERO-Plus: In-Depth Robustness Analysis of Vision-Language-Action Models. *arXiv preprint arXiv:2510.13626* **2025**.
15. Pumacay, W.; Singh, I.; Duan, J.; Krishna, R.; Thomason, J.; Fox, D. THE COLOSSEUM: A Benchmark for Evaluating Generalization for Robotic Manipulation. In Proceedings of the Robotics: Science and Systems (RSS), 2024.
16. Hu, A.; et al.. GAIA-1: A Generative World Model for Autonomous Driving. *arXiv preprint arXiv:2309.17080* **2023**.
17. Tan, S.; Dou, K.; Zhao, Y.; Krähenbühl, P. Interactive Post-Training for Vision-Language-Action Models. *arXiv preprint arXiv:2505.17016* **2025**.
18. Gemini Robotics Team.; et al.. Gemini Robotics: Bringing AI into the Physical World. *arXiv preprint arXiv:2503.20020* **2025**.

 doi:10.20944/preprints202606.1149.v1

19. Hu, Y.; et al.. Toward General-Purpose Robots via Foundation Models: A Survey and Meta-Analysis. *arXiv preprint arXiv:2312.08782* **2023**.
20. Firoozi, R.; et al.. Foundation Models in Robotics: Applications, Challenges, and the Future. *The International Journal of Robotics Research* **2024**.
21. Ma, Y.; Song, Z.; Zhuang, Y.; Hao, J.; King, I. A Survey on Vision-Language-Action Models for Embodied AI. *arXiv preprint arXiv:2405.14093* **2024**.
22. Kawaharazuka, K.; et al.. Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications. *arXiv preprint arXiv:2510.07077* **2025**.
23. Yu, Z.; Wang, B.; Zeng, P.; et al.. A Survey on Efficient Vision-Language-Action Models. *arXiv preprint arXiv:2510.24795* **2025**.
24. Karcini, E.; Mehrban, F.; Ajoudani, A.; et al.. Robots Need More Than VLAs and World Models. *arXiv preprint arXiv:2606.06556* **2026**.
25. Ha, D.; Schmidhuber, J. Recurrent World Models Facilitate Policy Evolution. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2018.
26. Hafner, D.; Lillicrap, T.; Ba, J.; Norouzi, M. Dream to Control: Learning Behaviors by Latent Imagination. In Proceedings of the International Conference on Learning Representations (ICLR), 2020.
27. Hafner, D.; Lillicrap, T.; Norouzi, M.; Ba, J. Mastering Atari with Discrete World Models. In Proceedings of the International Conference on Learning Representations (ICLR), 2021.
28. Hansen, N.; Su, H.; Wang, X. TD-MPC2: Scalable, Robust World Models for Continuous Control. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.
29. NVIDIA.; Agarwal, N.; et al.. Cosmos World Foundation Model Platform for Physical AI. *arXiv preprint arXiv:2501.03575* **2025**.
30. Pertsch, K.; et al.. FAST: Efficient Action Tokenization for Vision-Language-Action Models. *arXiv preprint arXiv:2501.09747* **2025**.
31. Zhao, T.Z.; Kumar, V.; Levine, S.; Finn, C. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In Proceedings of the Robotics: Science and Systems (RSS), 2023.
32. Kim, M.J.; Finn, C.; Liang, P. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success. *arXiv preprint arXiv:2502.19645* **2025**.
33. Liu, S.; et al.. RDT-1B: A Diffusion Foundation Model for Bimanual Manipulation. *arXiv preprint arXiv:2410.07864* **2024**.
34. NVIDIA.; Bjorck, J.; et al.. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots. *arXiv preprint arXiv:2503.14734* **2025**.
35. Jang, E.; et al.. BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning. In Proceedings of the Conference on Robot Learning (CoRL), 2021.
36. Shridhar, M.; Manuelli, L.; Fox, D. CLIPort: What and Where Pathways for Robotic Manipulation. In Proceedings of the Conference on Robot Learning (CoRL), 2021.
37. Reed, S.; et al.. A Generalist Agent. *Transactions on Machine Learning Research* **2022**.
38. Bousmalis, K.; et al.. RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation. *Transactions on Machine Learning Research* **2023**.
39. Zhou, G.; Pan, H.; LeCun, Y.; Pinto, L. DINO-WM: World Models on Pre-Trained Visual Features Enable Zero-Shot Planning. *arXiv preprint arXiv:2411.04983* **2024**.
40. Jiang, Z.; et al.. World4RL: Diffusion World Models for Policy Refinement with Reinforcement Learning for Robotic Manipulation. *arXiv preprint arXiv:2509.19080* **2025**.
41. Guo, Y.; Shi, L.X.; Chen, J.; Finn, C. Ctrl-World: A Controllable Generative World Model for Robot Manipulation. *arXiv preprint arXiv:2510.10125* **2025**.
42. Quevedo, J.; et al.. WorldGym: World Model as an Environment for Policy Evaluation. *arXiv preprint arXiv:2506.00613* **2025**.
43. Jang, J.; et al.. DreamGen: Unlocking Generalization in Robot Learning through Video World Models. *arXiv preprint arXiv:2505.12705* **2025**.
44. Ahn, M.; et al.. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In Proceedings of the Conference on Robot Learning (CoRL), 2022.
45. Belkhale, S.; Ding, T.; et al.. RT-H: Action Hierarchies Using Language. In Proceedings of the Robotics: Science and Systems (RSS), 2024.
46. Physical Intelligence.; Shi, L.X.; et al.. Hi Robot: Open-Ended Instruction Following with Hierarchical Vision-Language-Action Models. *arXiv preprint arXiv:2502.19417* **2025**.

 doi:10.20944/preprints202606.1149.v1

47. Physical Intelligence.; Black, K.; et al.. $\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054* **2025**.
48. Ye, S.; et al.. World Action Models are Zero-Shot Policies. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
49. Zhang, Z.; et al.. Towards Practical World Model-Based R Einforcement Learning for Vision-Language Action Models. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
50. Dasari, S.; et al.. RoboNet: Large-Scale Multi-Robot Learning. In Proceedings of the Conference on Robot Learning (CoRL), 2019.
51. Mees, O.; Hermann, L.; Rosete-Beas, E.; Burgard, W. CALVIN: A Benchmark for Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks. *IEEE Robotics and Automation Letters* **2022**.
52. Li, X.; et al.. Evaluating Real-World Robot Manipulation Policies in Simulation. In Proceedings of the Conference on Robot Learning (CoRL), 2024.
53. Liu, Y.; et al.. World Action Verifier: Self-Improving World Models via Forward-Inverse Asymmetry. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
54. Bredis, G.; Balagansky, N.; Gavrilov, D.; Rakhimov, R. Next Embedding Prediction Makes World Models Stronger. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
55. Zawalski, M.; et al.. Robotic Control via Embodied Chain-of-Thought Reasoning. In Proceedings of the Conference on Robot Learning (CoRL), 2024.
56. Zhang, X. What Do World Models Learn in Rl? Probing Latent Representations in Learned Environment Simulators. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
57. Arghal, R.; et al.. A Behavioural and Representational Evaluation of Goal-Directedness in Language Model Agents. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
58. Farid, K. What Drives Compositional Generalization? The Importance of Continuous Training Objectives in Visual Generative Models. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
59. Li, X.; et al.. Towards Generalist Robot Policies: What Matters in Building Vision-Language-Action Models. *arXiv preprint arXiv:2412.14058* **2024**.
60. Octo Model Team.; Ghosh, D.; et al.. Octo: An Open-Source Generalist Robot Policy. In Proceedings of the Robotics: Science and Systems (RSS), 2024.
61. Wang, L.; Chen, X.; Zhao, J.; He, K. Scaling Proprioceptive-Visual Learning with Heterogeneous Pre-trained Transformers. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2024.
62. Doshi, R.; Walke, H.; Mees, O.; Dasari, S.; Levine, S. Scaling Cross-Embodied Learning: One Policy for Manipulation, Navigation, Locomotion and Aviation. In Proceedings of the Conference on Robot Learning (CoRL), 2024.
63. Li, Q.; et al.. CogACT: A Foundational Vision-Language-Action Model for Synergizing Cognition and Action in Robotic Manipulation. *arXiv preprint arXiv:2411.19650* **2024**.
64. Qu, D.; et al.. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Models. *arXiv preprint arXiv:2501.15830* **2025**.
65. Zheng, R.; et al.. TraceVLA: Visual Trace Prompting Improves Spatial-Temporal Awareness for Generalist Robotic Policies. *arXiv preprint arXiv:2412.10345* **2024**.
66. Wu, P.; Escontrela, A.; Hafner, D.; Abbeel, P.; Goldberg, K. DayDreamer: World Models for Physical Robot Learning. In Proceedings of the Conference on Robot Learning (CoRL), 2022.
67. Micheli, V.; Alonso, E.; Fleuret, F. Transformers are Sample-Efficient World Models. In Proceedings of the International Conference on Learning Representations (ICLR), 2023.
68. Schrittwieser, J.; et al.. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *Nature* **2020**, *588*, 604–609.
69. Bruce, J.; et al.. Genie: Generative Interactive Environments. In Proceedings of the International Conference on Machine Learning (ICML), 2024.
70. Parker-Holder, J.; et al.. Genie 2: A Large-Scale Foundation World Model. Google DeepMind Blog, 2024.

 doi:10.20944/preprints202606.1149.v1

71. Yang, M.; et al.. Learning Interactive Real-World Simulators. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.
72. Valevski, D.; Leviathan, Y.; Arar, M.; Fruchter, S. Diffusion Models are Real-Time Game Engines. *arXiv preprint arXiv:2408.14837* **2024**.
73. Alonso, E.; et al.. Diffusion for World Modeling: Visual Details Matter in Atari. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2024.
74. Kanervisto, A.; et al.. World and Human Action Models towards Gameplay Ideation. *Nature* **2025**, *638*, 656–663.
75. Du, Y.; et al.. Learning Universal Policies via Text-Guided Video Generation. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023.
76. Black, K.; Nakamoto, M.; et al.. Zero-Shot Robotic Manipulation with Pretrained Image-Editing Diffusion Models. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.
77. Wu, H.; et al.. Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.
78. Cheang, C.L.; et al.. GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation. *arXiv preprint arXiv:2410.06158* **2024**.
79. Kang, B.; et al.. How Far is Video Generation from World Model: A Physical Law Perspective. *arXiv preprint arXiv:2411.02385* **2024**.
80. Wu, S. Rigid Bench: Evaluating Rigid-Body Physics in Video Generation Models. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
81. Deng, Y.; et al.. Rethinking Video Generation Model for the Embodied World. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
82. Assran, M.; et al.. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.
83. Bardes, A.; et al.. Revisiting Feature Prediction for Learning Visual Representations from Video. *Transactions on Machine Learning Research* **2024**.
84. LeCun, Y. A Path Towards Autonomous Machine Intelligence. OpenReview preprint, 2022.
85. Mur-Labadia, L.; et al.. V-JEPA 2.1: Unlocking Dense Features in Video Self-Supervised Learning. *arXiv preprint arXiv:2603.14482* **2026**.
86. Huang, W.; Chao, Y.W.; Mousavian, A.; Liu, M.Y.; Fox, D.; Mo, K.; Fei-Fei, L. Point World: Scaling 3D World Models for I N The-Wild Robotic Manipulation. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
87. Wang, Y.; et al.. West World: a Knowledge-Encoded Scalable Trajectory World Model for Diverse Robotics. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
88. Abdulsalam, A. LaMO: a Latent Motion World Model for Long-Horizon Prediction. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
89. Kim, H.; et al.. Hier Archical Latent Action Model. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
90. Xiang, C.; et al.. Consistent Video World Model With Geometry-Aware Rotary Position Embedding. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
91. Sharma, R. Cross-View World Models. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
92. Chen, W.; et al.. H-wm: Robotic Task and Motion Planning Guided by Hier Archical World Model. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
93. Kumar, A.; et al.. World-Gymnast: Training Robots with Reinforcement Learning in a World Model. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
94. Li, C.; et al.. Uncertainty-Aware Robotic World Model Makes Offline Model-Based Reinforcement Learning Work on Realrobots. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.

 doi:10.20944/preprints202606.1149.v1

95. Kolling, A.H.; et al.. Evidential Latent World Models for Safe Model-Based Reinforcement Learning. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
96. Deb, R. Model Predictive Control with Differentiable World Models for Offline Reinforcement Learning. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
97. Levy, J.; et al.. Simulation Distillation: Pretraining World Models in Simulation for Rapid Real-World Adaptation. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
98. Ennadir, S. Understanding Early Collapse in Predictive World-Model Pretraining. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
99. Yang, K. Temporal Reversal Asymmetry: a Physics Inspired Metric for Evaluating World Models. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
100. Mairukh, N.; et al.. Phys Lang: a Small Diagnostic Framework for Language-Grounded World Modeling. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
101. Zhu, Y. Do LLMs Build Spatial World Models? Evidence From Grid-World Maze Tasks. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
102. Maes, L.; Lidec, Q.L.; Haramati, D.; Massaudi, N.; Scieur, D.; LeCun, Y.; Balestriero, R. Stable-worldmodel-V 1: Reproducible World Modeling Research and Evaluation. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
103. Yang, Z.; et al.. Physical Informed Driving World Model. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
104. Scott, D.; et al.. Coherence-Validated Causal World Models for Multi-Scale Alzheimer's Disease Progression and Pharmacologic Reversal. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
105. Scott, D.; et al.. Reinforcement Learning with World Models for Optimizing Alzheimer ' S Disease Treatment Timing and Dosing. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
106. Memon, Z.; et al.. Toward World Models for Epidemiology. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
107. Supreeth, M.; et al.. World Models as Execution Simulators for Automated Program Repair. In Proceedings of the International Conference on Learning Representations (ICLR), 2026.
108. Guan, Y.; et al.. Computer-Using World Model. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
109. Zhang, Y.; et al.. Cognitive Digital Twin Framework: Modeling and Real-Time Decision Making. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
110. Terver, B.; et al.. A Lightweight Library for Energy-Based Jointembedding Predictive Architectures. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
111. Zhang, Q.; et al.. Gridwm-Judge: Evaluating Vision-Language Model Judges in Grid Worlds via World Model Deficits. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
112. Imagination, L. Latent Imagination Thinking: Beyond Recursive Models for Reasoning. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
113. Niimi, J. The Mouth is Not the Brain: Bridging Energy Based World Models and Language Generation. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
114. Wu, J.; et al.. Visual Generation Unlocks Human-Like Reasoning Through Multimodal World Models. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.

 doi:10.20944/preprints202606.1149.v1

115. Wang, Y.; Bigelow, E.; Ullman, T.; Tang, Y.; Risi, S. Integrating Simulation and Chain-Of-Thought Reasoning in Multimodal-Language Models For Physical Reasoning. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
116. Yu, X.; et al.. DYNA-Think: Synergizing Reasoning, Acting, and World Model Simulation in AI Agents. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
117. Kobanda, A.; Radji, W. Intrinsic-Energy Joint Embedding Predictive Architectures Induce Quasimetric Spaces. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
118. Facury, L.; et al.. Learning Navigable World Models via Latent Energy Shaping. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
119. Zhang, J.; et al.. Reward-Forcing: Autoregressive Video Generation with Reward Feedback. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
120. Meng, X.; et al.. Identity-Grpo: Optimizing Multi-Human Identity-Preserving Video Generation via Reinforcement Learning. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
121. Cudlenco, N.; et al.. GEST-Engine: Controllable Multi-Actor Video Synthesis with Perfect Spatiotemporal Annotations. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
122. Wang, J.; et al.. Evoworld: Evolving Panoramic World Generation with Explicit 3D Memory. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
123. Lee, A. Dexsim: Real-Time Dexterous Simulation with Unified Causal Video Diffusion. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
124. Liu, S.; Wu, Z.; Yu, H.; Gao, J.; Alvarez, J.M. Structure From Diffusion: Taming Video Diffusion Models for Camera Pose Estimation in Dynamic Videos. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
125. Mun, H.; Jin, I.H.; Kim, S.; Kong, K. Fluidworld: Fluid-Like Interactive Dynamics for 4D Worlds. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
126. Duong, C.; et al.. Toward Pixel-Grounded World Models for Powered Descent: Arocket Landing Benchmark and Expertbaseline. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
127. Khalid, S.; et al.. Latentgs: Probabilistic Densification for Efficient, Compact, and Faster 3D Gaussian Splatting. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
128. Vaezpour, E.; Javadi, A.; Javidi, T. Active World-Model with 4D-informed Retrieval for Exploration and Awareness. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
129. Can, T.; et al.. Spa RRTA: a Synthetic Benchmark for Evaluating Spatial Intelligence in Visual Foundation Models. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
130. Ma, W.; Wang, C.; Yuan, R.; Chen, H.; Dai, N.; Yang, Y.; Qian, C.; Wang, Z.Y.; Yuille, A.; Chen, J. Causal Spatial: a Benchmark for Objectc Entric Causal Spatial Reasoning. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
131. Zhang, X.; et al.. Predicting Camera Posefrom Perspective D Escriptions for Spatial Reasoning. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
132. Ze, Y.; et al.. 3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations. In Proceedings of the Robotics: Science and Systems (RSS), 2024.
133. Huang, W.; et al.. Inner Monologue: Embodied Reasoning through Planning with Language Models. In Proceedings of the Conference on Robot Learning (CoRL), 2022.
134. Liang, J.; et al.. Code as Policies: Language Model Programs for Embodied Control. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2023.
135. Huang, W.; et al.. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. In Proceedings of the Conference on Robot Learning (CoRL), 2023.

 doi:10.20944/preprints202606.1149.v1

136. Fang, K.; Liu, F.; Abbeel, P.; Levine, S. MOKA: Open-World Robotic Manipulation through Mark-Based Visual Prompting. In Proceedings of the Robotics: Science and Systems (RSS), 2024.
137. Driess, D.; et al.. PaLM-E: An Embodied Multimodal Language Model. In Proceedings of the International Conference on Machine Learning (ICML), 2023.
138. Zeng, X.; et al.. Tree of Options: Temporally Extended World Modeling, Planning, and Execution with Large Language Models. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
139. Feng, Y.; et al.. Environment Maps: Structured Environmental Representations for Long-Horizon Agents. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
140. Wang, Q.; Huang, W.; Zhou, Y.; Yin, H.; Bao, T.; Lyu, J.; Liu, W.; Zhang, R.; Wu, J.; Fei-Fei, L.; et al. Enact: Evaluating Embodied Cognition with World Modeling of Egocentric Interaction. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
141. Zhang, J.; et al.. Progress Lm: Towards Progress Reasoning in Vision-Language Models. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
142. Lu, G.; et al.. VLA-RL: Towards Masterful and General Robotic Manipulation with Scalable Reinforcement Learning. *arXiv preprint arXiv:2505.18719* **2025**.
143. Chen, K.; et al.. π_{RL}: Online RL Fine-Tuning for Flow-Based Vision-Language-Action Models. *arXiv preprint arXiv:2510.25889* **2025**.
144. Wang, H.; et al.. RobustVLA: Robustness-Aware Reinforcement Post-Training for Vision-Language-Action Models. *arXiv preprint arXiv:2511.01331* **2025**.
145. Zhang, Z.; Zheng, K.; Chen, Z.; et al.. GRAPE: Generalizing Robot Policy via Preference Alignment. *arXiv preprint arXiv:2411.19309* **2024**.
146. Nakamoto, M.; Mees, O.; Kumar, A.; Levine, S. Steering Your Generalists: Improving Robotic Foundation Models via Value Guidance. In Proceedings of the Conference on Robot Learning (CoRL), 2024.
147. RLinf Team. RLinf: Reinforcement Learning Infrastructure for Embodied and Agentic AI. https://github.com/RLinf/RLinf, 2025.
148. Physical Intelligence. openpi: Open-Source Robot Foundation Models. https://github.com/Physical-Intelligence/openpi, 2025.
149. Lazzati, F. Robustness in the Face of Partial Identifiability in Reward Learning Problems. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
150. Gupta, P.; Gupta, V. Bootstrapped Mixed Rewards for RL Posttraining: Injecting Canonical Action Order. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
151. Wu, K.; et al.. RoboMIND: Benchmark on Multi-Embodiment Intelligence Normative Data for Robot Manipulation. *arXiv preprint arXiv:2412.13877* **2024**.
152. Wu, S.; et al.. RoboCOIN: An Open-Sourced Bimanual Robotic Data Collection for Integrated Manipulation. *arXiv preprint arXiv:2511.17441* **2025**.
153. Dexterous, E. D(r,o) Grasp: a Unified Representation of Robot and Object Interaction for Cross-Embodiment Dexterous Grasping. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
154. Hejna, J.; Bhateja, C.; Jiang, Y.; Pertsch, K.; Sadigh, D. Re-Mix: Optimizing Data Mixtures for Large Scale Imitation Learning. In Proceedings of the Conference on Robot Learning (CoRL), 2024.
155. Wang, R.; et al.. Afford-VLA: Action-Aligned Visual Planning via Internalized Affordance. *arXiv preprint arXiv:2605.24203* **2026**.
156. Multi-Object, P. Mask2Act: Predictive Multi-Object Tracking as Video Pre-Training for Robot Manipulation. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
157. Ke, T.W.; Gkanatsios, N.; Fragkiadaki, K. 3D Diffuser Actor: Policy Diffusion with 3D Scene Representations. In Proceedings of the Conference on Robot Learning (CoRL), 2024.
158. Goyal, A.; et al.. RVT: Robotic View Transformer for 3D Object Manipulation. In Proceedings of the Conference on Robot Learning (CoRL), 2023.
159. Shridhar, M.; Manuelli, L.; Fox, D. Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation. In Proceedings of the Conference on Robot Learning (CoRL), 2022.

 doi:10.20944/preprints202606.1149.v1

160. Imitation, C. Sacil: Size-aware Contrastive Imitation Learning for Language-conditioned Multi-task Robotics. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
161. Nair, S.; Rajeswaran, A.; Kumar, V.; Finn, C.; Gupta, A. R3M: A Universal Visual Representation for Robot Manipulation. In Proceedings of the Conference on Robot Learning (CoRL), 2022.
162. Ma, Y.J.; et al.. VIP: Towards Universal Visual Reward and Representation via Value-Implicit Pre-Training. In Proceedings of the International Conference on Learning Representations (ICLR), 2023.
163. Radosavovic, I.; Xiao, T.; James, S.; Abbeel, P.; Malik, J.; Darrell, T. Real-World Robot Learning with Masked Visual Pre-Training. In Proceedings of the Conference on Robot Learning (CoRL), 2022.
164. Majumdar, A.; et al.. Where are We in the Search for an Artificial Visual Cortex for Embodied Intelligence? In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023.
165. Karamcheti, S.; et al.. Language-Driven Representation Learning for Robotics. In Proceedings of the Robotics: Science and Systems (RSS), 2023.
166. Ziakas, C. Grounding Generated Videos in Feasible Plans via World Models. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
167. Wu, Z.; et al.. Speedup Patch: Learning a Plug-And-Play Policy to Accelerate Embodied Manipulation. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
168. Mestha, H. Block Mamba: Efficient Scalable Structured Sparsity for Mamba. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
169. Walke, H.; et al.. BridgeData V2: A Dataset for Robot Learning at Scale. In Proceedings of the Conference on Robot Learning (CoRL), 2023.
170. Khazatsky, A.; et al.. Droid: a Large-Scale In-The-Wild Robot Manipulation Dataset. In Proceedings of the Robotics: Science and Systems, 2024.
171. Fang, H.S.; et al.. RH20T: A Comprehensive Robotic Dataset for Learning Diverse Skills in One-Shot. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2024.
172. Fu, Z.; Zhao, T.Z.; Finn, C. Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation. In Proceedings of the Conference on Robot Learning (CoRL), 2024.
173. Jiang, S.; et al.. Kaiwu: a Multimodal Manipulation Dataset and Framework for Robot Learning. In Proceedings of the IEEE Robotics and Automation Letters, 2025.
174. Open-H-Embodiment Consortium.; Nelson, N.; et al.. Open-H-Embodiment: A Large-Scale Dataset for Enabling Foundation Models in Medical Robotics. *arXiv preprint arXiv:2604.21017* **2026**.
175. Grauman, K.; et al.. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
176. Grauman, K.; et al.. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.
177. Hoque, R.; et al.. EgoDex: Learning Dexterous Manipulation from Large-Scale Egocentric Video. *arXiv preprint arXiv:2505.11709* **2025**.
178. Li, Y.; et al.. EgoLive: A Large-Scale Egocentric Dataset from Real-World Human Tasks. *arXiv preprint arXiv:2604.23570* **2026**.
179. Yu, T.; et al.. Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning. In Proceedings of the Conference on Robot Learning (CoRL), 2019.
180. James, S.; Ma, Z.; Arrojo, D.R.; Davison, A.J. RLBench: The Robot Learning Benchmark and Learning Environment. *IEEE Robotics and Automation Letters* **2020**.
181. Gu, J.; et al.. ManiSkill2: A Unified Benchmark for Generalizable Manipulation Skills. In Proceedings of the International Conference on Learning Representations (ICLR), 2023.
182. Tao, S.; et al.. ManiSkill3: GPU Parallelized Robotics Simulation and Rendering for Generalizable Embodied AI. *arXiv preprint arXiv:2410.00425* **2024**.
183. Mandlekar, A.; et al.. MimicGen: A Data Generation System for Scalable Robot Learning Using Human Demonstrations. In Proceedings of the Conference on Robot Learning (CoRL), 2023.
184. Li, C.; et al.. BEHAVIOR-1K: A Human-Centered, Embodied AI Benchmark with 1,000 Everyday Activities and Realistic Simulation. *arXiv preprint arXiv:2403.09227* **2024**.
185. Nasiriany, S.; et al.. RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots. In Proceedings of the Robotics: Science and Systems (RSS), 2024.

 doi:10.20944/preprints202606.1149.v1

186. Jiang, Z.; et al.. DexMimicGen: Automated Data Generation for Bimanual Dexterous Manipulation via Imitation Learning. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2025.
187. Chi, C.; et al.. Universal Manipulation Interface: In-the-Wild Robot Teaching Without In-the-Wild Robots. In Proceedings of the Robotics: Science and Systems (RSS), 2024.
188. Wang, C.; et al.. DexCap: Scalable and Portable Mocap Data Collection System for Dexterous Manipulation. In Proceedings of the Robotics: Science and Systems (RSS), 2024.
189. Cheng, X.; et al.. Open-TeleVision: Teleoperation with Immersive Active Visual Feedback. In Proceedings of the Conference on Robot Learning (CoRL), 2024.
190. Fu, Z.; Zhao, Q.; Wu, Q.; Wetzstein, G.; Finn, C. HumanPlus: Humanoid Shadowing and Imitation from Humans. In Proceedings of the Conference on Robot Learning (CoRL), 2024.
191. Tobin, J.; et al.. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017.
192. Peng, X.B.; Andrychowicz, M.; Zaremba, W.; Abbeel, P. Sim-to-Real Transfer of Robotic Control with Dynamics Randomization. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2018.
193. Akkaya, I.; et al.. Solving Rubik's Cube with a Robot Hand. *arXiv preprint arXiv:1910.07113* **2019**.
194. Xue, Z.; Deng, S.; Chen, Z.; Wang, Y.; Yuan, Z.; Xu, H. DemoGen: Synthetic Demonstration Generation for Data-Efficient Visuomotor Policy Learning. In Proceedings of the Robotics: Science and Systems (RSS), 2025.
195. Yu, T.; et al.. Scaling Robot Learning with Semantically Imagined Experience. In Proceedings of the Robotics: Science and Systems (RSS), 2023.
196. Chen, Z.; Kiami, S.; Gupta, A.; Kumar, V. GenAug: Retargeting Behaviors to Unseen Situations via Generative Augmentation. In Proceedings of the Robotics: Science and Systems (RSS), 2023.
197. Bharadhwaj, H.; et al.. RoboAgent: Generalization and Efficiency in Robot Manipulation via Semantic Augmentations and Action Chunking. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2024.
198. Wang, L.; et al.. GenSim: Generating Robotic Simulation Tasks via Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.
199. Wang, Y.; et al.. RoboGen: Towards Unleashing Infinite Data for Automated Robot Learning via Generative Simulation. In Proceedings of the International Conference on Machine Learning (ICML), 2024.
200. Ha, H.; Florence, P.; Song, S. Scaling Up and Distilling Down: Language-Guided Robot Skill Acquisition. In Proceedings of the Conference on Robot Learning (CoRL), 2023.
201. Gao, N.; et al.. GenManip: LLM-Driven Simulation for Generalizable Instruction-Following Manipulation. *arXiv preprint arXiv:2506.10966* **2025**.
202. Gu, C.; et al.. IGen: Scalable Data Generation for Robot Learning from Open-World Images. *arXiv preprint arXiv:2512.01773* **2025**.
203. Ghosh, R.; et al.. Action Shapley: Atraining Dataselection Metric for Training World Models for Reinforcement Learning. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
204. Wu, X.; et al.. Motion Attribution for Video Generation. In Proceedings of the ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling, 2026.
205. Lin, F.; Hu, Y.; Sheng, P.; Wen, C.; You, J.; Gao, Y. Data Scaling Laws in Imitation Learning for Robotic Manipulation. In Proceedings of the International Conference on Learning Representations (ICLR), 2025.
206. Pearce, T.; et al.. Scaling Laws for Pre-Training Agents and World Models. *arXiv preprint arXiv:2411.04434* **2024**.
207. Mandlekar, A.; et al.. What Matters in Learning from Offline Human Demonstrations for Robot Manipulation. In Proceedings of the Conference on Robot Learning (CoRL), 2021.
208. Belkhale, S.; Cui, Y.; Sadigh, D. Data Quality in Imitation Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023.
209. Wang, G.; et al.. LIBERO-X: Robustness Litmus for Vision-Language-Action Models. *arXiv preprint arXiv:2602.06556* **2026**.
210. Guruprasad, P.; Sikka, H.; Song, J.; Wang, Y.; Liang, P.P. Benchmarking Vision, Language, & Action Models on Robotic Learning Tasks. *arXiv preprint arXiv:2411.05821* **2024**.

 doi:10.20944/preprints202606.1149.v1