The Evaluation Bottleneck of Vision-Language-Action Models: A Evaluation-Centric Survey

Zirui Song †, Huaxing Liu, Xiang Wang, Shuai Li, Xinye Li, Yuheng Ji, Lang Gao, Jinghui Zhang, Xianhui Meng, Xiaojun Chang, Xiuying Chen *

Posted Date: 5 June 2026

doi: 10.20944/preprints202606.0425.v1

Keywords: vision language model; evaluation

 doi:10.20944/preprints202606.0425.v1

Article

The Evaluation Bottleneck of Vision-Language-Action Models: A Evaluation-Centric Survey

Zirui Song [1,2,†], Huaxing Liu [2], Xiang Wang [2], Shuai Li [2], Xinye Li [1], Yuheng Ji [3], Lang Gao [1] Jinghui Zhang [1], Xianhui Meng [4], Xiaojun Chang [1] and Xiuying Chen [1,*]

[1] MBZUAI, United Arab Emirates
[2] AMAP, Alibaba Group, China
[3] CASIA, China
[4] USTC, China
[*] Correspondence: xiuying.chen@mbzuai.ac.ae
[†] Work done during internship at AMAP, Alibaba Group.

Abstract

Vision-Language-Action (VLA) models are advancing faster than the field can evaluate them reliably. Researchers use different metrics and lab-specific protocols, making it hard to tell whether reported gains reflect genuine progress or favorable evaluation choices. We present the first comprehensive benchmark-centric survey of VLA evaluation, covering 582 papers from 2023 to May 2026. We argue that a benchmark number supports a progress claim only if four conditions hold: the benchmark discriminates among top models; metrics capture the claimed capability; the procedure permits cross-paper comparison; and the inference from benchmark to deployment is valid. Current practice fails at all four. Benchmark choice is concentrated and saturated, with leading models clustering near the ceiling of dominant simulation suites. Metric reporting is one-dimensional, dominated by task success rate while efficiency, safety, and trajectory consistency remain underreported. Real-world evaluation is fragmented, with no widely adopted standard and few trials per task. Simulation scores are widely treated as evidence of real-world capability, yet standard task-centric suites correlate only weakly with real-robot performance unless explicitly calibrated to the target physical setup. These failures reveal a fundamental evaluation bottleneck: VLA models are advancing faster than our ability to measure that advance.

Keywords: vision language model; evaluation

1. Introduction

Vision-Language-Action (VLA) models which map visual observations and language instructions to robot actions have rapidly become the dominant paradigm for learning-based robotic manipulation (Black, Brown, et al., 2024; M. J. Kim et al., 2024; Zitkovich et al., 2023). Recent models report striking numbers: over 98% average success on LIBERO (H. Luo et al., 2026; Y. Yang, Zeng, et al., 2026), average sequence lengths above 4.0 on CALVIN (Mees et al., 2021), and increasingly strong zero-shot transfer across unseen objects, scenes, and embodiments (S. Liu et al., 2026; S. Ye et al., 2026). Taken at face value, these results suggest that embodied manipulation may be approaching a solved problem.

However, a closer look at the benchmarks on which these claims rest, including what they measure, how they measure it, and what they leave unmeasured, suggests a more cautious answer. Figure 1 illustrates this widening evaluation gap: VLA models are moving beyond single-arm tabletop manipulation toward dexterous, deformable, whole-body, and multi-robot collabration, while in evaluation, critical dimensions such as coordination quality, recovery, safety, contact stability, and trajectory consistency remain weakly measured. The question is not whether current VLAs are impressive. They clearly are. The real question is whether our evaluation practices are strong enough to tell us what these impressive numbers actually mean. Different research groups evaluate on different

 doi:10.20944/preprints202606.0425.v1

simulation suites, report different metrics, and rely on lab-specific real-world protocols. As a result, it is increasingly difficult to know whether a reported gain reflects a genuine improvement in robotic capability, or simply a favorable choice of benchmark, metric, or evaluation setup.

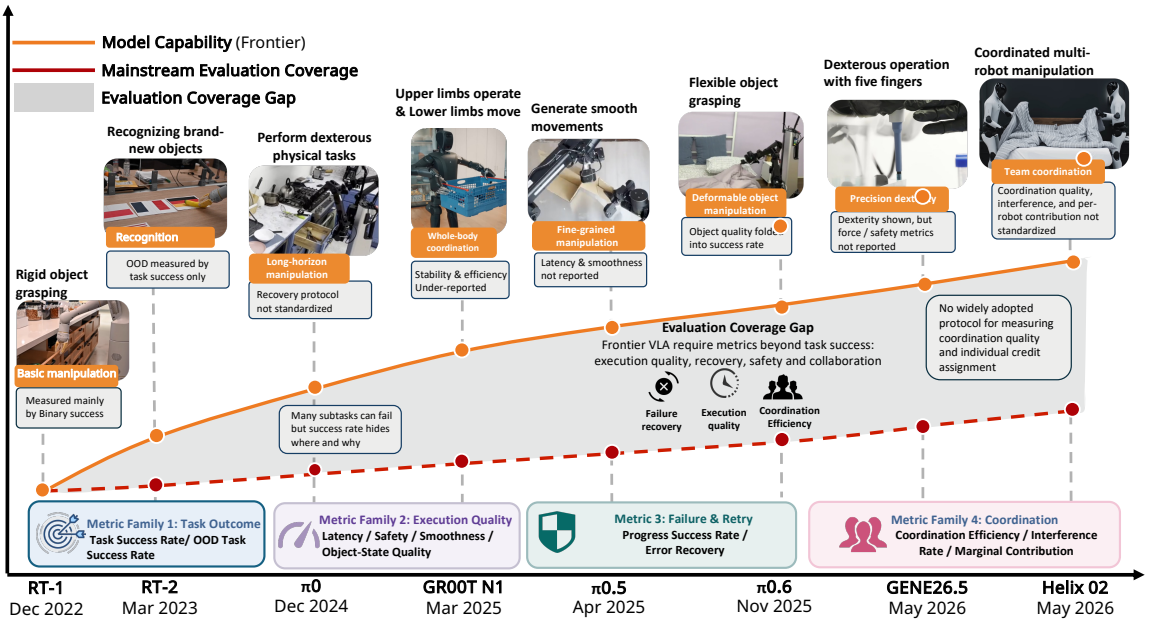

Figure 1. Conceptual timeline of the widening evaluation coverage gap in VLA models. The orange curve represents the advancing capability frontier, the red dashed curve represents mainstream evaluation coverage, and the shaded region marks capabilities that remain weakly measured by current benchmarks and metrics.

This survey examines that problem from a **benchmark-centric** perspective. Rather than focusing on how VLA models are built, we ask how they are evaluated. Across 582 VLA papers published between 2023 and May 2026, we find that evaluation shortcomings are not isolated but systematic. To diagnose them precisely, we ground this analysis in the principle that a benchmark number supports a progress claim only if four conditions hold jointly: the benchmark *discriminates* among top VLA models rather than clustering them near its ceiling; the reported *metrics* capture the capability actually being claimed, not only binary success; the real-world evaluation procedure is standardized enough to permit cross-paper *comparison*, rather than being locked to a single lab setup; and the *inference* from simulation score to real-world readiness is warranted by explicit calibration of the simulator against the target physical setup. Figure 2 shows that current practice fails at all four.

Empirically, our surveyed paper shows that current VLA evaluation falls short on each condition. First, benchmark evidence is concentrated in a few simulation suites: LIBERO, SimplerEnv, and CALVIN account for over 70% of papers with simulation evaluation, while recent models now cluster above 98% on LIBERO, reducing discriminative power. Second, metric coverage remains narrow: 98.3% of papers report task success rate, 57.9% report it as the sole main metric, and only 16.7% report confidence intervals, leaving efficiency, safety, smoothness, and recovery weakly measured. Third, real-robot evaluation is difficult to compare because most studies use lab-specific hardware, objects, scenes, reset procedures, and success criteria. Fourth, simulation-to-deployment inference is valid only when explicitly calibrated, yet task-centric simulation scores are often interpreted as evidence of real-world readiness without paired physical validation.

In response, this paper provides: (1) a two-axis taxonomy that maps the full benchmark landscape by evaluation setting and evaluation focus, exposing which capabilities are well covered and which remain untested; (2) a quantitative diagnosis of current evaluation practice against the four conditions, documenting benchmark saturation, metric narrowness, real-robot fragmentation, and the proxy-validity gap across 582 papers; and (3) an analysis of emerging capability frontiers where the mismatch between model ambition and evaluation capacity is widening most rapidly. To support

 doi:10.20944/preprints202606.0425.v1

more disciplined evaluation going forward, we also provide an evidence hierarchy for simulation benchmarks in Appendix E, minimum reporting protocols for simulation and real-robot evaluation in Appendix D, and a claim-metric alignment guide in Appendix D.3.

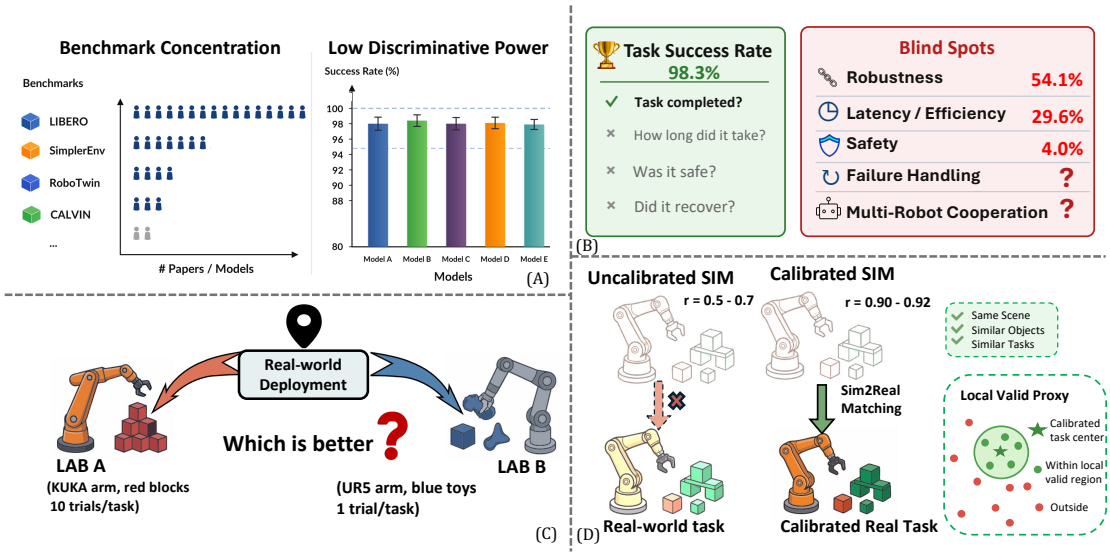

Figure 2. Four bottlenecks in current VLA evaluation: benchmark concentration and saturation, success-rate-dominated metrics, fragmented real-world protocols, and conditional sim-to-real validity.

2. Related Work

Recent VLA surveys have reviewed the field from different perspectives. Several works take a broad embodied-AI or application-oriented view, summarizing VLA concepts, model components, training paradigms, robot platforms, datasets, applications, and open challenges (Kawaharazuka et al., 2025; Y. Ma et al., 2026; Sapkota et al., 2025). Others are more model centric, organizing VLA models by architecture, learning paradigm, module design, or deployment pipeline, including autoregressive, diffusion-based, reinforcement-based, hierarchical, and large-VLM-based approaches (Kawaharazuka et al., 2025; C. Xu et al., 2025; D. Zhang et al., 2025).

A further group studies VLA models from more specialized angles. Y. Zhong et al. (2025) analyze VLA models through action tokenization; Xiang et al. (2026) connect VLA post-training with human motor learning; and S. Jiang et al. (2025) survey VLA models for autonomous driving. These surveys mainly focus on model design, training, applications, or domain-specific adaptation, while evaluation is usually treated as one component of a broader review.

The closest work Z. Wang, Wang, et al. (2026) to ours is the data-centric survey of VLA datasets, benchmarks, and data engines. While that work catalogs what benchmark and data resources exist for VLA research, our survey asks a different question: whether the results reported on these benchmarks can actually support the progress claims made from them. Specifically, we shift the focus from benchmark availability to benchmark validity, analyzing whether current evaluations discriminate among strong models, report metrics aligned with claimed capabilities, enable cross-paper comparison, and justify simulation-to-real-world inference. Thus, our work complements prior resource-centric surveys by auditing the evidential strength of VLA evaluation practice rather than only organizing its infrastructure.

3. A Taxonomy of VLA Benchmarks

Before diagnosing whether the four conditions are met in current practice, we first need a coordinate system: what benchmarks exist, what they cover, and what they leave out. We organize the benchmark landscape along two axes. As shown in Table 1, the first axis is the *evaluation setting*: closed-loop simulation, real-to-sim proxy evaluation, and real-robot evaluation. The second axis is the *evaluation focus*: the specific capability or failure mode that the benchmark is designed to probe,

 doi:10.20944/preprints202606.0425.v1

such as skill breadth, long-horizon composition, robustness, language grounding, safety, manipulation complexity, or deployment approximation.

This taxonomy exposes a common source of ambiguity in VLA evaluation: benchmark setting is often mistaken for evidential strength, blurring the distinction between what a benchmark tests and what conclusions it can support.

Simulation is not a single category of evidence. Task-centric suites such as LIBERO, CALVIN, RoboCasa, ManiSkill, and RLBench primarily support claims about benchmark-specific manipulation competence. Robustness-oriented suites test whether this competence survives controlled distribution shifts. Real-to-sim proxy benchmarks such as SimplerEnv and REALM ask a stronger and different question: whether simulated evaluation preserves real-world policy behavior, ranking, or sensitivity under calibrated visual and control conditions. These are different evidential claims, even when all are implemented in simulation.

Likewise, "real-robot evaluation" is not automatically comparable or conclusive. A custom hardware demonstration with few trials and vague success criteria may be less informative for cross-paper comparison than a controlled simulation benchmark, whereas centralized or protocol-standardized real-robot platforms provide stronger evidence because they fix more of the evaluation procedure. The relevant question, therefore, is not only where evaluation occurs, but what uncertainty the benchmark removes. The taxonomy below remains a two-axis map of evaluation setting and evaluation focus, whereas the benchmark evidence levels in Appendix E provide a separate claim-calibration layer. These evidence levels are not meant to rank benchmarks by quality, but to bound what kinds of progress or deployment claims a benchmark score can reasonably support.

Table 1. Taxonomy of VLA benchmarks. Each category is defined by its evaluation setting, evaluation focus, and evidential boundary, clarifying what claims a benchmark score can and cannot support. Examples are representative rather than exhaustive; full descriptions and citations appear in Appendix C.

Setting / focus	Evidential boundary	Representative examples
Simulation *Task competence*	Measures reproducible manipulation competence within the benchmark task distribution; it should not be read as physical deployment evidence.	LIBERO; Meta-World; RLBench; ManiSkill; RoboCasa.
Simulation *Sequential composition*	Tests chained subtask execution under predefined sequences; it does not establish open-ended long-horizon planning, memory, or recovery.	CALVIN; LIBERO-Long; BEHAVIOR-1K.
Simulation *Controlled robustness*	Supports robustness claims only for the specified perturbations; it does not imply general out-of-distribution robustness.	LIBERO-Plus/Pro; Colosseum; LIBERO-Para; LangGap.
Real-to-sim proxy *Calibrated deployment*	Supports local sim-to-real inference only when the robot, camera, controller, task family, and metric are explicitly calibrated.	SimplerEnv; REALM; Polaris, World-Env.
Real robot *Standardized physical*	Enables cross-paper physical comparison within a shared protocol; it does not imply general deployment superiority outside that protocol.	RoboChallenge; ManipArena; RoboArena; ManipulationNet; GM-100.
Real robot *Custom physical*	Provides feasibility evidence for the reported hardware, tasks, and lab setup; it is weak evidence for cross-paper model ranking.	Paper-specific lab evaluations.

4. Current Evaluation Practice

We test each of the four conditions against current practice. The picture that emerges is consistent: all four conditions fail. Before presenting the findings, we briefly summarize the corpus construction. Detailed collection, screening, and annotation procedures are provided in Appendix A.

 doi:10.20944/preprints202606.0425.v1

4.1. Benchmark Concentration and Saturation

The first condition requires that a benchmark discriminates among the models being compared. The most visible pattern in current VLA evaluation is a concentration of evidence around a small set of simulation benchmarks that increasingly fails this requirement.

The concentration of benchmark evidence is stark. The three most adopted simulation benchmarks, LIBERO B. Liu et al. (2023), SimplerEnv X. Li, Hsu, et al. (2024), and CALVIN Mees et al. (2021), together account for over 70% of all VLA papers with simulation evaluation. Further subset targets capabilities outside benchmark's scope(e.g., soft-body or deformable-object manipulation rather than rigid tabletop tasks). Among papers whose capability claims fall within table rigid tabletop coverage, the adoption rate is therefore substantially higher.

High adoption, however, carries a cost. When many papers optimize for the same benchmark and leading models approach the ceiling, small numerical gains become difficult to interpret as evidence of genuine progress.

Table 2 illustrates this issue. Several recent models report LIBERO averages above 98%, and the models cluster within roughly two percentage points of each other. At this level, a spread of roughly two percentage points may be driven by variation in checkpoint selection, random seeds, evaluation episodes, or implementation details, rather than by the advancing of model capability.

Table 2. Representative VLA performance on major simulation benchmarks, illustrating benchmark saturation among recent models.

Model	LIBERO	CALVIN	SimplerEnv
SimpleVLA-RL H. Li, Zuo, et al. (2025)	99.1	–	–
Being-H0.5 H. Luo et al. (2026)	98.9	–	–
Xiaomi-Robotics-0 R. Cai et al. (2026)	98.7	4.78	79.8
DiT4DiT T. Ma et al. (2026)	98.6	–	–
SimVLA Y. Luo, Chen, Liang, et al. (2026)	98.6	–	86.0
EO-1 Qu et al. (2025)	98.2	–	70.7
X-VLA J. Zheng, Li, Wang, et al. (2025)	98.1	4.43	84.0
VOTE J. Lin et al. (2025)	98.0	–	58.3
MMaDA-VLA Y. Liu et al. (2026)	98.0	4.78	–
FLOWER Reuss, Zhou, et al. (2025)	96.9	4.52	38.5
UniVLA Y. Wang, Li, et al. (2025)	95.5	4.63	69.8
RoboTron-Mani F. Yan et al. (2024)	91.7	3.51	60.0

This concentration also shapes which capabilities receive attention. Benchmarks that are easy to run, historically established, and widely reported tend to become the default evidence for progress, even when they no longer strongly separate models. Harder benchmarks that test richer visual scenes, longer task horizons, or systematic perturbations appear in fewer than 30% of surveyed papers, despite offering substantially more diagnostic headroom. For example, while many models exceed 98% on LIBERO, RoboCasa scores for strong recent models remain closer to the 40–55% range, suggesting that it exposes visual, semantic, and manipulation complexity that LIBERO do not.

Sequential benchmarks Mees et al. (2021), which measure how many chained few subtasks a policy completes, offer more headroom than saturated single-task suites. However, their performance both are sensitive to training recipes, checkpoint selection, and evaluation variants, so cross-paper numerical comparison remains harder than leaderboard-style tables suggest.

The broader lesson is that benchmark standardization and benchmark informativeness are not the same. A benchmark can be highly standardized yet weakly discriminative once top models saturate it. Conversely, a newer benchmark can be more informative about unsolved capabilities while still lacking enough adoption for broad cross-paper comparison. Current VLA evaluation therefore faces a coordination problem: the community benefits from shared benchmarks, but excessive reliance on saturated benchmarks narrows the evidence base for progress claims.

 doi:10.20944/preprints202606.0425.v1

4.2. Metric Coverage Gap

The second major pattern is the dominance of task success rate. Across our corpus, 98.3% of papers report task success rate as the main performance measure, and 57.9% report it as the *sole* metric in their experimental results. Success rate is simple and intuitive, but it collapses embodied behavior into a binary outcome and hides deployment-relevant properties such as speed, smoothness, safety, uncertainty, recovery, and robustness.

Several alternative metric families Ji et al. (2026) have begun to appear: partial-credit and progress scores for long-horizon tasks, sequential completion length for multi-step chaining, trajectory-level measures such as smoothness and reaction time, and efficiency metrics such as inference latency and control frequency. Each captures information that binary success rate discards, yet none has reached widespread adoption.

The metric gap matters because similar success rates can hide very different deployment behavior. A slow policy may succeed in static simulation but fail in dynamic manipulation; a jerky policy may complete a tabletop task while producing unsafe contacts; and a model may reach the goal through a path that would be unacceptable on hardware. Success rate alone cannot distinguish these cases, nor can it reveal rare but catastrophic failures under perturbation.

A number of recent papers J. Lin et al. (2025); J. Tang et al. (2025a); S. Xu et al. (2025) have begun reporting inference latency or control frequency alongside success rate, revealing that speed and accuracy are coupled in closed-loop control: improving benchmark success while reducing control frequency may not constitute a real deployment improvement. However, these remain isolated efforts rather than community practice. This pattern shows little change over time: from 2023 to 2026, 57.9% of papers report only success rate and out-of-distribution success rate, and the share of papers reporting latency, safety, or confidence intervals has not meaningfully grown. Full year-by-year statistics are provided in Appendix B.

Statistical reporting is also underdeveloped. Only 16.7% of surveyed papers report confidence intervals or comparable uncertainty estimates. This is especially problematic under benchmark saturation, where top LIBERO scores may differ by only one or two percentage points. Without trial counts, confidence intervals, or significance testing, small reported gains are difficult to distinguish from evaluation noise.

4.3. Fragmented Real-Robot Evaluation

Real-robot evaluation provides the strongest evidence of physical deployment capability, but current practice remains fragmented. Unlike simulation benchmarks, where task definitions and evaluation scripts are increasingly standardized, physical evaluations often differ in robots, objects, scenes, resets, success criteria, and trial budgets. As a result, a real-world result may be impressive within one laboratory setup but difficult to compare with another paper.

Current real-robot evaluation follows two main patterns. The most common is paper-specific evaluation: each study designs its own task suite around the capabilities it wishes to highlight, whether multi-embodiment long-horizon manipulation (Black, Brown, et al., 2024), transfer to unseen real-world scenes (S. Ye et al., 2026), or bimanual coordination (J. Liu et al., 2025). These evaluations can target realistic or novel capabilities, but because they differ in hardware, object sets, scene layouts, reset procedures, and success criteria, they rarely support direct cross-paper comparison. A second, newer pattern is the emergence of standardized real-robot benchmarks that attempt to fix one or more of these variables (Atreya et al., 2025; Y. Chen et al., 2026; Y. Sun et al., 2026a; Z. Wang, Liu, et al., 2026; Yakefu et al., 2025a). These efforts adopt different standardization strategies, centralizing hardware execution, shipping uniform object kits, or publishing shared task protocols and scoring rubrics, but adoption remains limited. A paper-level summary of real-robot platforms, scene scales, evaluation types, and task protocols is provided in Appendix I.

A pervasive concern across both patterns is statistical power. Many real-robot studies report only 10–20 trials per task, and some report fewer. Low trial counts also obscure failure structure: an

 doi:10.20944/preprints202606.0425.v1

aggregate success rate rarely reveals whether failures arise from perception, language grounding, grasp instability, controller latency, object slippage, unsafe contact, or task ambiguity.

The central problem is therefore not the absence of real-robot evaluation, but its limited comparability and statistical strength. Until standardized physical benchmarks become more widely adopted, custom real-world results are best interpreted as deployment demonstrations rather than definitive evidence of general model superiority.

4.4. Conditional Validity of Simulation Evidence

The fourth condition requires that the inference from benchmark setting to the claimed deployment context is warranted. This condition targets a specific and widespread practice: interpreting scores on task-centric simulation suites as evidence of real-world manipulation capability.

The issue is not that simulation is intrinsically misaligned with the real world, but that its evidential status is often underspecified. Recent real-to-sim benchmarks show that carefully designed simulation can provide a meaningful proxy for physical evaluation. For example, SimplerEnvconstructs simulated environments around specific real-robot manipulation setups and reports strong correlations with real-world policy performance on its original Google Robot and Bridge V2 configurations. However, its proxy validity is bounded: it does not support wrist-camera policies and has been shown to fail at predicting the performance of recent generalist VLAs (Jain et al., 2025), illustrating that even a well-designed proxy applies only within the scope of its calibration. Similarly, REALM explicitly targets real-to-sim validation by combining high-fidelity visuals, aligned robot control, systematic perturbations, and paired real-world comparisons.

These results change how simulation should be interpreted in VLA evaluation. They weaken the broad claim that simulation and real-robot performance are necessarily poorly correlated. At the same time, they strengthen a more precise claim: proxy validity is conditional. A simulator can predict real-world behavior only to the extent that the relevant embodiment, action interface, visual distribution, task family, perturbation space, and metric have been aligned or empirically validated. This validity does not automatically transfer from one benchmark to another.

This distinction is often blurred in current VLA reporting. Standard task-centric benchmarks such as LIBERO, CALVIN, RoboCasa, ManiSkill, and RLBench are valuable for measuring reproducible manipulation competence under controlled conditions. However, they are not designed primarily to answer whether a policy will preserve its behavior on physical hardware. Real-to-sim benchmarks such as SimplerEnv and REALM instead evaluate a different evidential claim: whether simulation can approximate real-world policy ranking, sensitivity, and failure modes under a calibrated setup. Both benchmark types are useful, but they support different conclusions.

The inference-validity condition is therefore unmet whenever a paper presents task-centric simulation scores as evidence of deployment readiness without providing either physical evaluation or an explicitly validated simulation proxy. The evaluation gap is no longer simply a sim-to-real gap, but a *proxy-validity gap*: VLA papers need to state whether each simulation benchmark is being used as a competence test, a robustness test, or a validated proxy for real-world behavior, and the strength of their conclusions should be bounded accordingly.

5. The Widening Evaluation Gap

Section 4 diagnoses four failures in current VLA evaluation. Figure 1 shows why these failures are not merely static shortcomings, but symptoms of a widening evaluation bottleneck: as VLA models move toward longer-horizon, dexterous, whole-body, and multi-robot tasks, the metrics and protocols needed to evaluate these capabilities remain underdeveloped. We therefore treat the evaluation gap as a dynamic problem: models are moving faster than evaluation infrastructure, and the distance between them is growing. We highlight four frontiers where the mismatch between model ambition and evaluation capacity is most acute.

 doi:10.20944/preprints202606.0425.v1

Long-horizon memory and reasoning

Current benchmarks Mees et al. (2021) cap task complexity at roughly five chained subtasks or short multi-step sequences. Real-world tasks of practical interest, such as tidying a room, preparing a full meal, or assembling a piece of furniture, involve tens to hundreds of sequential decisions, with intermediate failures that require error detection, recovery, and subgoal replanning.

At this scale, binary task success rate loses almost all diagnostic value. Evaluating long-horizon behavior requires fundamentally different metrics: partial-credit scoring that reflects how far the policy progressed, failure-locality measures that identify where in the task sequence errors concentrate, and explicit assessment of recovery capability after intermediate failures.

A further challenge is that long-horizon tasks amplify the compounding effect of per-step error rates. Even a policy with 99% per-step reliability reaches only $0.99^{100} \approx 37\%$ full-task success over 100 steps. This means that long-horizon evaluation must separately measure per-step reliability, compounding behavior, and recovery rate, rather than relying on end-to-end success alone.

Dexterous and contact-rich manipulation.

Dexterous manipulation sharpens the metric-coverage and proxy-validity problems discussed above. In contact-rich tasks, endpoint success is a weak proxy for the claimed capability: a policy may finish the task while relying on unstable grasps, excessive force, unsafe contacts, or simulator-specific contact artifacts. Conversely, a failed trial may still reveal meaningful partial dexterity, such as stable regrasping, compliant contact, tool alignment, or recovery after slip. Binary success therefore hides precisely the behaviors that distinguish dexterous manipulation from ordinary pick-and-place.

Comparability is also weaker than in standard tabletop manipulation. Existing dexterous benchmarks vary in hand morphology, action space, controller design, tactile sensing, contact modeling, and demonstration source. Recent efforts such as DexArt Bao et al. (2023), Ego Humanoid Manipulation R. Yang et al. (2025), RoboCasa Nasiriany et al. (2024), GR-1 tabletop tasks NVIDIA et al. (2025), and DexJoCo H. Wang et al. (2026) provide useful infrastructure, but their heterogeneity makes cross-paper ranking difficult. A score on one hand–controller–simulator stack rarely establishes transferable dexterity.

The resulting evaluation need is clear: dexterous VLA benchmarks should report not only task completion, but also grasp stability, slip, contact quality, force or torque violations, object damage, trajectory smoothness, recovery after failed contacts, and transfer across morphologies or controllers. Without these dimensions, dexterous manipulation risks becoming another frontier where benchmark success is mistaken for embodied capability.

Multi-robot coordination.

Bimanual manipulation, where benchmarks such as RoboTwin 2.0 have begun to provide evaluation support, represents the simplest case of multi-agent coordination: two arms controlled by a single policy with shared observations and a common objective. The next frontier involves genuinely distributed coordination: multiple robots with separate observation spaces collaborating on shared tasks, human-robot teams requiring dynamic role allocation, and multi-robots where individual objectives may conflict.

Evaluating these scenarios introduces dimensions that single-agent benchmarks do not address. Task success alone cannot capture whether coordination was efficient, whether communication overhead was acceptable, whether task allocation was reasonable, or whether one agent's failure was compensated by another's adaptation. Metrics must separately assess individual agent competence, coordination quality, communication efficiency, and robustness to single-agent failure. No existing VLA benchmark provides evaluation support for any of these dimensions, despite the growing number of multi-robot manipulation being proposed.

 doi:10.20944/preprints202606.0425.v1

Domain-specific fine manipulation.

The current benchmark landscape is built around a general-purpose manipulation paradigm: rigid or semi-rigid objects on tabletops, grasped and relocated by parallel-jaw or multi-finger grippers. However, some of the highest-value deployment scenarios for VLA models involve domain-specific manipulation where success criteria, precision requirements, and safety constraints are qualitatively different from general-purpose benchmarks.

Chemical synthesis requires sub-millimeter precision in liquid transfer, strict contamination control, and force-sensitive handling of fragile glassware. Surgical manipulation involves deformable soft tissue, real-time force feedback, and safety constraints that make any unintended contact a potential failure. Electronics assembly demands high-precision alignment, handling of small and fragile components, and tolerance specifications measured in micrometers. In each case, the relevant success criterion is not simply whether an object reached a target pose, but whether domain-specific constraints on force, purity, deformation, alignment, or biological safety were satisfied throughout the entire manipulation trajectory.

Current benchmarks largely fail to express these constraints because they lack the physical simulation fidelity such as fluid dynamics, soft-body deformation, contamination modeling, the domain-specific success criteria, and the safety metrics that these applications demand.

Recent efforts such as AutoBio Lan et al. (2025); R. Li et al. (2026) begin to address this gap by evaluating language-guided robotic manipulation in specialized scientific workflows, and their results suggest that current VLA models still struggle under precision-demanding domain constraints. These domains expose a broader limitation of current VLA evaluation: benchmark success on general tabletop manipulation does not necessarily imply readiness for high-precision, safety-critical deployment.

These frontiers share a common pattern: the evaluation challenge is not simply to build harder tasks within the existing benchmark paradigm, but to develop new evaluation dimensions, metrics, and protocols that the current paradigm does not support. Long-horizon reasoning requires temporal credit assignment and recovery measurement. Multi-agent coordination requires interaction-level metrics beyond individual task success. Domain-specific manipulation requires constraint-satisfaction evaluation throughout the trajectory rather than endpoint-only assessment. None of these needs can be met by adding more tasks to LIBERO or increasing trial counts on existing benchmarks.

The evaluation bottleneck documented in this survey is therefore not only a present-day problem but a forward-looking one. As VLA models move toward these harder capability frontiers, the gap between what models can attempt and what benchmarks can measure will continue to widen unless the community invests in evaluation infrastructure with the same urgency it currently devotes to model development. For capabilities where standardized evaluation already exists, Appendix D provides minimum reporting protocols, including separate simulation and real-robot checklists in Appendices D.1 and D.2, while Appendix E provides an evidence hierarchy to guide benchmark selection and claim calibration. For the emerging frontiers discussed here, the first step is recognizing that evaluation design is itself a research problem.

6. Conclusions

This survey identifies evaluation validity as a central bottleneck for VLA progress. Across surveyed papers, we find that current evaluation relies on saturated simulation benchmarks, success-rate-dominated metrics, fragmented real-robot protocols, and weakly justified sim-to-real inference. Future VLA research needs stronger evaluation infrastructure, not only stronger models. Benchmarks, metrics, and real-world protocols should be matched to the claims they support, especially as models move toward long-horizon, dexterous, multi-robot, and domain-specific manipulation.

Limitations

This survey focuses on VLA models for robotic manipulation and does not comprehensively cover navigation-only or locomotion-only benchmarks. Independent verification across all benchmarks was

 doi:10.20944/preprints202606.0425.v1

not feasible. Additionally, our real-robot evaluation analysis may undercount proprietary evaluation setups that are not fully described in publications. For real-to-sim proxy benchmarks, our discussion relies on the validation protocols and correlations reported by the benchmark authors rather than independent re-execution across all evaluated policies and physical setups.

Appendix A. Survey Methodology and Paper Corpus

This section describes how the paper corpus was constructed and how the paper-level evaluation statistics reported in the main text were computed. The goal of the corpus is not to count every robotics paper that mentions vision, language, or action, but to characterize evaluation practice in recent VLA research for robotic manipulation.

Survey scope.

We focus on VLA models and evaluation resources for robotic manipulation. In this survey, a VLA model refers to a policy, policy component, or embodied foundation model that conditions on visual observations and language instructions and produces robot actions, action tokens, motion plans, or low-level control commands. The survey covers papers released between January 2023 and May 2026. We include model papers, benchmark papers, dataset papers, evaluation papers, real-to-sim studies, and real-robot deployment papers when they report, introduce, or analyze evaluation for language-conditioned manipulation. We exclude navigation-only, locomotion-only, autonomous-driving-only, pure vision-language understanding, and purely open-loop action-prediction papers unless they are explicitly evaluated as part of a robotic manipulation policy.

Paper collection.

We constructed the corpus in three stages. First, we formed an initial candidate set from publicly available paper collections and community-maintained Awesome-style repositories related to VLA models, robot learning, embodied AI, language-conditioned manipulation, robot foundation models, and manipulation benchmarks. These public collections were used as the starting point because they aggregate many of the central model, dataset, and benchmark papers that define the VLA evaluation landscape.

Second, we supplemented this initial set using Hugging Face Daily Papers. We searched for recent papers using keywords associated with VLA models, robotic manipulation, robot foundation models, and commonly used evaluation benchmarks. This step was used to capture newly released preprints and fast-moving work that may not yet have appeared in curated paper lists or formal conference proceedings.

Third, we searched Google Scholar using the same keyword families and benchmark names. The search terms included “vision-language-action”, “VLA”, “language-conditioned manipulation”, “robot foundation model”, “generalist robot policy”, “robotic manipulation benchmark”, “embodied manipulation”, “real-to-sim”, “sim-to-real”, and benchmark names such as LIBERO, CALVIN, SimplerEnv, RoboCasa, RLBench, ManiSkill, Meta-World, RoboTwin, ManipArena, RoboArena, and RoboChallenge. Google Scholar was used both to identify additional relevant papers and to recover earlier or related works cited by recent VLA papers. When multiple versions of the same work were available, such as an arXiv preprint and a later conference version, we treated them as one paper and used the most recent public version available before the May 2026 cutoff.

Screening criteria.

After collecting candidate papers, we manually screened each paper for relevance to VLA evaluation. A paper was retained if it proposed or evaluated a VLA model for robotic manipulation, introduced a benchmark or dataset used for VLA manipulation evaluation, analyzed evaluation practice for VLA systems, studied robustness, language grounding, simulation-to-real transfer, safety, or real-robot deployment, or reported closed-loop manipulation results conditioned on language and visual observations. A paper was removed if it did not involve robot manipulation, did not evaluate

 doi:10.20944/preprints202606.0425.v1

action-producing policies, only reported offline prediction metrics without closed-loop evaluation, or lacked enough experimental detail to identify the evaluation setting or benchmark usage. Duplicate versions were merged before computing corpus-level statistics.

Annotation procedure.

For each retained paper, we annotated bibliographic information, evaluation setting, benchmark usage, metric reporting, and the type of capability claim supported by the experiments. The evaluation setting was categorized as closed-loop simulation, real-to-sim proxy evaluation, real-robot evaluation, or a combination of these. Benchmark usage was recorded at the paper level: a paper was counted as using a benchmark if it reported results on that benchmark in the main text, appendix, or official supplementary material. For real-robot evaluation, we further recorded whether the evaluation used a standardized physical benchmark or a paper-specific laboratory setup.

We also annotated the metrics reported by each paper. Task success rate was counted when a paper reported binary task completion, episode success, or an equivalent pass/fail manipulation measure. Inference latency was counted only when the paper provided a numerical runtime, inference-time, control-frequency, or closed-loop execution-rate measurement. Safety-related metrics included collision rate, unsafe-contact rate, constraint-violation rate, intervention rate, failure-detection accuracy, or other explicitly defined safety quantities. Uncertainty reporting was counted when a paper provided confidence intervals, standard errors, standard deviations across seeds or trials, significance tests, or comparable uncertainty estimates. Long-horizon metrics included average sequence length, subtask completion rate, progress score, or failure-position analysis for multi-step tasks.

Appendix B. Temporal Trends in Metric Reporting

Table A1 reports year-by-year metric coverage in the surveyed corpus. The purpose of this analysis is not to claim a precise causal trend for each individual metric, but to test whether VLA evaluation practice shows evidence of broadening over time. A healthy maturation pattern would suggest increasing adoption of deployment-relevant and uncertainty-aware metrics as VLA systems become more capable and are increasingly framed as real-world robotic agents.

Table A1. Year-by-year metric coverage in the surveyed VLA corpus. Values are percentages of papers in each year. "Any L/S/CI" denotes papers reporting at least one of latency, safety-related metrics, or confidence intervals. "Success/OOD only without L/S/CI" denotes papers that report success rate and/or OOD success but none of latency, safety, or confidence-interval metrics. The 2026* row includes papers through May 2026.

Year	N	Success	OOD success	Latency	Safety	CI	Any L/S/CI	Success/OOD only without L/S/CI
2023	48	100.0	52.1	25.0	0.0	33.3	43.8	56.3
2024	87	97.7	63.2	29.9	4.6	24.1	46.0	52.9
2025	319	98.1	52.7	28.5	3.8	11.0	36.7	62.1
2026*	128	98.4	52.3	33.6	5.5	19.5	47.7	51.6
All	582	98.3	54.1	29.6	4.0	16.7	41.1	57.9

The observed pattern does not show such convergence. Task success remains nearly universal in every year, appearing in 97.7–100.0% of papers. By contrast, latency reporting increases only modestly, from 25.0% in 2023 to 33.6% in 2026*. Safety-related metrics remain extremely rare throughout the period, rising from 0.0% in 2023 to only 5.5% in 2026*. Confidence-interval reporting does not improve over time; it is 33.3% in 2023, 24.1% in 2024, 11.0% in 2025, and 19.5% in 2026*.

The combined metric column further supports this conclusion. The fraction of papers reporting at least one of latency, safety, or confidence intervals fluctuates rather than increasing monotonically: 43.8% in 2023, 46.0% in 2024, 36.7% in 2025, and 47.7% in 2026*. Thus, the metric-coverage gap is not simply an artifact of early VLA research. Instead, it reflects a persistent lack of community

 doi:10.20944/preprints202606.0425.v1

consensus around what evidence is needed to support increasingly ambitious claims about deployment, robustness, safety, and real-world readiness.

The 2026* row covers papers published through May 2026 and should therefore be interpreted as year-to-date evidence rather than a complete annual estimate. Nevertheless, the year-to-date pattern is consistent with the broader conclusion: evaluation practice remains dominated by success-rate reporting, with limited adoption of broader deployment-relevant metrics.

Appendix C. Full Benchmark Taxonomy

Table A2 provides the complete taxonomy introduced in Section 3, with detailed protocol descriptions, key coverage notes, and per-benchmark summaries. For real-robot benchmarks and dataset-backed evaluation resources, the table further distinguishes standardization paradigms: centralized online, federated hardware-standardized, federated kit-standardized, federated protocol-standardized, and paper-specific setups.

Benchmark assignment criteria.

We assign each benchmark according to its dominant evaluation bottleneck rather than every property it contains. Many benchmarks combine broad task coverage, long-horizon variants, diverse objects, difficult physical interactions, or real-robot validation. To keep the taxonomy interpretable, secondary properties are reported in the coverage/protocol or summary column, while the evaluation focus column records the benchmark's primary evaluation target.

Operational definition of long-horizon benchmarks.

Because benchmarks report horizon length at different granularities, including control timesteps, primitive actions, annotated subtasks, and high-level skills, we classify *Long Horizon* by high-level temporal structure rather than raw simulator timesteps alone. A benchmark is categorized as *Long Horizon* if its primary protocol satisfies one of the following conditions: (i) it requires the ordered composition of at least five semantically distinct high-level subtasks or primitive skills; (ii) it explicitly evaluates extended planning, re-planning, or accumulated execution errors across multiple task stages; or (iii) it evaluates *temporal state tracking*, where the policy must retain, update, or retrieve task-relevant historical state across delayed, occluded, repeated, or multi-stage interactions. Benchmarks with ten or more high-level subtasks are treated as stronger or very-long-horizon cases, but ten subtasks is not used as the minimum inclusion threshold. Thus, *Long Horizon* includes both subtask-chain benchmarks and memory-intensive or history-conditioned benchmarks whose main bottleneck is temporal state tracking.

Distinguishing skill breadth from manipulation complexity.

We distinguish *Skill Breadth* from *Manipulation Complexity* by whether the main difficulty is horizontal coverage or vertical execution difficulty. *Skill Breadth* refers to benchmarks that primarily evaluate coverage across many task families, goals, objects, scenes, language prompts, task splits, demonstrations, or evaluation protocols. A benchmark can still be assigned to Skill Breadth if some tasks are multi-step or physically nontrivial, provided that the dominant protocol stresses coverage and generalization rather than physical execution difficulty. In contrast, *Manipulation Complexity* refers to benchmarks that primarily stress difficult physical execution, such as contact-rich dynamics, dexterous or high-DoF hand control, bimanual or multi-arm coordination, articulated or deformable objects, fluids, tool or appliance use, embodiment variation, controller-level difficulty, or precise object-state changes. Task count, object count, language diversity, scene diversity, or dataset size alone is not sufficient for Manipulation Complexity unless these variations primarily increase the difficulty of physical execution.

 doi:10.20944/preprints202606.0425.v1

Table A2. Taxonomy of VLA benchmarks along two axes: the *evaluation setting* and the *evaluation focus*. For real-robot benchmarks we further distinguish standardization paradigms, including centralized online, federated hardware-standardized, federated kit-standardized, federated protocol-standardized, and paper-specific setups. The table separates *where* evaluation happens from *what* capability is measured.

Evaluation Setting	Evaluation Focus	Benchmark	Key Coverage / Protocol	Summary
Closed-loop Simulation	Skill Breadth	LIBERO family B. Liu et al. (2023)	Spatial, Object, Goal, Long, and LIBERO-90/100-style variants	Canonical standardized suite for reproducible language-conditioned tabletop manipulation.
	Skill Breadth	Meta-World T. Yu et al. (2021)	50-task multi-task / multi-skill suite (ML10/ML50 splits)	Long-running multi-task baseline; widely reused for efficiency and RL post-training studies.
	Skill Breadth	VIMA-Bench Y. Jiang et al. (2023)	Multimodal-prompt manipulation benchmark covering object rearrangement, visual goals, and compositional task instructions.	Evaluates whether a policy can generalize across diverse prompted manipulation tasks rather than only fixed language templates.
	Skill Breadth	Ravens Zeng et al. (2022)	PyBullet-based tabletop rearrangement benchmark with 10 pick-and-place-oriented tasks, oracle demonstrations, and partial-credit reward functions.	Tests broad spatial rearrangement and manipulation-skill coverage across tabletop tasks, rather than explicit long-horizon subtask-chain planning.
	Skill Breadth	CLIPort Shridhar et al. (2021)	Ravens-based language-conditioned tabletop manipulation benchmark with 10 simulated tasks, 1000s of unique instances per task, seen/unseen splits over colors, shapes, and object categories, and 0–100 task-score evaluation over 100 rollout episodes	Evaluates horizontal breadth in language-conditioned manipulation, covering semantic grounding, attribute/object generalization, and spatial pick-and-place compositions rather than explicit long-horizon task-chain execution.
	Skill Breadth	RLBench James et al. (2020)	V-REP/PyRep-based simulated manipulation benchmark with a Franka Panda, 100 unique hand-designed tasks, task variations with textual descriptions, RGB-D/mask/proprioceptive observations, sparse rewards, multiple action spaces, and motion-planner-generated demonstrations	Provides a broad robot-learning task bank for evaluating visually guided manipulation across diverse task families, variations, demonstrations, and few-shot/multi-task settings; includes harder long-horizon cases, but its primary focus is task breadth rather than physical manipulation complexity.
	Skill Breadth	RoboCasa Nasiriany et al. (2024)	Large-scale kitchen simulation benchmark with 120 realistic scenes, 100 everyday tasks, 150+ object categories, 2,500+ 3D assets, interactable appliances, and 100K+ demonstration trajectories	Evaluates broad household manipulation coverage across tasks, objects, scenes, and datasets for generalist robot learning, rather than primarily targeting low-level physical manipulation complexity.
	Skill Breadth	RoboCasa365 Nasiriany et al. (2026)	RoboCasa-based household mobile-manipulation benchmark with 365 everyday kitchen tasks, including 65 atomic and 300 composite tasks across 60 activities, 2,500 pretraining kitchen scenes, 10 target kitchen scenes, 612 hours of human demonstrations, 1,615 hours of synthetic demonstrations, and Atomic / Composite-Seen / Composite-Unseen target splits	Evaluates generalist robot policies across large-scale task, scene, and dataset diversity for multi-task learning, robot foundation model training, and lifelong learning; includes long-horizon composite tasks, but its dominant focus is scalable skill breadth and generalization rather than dedicated long-horizon planning alone.
	Skill Breadth	OctoGibson J. Yang et al. (2024)	OmniGibson-based vision-language programming environment with 476 formulated household tasks, 367 routine and 109 reasoning tasks, 16 executable function calls, 3,776 instructional subtasks, 37,760 image-instruction pairs, and task-completion / plan-score evaluation over seen and unseen environments	Evaluates broad household embodied task coverage for vision-grounded planning and executable code generation; includes reasoning and multi-step planning tasks, but primarily targets VLM program generation and task breadth rather than low-level physical manipulation complexity.
	Long Horizon	CALVIN Mees et al. (2021)	ABC→D split; five chained subtasks; scene-level distribution shift	Measures whether a policy can compose multiple learned skills over extended horizons.
	Long Horizon	RoboCerebra Han et al. (2026)	100 household task variants with 1,000 human-annotated simulated trajectories; random disturbance, and mixed settings; 600-rollout protocol over 60 tasks × 10 trials	Evaluates deliberative long-horizon manipulation, targeting planning, reflection, and memory beyond reactive execution.
	Long Horizon	BEHAVIOR-1K C. Li et al. (2024)	1,000 household activities with realistic scenes, articulated objects, and long task chains	Pushes long-horizon evaluation toward open-vocabulary household activities at large scale.
	Long Horizon	Franka Kitchen Gupta et al. (2019)	Sequential kitchen manipulation tasks involving multiple object interactions, appliance operations, and temporally extended control	Provides a classic benchmark for testing whether policies can execute long-horizon manipulation beyond isolated pick-and-place actions.
	Long Horizon	LoHoSet Y. Yang, Sun, et al. (2025)	Ravens-based long-horizon tabletop manipulation suite with a UR5e suction-gripper setup, RGB-D top-down observations, primitive pick-and-place tasks, and 20 long-horizon rearrangement tasks with sub-task and action annotations	Evaluates whether policies can decompose high-level language goals into executable sub-tasks and complete multi-step tabletop rearrangement under closed-loop visual feedback.
	Long Horizon	LabUtopia R. Li et al. (2026)	High-fidelity laboratory simulation and hierarchical benchmark with physical and chemical interactions, diverse scientific instruments, and tasks from atomic manipulation to long-horizon mobile manipulation	Evaluates whether embodied agents can execute scientific laboratory workflows requiring perception, planning, instrument control, and robustness to long-horizon execution errors.
	Long Horizon	BiCoord Peng et al. (2026)	Long-horizon bimanual manipulation benchmark with tightly coordinated dual-arm tasks, stage-wise annotations, and spatial-temporal coordination metrics such as SMT, SMP, MRD, ARD, STI, and SSR	Evaluates whether policies can execute long-duration dual-arm manipulation requiring phased coupling, spatial-temporal constraints, predictive coordination, and fine-grained stage-wise task completion.
	Long Horizon	LongVILBench Q. Chen et al. (2025)	Real-world human-demonstration benchmark for long-horizon visual imitation learning, covering 150 tabletop manipulation tasks, 300 RGB videos, 1–18 atomic actions, six spatial relations, and executable plan/code annotations	Evaluates whether agents can infer temporally ordered action sequences and spatial object relations from human demonstrations, with generated programs verified through simulation and real-robot execution.
	Long Horizon	HiMan-Bench Y. Chen, Chen, et al. (2025)	RLBench-style long-horizon manipulation benchmark with atomic and compositional tasks under diverse perturbations, including object appearance, size, lighting, distractors, background, and camera-pose changes	Evaluates whether policies can compose learned atomic skills into robust long-horizon behaviors under systematic environmental perturbations.
	Long Horizon	VLABench S. Zhang et al. (2024)	MuJoCo/dm_control-based language-conditioned manipulation benchmark with 100 task categories, 60 primitive and 40 composite tasks, 163 object categories / 2164 objects, semantically rich implicit instructions, seen/unseen object evaluation, and Progress Score based on target selection and sub-step completion	Evaluates VLA policies, foundation-model workflows, and VLMs on long-horizon multi-step reasoning, combining semantic intent grounding, common-sense/world-knowledge transfer, spatial/physical understanding, and task planning across diverse manipulation skills and scenes.
	Long Horizon	MIKASA-Robo Cherepanov et al. (2026)	Tabletop manipulation benchmark with 32 memory-intensive tasks involving occluded objects, spatial recall, sequential cues, and multi-object state tracking under partial observability	Evaluates temporal state tracking in tabletop manipulation, requiring policies to retain and use task-relevant historical observations under delayed or occluded evidence rather than relying only on current visual input.

 doi:10.20944/preprints202606.0425.v1

Table A2. *Cont.*

Evaluation Setting	Evaluation Focus	Benchmark	Key Coverage / Protocol	Summary
Closed-loop Simulation (cont.)	Long Horizon	RMBench T. Chen et al. (2026)	RoboTwin 2.0/SAPIEN-based dual-arm manipulation benchmark with 9 history-conditioned tasks spanning M(1) and M(n) task-memory complexity, including observe-and-pick, rearrange, swap, cover, battery, ranking, and button-pressing tasks	Evaluates temporal state tracking for manipulation policies, especially whether they can accumulate, retrieve, and update historical observations during repeated exploration, trial-and-error, or multi-step execution.
	Environment Robustness	LIBERO-Plus Fei, Wang, Shi, et al. (2025) / LIBERO-Pro X. Zhou et al. (2025) / The Colosseum Pumacay et al. (2024)	LIBERO-Plus: controlled perturbations over layout, camera, robot initial state, language, lighting, background, and sensor noise; LIBERO-Pro: object, initial-state, instruction, and environment perturbations targeting memorization; The Colosseum: multi-task manipulation under controlled visual and environmental perturbations.	Robustness extensions that test whether standard benchmark success survives systematic distribution shifts rather than reflecting trajectory, scene, or template memorization.
	Environment Robustness	RoboVerse H. Geng et al. (2025)	Unified simulation benchmark with standardized generalization levels over task, object, environment, camera, lighting, and embodiment variations	Tests whether robot policies remain effective under systematic distribution shifts across assets, scenes, embodiments, and simulator-backed task sources.
	Environment Robustness	VLA-Arena B. Zhang et al. (2025)	Structured simulation benchmark for VLA models with 170 tasks across Safety, Distractor, Extrapolation, and Long-Horizon dimensions, three task difficulty levels, and graded language and visual perturbations	Evaluates whether VLA policies generalize beyond memorized training tasks under task-structure shifts, safety constraints, distractors, long-horizon composition, and multimodal perturbations.
	Language Robustness	LIBERO-Para C. Kim et al. (2026), LangGap Y. Hou and Zhao (2026), RAMA-Bench W. Song, Chen, Li, et al. (2025)	Paraphrase, language-gap, and defective-instruction perturbations on LIBERO-style or custom tasks	Diagnoses whether a VLA actually understands instructions versus pattern-matching template phrasings.
	Intention	INT-ACT I. Fang et al. (2025)	Intention-aware manipulation settings with potential action-intention mismatch	Probes whether the model's executed action matches the intended goal rather than only surface instructions.
	Safety	SAFE Ying et al. (2024)	Failure detection, unsafe-condition recognition, and robustness to invalid execution states	Evaluates whether a VLA can recognize failures or unsafe situations instead of blindly acting.
	Safety	LABSHIELD Q. Sun et al. (2026)	Multimodal laboratory safety benchmark with multi-view robotic observations, 164 operational tasks, four operation levels, and four safety tiers grounded in OSHA and GHS standards	Evaluates whether embodied agents can identify laboratory hazards, reason over safety-critical risks, refuse unsafe instructions, and generate safe action plans before physical execution.
	Safety	ResponsibleRobotBench L. Zhang et al. (2025)	Safety-critical robotic manipulation benchmark with 23 multi-stage tasks spanning electrical, fire/chemical, and human-related hazards, adversarial/defensive instructions, scene and planning complexity, and multiple action representations	Evaluates whether embodied agents can detect hazards, reason about safety constraints, plan risk-mitigating actions, request human assistance when necessary, and complete manipulation tasks responsibly.
	Manipulation Complexity	RoboTwin Y. Mu et al. (2025)	Dual-arm robotic manipulation benchmark with generative digital twins, synthetic expert demonstrations, real-world-aligned task scenarios, and validation on open-source bimanual robot platforms	Evaluates dual-arm coordination and complex object manipulation with simulation-generated data aligned to real-world execution.
	Manipulation Complexity	RoboTwin 2.0 T. Chen et al. (2025)	Scalable bimanual manipulation benchmark with 50 dual-arm tasks, five robot embodiments, 731 object instances across 147 categories, and structured domain randomization over clutter, lighting, background, tabletop height, and language	Extends RoboTwin toward robust bimanual manipulation by testing policy generalization across task diversity, object variation, embodiment changes, and sim-to-real perturbations.
	Manipulation Complexity	ManiSkill T. Mu et al. (2021)	Scalable physics-based manipulation tasks with varied objects, contacts, and scene layouts	Tests contact-rich manipulation in a more physically diverse simulation framework.
	Manipulation Complexity	ManiSkill2 Gu et al. (2023)	Unified SAPIEN-based manipulation benchmark with 20 task families, 2,000+ object models, 4M+ demonstration frames, and support for rigid/soft-body, single/dual-arm, stationary/mobile manipulation	Tests generalizable manipulation skills under diverse object geometry, task types, observations, and controller settings in fully dynamic simulation.
	Manipulation Complexity	ManiSkill3 Tao et al. (2025)	GPU-parallelized robotics simulator and benchmark with contact-rich physics, heterogeneous simulation, visual rendering, point-cloud/voxel observations, and tasks spanning multiple manipulation domains	Extends scalable manipulation evaluation by enabling high-throughput visual robot learning across diverse scenes, embodiments, and contact-rich task settings.
	Manipulation Complexity	robosuite Y. Zhu et al. (2020)	MuJoCo-based robot manipulation simulation framework with benchmark environments, multiple robot embodiments, controllers, and reusable task definitions	Provides a standardized simulation substrate for reproducible manipulation experiments, though benchmark comparability depends on the selected tasks, robots, and control settings.
	Manipulation Complexity	Adroit Kumar (2016)	Dexterous-hand manipulation benchmark with Shadow Hand tasks such as door opening, hammering, object relocation, and pen manipulation	Tests high-DoF dexterous control and contact-rich manipulation beyond parallel-jaw tabletop pick-and-place tasks.
	Manipulation Complexity	DexMimicGen Z. Jiang et al. (2025)	Bimanual dexterous manipulation suite and data-generation benchmark that synthesizes large-scale demonstrations from a small number of human demonstrations	Tests whether policies can learn coordinated dexterous manipulation from scalable demonstration data across tasks, simulators, and real-world variants.
	Manipulation Complexity	DexArt Bao et al. (2023)	Dexterous manipulation benchmark with articulated objects, multi-finger robot hands, 3D observations, and unseen-object generalization splits	Tests high-DoF dexterous control and articulated-object manipulation beyond parallel-jaw tabletop pick-and-place tasks.
	Manipulation Complexity	RoboMimic Mandlekar et al. (2021)	Imitation-learning benchmark and dataset suite built around robosuite-style manipulation tasks, demonstrations, and standardized policy-learning protocols	Provides a reusable benchmark for evaluating manipulation policies learned from demonstrations, with comparability depending on the selected tasks, observations, and control settings.
	Manipulation Complexity	RoboFactory Qin et al. (2025)	Multi-agent collaborative manipulation benchmark with diverse multi-arm tasks involving two to four robotic arms, coordination constraints, and physical interaction patterns	Evaluates whether policies can coordinate multiple embodied agents for manipulation tasks requiring synchronized execution, collision avoidance, and shared scene understanding.
	Manipulation Complexity	ARNOLD Gong et al. (2023)	Isaac Sim / PhysX-based language-grounded manipulation benchmark with 8 continuous-state tasks involving rigid objects, articulated objects, and fluids in realistic 3D scenes	Evaluates whether policies can ground language instructions into continuous object states and execute fine-grained physical interactions with rigid, articulated, and fluid objects.

 doi:10.20944/preprints202606.0425.v1

Table A2. *Cont.*

Evaluation Setting	Evaluation Focus	Benchmark	Key Coverage / Protocol	Summary
Real-to-sim Proxy	Deployment approximation	SimplerEnv X. Li, Hsu, et al. (2024)	Real-robot setups reproduced in simulation with visual matching and variant aggregation	Bridges simulation and hardware by testing whether calibrated simulation predicts real-world policy behavior.
	Deployment approximation	Language Table Lynch et al. (2022)	Real-robot language-conditioned tabletop manipulation data with associated simulated evaluation, visually grounded instructions, object-centric goals, and extended interaction sequences	Bridges real robot interaction and simulation by testing whether policies can ground natural-language commands into sustained tabletop manipulation under a reproducible proxy setting.
	Deployment approximation	RobotArena ∞ Jangir et al. (2025)	Real-to-sim robot policy evaluation benchmark that automatically translates real robot demonstration videos into simulated environments, covering BridgeSim, DROIDSim, and RH20TSim with controlled background, color, and object-pose perturbations	Provides scalable proxy evaluation for real-world-trained VLA policies through simulated rollouts, VLM-based progress scoring, and human preference comparisons under reproducible distribution shifts.
	Deployment approximation	World-Env Xiao et al. (2026)	World-model-based VLA post-training framework with a physically consistent future-observation simulator and a VLM-guided instant reflector for reward and termination signals	Replaces physical interaction with imagined world-model rollouts for safe, low-cost VLA policy optimization under data-scarce manipulation settings.
	Deployment approximation	FurnitureSim Heo et al. (2023)	Simulated version of FurnitureBench built on Isaac Gym / Factory with the same 3D-printable furniture models and robot controller as the real-world setup, camera observations, proprioceptive states, and phase-completion / success-rate evaluation	Provides a fast simulation proxy for debugging and evaluating long-horizon furniture-assembly policies before real-robot deployment, while retaining visual and physical sim-to-real gaps.
	Generalization proxy	REALM L. Geng and Chang (2025)	Real-to-sim validated manipulation benchmark with high-fidelity visuals, aligned robot control, 7 manipulation skills, and 15 perturbation factors.	Uses calibrated simulation to evaluate whether VLA generalization and robustness trends transfer to physical setups.
	Generalization proxy	WorldSimBench Qin et al. (2024)	Dual evaluation benchmark for video generation models as World Simulators, combining HF-Embodied perceptual scoring with video-to-action closed-loop evaluation across Minecraft-style open-ended tasks, CARLA driving, and CALVIN robot manipulation	Tests whether generated videos are visually plausible, instruction-aligned, physically consistent, and actionable enough to proxy embodied world simulation.
	Generalization proxy	RoboWM-Bench F. Jiang et al. (2026)	Manipulation-centric benchmark for video world models using simulation-native and real-to-sim scenes, video-to-action conversion, step-level checks, and task-level success metrics	Tests whether generated manipulation videos are physically executable and action-consistent under embodied robot-control constraints.
Real Robot	Centralized online platform	RoboChallenge Yakefu et al. (2025b)	Online-hosted real-robot evaluation system with Table30, 30 fixed-table tasks across four hosted robot types (UR5, Franka Panda, Cobot Magic ALOHA, and ARX-5), a 10-machine fleet, RealSense RGB-D sensing, and asynchronous observation/action APIs	Organizer hosts the robots and evaluation process, while participants keep models local and control the machines through remote robot APIs; Table30 is the initial multi-embodiment real-robot suite.
	Centralized online platform	ManipArena Y. Sun et al. (2026b)	Server-side real-robot evaluation platform where participants expose a single HTTP policy endpoint; 20 reasoning-oriented tasks on a unified X2Robot bimanual embodiment, with controlled OOD trials over appearance, layout, material, object, and semantic shifts	Organizer-run platform that fixes hardware, control infrastructure, and scoring while evaluating one unified policy across execution reasoning, semantic reasoning, and mobile manipulation tasks.
	Centralized online platform	AutoEval Z. Zhou et al. (2025)	Queue-based policy submission to real WidowX/BridgeData-style cells with automatic success detection, reset, safety checks, episode logs, and videos	Hosted real-robot evaluation platform that automates repeated trials and scoring, but currently covers a narrower set of public scenes than broader challenge suites.
	Federated, hardware-standardized	RoboArena Atreya et al. (2025)	Distributed DROID-based real-world evaluation network across 7 universities, using remote policy inference, local evaluator clients, double-blind A/B comparisons, evaluator-chosen tasks and scenes, and 612 pairwise comparisons across 7 generalist policies	Standardizes the DROID hardware platform and pairwise comparison protocol rather than fixed tasks or scenes, trading strict task comparability for scalable preference-based ranking across diverse real-world settings.
	Federated, kit-standardized	ManipulationNet Y. Chen et al. (2026)	Distributed real-world manipulation benchmarking infrastructure with standardized task/object kits, participant-owned robots, mnet-client / mnet-server submission, real-time task instructions, video/log integrity checks, and Physical Skills plus Embodied Reasoning tracks	Standardizes task setups and verification rather than robot embodiment, enabling heterogeneous robots to run shared tasks locally while scoring remains centrally audited and comparable.
	Federated, hardware-standardized	FurnitureBench Heo et al. (2023)	Reproducible real-world furniture assembly benchmark with a standardized Franka Panda setup, three RealSense D435 RGB-D cameras, 8 3D-printable furniture models, task initialization tools, 219.6h / 5,100 demonstrations, and a matched simulator, FurnitureSim	Tests long-horizon real-world furniture assembly under reproducible hardware, object, initialization, and software settings.
	Federated, protocol-standardized	GM-100 Z. Wang, Liu, et al. (2026)	Real-world task-list benchmark with 100 detail-oriented manipulation tasks covering diverse and long-tail human–object interactions; 13K+ teleoperated trajectories across Agilex Cobot Magic and Dobot Xtrainer, baseline results, evaluation videos, and SR / PSR / action-prediction-error metrics	Standardizes task design, completion criteria, reference data, and evaluation metrics rather than robot hardware or physical kits, enabling adopter-run real-world evaluation of long-tail manipulation behaviors.
	Paper-specific physical setups	Custom real-robot setups	Lab-specific robots, objects, scenes, tasks, and evaluation protocols	No standardization at any layer; high deployment validity but minimal cross-paper comparability.

Deployment approximation versus generalization proxy.

Within the *Real-to-sim Proxy* setting, *Deployment approximation* uses simulation or world models to approximate real-world policy deployment, policy ranking, checkpoint selection, or pre-deployment debugging. Its central question is whether simulated or world-model rollouts predict how a policy would behave on physical hardware. By contrast, *Generalization proxy* uses calibrated simulation, reconstructed environments, or world-model-based execution to diagnose whether robustness, out-of-distribution generalization, physical consistency, or action-level executability trends transfer to physical settings. Its central question is whether the proxy exposes reliable generalization or embodiment failures, rather than whether a specific policy score directly matches real-world deployment.

 doi:10.20944/preprints202606.0425.v1

Mixed benchmark suites.

For mixed benchmark suites, we use the following rule. If the benchmark primarily measures broad task, object, scene, prompt, dataset, or protocol coverage, it is categorized as *Skill Breadth*, with any long-horizon or physically difficult subset described in the coverage/protocol column. If the benchmark primarily targets chained subtask execution, long-term planning, temporal state tracking, re-planning, or stage-wise error accumulation, it is categorized as *Long Horizon*. If the benchmark primarily targets difficult physical interaction or control, it is categorized as *Manipulation Complexity*. For example, broad suites such as RLBench, RoboCasa, and RoboCasa365 are assigned to Skill Breadth when their dominant contribution is task, scene, or dataset coverage, whereas LoHoSet and VLABench are assigned to Long Horizon because their primary protocols target multi-step reasoning, subtask decomposition, or extended task execution. Memory-intensive benchmarks such as MIKASA-Robo and RMBench are also assigned to Long Horizon through the temporal-state-tracking criterion rather than to Manipulation Complexity.

Appendix D. Minimum Reporting Protocol

The checklists below operationalize the *claim-bounded reporting*: the conclusions a paper may draw should be bounded by the metrics it actually reports. Each checklist specifies the minimum set of quantities needed to make a given class of evaluation interpretable and comparable. Items marked **[R]** are *required* for any paper that includes the corresponding evaluation type; items marked **[C]** are *conditional*, required only when the paper makes a specific capability claim that the item is designed to substantiate.

Appendix D.1. Simulation Evaluation

Core reporting (R).

1. Task success rate per benchmark suite, reported at the individual-task-family level rather than only as an aggregate average.
2. Number of evaluation episodes per task.
3. Number of random seeds, with per-seed results or standard deviation.
4. 95% confidence intervals or standard errors for all reported success rates.
5. Inference latency (mean and p_{95}, measured in milliseconds on specified hardware).
6. Effective control frequency (Hz) during closed-loop rollout.
7. Total wall-clock rollout time per episode.
8. Benchmark version, evaluation script commit hash or release tag, and any non-default configuration.

Long-horizon tasks (C).

Required when the paper claims multi-step or compositional capability.

1. Subtask completion rate (fraction of individual subtasks completed, independent of full-sequence success).
2. Average completed sequence length (e.g., CALVIN-style metric).
3. Progress score or partial-credit metric, if supported by the benchmark.
4. Distribution of failure positions along the subtask sequence (e.g., histogram showing at which step failures concentrate).

Robustness (C).

Required when the paper claims robustness or generalization under distribution shift.

1. Absolute success rate under each perturbation type (object, layout, camera, lighting, background, initial state, sensor noise).
2. Relative success-rate drop ($\Delta\%$) from the unperturbed baseline for each perturbation type.

 doi:10.20944/preprints202606.0425.v1

3. Whether perturbations are applied individually or in combination, and if combined, which combination protocol is used.

Language grounding (C).

Required when the paper claims language understanding beyond template matching.

1. Success rate under the original (template) instructions.
2. Success rate under paraphrased instructions.
3. Success rate under semantically equivalent but syntactically different instructions.
4. Success rate under ambiguous or underspecified instructions.
5. Success rate under distractor or irrelevant instructions.

Appendix D.2. Real-Robot Evaluation

Setup description (R).

1. Robot embodiment (manufacturer, model, DoF, end-effector type).
2. Camera configuration (number, mounting positions, resolution, whether wrist cameras are used).
3. Action space definition (joint position, joint velocity, end-effector pose, or hybrid).
4. Control frequency (Hz) on the deployed hardware.
5. Object set (number of objects, material categories, whether objects are seen or unseen during training).
6. Scene layout (workspace dimensions, fixture positions, lighting conditions).
7. Reset procedure (manual, semi-automatic, or automatic; degree of randomization per trial).
8. Success criteria (explicit definition, including tolerance thresholds).

Statistical reporting (R).

1. Number of trials per task (minimum recommended: 50 for primary claims; see discussion below).
2. Success rate with 95% binomial confidence intervals.
3. Task completion time (mean and standard deviation, in seconds).
4. Inference latency on the deployed hardware (mean and p_{95}, in milliseconds).

Operational metrics (R).

1. Human intervention rate (fraction of trials requiring manual correction during execution).
2. Collision rate (fraction of trials in which unintended contact occurs).
3. Constraint-violation rate (fraction of trials violating predefined workspace, force, or safety constraints).
4. Unsafe-contact rate (fraction of trials producing contacts that could damage the robot, object, or environment).
5. Reset failure rate (fraction of trials in which the reset procedure itself fails or introduces bias).

Failure analysis (R).

Papers should report a failure-mode breakdown that, at minimum, distinguishes the following categories:

1. *Perception errors*: the policy misidentifies or fails to locate the target object.
2. *Language-understanding errors*: the executed behavior does not match the instruction semantics.
3. *Planning errors*: the policy selects an incorrect or suboptimal action sequence.
4. *Grasping errors*: the end-effector fails to acquire or maintain a stable grasp.
5. *Control errors*: the low-level controller produces jerky, oscillatory, or divergent motions.
6. *Recovery failures*: the policy fails to recover after an intermediate error.
7. *Hardware or calibration failures*: the trial fails due to mechanical, sensor, or calibration issues unrelated to the policy.

 doi:10.20944/preprints202606.0425.v1

We recommend a minimum of 50 trials per task for primary performance claims, with the acknowledgment that hardware constraints may limit this in practice. When fewer than 50 trials are reported, papers should explicitly state the resulting confidence-interval width and refrain from drawing fine-grained comparative conclusions.

Appendix D.3. Claim-Metric Alignment Guide

Table A3 maps common capability claims to the minimum benchmark evidence and metric set required to substantiate them. A paper whose reported metrics do not cover the right-hand column for a given claim should not draw conclusions about that capability. The table is intended as a practical reference: authors can identify which claims their paper makes, then verify that the corresponding metrics are reported.

Table A3. Claim-metric alignment guide. Each row specifies the minimum benchmark and metric requirements for a common VLA capability claim. Papers that do not report the quantities in the right column should refrain from drawing conclusions about the corresponding capability.

Capability Claim	Minimum Benchmark Evidence	Required Metrics
Task competence	At least one standardized simulation suite (e.g., LIBERO, ManiSkill, RLBench); results on saturated suites should be supplemented with a more diagnostic benchmark.	Success rate with CI, trials/seeds, task-level breakdown.
Robustness / generalization	A robustness-oriented suite (e.g., LIBERO-Plus, The Colosseum, REALM perturbation splits) or controlled perturbation experiments on a standard benchmark.	Per-perturbation success rate, relative drop from baseline, perturbation protocol description.
Language understanding	Language-perturbation benchmark (e.g., LIBERO-Para, LangGap, RAMA-Bench) or controlled instruction-variation experiments.	Success under original, paraphrased, semantically equivalent, ambiguous, and distractor instructions.
Long-horizon reasoning	A sequential or compositional benchmark (e.g., CALVIN, LIBERO-Long, BEHAVIOR-1K).	Subtask completion rate, average sequence length, progress score, failure-position distribution.
Inference efficiency	Any benchmark, but latency must be measured on specified hardware under closed-loop control.	Inference latency (mean, p_{95}), control frequency, FLOPs or model size, rollout time.
Safety awareness	A safety-oriented benchmark (e.g., SAFE) or experiments with explicit constraint-violation conditions.	Collision rate, constraint-violation rate, unsafe-contact rate, failure-detection accuracy.
Real-world readiness	Either (a) a standardized real-robot benchmark, or (b) a real-to-sim validated proxy with stated calibration scope, or (c) custom real-robot evaluation with $\geq$50 trials per task and full operational reporting.	All items in Appendix D.2: success rate with CI, completion time, latency, intervention rate, collision rate, failure breakdown.
Cross-embodiment transfer	Evaluation on $\geq$2 distinct embodiments (different morphologies, not just different instances of the same robot).	Per-embodiment success rate with CI, description of embodiment differences, whether the policy is shared or separately adapted.

Appendix E. Evidence Hierarchy for Simulation Benchmarks

Section 4.4 argues that the evidential status of a simulation benchmark depends not on whether it runs in simulation, but on what proxy-validity relationship has been established between the simulated evaluation and the target deployment context. Table A4 formalizes this argument as a five-level evidence hierarchy.

 doi:10.20944/preprints202606.0425.v1

Each level specifies (1) the type of benchmark, (2) the class of claims it can support, (3) the boundary conditions beyond which those claims do not extend, and (4) representative examples from the current benchmark landscape. The hierarchy is ordered by the strength of inference from benchmark result to real-world deployment behavior, from weakest (Level 1) to strongest (Level 5).

Two principles govern its use. First, *claims should not exceed the evidence level*. A high score on a Level 1 benchmark supports a competence claim under that benchmark's conditions, but not a robustness claim (Level 2), a deployment-approximation claim (Level 3), or a general real-world superiority claim (Levels 4 and 5). Second, *higher levels do not invalidate lower levels*. A Level 3 proxy benchmark does not replace the need for Level 1 competence checks; rather, each level answers a different question, and a thorough evaluation draws on multiple levels.

Table A4. Evidence hierarchy for VLA benchmarks. Levels are ordered by the strength of inference from benchmark result to real-world deployment capability. Each level specifies the claims it supports, the conditions that bound those claims, and representative benchmarks. Claims drawn from a benchmark should not exceed the evidence level that benchmark provides.

Evidence Level	Benchmark Type	Supported Claims	Boundary Conditions	Representative Examples
I	Task-centric simulation suite	The policy achieves a specified level of manipulation competence under the benchmark's task definitions, physics, and visual conditions.	Claims are bounded by the benchmark's task distribution, object set, physics engine, and rendering pipeline. No inference to other task families, visual domains, or physical hardware is warranted without additional evidence.	LIBERO (Spatial/Object/Goal/Long), Meta-World (ML10/ML50), RLBench, ManiSkill.
II	Robustness-oriented simulation suite	The policy maintains competence under controlled distribution shifts in object properties, layout, camera pose, lighting, background, initial state, or sensor noise.	Claims are bounded by the perturbation types and magnitudes implemented in the suite. Robustness to one perturbation type does not imply robustness to others. The degree to which simulated perturbations match real-world variability is typically unvalidated.	LIBERO-Plus, LIBERO-Pro, The Colosseum, LIBERO-Para, LangGap, RAMA-Bench, RoboCasa (randomized settings), RoboTwin 2.0 (randomized settings).
III	Real-to-sim validated proxy	Simulated policy rankings, sensitivity patterns, or performance trends approximate those observed on the specific physical setup against which the proxy was calibrated.	Proxy validity is local: it holds only for the embodiment, control interface, visual distribution, task family, and metric against which calibration was performed. It does not transfer automatically to other robots, cameras, action spaces, or task types not included in the validation.	SimplerEnv (Google Robot, Bridge V2 configurations), REALM (7 skills, 15 perturbation factors, paired real-world comparison).
IV	Standardized real-robot benchmark	The policy achieves a specified level of physical manipulation capability under conditions that permit cross-lab or cross-paper comparison, because the benchmark standardizes hardware, objects, tasks, scoring, or execution.	Claims are bounded by the benchmark's embodiment, object kit, task suite, and scoring protocol. Generalization beyond the standardized conditions requires separate evidence. Cross-benchmark comparison (e.g., RoboChallenge vs. ManipArena) remains nontrivial due to differences in embodiment and task design.	RoboChallenge (centralized hardware), ManipArena (hosted platform, single-policy rule), RoboArena (federated, DROID-standardized), ManipulationNet (standardized object kits), GM-100 (protocol-standardized, community leaderboard).
V	Custom real-robot evaluation with full reporting	The policy performs successfully on a specific physical setup under conditions described in sufficient detail for independent assessment.	Claims are deployment demonstrations for the reported setup. Cross-paper comparison is limited unless another paper replicates the same robot, objects, scenes, and protocol. Statistical strength depends on trial count and reporting completeness (see Appendix D.2).	Paper-specific setups (e.g., $\pi 0$ multi-embodiment demos, HybridVLA single/dual-arm evaluation, DreamZero seen/unseen tasks).

 doi:10.20944/preprints202606.0425.v1

Using the hierarchy in practice.

A paper that reports strong results on LIBERO (Level 1) and LIBERO-Plus (Level 2) can claim task competence and robustness to the tested perturbation types, but should not infer deployment readiness. Adding REALM evaluation (Level 3) strengthens the case for real-world transfer, but only within REALM's calibrated scope (the specific embodiment, task set, and visual conditions that REALM validates). A standardized real-robot result (Level 4) provides the strongest cross-lab evidence, while a custom real-robot demonstration (Level 5) provides deployment validity for the specific setup but limited comparability.

The recommended practice is to combine multiple levels. For example, a paper claiming deployment-ready manipulation might report Level 1 for baseline competence, Level 2 for perturbation resilience, Level 3 or Level 4 for real-world evidence, and the corresponding metrics from Appendix D.3 at each level. No single level is sufficient for broad deployment claims; conversely, no level is dispensable, as each answers a distinct evaluative question.

Relationship to the four conditions.

The hierarchy connects directly to the four evaluative conditions introduced in Section 1. Level 1 benchmarks primarily serve the *discrimination* condition (provided they are not saturated). Level 2 benchmarks serve the *metric coverage* condition by testing dimensions beyond aggregate success. Levels 3 and 4 serve the *inference validity* condition by providing calibrated or standardized physical evidence. All levels require proper statistical reporting to satisfy the *comparability* condition.

 doi:10.20944/preprints202606.0425.v1

Appendix F. Simulation Performance Across Models

Table A5. Performance snapshot of VLA models on major simulation benchmarks. The BEHAVIOR-1K column is omitted, and rows without any remaining benchmark result are removed. For compact presentation, cells report the extracted numerical score for each benchmark; when multiple scores are reported in the source table, their arithmetic mean is used. Dashes indicate unreported results. Part 1 of 3.

model name	year	LIBERO	CALVIN	Simpler Env	RoboCasa	RoboTwin	Meta-World	RLBench	ManiSkill
TTF-VLA C. Liu et al. (2026)	2025	72.4	-	34.9	-	-	-	-	-
EO-1 Qu et al. (2025)	2025	98.2	-	72.7/76.5/63.0	-	-	-	-	-
FPC-VLA Y. Yang, Duan, et al. (2025)	2025	86.9	-	64.6/78	-	-	-	-	-
F1 Lv et al. (2025)	2025	77.5	-	-	-	-	-	-	-
LLaDA-VLA Y. Wen et al. (2025)	2025	-	4.01	55.1	-	-	-	-	-
SpecPrune-VLA H. Wang et al. (2025)	2025	96.1	-	-	-	-	-	-	-
SimpleVLA-RL H. Li, Zuo, et al. (2025)	2025	99.1	-	-	-	69.6	-	-	-
VLA-ADAPTER Y. Wang, Ding, et al. (2025)	2025	95.4	-	-	-	-	-	-	-
VLA-RFT H. Li, Ding, et al. (2025)	2025	91.1	-	-	-	-	-	-	-
X-VLA J. Zheng, Li, Wang, et al. (2025)	2025	98.1	4.43	80.4/75.7/95.8	-	53.37	-	-	-
InternVLA-M1 X. Chen, Chen, et al. (2025)	2025	-	-	80.7/76	-	-	-	-	-
Spatial Forcin F. Li et al. (2025)	2025	98.5	-	-	-	-	-	-	-
NORA-1.5 Hung et al. (2025)	2025	95	-	82.8/71.9	-	-	-	-	-
DUALVL Z. Fang et al. (2025)	2025	-	-	61.0	-	-	-	-	-
VideoVLA Y. Shen et al. (2026)	2025	-	-	63.0	-	-	-	-	-
InternVLA-A1 J. Cai et al. (2026)	2025	-	-	-	-	89.5	-	-	-
X. Li et al. (2026)	2024	-	4.25	-	-	-	-	-	-
BAKU Haldar et al. (2024)	2024	88	-	-	-	-	79	-	-
MDT Reuss et al. (2024)	2024	-	4.52	-	-	-	-	-	-
TinyVLA J. Wen, Zhu, Li, Zhu, et al. (2025)	2024	-	-	-	-	-	31.6	-	-
RoboTron-Mani F. Yan et al. (2024)	2024	91.7	3.51	60	47.4	-	80.1	-	-
Pang et al. (2024)	2024	63.15	-	-	-	-	-	-	-
RoboUniView F. Liu et al. (2024)	2024	-	3.855/3.647	-	-	-	-	-	-
GR-2 Cheang et al. (2024)	2024	-	4.64	-	-	-	-	-	-
TraceVLA R. Zheng et al. (2024)	2024	74.8	-	-	-	-	-	-	-
Mot Y. Chen, Ge, et al. (2025)	2024	-	-	78.3	-	-	-	-	-
ATM C. Wen et al. (2023)	2024	63	-	-	-	-	-	-	-
Astra Y. Ma et al. (2025)	2024	-	3.29	-	-	-	-	-	84.43
GR-1 H. Wu et al. (2024)	2023	-	4.21/3.06	-	-	-	-	-	-
SuSIE Black, Nakamoto, et al. (2024)	2023	-	2.69	-	-	-	-	-	-
FALCON Z. Zhang et al. (2025)	2025	-	4.40/4.53	56.3/62.9	-	-	-	-	-
A1 K. Zhang et al. (2026)	2026	96.6	-	-	-	-	-	-	-
AnoleVLA Takagi et al. (2026)	2026	-	-	-	-	-	67.85	-	-
Xiaomi-Robotics-0 R. Cai et al. (2026)	2026	98.7	4.80/4.75	85.5/74.4/79.2	-	-	-	-	-
EveryDayVLA Chopra et al. (2025)	2025	91.4	-	-	-	-	-	-	-
VLM with EmbodiedMidtrain Du et al. (2026)	2026	54.2	3.714	56.3	-	-	-	-	-
FASTER Y. Lu et al. (2026)	2026	96.5	4.292	-	-	-	-	-	-
SimVLA Y. Luo, Chen, Liang, et al. (2026)	2026	98.6	-	95.8/76.1	-	-	-	-	-
AtomVLA X. Sun et al. (2026)	2026	97	-	-	-	-	-	-	-
AIM L. Fan et al. (2026)	2026	-	-	-	-	93.05	-	-	-
W. Yu et al. (2026)	2026	-	-	78.6/54.5/61.6	-	-	-	-	-
GeoPredict J. Qian et al. (2025)	2026	96.5	-	-	52.4	-	-	-	-
ActiveVLA Z. Liu et al. (2026)	2026	-	-	-	-	-	-	91.8	-
Robotics et al. (2026)	2026	-	-	-	-	90.5	-	-	-
Driess et al. (2026)	2025	94.64	-	-	-	-	-	-	-
GeoManip W. Tang et al. (2025)	2025	-	-	-	-	-	71.1	-	-
DexVLA J. Wen, Zhu, Li, Tang, et al. (2025)	2025	97.3	-	-	-	-	-	-	-
HAMSTER Y. Li et al. (2025)	2025	-	-	-	-	-	-	46	-
NORA Hung et al. (2025)	2025	87.9	-	-	-	-	-	-	-
UniVLABu et al. (2025)	2025	95.2	3.80	42.7	-	-	-	-	-
A0 R. Xu et al. (2025)	2025	-	-	-	-	-	-	-	5.5(MAE)
Evo-0 T. Lin et al. (2025)	2025	-	-	-	-	-	-	56	-
VLA-Reasoner W. Guo et al. (2025)	2025	81.0	-	37.3/41.8	-	-	-	-	-
MLA Z. Liu, Liu, et al. (2025)	2025	-	-	-	-	-	-	81	-
OpenVLA M. J. Kim et al. (2024)	2024	76.5	-	-	-	-	-	-	-
CapVector W. Song, Zhao, et al. (2026)	2026	91.7	-	-	-	-	-	-	-
D. Kim, Jang, et al. (2026)	2026	97.8	-	81.5/77.4/71.9	70.6/32.1	-	-	-	-
J. Guo et al. (2026)	2026	-	-	-	79.2	90.25	-	-	-
W. Zhang et al. (2026)	2026	-	4.51	-	-	-	-	-	-
Being-H0.5 H. Luo et al. (2026)	2026	98.9	-	-	53.9	-	-	-	-
HiVLA T. Yang et al. (2026)	2026	-	-	-	-	83.3	-	-	-
MINT R. Huang et al. (2026)	2026	98.3	4.58	-	-	-	67.2	-	-

 doi:10.20944/preprints202606.0425.v1

Table A6. Performance snapshot of VLA models on major simulation benchmarks, continued. Part 2 of 3.

model name	year	LIBERO	CALVIN	Simpler Env	RoboCasa	RoboTwin	Meta-World	RLBench	ManiSkill
WristWorld Z. Qian et al. (2025)	2025	-	3.81	-	-	-	-	-	-
StaMo M. Liu et al. (2025)	2025	87.6	-	-	-	-	-	-	-
Cao et al. (2025)	2025	-	-	-	-	86	-	-	-
J. Zhao et al. (2025)	2025	94.5	-	-	-	-	-	-	-
FLOWER Reuss, Zhou, et al. (2025)	2025	96.9	4.53/4.67/4.35	45/31.9	-	-	-	-	-
Hsieh et al. (2025)	2025	-	-	-	-	-	-	42.67	-
AimBot Dai, Lee, Zhang, et al. (2025)	2025	95.9	-	-	-	-	-	-	-
ToBo T. Kim et al. (2025)	2025	-	-	-	-	-	-	43.5	-
DreamVLA J. Ye et al. (2025)	2025	92.6	4.44	-	-	-	-	-	-
EfficientVLA Y. Yang, Wang, et al. (2026)	2025	-	-	76.1	-	-	-	-	-
RIPT-VLA S. Tan et al. (2025)	2025	97.5	-	-	-	-	-	-	-
H3DP Y. Lu et al. (2025)	2025	-	-	-	-	57.4	93.97	-	59.3/65.3
TesserAct Zhen et al. (2025)	2025	-	-	-	-	-	-	64	-
FLAME Betran et al. (2025)	2025	-	-	-	-	-	-	86	-
S. Li et al. (2025)	2025	90	-	-	-	-	-	-	-
VideoWorld Ren et al. (2025)	2025	-	-	-	-	-	-	67.1/62.5	-
MoDE Reuss, Pari, et al. (2025)	2024	92	3.39/4.30	26.30	-	-	-	-	-
Code-as-Monitor E. Zhou et al. (2025)	2024	-	-	-	-	-	-	97.08	-
G. Jiang et al. (2025)	2024	-	-	-	75	-	-	-	-
SGRv2 T. Zhang et al. (2024)	2024	-	-	-	-	-	-	53.2/63.3	55.8
RoboDual Bu et al. (2024)	2024	-	3.66	-	-	-	-	-	-
RACER Dai, Lee, Fazeli, and Chai (2025)	2024	-	-	-	-	-	-	70.2	-
Shang et al. (2024)	2024	-	-	-	-	-	79.79	-	-
RVT-2 Goyal et al. (2024)	2024	-	-	-	-	-	-	81.4	-
Ag2Manip P. Li et al. (2024)	2024	-	-	-	-	-	-	-	65.3
Hejna et al. (2024)	2023	-	-	-	-	-	80	-	-
RVT Goyal et al. (2023)	2023	-	-	-	-	-	-	62.9	-
VP-VLA Z. Wang, Chen, et al. (2026)	2026	-	-	58.3	53.8	-	-	-	-
RoboAlign D. Kim, Park, et al. (2026)	2026	86.8	2.57	-	-	-	-	-	-
Y. Luo, Chen, Wu, et al. (2026)	2026	-	-	-	-	-	-	83	-
GigaWorld-Policy A. Ye et al. (2026)	2026	-	-	-	-	85	-	-	-
Z. Sun and Song (2026)	2026	97.8	-	-	55.6	-	-	-	-
MolmoB0T Deshpande et al. (2026)	2026	36.6	-	11.5	-	-	-	-	-
J. Hu et al. (2026)	2026	88.07	-	-	-	-	-	-	-
OmniStream Y. Yan et al. (2026)	2026	-	3.885	45.8	-	-	-	-	-
SeedPolicy Gui et al. (2026)	2026	-	-	-	-	23.52	-	-	-
pi-StepNFT S. Wang et al. (2026)	2026	94.0	-	-	-	-	-	-	59.5
VLANeXt X.-M. Wu et al. (2026)	2026	97.4	-	-	-	-	-	-	-
GeneralVLA G. Ma et al. (2026)	2026	-	-	-	-	-	-	91	-
VLA-JEPA J. Sun et al. (2026)	2026	97.2	-	65.2/57.3	-	-	-	-	-
SCALE Choi et al. (2026)	2026	81.5	-	49.0	-	-	-	-	-
BagelVLA Y. Hu et al. (2026)	2026	-	4.405	-	-	48.07	-	-	-
Tur et al. (2026)	2026	93	3.39	-	-	-	-	-	-
Green-VLA Apanasevich et al. (2026)	2026	-	-	71.8/94.6	-	-	-	-	-
Shallow-pi Jeon et al. (2026)	2026	96	-	-	-	-	-	-	-
CLARE Römer et al. (2026)	2026	70.91	-	-	-	-	-	-	-
Ranasinghe et al. (2026)	2026	-	4.48	-	-	68.6	-	-	-
ACoT-VLA L. Zhong et al. (2026)	2026	98.5	-	-	-	-	-	-	-
Fast-ThinkAct C.-P. Huang, Man, et al. (2026)	2026	-	-	-	-	32.8	-	-	-
Robo-Dopamine H. Tan et al. (2025)	2025	81	-	-	-	-	-	-	-
Dream-VLA J. Ye et al. (2025)	2025	59.0	-	-	-	-	-	-	-
PhysBrain X. Lin et al. (2025)	2025	-	-	67.4/65.9	55.25/49.75	-	-	-	-
VLSA S. Hu et al. (2025)	2025	72.99	-	-	-	-	-	-	-
Feng et al. (2025)	2025	-	-	-	45.8	-	-	-	-
HiF-VLA M. Lin et al. (2025)	2025	95.4	-	-	-	-	-	-	-
S. Yang et al. (2025)	2025	96.6	-	55.5	-	41.3/64	-	-	-
SwiftVLA Ni et al. (2025)	2025	94.7/95.1	-	-	-	-	-	-	-
VLA-4D H. Zhou et al. (2025)	2025	97.85	-	-	-	-	-	-	-
Mantis Y. Yang, Li, et al. (2025)	2025	96.7	-	-	-	-	-	-	-
SRPO Fei, Wang, Ji, et al. (2025)	2025	82.1	-	-	-	-	-	-	-
RoboOmni S. Wang, Fu, et al. (2025)	2025	81.1	-	-	-	-	-	-	-
Park et al. (2025)	2025	-	-	-	39.3	-	-	-	-
H. Zhao, Zhang, et al. (2025)	2025	80.1	-	-	-	-	-	-	-
W. Shen et al. (2025)	2025	96	-	-	-	49.7	-	-	-
D. Niu, Sharma, Shi, et al. (2025)	2025	-	-	-	-	-	-	-	67
Vlaser G. Yang et al. (2025)	2025	-	-	65.1	-	-	-	-	-
HiRT J. Zhang et al. (2024)	2024	-	-	-	-	-	76.4	-	-
MMaDA-VLA Y. Liu et al. (2026)	2026	98.0	4.78	-	-	-	-	-	-

 doi:10.20944/preprints202606.0425.v1

Table A7. Performance snapshot of VLA models on major simulation benchmarks, continued. Part 3 of 3.

model name	year	LIBERO	CALVIN	Simpler Env	RoboCasa	RoboTwin	Meta-World	RLBench	ManiSkill
ABot-M0 Y. Yang, Zeng, et al. (2026)	2026	98.6	-	-	-	85.57	-	-	-
LingBot-VA L. Li et al. (2026)	2026	98.5	-	-	-	92.24	-	-	-
VLM Backbone J. Zhang et al. (2026)	2026	-	4.06	-	-	-	-	-	-
Cosmos Policy M. J. Kim et al. (2026)	2026	98.5	-	-	67.1	-	-	-	-
TwinBrainVLA B. Yu et al. (2026)	2026	97.6	-	64.5	54.6	-	-	-	-
LangForce Lian et al. (2026)	2026	98.4	-	66.5	52.6	-	-	-	-
CycleVLA C. Ma et al. (2026)	2026	95.3	-	-	-	-	-	-	-
DiT4DiT T. Ma et al. (2026)	2026	98.6	-	-	50.8	-	-	-	-
ThinkAct C.-P. Huang, Wu, et al. (2026)	2025	84.4	-	71.5 / 65.1 / 43.8	-	-	-	-	-
VLA-0 Goyal et al. (2025)	2025	95.7	-	-	-	-	-	-	-
GR00T N1 Bjorck et al. (2025)	2025	-	-	-	32.1	-	-	-	-
OpenVLA-OFT M. J. Kim et al. (2025)	2025	97.1	-	-	-	-	-	-	-
DREAMGEN Jang et al. (2025)	2025	-	-	-	20.6	-	-	-	-
SAM2Act H. Fang et al. (2025)	2025	-	-	-	-	-	-	86.8	-
VLA-Cache S. Xu et al. (2025)	2025	74.7 / 97.4	-	74.4 / 62.3	-	-	-	-	-
GRAPE Z. Zhang et al. (2024)	2025	57.2	-	43.1	-	-	-	-	-
RoboBERT S. Wang, Liu, et al. (2025)	2025	-	4.52/3.79	-	-	-	-	-	-
DiT Policy Z. Hou et al. (2024)	2025	82.4	3.61	-	-	-	-	-	65.8
OTTER H. Huang et al. (2025)	2025	84	-	-	-	-	-	-	-
Jülg et al. (2025)	2025	-	-	-	-	-	-	-	80
HybridVLA J. Liu et al. (2025)	2025	-	-	-	-	-	-	74.0	-
MoLe-VLA R. Zhang et al. (2026)	2025	-	-	-	-	-	-	60.8	-
Y. Yang, Cai, et al. (2025)	2025	-	-	-	-	-	-	80.8	-
ViSA-Flow C. Chen et al. (2025)	2025	-	2.96	-	-	-	-	-	-
GR-MG P. Li et al. (2025)	2025	-	4.04	-	-	-	-	-	-
OpenHelix Cui et al. (2025)	2025	-	3.53	-	-	-	-	-	-
CLIP-RT+ Kang et al. (2024)	2025	93.1	-	-	-	-	-	-	-
W. Chen et al. (2025)	2025	90.8	-	-	-	-	-	-	-
InSpire J. Zhang et al. (2025)	2025	89.5	-	-	-	-	-	-	-
Hume H. Song et al. (2025)	2025	98.6	-	-	-	-	-	-	-
DiT Policy Z. Hou et al. (2024)	2025	82.4	3.61	-	-	-	-	-	-
SmolVLA Shukor et al. (2025)	2025	87.3	-	-	-	-	57.3	-	-
VLA-RL G. Lu et al. (2025)	2025	81.0	-	-	-	-	-	-	-
Fast ECoT Duan et al. (2025)	2025	80.0	-	-	-	-	-	-	-
BridgeVLA P. Li et al. (2026)	2025	-	-	-	-	-	-	88.2	-
TUDP Y. Niu et al. (2025)	2025	-	-	-	-	-	-	83.2	-
Fast-in-Slow H. Chen, Liu, et al. (2025)	2025	-	-	-	-	-	-	69	-
CEED-VLA W. Song, Chen, Ding, et al. (2025)	2025	-	3.67	-	-	-	-	-	-
MinD Chi et al. (2025)	2025	-	-	-	-	-	-	63	-
WorldVLA Cen et al. (2025)	2025	78.1	-	-	-	-	-	-	-
TriVLA Z. Liu, Gu, et al. (2025)	2025	87.0	4.37/3.46	-	-	-	71.4	-	-
VQ-VLA Y. Wang, Zhu, et al. (2025)	2025	86.61	-	-	-	-	-	-	-
AC-DiT S. Chen et al. (2026)	2025	-	-	-	-	90.1	-	-	55.6
VOTE J. Lin et al. (2025)	2025	98.0	-	58.3	-	-	-	-	-
LTM-H Ranasinghe et al. (2025)	2025	-	3.81	-	-	-	57.7	-	-
villa-X X. Chen, Wei, et al. (2025)	2025	-	-	58.5/40.8	-	-	-	-	-
ReconVLA W. Song, Zhou, et al. (2026)	2025	-	3.95/4.23	-	-	-	-	-	-
GeoVLA L. Sun et al. (2025)	2025	97.7	-	-	-	-	-	-	77
T. Zhang, Duan, et al. (2026)	2025	-	-	-	-	-	-	-	53.2
Embodied-R1 Yuan et al. (2025)	2025	79.4	-	56.2	-	-	-	-	-
FlowVLA Z. Zhong et al. (2025)	2025	88.1	-	74.0	-	-	-	-	-
CogVLA W. Li et al. (2026)	2025	97.4	-	-	-	-	-	-	-
Discrete Diffusion VLA Liang et al. (2025)	2025	96.3	-	-	-	-	-	-	-
pi0.5 with MoH Jing et al. (2025)	2025	-	-	-	-	2.0	-	-	-
3D-VLA Zhen et al. (2024)	2024	-	44.7	-	-	-	-	-	-
DeeR-VLA Yue et al. (2024)	2024	-	2.92/4.13/2.9	-	-	-	-	-	-
RoboFlaming X. Li, Liu, et al. (2024)	2023	-	4.09	-	-	-	-	-	-
AVDC Ko et al. (2024)	2023	-	-	-	-	-	43.1	-	-
FAST Pertsch et al. (2025)	2025	82.5	-	-	-	-	-	-	-
Interleave-VLA C. Fan et al. (2025)	2025	-	-	71	-	-	-	-	-
R. Wang et al. (2026)	2026	-	-	-	-	88.55	-	-	-
FRAPPE H. Zhao et al. (2026)	2026	-	-	-	-	41.5	-	-	-
EVOLVE-VLA Bai et al. (2025)	2025	95.8	-	-	-	-	-	-	-
VLASH J. Tang et al. (2025b)	2025	97.2	-	-	-	-	-	-	-
HyCodePolicy Y. Liu et al. (2025)	2025	-	-	-	-	71.3	-	-	-
UWM C. Zhu et al. (2025)	2025	79.0	-	-	-	-	-	-	-
SlotMIM X. Wen et al. (2025)	2025	-	-	-	-	-	78.4	-	-
JoyAI-RA 0.1 T. Zhang, Yuan, et al. (2026)	2026	-	-	-	63.2	89.88	-	-	-

 doi:10.20944/preprints202606.0425.v1

Appendix H. Real-Robot Evaluation Across Models

Table A8. Real-robot evaluation summary across models. Scene level denotes the physical scale of the real-world evaluation: T = tabletop-level, R = room-level, H = house-level, M = mixed-level, and U = unclear. Evaluation type denotes the source and strength of real-robot evidence: PB = public benchmark/protocol, CT = custom tasks, DO = demonstration only, and UC = unclear. Because real-robot evaluations often use different platforms, scenes, tasks, and protocols, the entries are summarized as deployment evidence rather than directly comparable leaderboard scores.

model name	year	robot / platform	scene level	eval. type	task description
VLA-0 Goyal et al. (2025)	2025	SO-100	T	CT	SO-100 tabletop manipulation; 4 tasks; block reorientation, apple pushing, banana/cupcake pick-place
GraspVLA Deng et al. (2025)	2025	Franka Panda	T	CT	open-vocabulary tabletop grasping; 2 object groups × 5 test settings; 300 real-world trials
L. Li et al. (2026)	2026	AgileX / Agibot G1 / Galaxea R1Pro	T	PB	GM-100 tabletop manipulation; 100 tasks × 3 platforms; 15 trials/task-platform
Cosmos Policy M. J. Kim et al. (2026)	2026	ALOHA	T	CT	bimanual tabletop manipulation; 4 tasks / 101 trials; plate placing, shirt folding, candy collection, ziploc insertion
LingBot-VLA W. Wu et al. (2026)	2026	AgileX / Agibot G1 / Galaxea R1Pro	T	PB	GM-100 tabletop manipulation; 100 tasks × 3 platforms; 15 trials/task-platform
TwinBrainVLA B. Yu et al. (2026)	2026	Franka Research 3	T	CT	Franka tabletop pick-and-place; 3 protocols × 30 trials; ID/OOD/Pick-All block-to-box tasks
DiT4DiT T. Ma et al. (2026)	2026	Unitree G1	T	CT	Unitree G1 tabletop household manipulation; 7 tasks / 20 rollouts each; pick-place, flower arrangement, cup stacking, plate insertion, box packing, spoon moving, drawer interaction
GE-Act Liao et al. (2025)	2025	AgiBot G1	T	CT	AgiBot G1 tabletop placement; red cylinder to paper cup; 305 demonstrations
Barreiros et al. (2026)	2025	Dual Franka FR3 arms w/ parallel grippers	T	CT	bimanual tabletop manipulation; seen short tasks: kiwi-to-center, mug-rightside-up, coaster-to-mug; unseen long-horizon tasks: clear kitchen counter, bike rotor install, cut apple slices, clean litter box, set breakfast table; 50 real-world rollouts/task/policy/condition
GR00T-N1 Bjorck et al. (2025)	2025	Fourier GR-1 humanoid robot	T	CT	GR-1 tabletop bimanual manipulation; pre-training evals: left-to-right handover and novel-object-to-container; post-training evals: object-to-container pick-place, articulated storage insertion/closing, machinery packing, mesh-cup pouring, cylinder handover, and two-robot coordination
OpenVLA-OFT M. J. Kim et al. (2025)	2025	ALOHA; dual ViperX 300 S arms	T	CT	bimanual ALOHA tabletop manipulation; fold shorts, fold T-shirt, scoop specified ingredient into bowl with spoon, open pot and put specified item into pot; 20/30/45/300 demos and 10/10/12/24 eval trials
DREAMGEN Jang et al. (2025)	2025	Fourier GR1 / Franka Emika / SO-100	M: T/R	CT	video-world-model neural trajectories for real robots; GR1 hammering, wiping, folding, stacking and novel behavior/environment tasks including microwave/MacBook/lunchbox, pouring/watering, vacuuming/ironing, spoon/whisk/pot tasks; Franka milk-to-bowl, cube stacking, M&M scooping; SO-100 strawberry pick-place and tic-tac-toe
ENERVERSE S. Huang et al. (2026)	2025	Commercial robot arm (unspecified)	T	CT	foam-worktable block insertion to instructed row/column compartments; 9 target compartments × 5 trials; additional transparent plastic-object sorting and fruit sorting tasks
FLaRe J. Hu et al. (2025)	2024	Stretch RE-1 mobile manipulator	H	CT	real-world multi-room apartment mobile manipulation/navigation; ObjectNav, Fetch, PickUp, and RoomVisit across bedroom, office, bathroom, kitchen, storage, corridor, and living-room areas; 46 real-world tasks with direct sim-to-real deployment
UniAct J. Zheng, Li, Liu, et al. (2025)	2025	WidowX / AIRBOT / Bimanual AIRBOT	T	CT	WidowX tabletop eval: 19 tasks / 190 rollouts covering cup-to-sink, carrot-to-plate with height change, flip pot upright, doll-to-drying-rack, green/red-cup-to-plate; AIRBOT adaptation: cube-on-cuboid stacking under relative/absolute EEF and joint control; Bimanual AIRBOT: sweep plate, fold towel, cup on plate, transport pen
SAM2Act H. Fang et al. (2025)	2025	Franka Emika Panda w/ Robotiq 2F-85 gripper	T	CT	real-world Franka tabletop tasks with D455 depth sensing; turn on lamp, press red-then-blue button sequence, stack blue block on orange block, and memory task: press button that previously had block in front twice; 10–15 demos/task and 10 ID + 10 OOD trials/task
iRe-VLA Y. Guo et al. (2025)	2025	Franka Panda arm w/ wrist camera	T	CT	Panda tabletop manipulation; 10 expert tasks from pick/grasp, place, button-press, cable-route, drawer-open; online RL on irregular-object picking such as eggplant/carrot; 8 unseen-object pick tasks; 2,000 expert demos + 20 successful RL trajectories buffer
VLA-Cache S. Xu et al. (2025)	2025	Kinova Jaco2 arm	T	CT	Jaco2 tabletop manipulation with front-facing camera; PickPot, PlaceCube, PutSausage, and WipeTable; 100 real-world trials total with randomized robot/object initial states
GRAPE Z. Zhang et al. (2024)	2024	Franka arm w/ Robotiq gripper	T	CT	Franka tabletop manipulation; 30 tasks / 300 executions across in-domain, visual, subject, action, semantic, and language-grounding generalization; pick-place to bowls/plates, push-button, knock-down, towel folding, carrot stacking, and obstacle-aware safety variants
TRA Myers et al. (2026)	2025	WidowX250 7-DoF manipulator	T	CT	BridgeData-style tabletop compositional manipulation; 4 scenes / 13 tasks: drawer open/mushroom-in/close, spoon-on-plate/towel and cloth folding/sweeping, food-item placement to plate/bowl, dependency tasks incl. mushroom-out-of-drawer, bell-pepper move + towel sweep, corn-to-plate then sushi-to-pot; 5–10 trials/task
RoboBERT S. Wang, Liu, et al. (2025)	2025	REALMAN RM65B 6-DoF arm	T	CT	tabletop manipulation with static and gripper cameras; individual tasks: stacking cubes, pen-to-holder, cabinet-door opening; sequential CALVIN-style tasks: object transfer plus drawer opening/closing; 25–30 teleoperated trajectories/task
DiT Policy Z. Hou et al. (2024)	2025	Franka arm w/ single third-person RGB camera	T	CT	real Franka tabletop setup with black background and 9-grid object placement; 10-shot tasks: PickBlock, PickBanana, bowl-to-box ball pouring, open box then pick block, banana-to-pen-container insertion, bowl stacking, cup-to-bowl coffee-bean pouring; in-domain/OOD pick-place tasks: green block, kiwifruit, banana, tiny green block, pink block
SOFAR Qi et al. (2026)	2025	Franka Panda w/ gripper / UR arm w/ LeapHand / Flexiv arm w/ suction tool	T	CT	real-world language-grounded 6-DoF rearrangement and articulated-object manipulation; 60 tasks over 100+ objects: 3×3 bottle arrangement, knife-handle grasp and bread cutting, highest-box placement, flashlight rotation, tissue pulling, camera aiming, microwave/cabinet opening, bottle/glass righting, toy/lego spatial placement
ARM4R D. Niu, Sharma, Xue, et al. (2025)	2025	Kinova Gen3 7-DoF w/ Robotiq 2F-85; Franka Emika Panda	T	CT	Kinova tabletop manipulation: pick yellow/cyan/green cube, destack yellow/cyan cubes, stack yellow-on-cyan/cyan-on-yellow, pick spiderman/penguin/pig toys to target, play basketball, push red button, push red-then-blue buttons; additional Franka cross-robot pick/stack/destack transfer
OTTER H. Huang et al. (2025)	2025	Franka robot / DROID-style setup	T	CT	tabletop manipulation across 4 primitives: pick-place objects into bowls/pots, poke blocks/objects/bowls, pour cup contents into bowls, and open/close drawer with objects inside/on top; 19 in-distribution + 15 unseen tasks, 10 trials/task
PointVLA C. Li et al. (2026)	2025	Bimanual UR5e / Bimanual AgileX	T	CT	bimanual tabletop manipulation with L515 top camera and dual wrist D435i cameras; AgileX few-shot tasks: ChargePhone, WipePlate, PlaceBread, TransportFruit with 20 demos/task; UR5e long-horizon moving-conveyor packing: pick two laundry-detergent bottles from conveyor, place into box, straighten box, close two hinges; real-vs-photo detergent discrimination and table-height adaptation
VidBot H. Chen, Sun, et al. (2025)	2025	Hello Robot Stretch 3 / Boston Dynamics Spot	R	CT	zero-shot household mobile manipulation across 3 human-suited environments; 55 real-world trials covering cabinet opening/closing, drawer pushing/pulling, tissue taking, paperball dropping, toy pickup, microwave/fridge/cupboard interactions
iManip Z. Zheng et al. (2025)	2025	Franka Panda w/ RealSense D455 RGB-D camera	T	CT	skill-incremental tabletop manipulation under B1-4N1 setup; 5 real-world skills learned sequentially: slide toy to instructed color target, open top/middle/bottom drawer, pick specified object and place at target location, pour water from specified-color cup into specified-color mug, screw cap onto bottle/jar; 20 demonstrations/training step and 10 test runs/learned skill
HybridVLA J. Liu et al. (2025)	2025	Franka Research 3 / AgileX dual-arm robot	T	CT	Franka FR3 two-view tabletop tasks: colored-block pick-place, charger unplugging, bottle-to-cup pouring, blackboard wiping, drawer open/place/close; AgileX three-view dual-arm tasks: bimanual pick-place, lift ball to container, place two bottles at rack, one-arm-hold/one-arm-wipe blackboard, fold shorts; 100 demos/task and 20 rollouts/task
MoRE H. Zhao, Song, et al. (2025)	2025	Unitree Go2 quadruped w/ front RealSense D435 camera	R	DO	indoor quadruped VLA deployment; real-world fine-tuning on colored-ball navigation and letter-distinguishing tasks, then demoed go-to-computer navigation, crawl-under-bar posture locomotion, and unload-yellow-ball-into-pink-block whole-body manipulation
MoManipVLA Z. Wu et al. (2025)	2025	Hexman Echo Plus base + RM65 arm	R	CT	room-scale mobile manipulation in a large workspace; stack block, open drawer, put object in bowl; base–arm bi-level trajectory optimization from fixed-base VLA waypoints; 50 fine-tuning samples and 10 trials/task
MoLe-VLA R. Zhang et al. (2026)	2025	Franka Research 3 w/ 3D-printed UMI gripper and GoPro 9 wrist camera	T	CT	FR3 tabletop/workspace manipulation; detach charger, pull drawer, pour water; 50 demonstrations/task and 10 real-world trials/task
Y. Yang, Cai, et al. (2025)	2025	Dual Franka Research 3 robots w/ RealSense D435i cameras; Robotiq-2F-85 grippers for shelf task	T	CT	bimanual desktop/shelf manipulation; carry tray from lower platform to small table, handover plate and insert into rack, sponge-wipe plate with curved motion, scan bottle from shelf and place into box; 10 settings × 3 trials/task
ViSA-Flow C. Chen et al. (2025)	2025	7-DoF Franka Emika Panda arm w/ eye-in-hand and eye-to-hand RGB cameras	T	CT	Franka tabletop manipulation; single-stage MoveContainer and PickEggplant with 46/54 demos, plus long-horizon MoveContainer → PickEggplant sequence; 12 initial positions per policy
GR-MG P. Li et al. (2025)	2024	Kinova Gen-3 arm w/ Robotiq 2F-85 gripper, static camera + end-effector camera	T	CT	real-robot tabletop manipulation; 58 evaluated tasks across Simple, Unseen Distractors, Unseen Instructions, Unseen Backgrounds, and Unseen Objects; 37 training tasks incl. 23 pick-place tasks and 14 non-pick-place tasks such as pouring, flipping, rotating
CrayonRobo X. Li et al. (2025)	2025	Franka Emika arm w/ default parallel gripper and RealSense 415 camera	T	CT	prompt-driven real-world tabletop manipulation without sim-to-real finetuning; open trashcan, open microwave, lift lid, wipe table, heat toaster; manual/automatic crayon prompts, 5 trials per object shape with varied camera view and initial pose
CLIP-RT Kang et al. (2024)	2025	UR5 6-DoF arm w/ two-finger gripper	T	CT	18 tabletop manipulation tasks collected via language-based teleoperation; common tasks: point, pull, place, pick, push, flip, knock-over, slide, move; novel tasks: draw line, pour dog food, open cabinet, play toy car, close laptop, erase whiteboard, open trashcan, stamp, hide Pooh; 10 episodes/task with STA augmentation and 10 trials/task
W. Chen et al. (2025)	2025	BridgeData WidowX station	T	CT	real-world BridgeData V2 tabletop manipulation; in-distribution, motion, spatial, and semantic generalization tasks: mushroom/corn/eggplant to pot/bowl, spoon/carrot to towel/plate, green-object challenge, elevated carrot/mushroom placement, banana/tomato to left/right bowl, toy to left/right pot, relative object placement, unseen objects incl. watermelon, toothbrush, screw, ketchup, wrench, mallet; 444 real-robot trials
VTLA C. Zhang et al. (2025)	2025	UR3 arm w/ Robotiq 2F-85 gripper, wrist RealSense D405, dual GelStereo 2.0 tactile sensors	T	CT	contact-rich tabletop peg-in-hole insertion trained only on simulation; real-world 1:1 setup with randomized peg-hole misalignment; square peg clearances 1.6/1.0/0.6 mm and 0.6 mm square/triangle/hexagon/pentagon/round pegs; 20 trials/task
InSpire J. Zhang et al. (2025)	2025	AGILEX PiPER 6-DoF arm	T	CT	tabletop manipulation with intrinsic spatial-reasoning prompts; 10 seen + 5 unseen real-world tasks targeting spatial relations, novel object/scene interaction, and unseen instructions; 10 demos/seen task and 10 trials/task
Hume H. Song et al. (2025)	2025	WidowX 250 / Franka Emika Panda / AgiBot G-1 humanoid	M: T/R	CT	WidowX zero-shot tabletop tasks with 10 tasks/15 objects: close microwave, lift red pepper, put green cup on cloth/stove, purple cup on plate, eggplant/carrot placement; Franka 7-DoF tasks: banana-to-basket, pot-to-cutting-board, teapot-handle press, plush-toy and cube-to-car placement; AgiBot G-1 whole-upper-body tasks: pour water, fold shorts, pass water to human, restock hanging-basket snack shelf
TrackVLA S. Wang, Zhang, et al. (2025)	2025	Unitree GO2 quadruped w/ Intel RealSense D455 RGB camera	R	CT	real-world embodied visual tracking; quadruped follows specified humans/targets under clutter, low illumination, pursuit-evasion, multi-person recognition, occlusion, and high-speed target motion; Easy/Medium/Hard tracking scenarios with 10 trials each

 doi:10.20944/preprints202606.0425.v1

References

Apanasevich, I., Artemyev, M., Babakyan, R., Fedotova, P., Grankin, D., Kupryashin, E., Misailidi, A., Nerus, D., Nutalapati, A., Sidorov, G., et al. (2026). Green-VLA: Staged Vision-Language-Action Model for Generalist Robots. *arXiv preprint arXiv:2602.00919*. Available online: https://arxiv.org/abs/2602.00919 (accessed on).

Atreya, P., Pertsch, K., Lee, T., Kim, M. J., Jain, A., Kuramshin, A., Eppner, C., Neary, C., Hu, E., Ramos, F., et al. (2025). Roboarena: Distributed real-world evaluation of generalist robot policies. *arXiv preprint arXiv:2506.18123*.

Bai, Z., Gao, C., & Shou, M. Z. (2025). EVOLVE-VLA: Test-Time Training from Environment Feedback for Vision-Language-Action Models. *arXiv preprint arXiv:2512.14666*. Available online: https://arxiv.org/abs/2512.14666 (accessed on).

Bao, C., Xu, H., Qin, Y., & Wang, X. (2023). DexArt: Benchmarking Generalizable Dexterous Manipulation with Articulated Objects. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21190-21200.

Barreiros, J., Beaulieu, A., Bhat, A., Cory, R., Cousineau, E., Dai, H., Fang, C.-H., Hashimoto, K., Irshad, M. Z., Itkina, M., et al. (2026). A careful examination of large behavior models for multitask dexterous manipulation. *Science Robotics*, *11*, eaea6201. Available online: https://arxiv.org/pdf/2507.05331 (accessed on).

Betran, S. B., Longhini, A., Vasco, M., Zhang, Y., & Kragic, D. (2025). FLAME: A Federated Learning Benchmark for Robotic Manipulation. *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2494-2500.

Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al. (2025). Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*. Available online: https://arxiv.org/abs/2503.14734 (accessed on).

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al. (2024). π0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*.

Black, K., Nakamoto, M., Atreya, P., Walke, H., Finn, C., Kumar, A., & Levine, S. (2024). Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In *International conference on learning representations* (Vol. 2024, pp. 33431–33452). Available online: https://arxiv.org/abs/2310.10639 (accessed on).

Bu, Q., Li, H., Chen, L., Cai, J., Zeng, J., Cui, H., Yao, M., & Qiao, Y. (2024). Towards synergistic, generalized, and efficient dual-system for robotic manipulation. *arXiv preprint arXiv:2410.08001*. Available online: https://arxiv.org/abs/2410.08001 (accessed on).

Bu, Q., Yang, Y., Cai, J., Gao, S., Ren, G., Yao, M., Luo, P., & Li, H. (2025). UniVLA: Learning to Act Anywhere with Task-centric Latent Actions. *ArXiv*, *abs/2505.06111*.

Cai, J., Cai, Z., Cao, J., Chen, Y., He, Z., Jiang, L., Li, H., Li, H., Li, Y., Liu, Y., et al. (2026). InternVLA-A1: Unifying Understanding, Generation and Action for Robotic Manipulation. *arXiv preprint arXiv:2601.02456*. Available online: https://arxiv.org/pdf/2601.02456 (accessed on).

Cai, R., Guo, J., He, X., Jin, P., Li, J., Lin, B., Liu, F., Liu, W., Ma, F., Ma, K., Qiu, F., Qu, H., Su, Y., Sun, Q., Wang, D., Wang, D., Wang, Y., Wu, R., Xiang, D., … Zhou, Q. (2026). *Xiaomi-robotics-0: An open-sourced vision-language-action model with real-time execution.* Available online: https://arxiv.org/abs/2602.12684 (accessed on).

Cao, J., Huang, Y., Guo, H., Zhang, R., Nan, M., Mai, W., Wang, J., Cheng, H., Sun, J., Han, G., et al. (2025). Compose Your Policies! Improving Diffusion-based or Flow-based Robot Policies via Test-time Distribution-level Composition. *arXiv preprint arXiv:2510.01068*. Available online: https://arxiv.org/abs/2510.01068 (accessed on).

Cen, J., Yu, C., Yuan, H., Jiang, Y., Huang, S., Guo, J., Li, X., Song, Y., Luo, H., Wang, F., et al. (2025). Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*.

Cheang, C.-L., Chen, G., Jing, Y., Kong, T., Li, H., Li, Y., Liu, Y., Wu, H., Xu, J., Yang, Y., et al. (2024). Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*. Available online: https://arxiv.org/pdf/2410.06158 (accessed on).

Chen, C., Yang, Q., Xu, X., Fazeli, N., & Andersson, O. (2025). Visa-flow: Accelerating robot skill learning via large-scale video semantic action flow. *arXiv preprint arXiv:2505.01288*. Available online: https://arxiv.org/abs/2505.01288 (accessed on).

 doi:10.20944/preprints202606.0425.v1

Chen, H., Liu, J., Gu, C., Liu, Z., Zhang, R., Li, X., He, X., Guo, Y., Fu, C.-W., Zhang, S., et al. (2025). Fast-in-slow: A dual-system foundation model unifying fast manipulation within slow reasoning. *arXiv preprint arXiv:2506.01953*. Available online: https://arxiv.org/abs/2506.01953 (accessed on).

Chen, H., Sun, B., Zhang, A., Pollefeys, M., & Leutenegger, S. (2025). Vidbot: Learning generalizable 3d actions from in-the-wild 2d human videos for zero-shot robotic manipulation. In *Proceedings of the computer vision and pattern recognition conference* (pp. 27661–27672). Available online: https://arxiv.org/abs/2503.07135 (accessed on).

Chen, Q., Shi, C., Chen, Q., Wu, Y., Gao, Z., Zhang, X., Gao, R., Wu, K., & Jia, Y. (2025). *Long-horizon visual imitation learning via plan and code reflection*. Available online: https://arxiv.org/abs/2509.05368 (accessed on).

Chen, S., Liu, J., Qian, S., Jiang, H., Liu, Z., Gu, C., Li, X., Hou, C., Wang, P., Wang, Z., et al. (2026). Ac-dit: Adaptive coordination diffusion transformer for mobile manipulation. *Advances in Neural Information Processing Systems*, *38*, 64008–64036. Available online: https://arxiv.org/abs/2507.01961 (accessed on).

Chen, T., Chen, Z., Chen, B., Cai, Z., Liu, Y., Li, Z., Liang, Q., Lin, X., Ge, Y., Gu, Z., et al. (2025). Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*.

Chen, T., Wang, Y., Li, M., Qin, Y., Shi, H., Li, Z., Hu, Y., Zhang, Y., Wang, K., Chen, Y., Wang, H., Xu, R., Wu, R., Mu, Y., Yang, Y., Dong, H., & Luo, P. (2026). *Rmbench: Memory-dependent robotic manipulation benchmark with insights into policy design*. Available online: https://arxiv.org/abs/2603.01229 (accessed on).

Chen, W., Belkhale, S., Mirchandani, S., Mees, O., Driess, D., Pertsch, K., & Levine, S. (2025). Training strategies for efficient embodied reasoning. *arXiv preprint arXiv:2505.08243*. Available online: https://arxiv.org/pdf/2505.08243 (accessed on).

Chen, X., Chen, Y., Fu, Y., Gao, N., Jia, J., Jin, W., Li, H., Mu, Y., Pang, J., Qiao, Y., et al. (2025). Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy. *arXiv preprint arXiv:2510.13778*. Available online: https://arxiv.org/abs/2510.13778 (accessed on).

Chen, X., Wei, H., Zhang, P., Zhang, C., Wang, K., Guo, Y., Yang, R., Wang, Y., Xiao, X., Zhao, L., et al. (2025). Villa-x: enhancing latent action modeling in vision-language-action models. *arXiv preprint arXiv:2507.23682*. Available online: https://arxiv.org/abs/2507.23682 (accessed on).

Chen, Y., Chen, Z., Chan, N. T., Chen, J., Yin, J., Shi, J., Gao, Y., Li, Y.-L., & Huo, J. (2025). RoboHiMan: A hierarchical evaluation paradigm for compositional generalization in long-horizon manipulation. *arXiv preprint arXiv:2510.13149*.

Chen, Y., Ge, Y., Tang, W., Li, Y., Ge, Y., Ding, M., Shan, Y., & Liu, X. (2025). Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 19752–19763). Available online: https://arxiv.org/abs/2412.04445 (accessed on).

Chen, Y., Kimble, K., Adelson, E. H., Asfour, T., Chanrungmaneekul, P., Chitta, S., Chitambar, Y., Chen, Z., Goldberg, K., Kragic, D., et al. (2026). ManipulationNet: An Infrastructure for Benchmarking Real-World Robot Manipulation with Physical Skill Challenges and Embodied Multimodal Reasoning. *arXiv preprint arXiv:2603.04363*.

Cherepanov, E., Kachaev, N., Kovalev, A. K., & Panov, A. I. (2026). *Memory, benchmark & robots: A benchmark for solving complex tasks with reinforcement learning*. Available online: https://arxiv.org/abs/2502.10550 (accessed on).

Chi, X., Ge, K., Liu, J., Zhou, S., Jia, P., He, Z., Liu, Y., Li, T., Han, L., Han, S., et al. (2025). MinD: Learning A Dual-System World Model for Real-Time Planning and Implicit Risk Analysis. *arXiv preprint arXiv:2506.18897*. Available online: https://arxiv.org/abs/2506.18897 (accessed on).

Choi, H., Ahn, D., Lee, Y., Kang, T., Cho, S., & Choi, J. (2026). SCALE: Self-uncertainty Conditioned Adaptive Looking and Execution for Vision-Language-Action Models. *arXiv preprint arXiv:2602.04208*. Available online: https://arxiv.org/abs/2602.04208 (accessed on).

Chopra, S., McMoil, A., Carnovale, B., Sokolson, E., Kubendran, R., & Dickerson, S. (2025). EveryDayVLA: A Vision-Language-Action Model for Affordable Robotic Manipulation. *arXiv preprint arXiv:2511.05397*. Available online: https://arxiv.org/abs/2511.05397 (accessed on).

Cui, C., Ding, P., Song, W., Bai, S., Tong, X., Ge, Z., Suo, R., Zhou, W., Liu, Y., Jia, B., et al. (2025). Openhelix: A short survey, empirical analysis, and open-source dual-system vla model for robotic manipulation. *arXiv preprint arXiv:2505.03912*.

 doi:10.20944/preprints202606.0425.v1

Dai, Y., Lee, J., Fazeli, N., & Chai, J. (2025). Racer: Rich language-guided failure recovery policies for imitation learning. In *2025 ieee international conference on robotics and automation (icra)* (pp. 15657–15664). Available online: https://arxiv.org/abs/2409.14674 (accessed on).

Dai, Y., Lee, J., Zhang, Y., Ma, Z., Yang, J., Zadeh, A., Li, C., Fazeli, N., & Chai, J. (2025). Aimbot: A simple auxiliary visual cue to enhance spatial awareness of visuomotor policies. *arXiv preprint arXiv:2508.08113*. Available online: https://arxiv.org/abs/2508.08113 (accessed on).

Deng, S., Yan, M., Wei, S., Ma, H., Yang, Y., Chen, J., Zhang, Z., Yang, T., Zhang, X., Zhang, W., et al. (2025). Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*. Available online: https://arxiv.org/pdf/2505.03233 (accessed on).

Deshpande, A., Guru, M., Hendrix, R., Jauhri, S., Fang, H., Pumacay, W., Kim, Y., Pfeifer, Q., Lee, Y.-C., Wolters, P., Rayyan, O., Zhang, M., Farley, K., Han, W., VanderBilt, E., Fox, D., Farhadi, A., Chalvatzaki, G., Shah, D., & Krishna, R. (2026). MolmoB0T: Large-Scale Simulation Enables Zero-Shot Manipulation.

Driess, D., Springenberg, J., Ichter, B., Yu, L., Li-Bell, A., Pertsch, K., Ren, A., Walke, H., Vuong, Q., Shi, L. X., et al. (2026). Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *Advances in Neural Information Processing Systems*, *38*, 102867–102888.

Du, Y., Guo, Z., Ye, X., Ren, L., & Xiong, C. (2026). EmbodiedMidtrain: Bridging the Gap between Vision-Language Models and Vision-Language-Action Models via Mid-training. *arXiv preprint arXiv:2604.20012*. Available online: https://arxiv.org/abs/2604.20012 (accessed on).

Duan, Z., Zhang, Y., Geng, S., Liu, G., Boedecker, J., & Lu, C. X. (2025). Fast ecot: Efficient embodied chain-of-thought via thoughts reuse. *arXiv preprint arXiv:2506.07639*. Available online: https://arxiv.org/abs/2506.07639 (accessed on).

Fan, C., Jia, X., Sun, Y., Wang, Y., Wei, J., Gong, Z., Zhao, X., Tomizuka, M., Yang, X., Yan, J., et al. (2025). Interleave-vla: Enhancing robot manipulation with interleaved image-text instructions. *arXiv preprint arXiv:2505.02152*. Available online: https://arxiv.org/abs/2505.02152 (accessed on).

Fan, L., Xu, Z., Cao, C., Zhang, W., Yuan, M., & Chen, J. (2026). AIM: Intent-Aware Unified world action Modeling with Spatial Value Maps. *arXiv preprint arXiv:2604.11135*. Available online: https://arxiv.org/abs/2604.11135 (accessed on).

Fang, H., Grotz, M., Pumacay, W., Wang, Y. R., Fox, D., Krishna, R., & Duan, J. (2025). Sam2act: Integrating visual foundation model with a memory architecture for robotic manipulation. *arXiv preprint arXiv:2501.18564*.

Fang, I., Zhang, J., Tong, S., & Feng, C. (2025). *From intention to execution: Probing the generalization boundaries of vision-language-action models*. Available online: https://arxiv.org/abs/2506.09930 (accessed on).

Fang, Z., Liu, Z., Liu, J., Chen, H., Zeng, Y., Huang, S., Chen, Z., Chen, L., Zhang, S., & Zhao, F. (2025). Dualvla: Building a generalizable embodied agent via partial decoupling of reasoning and action. *arXiv preprint arXiv:2511.22134*. Available online: https://arxiv.org/pdf/2511.22134 (accessed on).

Fei, S., Wang, S., Ji, L., Li, A., Zhang, S., Liu, L., Hou, J., Gong, J., Zhao, X., & Qiu, X. (2025). SRPO: Self-Referential Policy Optimization for Vision-Language-Action Models. *arXiv preprint arXiv:2511.15605*. Available online: https://arxiv.org/abs/2511.15605 (accessed on).

Fei, S., Wang, S., Shi, J., Dai, Z. G., Cai, J., Qian, P., Ji, L., He, X., Zhang, S., Fei, Z., Fu, J., Gong, J., & Qiu, X. (2025). LIBERO-Plus: In-depth Robustness Analysis of Vision-Language-Action Models. *ArXiv*, *abs/2510.13626*.

Feng, Y., Zhang, W., Wang, Y., Luo, H., Yuan, H., Zheng, S., & Lu, Z. (2025). Spatial-Aware VLA Pretraining through Visual-Physical Alignment from Human Videos. *arXiv preprint arXiv:2512.13080*. Available online: https://arxiv.org/abs/2512.13080 (accessed on).

Geng, H., Wang, F., Wei, S., Li, Y., Wang, B., An, B., Cheng, C. T., Lou, H., Li, P., Wang, Y.-J., et al. (2025). Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning. *arXiv preprint arXiv:2504.18904*.

Geng, L., & Chang, E. Y. (2025). *Realm-bench: A benchmark for evaluating multi-agent systems on real-world, dynamic planning and scheduling tasks*. Available online: https://arxiv.org/abs/2502.18836 (accessed on).

Gong, R., Huang, J., Zhao, Y., Geng, H., Gao, X., Wu, Q., Ai, W., Zhou, Z., Terzopoulos, D., Zhu, S.-C., et al. (2023). ARNOLD: A Benchmark for Language-Grounded Task Learning With Continuous States in Realistic 3D Scenes. In *Proceedings of the ieee/cvf international conference on computer vision (iccv)*.

Goyal, A., Blukis, V., Xu, J., Guo, Y., Chao, Y.-W., & Fox, D. (2024). RVT-2: Learning Precise Manipulation from Few Demonstrations. *ArXiv*, *abs/2406.08545*.

Goyal, A., Hadfield, H., Yang, X., Blukis, V., & Ramos, F. (2025). Vla-0: Building state-of-the-art vlas with zero modification. *arXiv preprint arXiv:2510.13054*. Available online: https://arxiv.org/pdf/2510.13054 (accessed on).

 doi:10.20944/preprints202606.0425.v1

Goyal, A., Xu, J., Guo, Y., Blukis, V., Chao, Y.-W., & Fox, D. (2023). Rvt: Robotic view transformer for 3d object manipulation. In *Conference on robot learning* (pp. 694–710). Available online: https://arxiv.org/abs/2306.14896 (accessed on).

Gu, J., Xiang, F., Li, X., Ling, Z., Liu, X., Mu, T., Tang, Y., Tao, S., Wei, X., Yao, Y., Yuan, X., Xie, P., Huang, Z., Chen, R., & Su, H. (2023). ManiSkill2: A Unified Benchmark for Generalizable Manipulation Skills. In *International conference on learning representations*.

Gui, Y., Zhou, Y., Cheng, S., Yuan, X., Fan, H., Cheng, P., & Liu, S. (2026). SeedPolicy: Horizon Scaling via Self-Evolving Diffusion Policy for Robot Manipulation. *arXiv preprint arXiv:2603.05117*.

Guo, J., Li, Q., Li, P., Chen, Z., Sun, N., Su, Y., Wang, H., Zhang, Y., Li, X., & Liu, H. (2026). Unified 4D World Action Modeling from Video Priors with Asynchronous Denoising. *arXiv preprint arXiv:2604.26694*. Available online: https://arxiv.org/abs/2604.26694 (accessed on).

Guo, W., Lu, G., Deng, H., Wu, Z., Tang, Y., & Wang, Z. (2025). Vla-reasoner: Empowering vision-language-action models with reasoning via online monte carlo tree search. *arXiv preprint arXiv:2509.22643*. Available online: https://arxiv.org/pdf/2509.22643 (accessed on).

Guo, Y., Zhang, J., Chen, X., Ji, X., Wang, Y.-J., Hu, Y., & Chen, J. (2025). Improving vision-language-action model with online reinforcement learning. In *2025 ieee international conference on robotics and automation (icra)* (pp. 15665–15672). Available online: https://arxiv.org/abs/2501.16664 (accessed on).

Gupta, A., Kumar, V., Lynch, C., Levine, S., & Hausman, K. (2019). *Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning*. Available online: https://arxiv.org/abs/1910.11956 (accessed on).

Haldar, S., Peng, Z., & Pinto, L. (2024). Baku: An efficient transformer for multi-task policy learning. *Advances in Neural Information Processing Systems*, *37*, 141208–141239. Available online: https://arxiv.org/pdf/2406.07539 (accessed on).

Han, S., Qiu, B., Liao, Y., Huang, S., Gao, C., Yan, S., & Liu, S. (2026). Robocerebra: A large-scale benchmark for long-horizon robotic manipulation evaluation. *Advances in Neural Information Processing Systems*, *38*.

Hejna, J., Rafailov, R., Sikchi, H., Finn, C., Niekum, S., Knox, W. B., & Sadigh, D. (2024). Contrastive Preference Learning: Learning from Human Feedback without Reinforcement Learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, & Y. Sun (Eds.), *International conference on learning representations* (Vol. 2024, pp. 18770–18798). Available online: https://proceedings.iclr.cc/paper_files/paper/2024/file/51f547584cd1fcb87114ea022822a60d-Paper-Conference.pdf (accessed on).

Heo, M., Lee, Y., Lee, D., & Lim, J. J. (2023). *Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation*. Available online: https://arxiv.org/abs/2305.12821 (accessed on).

Hou, Y., & Zhao, L. (2026). LangGap: Diagnosing and Closing the Language Gap in Vision-Language-Action Models. *arXiv preprint arXiv:2603.00592*. Available online: https://arxiv.org/abs/2603.00592 (accessed on).

Hou, Z., Zhang, T., Xiong, Y., Pu, H., Zhao, C., Tong, R., Qiao, Y., Dai, J., & Chen, Y. (2024). Diffusion transformer policy. *arXiv preprint arXiv:2410.15959*. Available online: https://arxiv.org/abs/2410.15959 (accessed on).

Hsieh, W.-H., Hsieh, E., Niu, D., Darrell, T., Herzig, R., & Chan, D. M. (2025). Do what? Teaching vision-language-action models to reject the impossible. *arXiv preprint arXiv:2508.16292*, *2*. Available online: https://arxiv.org/abs/2508.16292 (accessed on).

Hu, J., Hendrix, R., Farhadi, A., Kembhavi, A., Martín-Martín, R., Stone, P., Zeng, K.-H., & Ehsani, K. (2025). Flare: Achieving masterful and adaptive robot policies with large-scale reinforcement learning fine-tuning. In *2025 ieee international conference on robotics and automation (icra)* (pp. 3617–3624). Available online: https://arxiv.org/abs/2409.16578 (accessed on).

Hu, J., Shim, J. J., Tang, C., Sung, Y., Liu, B., Stone, P., & Martín-Martín, R. (2026). Simple Recipe Works: Vision-Language-Action Models are Natural Continual Learners with Reinforcement Learning.

Hu, S., Liu, Z., Liu, S., Cen, J., Meng, Z., & He, X. (2025). VLSA: Vision-Language-Action Models with Plug-and-Play Safety Constraint Layer. *arXiv preprint arXiv:2512.11891*. Available online: https://arxiv.org/abs/2512.11891 (accessed on).

Hu, Y., Zhang, J., Luo, Y., Guo, Y., Chen, X., Sun, X., Feng, K., Lu, Q., Chen, S., Zhang, Y., et al. (2026). Bagelvla: Enhancing long-horizon manipulation via interleaved vision-language-action generation. *arXiv preprint arXiv:2602.09849*.

Huang, C.-P., Man, Y., Yu, Z., Chen, M.-H., Kautz, J., Wang, Y.-C. F., & Yang, F.-E. (2026). Fast-ThinkAct: Efficient Vision-Language-Action Reasoning via Verbalizable Latent Planning. *arXiv preprint arXiv:2601.09708*. Available online: https://arxiv.org/abs/2601.09708 (accessed on).

Huang, C.-P., Wu, Y.-H., Chen, M.-H., Wang, F., & Yang, F.-E. (2026). Thinkact: Vision-language-action reasoning via reinforced visual latent planning. *Advances in Neural Information Processing Systems*, *38*, 82782–82802.

 doi:10.20944/preprints202606.0425.v1

Huang, H., Liu, F., Fu, L., Wu, T., Mukadam, M., Malik, J., Goldberg, K., & Abbeel, P. (2025). OTTER: A Vision-Language-Action Model with Text-Aware Visual Feature Extraction. *ArXiv*, *abs/2503.03734*.

Huang, R., Zeng, C., Tang, W., Cai, J., Lu, C., & Cai, P. (2026). Mimic intent, not just trajectories. *arXiv preprint arXiv:2602.08602*.

Huang, S., Chen, L., Zhou, P., Chen, S., Liao, Y., Jiang, Z., Hu, Y., Gao, P., Li, H., Yao, M., et al. (2026). Enerverse: Envisioning embodied future space for robotics manipulation. *Advances in Neural Information Processing Systems*, *38*, 37693–37720.

Hung, C.-Y., Majumder, N., Deng, H., Renhang, L., Ang, Y., Zadeh, A., Li, C., Herremans, D., Wang, Z., & Poria, S. (2025). Nora-1.5: A vision-language-action model trained using world model-and action-based preference rewards. *arXiv preprint arXiv:2511.14659*. Available online: https://arxiv.org/abs/2511.14659 (accessed on).

Jain, A., Zhang, M., Arora, K., Chen, W., Torne, M., Irshad, M. Z., Zakharov, S., Wang, Y., Levine, S., Finn, C., et al. (2025). Polaris: Scalable real-to-sim evaluations for generalist robot policies. *arXiv preprint arXiv:2512.16881*.

James, S., Ma, Z., Rovick Arrojo, D., & Davison, A. J. (2020). RLBench: The Robot Learning Benchmark & Learning Environment. *IEEE Robotics and Automation Letters*.

Jang, J., Ye, S., Lin, Z., Xiang, J., Bjorck, J., Fang, Y., Hu, F., Huang, S., Kundalia, K., Lin, Y.-C., et al. (2025). Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*. Available online: https://arxiv.org/abs/2505.12705 (accessed on).

Jangir, Y., Zhang, Y., Lo, P.-C., Yamazaki, K., Zhang, C., Tu, K.-H., Ke, T.-W., Ke, L., Bisk, Y., & Fragkiadaki, K. (2025). *Robotarena ∞: Scalable robot benchmarking via real-to-sim translation*. Available online: https://arxiv.org/abs/2510.23571 (accessed on).

Jeon, B., Choi, Y., & Kim, T. (2026). Shallow-π: Knowledge Distillation for Flow-based VLAs. *arXiv preprint arXiv:2601.20262*. Available online: https://arxiv.org/abs/2601.20262 (accessed on).

Ji, Y., Liu, Y., Tan, H., Huang, X., Huang, F., Xu, Y., Chi, C., Zhao, Y., Lyu, H., Co, P., et al. (2026). PRM-as-a-Judge: A Dense Evaluation Paradigm for Fine-Grained Robotic Auditing. *arXiv preprint arXiv:2603.21669*.

Jiang, F., Chen, Y., Xu, K., Liu, Y., Wang, H., Shen, Z., Lu, J., Huang, S., Wang, Y., Xie, C., & Wu, R. (2026). *Robowm-bench: A benchmark for evaluating world models in robotic manipulation*. Available online: https://arxiv.org/abs/2604.19092 (accessed on).

Jiang, G., Sun, Y., Huang, T., Li, H., Liang, Y., & Xu, H. (2025). Robots pre-train robots: Manipulation-centric robotic representation from large-scale robot datasets. In *International conference on learning representations* (Vol. 2025, pp. 81885–81905). Available online: https://arxiv.org/abs/2410.22325 (accessed on).

Jiang, S., Huang, Z., Qian, K., Luo, Z., Zhu, T., Zhong, Y., Tang, Y., Kong, M., Wang, Y., Jiao, S., Ye, H., Sheng, Z., Zhao, X., Wen, T., Fu, Z., Chen, S., Jiang, K., Yang, D., Choi, S., & Sun, L. (2025). *A survey on vision-language-action models for autonomous driving*. Available online: https://arxiv.org/abs/2506.24044 (accessed on).

Jiang, Y., Gupta, A., Zhang, Z., Wang, G., Dou, Y., Chen, Y., Fei-Fei, L., Anandkumar, A., Zhu, Y., & Fan, L. (2023). *Vima: General robot manipulation with multimodal prompts*. Available online: https://arxiv.org/abs/2210.03094 (accessed on).

Jiang, Z., Xie, Y., Lin, K., Xu, Z., Wan, W., Mandlekar, A., Fan, L., & Zhu, Y. (2025). *Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning*. Available online: https://arxiv.org/abs/2410.24185 (accessed on).

Jing, D., Wang, G., Liu, J., Tang, W., Sun, Z., Yao, Y., Wei, Z., Liu, Y., Lu, Z., & Ding, M. (2025). Mixture of Horizons in Action Chunking. *arXiv preprint arXiv:2511.19433*. Available online: https://arxiv.org/pdf/2511.19433 (accessed on).

Jülg, T., Burgard, W., & Walter, F. (2025). Refined policy distillation: From vla generalists to rl experts. In *2025 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 11677–11684). Available online: https://arxiv.org/abs/2503.05833 (accessed on).

Kang, G.-C., Kim, J., Shim, K., Lee, J. K., & Zhang, B.-T. (2024). CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision. *ArXiv*, *abs/2411.00508*.

Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., & Zhu, Y. (2025). Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications. *IEEE Access*, *13*, 162467-162504.

Kim, C., Kim, M., Kang, M., Kim, H., & Jung, D. (2026). LIBERO-Para: A Diagnostic Benchmark and Metrics for Paraphrase Robustness in VLA Models. *arXiv preprint arXiv:2603.28301*. Available online: https://arxiv.org/abs/2603.28301 (accessed on).

 doi:10.20944/preprints202606.0425.v1

Kim, D., Jang, H., Koo, M., Jang, S., Kim, T., Kim, B., Yoon, B., Jang, C., Choi, D., Han, D., et al. (2026). RLDX-1 Technical Report. *arXiv preprint arXiv:2605.03269*. Available online: https://arxiv.org/abs/2605.03269 (accessed on).

Kim, D., Park, S., Song, W., Kim, S., Kim, T., Jang, H., Shin, J., Kim, J., & Seo, Y. (2026). RoboAlign: Learning Test-Time Reasoning for Language-Action Alignment in Vision-Language-Action Models. *arXiv preprint arXiv:2603.21341*. Available online: https://arxiv.org/abs/2603.21341 (accessed on).

Kim, M. J., Finn, C., & Liang, P. (2025). Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success. *ArXiv*, *abs/2502.19645*.

Kim, M. J., Gao, Y., Lin, T.-Y., Lin, Y.-C., Ge, Y., Lam, G., Liang, P., Song, S., Liu, M.-Y., Finn, C., et al. (2026). Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al. (2024). Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*. Available online: https://arxiv.org/pdf/2406.09246 (accessed on).

Kim, T., Han, D., Heo, B., Park, J., & Yun, S. (2025). Token Bottleneck: One Token to Remember Dynamics. *ArXiv*, *abs/2507.06543*. Available online: https://api.semanticscholar.org/CorpusID:280069025 (accessed on).

Ko, P.-C., Mao, J., Du, Y., Sun, S.-H., & Tenenbaum, J. B. (2024). Learning to act from actionless videos through dense correspondences. In *International conference on learning representations* (Vol. 2024, pp. 40938–40958). Available online: https://arxiv.org/abs/2310.08576 (accessed on).

Kumar, V. (2016). *Manipulators and manipulation in high dimensional spaces* (Doctoral dissertation, University of Washington, Seattle). Available online: https://digital.lib.washington.edu/researchworks/handle/1773/38104 (accessed on).

Lan, Z., Jiang, Y., Wang, R., Xie, X., Zhang, R., Zhu, Y., Li, P., Yang, T., Chen, T., Gao, H., Yang, X., Li, X., Zhang, H., Mu, Y., & Luo, P. (2025). *Autobio: A simulation and benchmark for robotic automation in digital biology laboratory.* Available online: https://arxiv.org/abs/2505.14030 (accessed on).

Li, C., Wen, J., Peng, Y., Peng, Y., & Zhu, Y. (2026). Pointvla: Injecting the 3d world into vision-language-action models. *IEEE Robotics and Automation Letters*, *11*, 2506–2513. Available online: https://arxiv.org/abs/2503.07511 (accessed on).

Li, C., Zhang, R., Wong, J., Gokmen, C., Srivastava, S., Martín-Martín, R., Wang, C., Levine, G., Ai, W., Martinez, B., Yin, H., Lingelbach, M., Hwang, M., Hiranaka, A., Garlanka, S., Aydin, A., Lee, S., Sun, J., Anvari, M., … Fei-Fei, L. (2024). *Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation.* Available online: https://arxiv.org/abs/2403.09227 (accessed on).

Li, F., Song, W., Zhao, H., Wang, J., Ding, P., Wang, D., Zeng, L., & Li, H. (2025). Spatial forcing: Implicit spatial representation alignment for vision-language-action model. *arXiv preprint arXiv:2510.12276*. Available online: https://arxiv.org/abs/2510.12276 (accessed on).

Li, H., Ding, P., Suo, R., Wang, Y., Ge, Z., Zang, D., Yu, K., Sun, M., Zhang, H., Wang, D., et al. (2025). Vla-rft: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators. *arXiv preprint arXiv:2510.00406*. Available online: https://arxiv.org/abs/2510.00406 (accessed on).

Li, H., Zuo, Y., Yu, J., Zhang, Y., Yang, Z., Zhang, K., Zhu, X., Zhang, Y., Chen, T., Cui, G., et al. (2025). Simplevla-rl: Scaling vla training via reinforcement learning. *arXiv preprint arXiv:2509.09674*.

Li, L., Zhang, Q., Luo, Y., Yang, S., Wang, R., Han, F., Yu, M., Gao, Z., Xue, N., Zhu, X., et al. (2026). Causal World Modeling for Robot Control. *arXiv preprint arXiv:2601.21998*. Available online: https://arxiv.org/pdf/2601.21998 (accessed on).

Li, P., Chen, Y., Wu, H., Ma, X., Wu, X., Huang, Y., Wang, L., Kong, T., & Tan, T. (2026). Bridgevla: Input-output alignment for efficient 3d manipulation learning with vision-language models. *Advances in Neural Information Processing Systems*, *38*, 63635–63673.

Li, P., Liu, T., Li, Y., Han, M., Geng, H., Wang, S., Zhu, Y., Zhu, S.-C., & Huang, S. (2024). Ag2manip: Learning novel manipulation skills with agent-agnostic visual and action representations. In *2024 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 573–580). Available online: https://arxiv.org/abs/2404.17521 (accessed on).

Li, P., Wu, H., Huang, Y., Cheang, C., Wang, L., & Kong, T. (2025). Gr-mg: Leveraging partially-annotated data via multi-modal goal-conditioned policy. *IEEE Robotics and Automation Letters*, *10*, 1912–1919. Available online: https://arxiv.org/abs/2408.14368 (accessed on).

Li, R., Hu, Z., Qu, W., Zhang, J., Yin, Z., Zhang, S., Huang, X., Wang, H., Wang, T., Pang, J., et al. (2026). Labutopia: High-fidelity simulation and hierarchical benchmark for scientific embodied agents. *Advances in Neural Information Processing Systems*, *38*.

 doi:10.20944/preprints202606.0425.v1

Li, S., Gao, Y., Sadigh, D., & Song, S. (2025). Unified video action model. *arXiv preprint arXiv:2503.00200*. Available online: https://arxiv.org/abs/2503.00200 (accessed on).

Li, W., Zhang, R., Shao, R., He, J., & Nie, L. (2026). CogVLA: Cognition-Aligned Vision-Language-Action Models via Instruction-Driven Routing & Sparsification. *Advances in neural information processing systems*, *38*, 137646–137675.

Li, X., Hsu, K., Gu, J., Pertsch, K., Mees, O., Walke, H. R., Fu, C., Lunawat, I., Sieh, I., Kirmani, S., et al. (2024). Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*.

Li, X., Li, P., Qian, L., Liu, M., Wang, D., Liu, J., Kang, B., Ma, X., Wang, X., Guo, D., Kong, T., Zhang, H., & Liu, H. (2026). *What matters in building vision-language-action models for generalist robots*. Available online: https://arxiv.org/abs/2412.14058 (accessed on).

Li, X., Liu, M., Zhang, H., Yu, C., Xu, J., Wu, H., Cheang, C., Jing, Y., Zhang, W., Liu, H., et al. (2024). Vision-language foundation models as effective robot imitators. In *International conference on learning representations* (Vol. 2024, pp. 26703–26721). Available online: https://arxiv.org/abs/2311.01378 (accessed on).

Li, X., Xu, L., Zhang, M., Liu, J., Shen, Y., Ponomarenko, I., Xu, J., Heng, L., Huang, S., Zhang, S., et al. (2025). Crayonrobo: Object-centric prompt-driven vision-language-action model for robotic manipulation. *arXiv preprint arXiv:2505.02166*. Available online: https://arxiv.org/abs/2505.02166 (accessed on).

Li, Y., Deng, Y., Zhang, J., Jang, J., Memmel, M., Garrett, C., Ramos, F., Fox, D., Li, A., Gupta, A., et al. (2025). Hamster: Hierarchical action models for open-world robot manipulation. In *International conference on learning representations* (Vol. 2025, pp. 24040–24068). Available online: https://arxiv.org/abs/2502.05485 (accessed on).

Lian, S., Yu, B., Lin, X., Yang, L. T., Shen, Z., Wu, C., Miao, Y., Huang, C., & Chen, K. (2026). BayesianVLA: Bayesian Decomposition of Vision Language Action Models via Latent Action Queries. *arXiv preprint arXiv:2601.15197*. Available online: https://arxiv.org/pdf/2601.15197 (accessed on).

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

Liao, Y., Zhou, P., Huang, S., Yang, D., Chen, S., Jiang, Y., Hu, Y., Cai, J., Liu, S., Luo, J., et al. (2025). Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*. Available online: https://arxiv.org/pdf/2508.05635 (accessed on).

Lin, J., Taherin, A., Akbari, A., Akbari, A., Lu, L., Chen, G., Padir, T., Yang, X., Chen, W., Li, Y., et al. (2025). Vote: vision-language-action optimization with trajectory ensemble voting. *arXiv preprint arXiv:2507.05116*. Available online: https://arxiv.org/abs/2507.05116 (accessed on).

Lin, M., Ding, P., Wang, S., Zhuang, Z., Liu, Y., Tong, X., Song, W., Lyu, S., Huang, S., & Wang, D. (2025). HiF-VLA: Hindsight, Insight and Foresight through Motion Representation for Vision-Language-Action Models. *arXiv preprint arXiv:2512.09928*. Available online: https://arxiv.org/abs/2512.09928 (accessed on).

Lin, T., Li, G., Zhong, Y., Zou, Y., Du, Y., Liu, J., Gu, E., & Zhao, B. (2025). Evo-0: Vision-language-action model with implicit spatial understanding. *arXiv preprint arXiv:2507.00416*. Available online: https://arxiv.org/abs/2507.00416 (accessed on).

Lin, X., Lian, S., Yu, B., Yang, R., Shen, Z., Wu, C., Miao, Y., Jin, Y., Shi, Y., He, J., et al. (2025). Physbrain: Human egocentric data as a bridge from vision language models to physical intelligence. *arXiv preprint arXiv:2512.16793*. Available online: https://arxiv.org/abs/2512.16793 (accessed on).

Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., & Stone, P. (2023). LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. *ArXiv*, *abs/2306.03310*. Available online: https://api.semanticscholar.org/CorpusID:259089508 (accessed on).

Liu, C., Zhang, J., Li, C., Zhou, Z., Wu, S., Huang, S., & Duan, H. (2026). Ttf-vla: Temporal token fusion via pixel-attention integration for vision-language-action models. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 40, pp. 18452–18459). Available online: https://arxiv.org/abs/2508.19257 (accessed on).

Liu, F., Yan, F., Zheng, L., Feng, C., Huang, Y., & Ma, L. (2024). Robouniview: Visual-language model with unified view representation for robotic manipulation. *arXiv preprint arXiv:2406.18977*. Available online: https://arxiv.org/pdf/2406.18977 (accessed on).

Liu, J., Chen, H., An, P., Liu, Z., Zhang, R., Gu, C., Li, X., Guo, Z., Chen, S., Liu, M., et al. (2025). Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*.

 doi:10.20944/preprints202606.0425.v1

Liu, M., Shu, J., Chen, H., Li, Z., Zhao, C., Yang, J., Gao, S., Chen, H., & Shen, C. (2025). StaMo: Unsupervised Learning of Generalizable Robot Motion from Compact State Representation. *arXiv preprint arXiv:2510.05057*. Available online: https://arxiv.org/abs/2510.05057 (accessed on).

Liu, S., Li, B., Ma, K., Wu, L., Tan, H., Ouyang, X., Su, H., & Zhu, J. (2026). RDT2: Exploring the Scaling Limit of UMI Data Towards Zero-Shot Cross-Embodiment Generalization. *arXiv preprint arXiv:2602.03310*.

Liu, Y., Ding, P., Jiang, T., Wang, X., Song, W., Lin, M., Zhao, H., Zhang, H., Zhuang, Z., Zhao, W., et al. (2026). MMaDA-VLA: Large Diffusion Vision-Language-Action Model with Unified Multi-Modal Instruction and Generation. *arXiv preprint arXiv:2603.25406*. Available online: https://arxiv.org/abs/2603.25406 (accessed on).

Liu, Y., Liang, Z., Chen, Z., Chen, T., Hu, M., Dong, W., Xu, C., Han, Z., Qin, Y., & Mu, Y. (2025). Hycodepolicy: Hybrid language controllers for multimodal monitoring and decision in embodied agents. *arXiv preprint arXiv:2508.02629*. Available online: https://arxiv.org/abs/2508.02629 (accessed on).

Liu, Z., Gu, Y., Wang, Y., Xue, X., & Fu, Y. (2026). ActiveVLA: Injecting Active Perception into Vision-Language-Action Models for Precise 3D Robotic Manipulation. *arXiv preprint arXiv:2601.08325*. Available online: https://arxiv.org/abs/2601.08325 (accessed on).

Liu, Z., Gu, Y., Zheng, S., Fu, Y., Xue, X., & Jiang, Y.-G. (2025). TriVLA: A Triple-System-Based Unified Vision-Language-Action Model with Episodic World Modeling for General Robot Control. *arXiv preprint arXiv:2507.01424*. Available online: https://arxiv.org/abs/2507.01424 (accessed on).

Liu, Z., Liu, J., Xu, J., Han, N., Gu, C., Chen, H., Zhou, K., Zhang, R., Hsieh, K. C., Wu, K., et al. (2025). Mla: A multisensory language-action model for multimodal understanding and forecasting in robotic manipulation. *arXiv preprint arXiv:2509.26642*. Available online: https://arxiv.org/abs/2509.26642 (accessed on).

Lu, G., Guo, W., Zhang, C., Zhou, Y., Jiang, H., Gao, Z., Tang, Y., & Wang, Z. (2025). Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning. *arXiv preprint arXiv:2505.18719*. Available online: https://arxiv.org/abs/2505.18719 (accessed on).

Lu, Y., Liu, Z., Fan, X., Yang, Z., Hou, J., Li, J., Ding, K., & Zhao, H. (2026). FASTER: Rethinking Real-Time Flow VLAs. *arXiv preprint arXiv:2603.19199*. Available online: https://arxiv.org/abs/2603.19199 (accessed on).

Lu, Y., Tian, Y., Yuan, Z., Wang, X., Hua, P., Xue, Z., & Xu, H. (2025). H^3DP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning. *arXiv preprint arXiv:2505.07819*. Available online: https://arxiv.org/abs/2505.07819 (accessed on).

Luo, H., Wang, Y., Zhang, W., Zheng, S., Xi, Z., Xu, C., Xu, H., Yuan, H., Zhang, C., Wang, Y., et al. (2026). Being-H0. 5: Scaling Human-Centric Robot Learning for Cross-Embodiment Generalization. *arXiv preprint arXiv:2601.12993*.

Luo, Y., Chen, H., Wu, Z., Sui, B., Liu, J., Gu, C., Liu, Z., Feng, Q., Yu, J., Gu, S., Jia, P., Heng, P.-A., & Zhang, S. (2026). Look Before Acting: Enhancing Vision Foundation Representations for Vision-Language-Action Models. Available online: https://api.semanticscholar.org/CorpusID:286571570 (accessed on).

Luo, Y., Chen, W., Liang, T., Wang, B., & Li, Z. (2026). SimVLA: A Simple VLA Baseline for Robotic Manipulation. *arXiv preprint arXiv:2602.18224*.

Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., Song, H., Chen, Q., Deng, X., & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951*.

Lynch, C., Wahid, A., Tompson, J., Ding, T., Betker, J., Baruch, R., Armstrong, T., & Florence, P. (2022). *Interactive language: Talking to robots in real time.* Available online: https://arxiv.org/abs/2210.06407 (accessed on).

Ma, C., Yang, G., Lu, K., Xu, S., Byrne, B., Trigoni, N., & Markham, A. (2026). CycleVLA: Proactive Self-Correcting Vision-Language-Action Models via Subtask Backtracking and Minimum Bayes Risk Decoding. *arXiv preprint arXiv:2601.02295*. Available online: https://arxiv.org/pdf/2601.02295 (accessed on).

Ma, G., Wang, S., Zhang, Z., Yu, S., & Tang, H. (2026). GeneralVLA: Generalizable Vision-Language-Action Models with Knowledge-Guided Trajectory Planning. *arXiv preprint arXiv:2602.04315*.

Ma, T., Zheng, J., Wang, Z., Jiang, C., Cui, A., Liang, J., & Yang, S. (2026). DiT4DiT: Jointly Modeling Video Dynamics and Actions for Generalizable Robot Control.

Ma, Y., Chi, D., Wu, S., Liu, Y., Zhuang, Y., & King, I. (2025, November). Astra: Efficient Transformer Architecture and Contrastive Dynamics Learning for Embodied Instruction Following. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 13621–13639). Suzhou, China: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.emnlp-main.688/ (accessed on). https://doi.org/10.18653/v1/2025.emnlp-main.688.

 doi:10.20944/preprints202606.0425.v1

Ma, Y., Song, Z., Zhuang, Y., Hao, J., & King, I. (2026). A Survey on Vision–Language–Action Models for Embodied AI. *IEEE Transactions on Neural Networks and Learning Systems*.

Mandlekar, A., Xu, D., Wong, J., Nasiriany, S., Wang, C., Kulkarni, R., Fei-Fei, L., Savarese, S., Zhu, Y., & Martín-Martín, R. (2021). What Matters in Learning from Offline Human Demonstrations for Robot Manipulation. In *arxiv preprint arxiv:2108.03298*.

Mees, O., Hermann, L., Rosete-Beas, E., & Burgard, W. (2021). CALVIN: A Benchmark for Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks. *IEEE Robotics and Automation Letters*, *7*, 7327-7334.

Mu, T., Ling, Z., Xiang, F., Yang, D., Li, X., Tao, S., Huang, Z., Jia, Z., & Su, H. (2021). *Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations.* Available online: https://arxiv.org/abs/2107.14483 (accessed on).

Mu, Y., Chen, T., Chen, Z., Peng, S., Lan, Z., Gao, Z., Liang, Z., Yu, Q., Zou, Y., Xu, M., Lin, L., Xie, Z., Ding, M., & Luo, P. (2025). RoboTwin: Dual-Arm Robot Benchmark with Generative Digital Twins. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27649-27660.

Myers, V., Zheng, B., Dragan, A., Fang, K., & Levine, S. (2026). Temporal representation alignment: Successor features enable emergent compositionality in robot instruction following. *Advances in Neural Information Processing Systems*, *38*, 149934–149961. Available online: https://arxiv.org/abs/2502.05454 (accessed on).

Nasiriany, S., Maddukuri, A., Zhang, L., Parikh, A., Lo, A., Joshi, A., Mandlekar, A., & Zhu, Y. (2024). RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots. *ArXiv*, *abs/2406.02523*. Available online: https://api.semanticscholar.org/CorpusID:270226600 (accessed on).

Nasiriany, S., Nasiriany, S., Maddukuri, A., & Zhu, Y. (2026). Robocasa365: A large-scale simulation framework for training and benchmarking generalist robots. *arXiv preprint arXiv:2603.04356*.

Ni, C., Chen, C., Wang, X., Zhu, Z., Zheng, W., Wang, B., Chen, T., Zhao, G., Li, H., Dong, Z., et al. (2025). SwiftVLA: Unlocking Spatiotemporal Dynamics for Lightweight VLA Models at Minimal Overhead. *arXiv preprint arXiv:2512.00903*. Available online: https://arxiv.org/abs/2512.00903 (accessed on).

Niu, D., Sharma, Y., Shi, B., Ding, R., Gioia, M., Xue, H., Tsai, H., Kallidromitis, K., Pai, A., Regan, C., et al. (2025). Learning to Grasp Anything by Playing with Random Toys. *arXiv preprint arXiv:2510.12866*. Available online: https://arxiv.org/abs/2510.12866 (accessed on).

Niu, D., Sharma, Y., Xue, H., Biamby, G., Zhang, J., Ji, Z., Darrell, T., & Herzig, R. (2025). Pre-training auto-regressive robotic models with 4d representations. *arXiv preprint arXiv:2502.13142*. Available online: https://arxiv.org/abs/2502.13142 (accessed on).

Niu, Y., Zhou, S., Li, Y., Den, Y., & Wang, L. (2025). Time-unified diffusion policy with action discrimination for robotic manipulation. *arXiv preprint arXiv:2506.09422*. Available online: https://arxiv.org/pdf/2506.09422 (accessed on).

NVIDIA, Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L. J., Fang, Y., Fox, D., Hu, F., Huang, S., Jang, J., Jiang, Z., Kautz, J., Kundalia, K., Lao, L., Li, Z., Lin, Z., Lin, K., … Zhu, Y. (2025). *Gr00t n1: An open foundation model for generalist humanoid robots.* Available online: https://arxiv.org/abs/2503.14734 (accessed on).

Pang, X., Xia, W., Wang, Z., Zhao, B., Hu, D., Wang, D., & Li, X. (2024). Depth helps: Improving pre-trained rgb-based policy with depth information injection. In *2024 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 7251–7256). Available online: https://arxiv.org/pdf/2408.05107 (accessed on).

Park, M., Kim, K., Hyung, J., Jang, H., Jin, H., Yun, J., Lee, H., & Choo, J. (2025). ACG: Action Coherence Guidance for Flow-based VLA models. *arXiv preprint arXiv:2510.22201*. Available online: https://arxiv.org/abs/2510.22201 (accessed on).

Peng, X., Gao, C., Jin, L., Li, A., & Liu, S. (2026). BiCoord: A Bimanual Manipulation Benchmark towards Long-Horizon Spatial-Temporal Coordination. *arXiv preprint arXiv:2604.05831*.

Pertsch, K., Stachowicz, K., Ichter, B., Driess, D., Nair, S., Vuong, Q., Mees, O., Finn, C., & Levine, S. (2025). FAST: Efficient Action Tokenization for Vision-Language-Action Models. *ArXiv*, *abs/2501.09747*.

Pumacay, W., Singh, I., Duan, J., Krishna, R., Thomason, J., & Fox, D. (2024). The colosseum: A benchmark for evaluating generalization for robotic manipulation. *arXiv preprint arXiv:2402.08191*.

Qi, Z., Zhang, W., Ding, Y., Dong, R., Yu, X., Li, J., Xu, L., Li, B., He, X., Fan, G., et al. (2026). Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. *Advances in neural information processing systems*, *38*, 76367–76412.

 doi:10.20944/preprints202606.0425.v1

Qian, J., Han, B., Shi, C., Xiao, L., Yang, L., Shi, S., & Jiang, L. (2025). GeoPredict: Leveraging Predictive Kinematics and 3D Gaussian Geometry for Precise VLA Manipulation. *arXiv preprint arXiv:2512.16811*. Available online: https://arxiv.org/abs/2512.16811 (accessed on).

Qian, Z., Chi, X., Li, Y., Wang, S., Qin, Z., Ju, X., Han, S., & Zhang, S. (2025). Wristworld: Generating wrist-views via 4d world models for robotic manipulation. *arXiv preprint arXiv:2510.07313*. Available online: https://arxiv.org/abs/2510.07313 (accessed on).

Qin, Y., Kang, L., Song, X., Yin, Z., Liu, X., Liu, X., Zhang, R., & Bai, L. (2025). *Robofactory: Exploring embodied agent collaboration with compositional constraints.* Available online: https://arxiv.org/abs/2503.16408 (accessed on).

Qin, Y., Shi, Z., Yu, J., Wang, X., Zhou, E., Li, L., Yin, Z., Liu, X., Sheng, L., Shao, J., Bai, L., Ouyang, W., & Zhang, R. (2024). *Worldsimbench: Towards video generation models as world simulators.* Available online: https://arxiv.org/abs/2410.18072 (accessed on).

Qu, D., Song, H., Chen, Q., Chen, Z., Gao, X., Ye, X., Lv, Q., Shi, M., Ren, G., Ruan, C., Yao, M., Yang, H., Bao, J., Zhao, B., & Wang, D. (2025). EO-1: An Open Unified Embodied Foundation Model for General Robot Control.

Ranasinghe, K., Li, X., Nguyen, E.-R., Mata, C., Park, J., & Ryoo, M. S. (2025). Pixel motion as universal representation for robot control. *arXiv preprint arXiv:2505.07817*. Available online: https://arxiv.org/abs/2505.07817 (accessed on).

Ranasinghe, K., Zhou, H., Fang, Y., Yang, L., Xue, L., Xu, R., Xiong, C., Savarese, S., Ryoo, M. S., & Niebles, J. C. (2026). Future Optical Flow Prediction Improves Robot Control & Video Generation. *arXiv preprint arXiv:2601.10781*. Available online: https://arxiv.org/abs/2601.10781 (accessed on).

Ren, Z., Wei, Y., Guo, X., Zhao, Y., Kang, B., Feng, J., & Jin, X. (2025). Videoworld: Exploring knowledge learning from unlabeled videos. In *Proceedings of the computer vision and pattern recognition conference* (pp. 29029–29039). Available online: https://arxiv.org/abs/2501.09781 (accessed on).

Reuss, M., Pari, J., Agrawal, P., & Lioutikov, R. (2025). Efficient diffusion transformer policies with mixture of expert denoisers for multitask learning. In *International conference on learning representations* (Vol. 2025, pp. 17247–17275). Available online: https://arxiv.org/abs/2412.12953 (accessed on).

Reuss, M., Yağmurlu, Ö. E., Wenzel, F., & Lioutikov, R. (2024). Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. *arXiv preprint arXiv:2407.05996*. Available online: https://arxiv.org/pdf/2407.05996 (accessed on).

Reuss, M., Zhou, H., Rühle, M., Yağmurlu, Ö. E., Otto, F., & Lioutikov, R. (2025). Flower: Democratizing generalist robot policies with efficient vision-language-action flow policies. *arXiv preprint arXiv:2509.04996*. Available online: https://arxiv.org/abs/2509.04996 (accessed on).

Robotics, T., Kang, J., Park, T., An, J., Kimm, S. M., Kim, J., Pahk, J., Kim, B., Lee, J., Baek, N., et al. (2026). Habilis-β: A Fast-Motion and Long-Lasting On-Device Vision-Language-Action Model. *arXiv preprint arXiv:2602.18813*. Available online: https://arxiv.org/abs/2602.18813 (accessed on).

Römer, R., Zhang, Y., Li, Y., & Schoellig, A. P. (2026). CLARE: Continual Learning for Vision-Language-Action Models via Autonomous Adapter Routing and Expansion. *IEEE Robotics and Automation Letters*. Available online: https://arxiv.org/abs/2601.09512 (accessed on).

Sapkota, R., Cao, Y., Roumeliotis, K. I., & Karkee, M. (2025). Vision-Language-Action (VLA) Models: Concepts, Progress, Applications and Challenges.

Shang, J., Schmeckpeper, K., May, B. B., Minniti, M. V., Kelestemur, T., Watkins, D., & Herlant, L. (2024). Theia: Distilling diverse vision foundation models for robot learning. *arXiv preprint arXiv:2407.20179*. Available online: https://arxiv.org/abs/2407.20179 (accessed on).

Shen, W., Liu, Y., Wu, Y., Liang, Z., Gu, S., Wang, D., Nian, T., Xu, L., Qin, Y., Pang, J., et al. (2025). Expertise need not monopolize: Action-Specialized Mixture of Experts for Vision-Language-Action Learning. *arXiv preprint arXiv:2510.14300*. Available online: https://arxiv.org/abs/2510.14300 (accessed on).

Shen, Y., Wei, F., Du, Z., Liang, Y., Lu, Y., Yang, J., Zheng, N., & Guo, B. (2026). Videovla: Video generators can be generalizable robot manipulators. *Advances in neural information processing systems*, *38*, 95597–95621. Available online: https://arxiv.org/pdf/2512.06963 (accessed on).

Shridhar, M., Manuelli, L., & Fox, D. (2021). CLIPort: What and Where Pathways for Robotic Manipulation. In *Proceedings of the 5th conference on robot learning (corl).*

Shukor, M., Aubakirova, D., Capuano, F., Kooijmans, P., Palma, S., Zouitine, A., Aractingi, M., Pascal, C., Russi, M., Marafioti, A., et al. (2025). Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*. Available online: https://arxiv.org/abs/2506.01844 (accessed on).

 doi:10.20944/preprints202606.0425.v1

Song, H., Qu, D., Yao, Y., Chen, Q., Lv, Q., Tang, Y., Shi, M., Ren, G., Yao, M., Zhao, B., et al. (2025). Hume: Introducing system-2 thinking in visual-language-action model. *arXiv preprint arXiv:2505.21432*. Available online: https://arxiv.org/abs/2505.21432 (accessed on).

Song, W., Chen, J., Ding, P., Huang, Y., Zhao, H., Wang, D., & Li, H. (2025). Ceed-vla: Consistency vision-language-action model with early-exit decoding. *arXiv preprint arXiv:2506.13725*. Available online: https://arxiv.org/abs/2506.13725 (accessed on).

Song, W., Chen, J., Li, W., He, X., Zhao, H., Cui, C., Su, P. D. S., Tang, F., Cheng, X., Wang, D., et al. (2025). Rationalvla: A rational vision-language-action model with dual system. *arXiv preprint arXiv:2506.10826*.

Song, W., Zhao, H., Li, F., Zhou, Z., Wang, X., Lyu, J., Ding, P., Wang, Y., Wang, D., & Li, H. (2026). CapVector: Learning Transferable Capability Vectors in Parametric Space for Vision-Language-Action Models. *arXiv preprint arXiv:2605.10903*. Available online: https://arxiv.org/abs/2605.10903 (accessed on).

Song, W., Zhou, Z., Zhao, H., Chen, J., Ding, P., Yan, H., Huang, Y., Tang, F., Wang, D., & Li, H. (2026). Reconvla: Reconstructive vision-language-action model as effective robot perceiver. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 40, pp. 18549–18557). Available online: https://arxiv.org/abs/2508.10333 (accessed on).

Sun, J., Zhang, W., Qi, Z., Ren, S., Liu, Z., Zhu, H., Sun, G., Jin, X., & Chen, Z. (2026). Vla-jepa: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*.

Sun, L., Xie, B., Liu, Y., Shi, H., Wang, T., & Cao, J. (2025). Geovla: Empowering 3d representations in vision-language-action models. *arXiv preprint arXiv:2508.09071*. Available online: https://arxiv.org/abs/2508.09071 (accessed on).

Sun, Q., Chi, X., Rui, Y., Li, Y., Ge, K., Li, J., Han, S., & Zhang, S. (2026). LABSHIELD: A Multimodal Benchmark for Safety-Critical Reasoning and Planning in Scientific Laboratories. *arXiv preprint arXiv:2603.11987*.

Sun, X., Xu, Z., Cao, C., Liu, Z., Sun, Y., Pang, J., Zhang, R., Yang, Z., Pang, K., He, D., et al. (2026). Atom-VLA: Scalable Post-Training for Robotic Manipulation via Predictive Latent World Models. *arXiv preprint arXiv:2603.08519*. Available online: https://arxiv.org/abs/2603.08519 (accessed on).

Sun, Y., Cao, M., Yang, P., Xu, R., Yan, Y., Xu, R., Ma, L., Gan, R., Zhai, A., Chen, Q., et al. (2026a). ManipArena: Comprehensive Real-world Evaluation of Reasoning-Oriented Generalist Robot Manipulation. *arXiv preprint arXiv:2603.28545*.

Sun, Y., Cao, M., Yang, P., Xu, R., Yan, Y., Xu, R., Ma, L., Gan, R., Zhai, A., Chen, Q., Xu, Z., Wang, H., Yu, J., Liang, L., Wang, Q., Laptev, I., Reid, I. D., & Liang, X. (2026b). *Maniparena: Comprehensive real-world evaluation of reasoning-oriented generalist robot manipulation*. Available online: https://arxiv.org/abs/2603.28545 (accessed on).

Sun, Z., & Song, S. (2026). From Prior to Pro: Efficient Skill Mastery via Distribution Contractive RL Finetuning.

Takagi, Y., Kambara, M., Yashima, D., Seno, K., Tokura, K., & Sugiura, K. (2026). AnoleVLA: Lightweight Vision-Language-Action Model with Deep State Space Models for Mobile Manipulation. *arXiv preprint arXiv:2603.15046*. Available online: https://arxiv.org/abs/2603.15046 (accessed on).

Tan, H., Chen, S., Xu, Y., Wang, Z., Ji, Y., Chi, C., Lyu, Y., Zhao, Z., Chen, X., Co, P., et al. (2025). Robo-Dopamine: General Process Reward Modeling for High-Precision Robotic Manipulation. *arXiv preprint arXiv:2512.23703*. Available online: https://arxiv.org/abs/2512.23703 (accessed on).

Tan, S., Dou, K., Zhao, Y., & Krähenbühl, P. (2025). Interactive post-training for vision-language-action models. *arXiv preprint arXiv:2505.17016*. Available online: https://arxiv.org/abs/2505.17016 (accessed on).

Tang, J., Sun, Y., Zhao, Y., Yang, S., Lin, Y., Zhang, Z., Hou, J., Lu, Y., Liu, Z., & Han, S. (2025a). VLASH: Real-Time VLAs via Future-State-Aware Asynchronous Inference. *ArXiv*, *abs/2512.01031*.

Tang, J., Sun, Y., Zhao, Y., Yang, S., Lin, Y., Zhang, Z., Hou, J., Lu, Y., Liu, Z., & Han, S. (2025b). Vlash: Real-time vlas via future-state-aware asynchronous inference. *arXiv preprint arXiv:2512.01031*. Available online: https://arxiv.org/abs/2512.01031 (accessed on).

Tang, W., Pan, J.-H., Liu, Y.-H., Tomizuka, M., Li, L. E., Fu, C.-W., & Ding, M. (2025). Geomanip: Geometric constraints as general interfaces for robot manipulation. *arXiv preprint arXiv:2501.09783*. Available online: https://arxiv.org/abs/2501.09783 (accessed on).

Tao, S., Xiang, F., Shukla, A., Qin, Y., Hinrichsen, X., Yuan, X., Bao, C., Lin, X., Liu, Y., kai Chan, T., Gao, Y., Li, X., Mu, T., Xiao, N., Gurha, A., Rajesh, V. N., Choi, Y. W., Chen, Y.-R., Huang, Z., … Su, H. (2025). ManiSkill3: GPU Parallelized Robotics Simulation and Rendering for Generalizable Embodied AI. *Robotics: Science and Systems*.

 doi:10.20944/preprints202606.0425.v1

Tur, Y., Naghiyev, J., Fang, H., Tsai, W.-C., Duan, J., Fox, D., & Krishna, R. (2026). Recurrent-depth vla: Implicit test-time compute scaling of vision-language-action models via latent iterative reasoning. *arXiv preprint arXiv:2602.07845*.

Wang, H., Xu, J., Xiang, Y., Pan, J., Zhou, Y., Li, Y.-L., & Dai, G. (2025). Specprune-vla: Accelerating vision-language-action models via action-aware self-speculative pruning. *arXiv preprint arXiv:2509.05614*.

Wang, H., Zhao, W., Wang, X., Huang, S., Lin, H., Zheng, B., Xu, R., Wang, G., Mu, Y., Wang, H., Fan, L., Li, H., Zhang, Z., & Tan, T. (2026). *Dexjoco: A benchmark and toolkit for task-oriented dexterous manipulation on mujoco.* Available online: https://arxiv.org/abs/2605.16257 (accessed on).

Wang, R., Zhang, Y., Lin, J., Luo, K., Wang, J., Wang, Z., & Qi, X. (2026). When to Trust Imagination: Adaptive Action Execution for World Action Models. *arXiv preprint arXiv:2605.06222*. Available online: https://arxiv.org/abs/2605.06222 (accessed on).

Wang, S., Fu, J., Liu, F., He, X., Wu, H., Shi, J., Huang, K., Fei, Z., Gong, J., Wu, Z., et al. (2025). RoboOmni: Proactive Robot Manipulation in Omni-modal Context. *arXiv preprint arXiv:2510.23763*. Available online: https://arxiv.org/abs/2510.23763 (accessed on).

Wang, S., Liu, S., Wang, W., Shan, J., & Fang, B. (2025). Robobert: An end-to-end multimodal robotic manipulation model. *arXiv preprint arXiv:2502.07837*. Available online: https://arxiv.org/abs/2502.07837 (accessed on).

Wang, S., Wang, X., Zhu, Z., Pei, M., Cui, X., Deng, C., Zhao, J., Huang, G., Zhang, H., & Wang, J. (2026). π-StepNFT: Wider Space Needs Finer Steps in Online RL for Flow-based VLAs. *arXiv preprint arXiv:2603.02083*.

Wang, S., Zhang, J., Li, M., Liu, J., Li, A., Wu, K., Zhong, F., Yu, J., Zhang, Z., & Wang, H. (2025). Trackvla: Embodied visual tracking in the wild. *arXiv preprint arXiv:2505.23189*. Available online: https://arxiv.org/abs/2505.23189 (accessed on).

Wang, Y., Ding, P., Li, L., Cui, C., Ge, Z., Tong, X., Song, W., Zhao, H., Zhao, W., Hou, P., Huang, S., Tang, Y., Wang, W., Zhang, R., Liu, J., & Wang, D. (2025). VLA-Adapter: An Effective Paradigm for Tiny-Scale Vision-Language-Action Model. *ArXiv*, *abs/2509.09372*.

Wang, Y., Li, X., Wang, W., Zhang, J., Li, Y., Chen, Y., Wang, X., & Zhang, Z. (2025). Unified vision-language-action model. *arXiv preprint arXiv:2506.19850*.

Wang, Y., Zhu, H., Liu, M., Yang, J., Fang, H.-S., & He, T. (2025). Vq-vla: Improving vision-language-action models via scaling vector-quantized action tokenizers. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 11089–11099). Available online: https://arxiv.org/abs/2507.01016 (accessed on).

Wang, Z., Chen, Y., Liu, Y., Ye, J., Chen, P., Lu, C., Liu, S., Yu, B., & Jia, J. (2026). VP-VLA: Visual Prompting as an Interface for Vision-Language-Action Models. *arXiv preprint arXiv:2603.22003*. Available online: https://arxiv.org/abs/2603.22003 (accessed on).

Wang, Z., Liu, C., Xiang, Y., Zhang, R., Hao, Q., Lu, H., Chen, H., Feng, Z., Zheng, K., Ye, D., et al. (2026). The Great March 100: 100 Detail-oriented Tasks for Evaluating Embodied AI Agents. *arXiv preprint arXiv:2601.11421*.

Wang, Z., Wang, B., Zhang, H., Du, T., Chen, T., Sun, G., He, Y., Shen, Z., Ye, W., & Li, A. (2026). Vision-Language-Action in Robotics: A Survey of Datasets, Benchmarks, and Data Engines. *arXiv preprint arXiv:2604.23001*.

Wen, C., Lin, X., So, J., Chen, K., Dou, Q., Gao, Y., & Abbeel, P. (2023). Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*. Available online: https://arxiv.org/abs/2401.00025 (accessed on).

Wen, J., Zhu, Y., Li, J., Tang, Z., Shen, C., & Feng, F. (2025). Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*. Available online: https://arxiv.org/abs/2502.05855 (accessed on).

Wen, J., Zhu, Y., Li, J., Zhu, M., Tang, Z., Wu, K., Xu, Z., Liu, N., Cheng, R., Shen, C., et al. (2025). Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*. Available online: https://arxiv.org/pdf/2409.12514 (accessed on).

Wen, X., Zhao, B., Chen, Y., Pang, J., & Qi, X. (2025). A data-centric revisit of pre-trained vision models for robot learning. In *Proceedings of the computer vision and pattern recognition conference* (pp. 12143–12154). Available online: https://arxiv.org/abs/2503.06960 (accessed on).

Wen, Y., Li, H., Gu, K., Zhao, Y., Wang, T., & Sun, X. (2025). Llada-vla: Vision language diffusion action models. *arXiv preprint arXiv:2509.06932*. Available online: https://arxiv.org/pdf/2509.06932 (accessed on).

Wu, H., Jing, Y., Cheang, C., Chen, G., Xu, J., Li, X., Liu, M., Li, H., & Kong, T. (2024). Unleashing large-scale video generative pre-training for visual robot manipulation. In *International conference on learning representations* (Vol. 2024, pp. 10641–10662). Available online: https://arxiv.org/abs/2312.13139 (accessed on).

Wu, W., Lu, F., Wang, Y., Yang, S., Liu, S., Wang, F., Zhu, Q., Sun, H., Wang, Y., Ma, S., et al. (2026). A Pragmatic VLA Foundation Model. *arXiv preprint arXiv:2601.18692*.

 doi:10.20944/preprints202606.0425.v1

Wu, X.-M., Fan, B., Liao, K., Jiang, J.-J., Yang, R., Luo, Y., Wu, Z., Zheng, W.-S., & Loy, C. C. (2026). VLANeXt: Recipes for Building Strong VLA Models. *arXiv preprint arXiv:2602.18532*.

Wu, Z., Zhou, Y., Xu, X., Wang, Z., & Yan, H. (2025). Momanipvla: Transferring vision-language-action models for general mobile manipulation. In *Proceedings of the computer vision and pattern recognition conference* (pp. 1714–1723). Available online: https://arxiv.org/pdf/2503.13446 (accessed on).

Xiang, T.-Y., Jin, A.-Q., Zhou, X.-H., Gui, M.-J., Xie, X.-L., Liu, S.-Q., Wang, S.-Y., Duan, S.-B., Xie, F.-C., Wang, W.-K., Wang, S.-C., Li, L.-Y., Tu, T., & Hou, Z.-G. (2026). *Parallels between vla model post-training and human motor learning: Progress, challenges, and trends.* Available online: https://arxiv.org/abs/2506.20966 (accessed on).

Xiao, J., Yang, Y., Chang, X., Chen, R., Xiong, F., Xu, M., Zheng, W.-S., & Zhang, Q. (2026). *World-env: Leveraging world model as a virtual environment for vla post-training.* Available online: https://arxiv.org/abs/2509.24948 (accessed on).

Xu, C., Zhang, S., Liu, Y., Sun, B., Chen, W., Xu, B., Liu, Q., Wang, J., Wang, S., Luo, S., et al. (2025). An Anatomy of Vision-Language-Action Models: From Modules to Milestones and Challenges. *arXiv preprint arXiv:2512.11362*.

Xu, R., Zhang, J., Guo, M., Wen, Y., Yang, H., Lin, M., Huang, J., Li, Z., Zhang, K., Wang, L., et al. (2025). A0: An affordance-aware hierarchical model for general robotic manipulation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 13491–13501). Available online: https://arxiv.org/abs/2504.12636 (accessed on).

Xu, S., Wang, Y., Xia, C., Zhu, D., Huang, T., & Xu, C. (2025). VLA-Cache: Efficient Vision-Language-Action Manipulation via Adaptive Token Caching. Available online: https://api.semanticscholar.org/CorpusID:276107365 (accessed on).

Yakefu, A., Xie, B., Xu, C., Zhang, E., Zhou, E., Jia, F., Yang, H., Fan, H., Zhang, H., Peng, H., Tan, J., Huang, J., Liu, K., Liu, K., Gu, K., Zhang, Q., Zhang, R., Huang, S., Cheng, S., … Yan, Z. (2025a). RoboChallenge: Large-scale Real-robot Evaluation of Embodied Policies. *ArXiv*, *abs/2510.17950*.

Yakefu, A., Xie, B., Xu, C., Zhang, E., Zhou, E., Jia, F., Yang, H., Fan, H., Zhang, H., Peng, H., et al. (2025b). RoboChallenge: Large-scale Real-robot Evaluation of Embodied Policies. *arXiv preprint arXiv:2510.17950*. Available online: https://arxiv.org/abs/2510.17950 (accessed on).

Yan, F., Liu, F., Zheng, L., Zhong, Y., Huang, Y., Guan, Z., Feng, C., & Ma, L. (2024). RoboTron-Mani: All-in-One Multimodal Large Model for Robotic Manipulation. *arXiv preprint arXiv:2412.07215*. Available online: https://arxiv.org/pdf/2412.07215 (accessed on).

Yan, Y., Xu, J., Di, S., Wu, H., & Xie, W. (2026). OmniStream: Mastering Perception, Reconstruction and Action in Continuous Streams.

Yang, G., Zhang, T., Hao, H., Wang, W., Liu, Y., Wang, D., Chen, G., Cai, Z., Chen, J., Su, W., Zhou, W., Qiao, Y., Dai, J., Pang, J., Luo, G., Wang, W., Mu, Y., & Hou, Z. (2025). Vlaser: Vision-Language-Action Model with Synergistic Embodied Reasoning. *ArXiv*, *abs/2510.11027*.

Yang, J., Dong, Y., Liu, S., Li, B., Wang, Z., Tan, H., Jiang, C., Kang, J., Zhang, Y., Zhou, K., et al. (2024). Octopus: Embodied vision-language programmer from environmental feedback. In *European conference on computer vision* (pp. 20–38).

Yang, R., Yu, Q., Wu, Y., Yan, R., Li, B., Cheng, A.-C., Zou, X., Fang, Y., Yin, H., Liu, S., Han, S., Lu, Y., & Wang, X. (2025). *Egovla: Learning vision-language-action models from egocentric human videos.* Available online: https://arxiv.org/abs/2507.12440 (accessed on).

Yang, S., Zhang, Y., He, H., Pan, L., Li, X., Bai, C., & Li, X. (2025). Steering Vision-Language-Action Models as Anti-Exploration: A Test-Time Scaling Approach. *arXiv preprint arXiv:2512.02834*. Available online: https://arxiv.org/abs/2512.02834 (accessed on).

Yang, T., Chen, G., Chen, Y., Liang, Z., Liu, Y., Chen, Z., Xu, C., Liang, H., Pang, J., Mu, Y., & Luo, P. (2026). HiVLA: A Visual-Grounded-Centric Hierarchical Embodied Manipulation System.

Yang, Y., Cai, Z., Tian, Y., Zeng, J., & Pang, J. (2025). Gripper keypose and object pointflow as interfaces for bimanual robotic manipulation. *arXiv preprint arXiv:2504.17784*. Available online: https://arxiv.org/abs/2504.17784 (accessed on).

Yang, Y., Duan, Z., Xie, T., Cao, F., Shen, P., Song, P., Jin, P., Sun, G., Xu, S., You, Y., et al. (2025). FPC-VLA: A Vision-Language-Action Framework with a Supervisor for Failure Prediction and Correction. *arXiv preprint arXiv:2509.04018*. Available online: https://arxiv.org/pdf/2509.04018 (accessed on).

Yang, Y., Li, X., Chen, Y., Song, J., Wang, Y., Xiao, Z., Su, J., You, Q., Liu, P., & Deng, Z. (2025). Mantis: A Versatile Vision-Language-Action Model with Disentangled Visual Foresight. *ArXiv*, *abs/2511.16175*.

 doi:10.20944/preprints202606.0425.v1

Yang, Y., Sun, J., Kou, S., Wang, Y., & Deng, Z. (2025). Lohovla: A unified vision-language-action model for long-horizon embodied tasks. *arXiv preprint arXiv:2506.00411*. Available online: https://arxiv.org/abs/2506.00411 (accessed on).

Yang, Y., Wang, Y., Wen, Z., Zhongwei, L., Zou, C., Zhang, Z., Wen, C., & Zhang, L. (2026). Efficientvla: Training-free acceleration and compression for vision-language-action models. *Advances in Neural Information Processing Systems*, *38*, 40891–40914.

Yang, Y., Zeng, S., Lin, T., Chang, X., Qi, D., Xiao, J., Liu, H., Chen, R., Chen, Y., Huo, D., et al. (2026). Abot-m0: Vla foundation model for robotic manipulation with action manifold learning. *arXiv preprint arXiv:2602.11236*.

Ye, A., Wang, B., Ni, C., Huang, G., Zhao, G., Li, H., Li, H., Li, J., Lv, J., Liu, J., et al. (2026). GigaWorld-Policy: An Efficient Action-Centered World–Action Model. *arXiv preprint arXiv:2603.17240*.

Ye, J., Gong, S., Gao, J., Fan, J., Wu, S., Bi, W., Bai, H., Shang, L., & Kong, L. (2025). Dream-VL & Dream-VLA: Open Vision-Language and Vision-Language-Action Models with Diffusion Language Model Backbone. *arXiv preprint arXiv:2512.22615*. Available online: https://arxiv.org/abs/2512.22615 (accessed on).

Ye, S., Ge, Y., Zheng, K., Gao, S., Yu, S., Kurian, G., Indupuru, S., Tan, Y. L., Zhu, C., Xiang, J., et al. (2026). World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*.

Ying, Z., Liu, A., Liang, S., Huang, L., Guo, J., Zhou, W., Liu, X., & Tao, D. (2024). *Safebench: A safety evaluation framework for multimodal large language models*. Available online: https://arxiv.org/abs/2410.18927 (accessed on).

Yu, B., Lian, S., Lin, X., Wei, Y., Shen, Z., Wu, C., Miao, Y., Wang, X., Wang, B., Huang, C., et al. (2026). TwinBrainVLA: Unleashing the Potential of Generalist VLMs for Embodied Tasks via Asymmetric Mixture-of-Transformers. *arXiv preprint arXiv:2601.14133*. Available online: https://arxiv.org/pdf/2601.14133 (accessed on).

Yu, T., Quillen, D., He, Z., Julian, R., Narayan, A., Shively, H., Bellathur, A., Hausman, K., Finn, C., & Levine, S. (2021). *Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning*. Available online: https://arxiv.org/abs/1910.10897 (accessed on).

Yu, W., Wang, T., Li, F., Li, J., & Zhu, L. (2026). AC^2-VLA: Action-Context-Aware Adaptive Computation in Vision-Language-Action Models for Efficient Robotic Manipulation. *arXiv preprint arXiv:2601.19634*. Available online: https://arxiv.org/abs/2601.19634 (accessed on).

Yuan, Y., Cui, H., Huang, Y., Chen, Y., Ni, F., Dong, Z., Li, P., Zheng, Y., & Hao, J. (2025). Embodied-r1: Reinforced embodied reasoning for general robotic manipulation. *arXiv preprint arXiv:2508.13998*. Available online: https://arxiv.org/abs/2508.13998 (accessed on).

Yue, Y., Wang, Y., Kang, B., Han, Y., Wang, S., Song, S., Feng, J., & Huang, G. (2024). Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. *Advances in Neural Information Processing Systems*, *37*, 56619–56643. Available online: https://arxiv.org/abs/2411.02359 (accessed on).

Zeng, A., Florence, P., Tompson, J., Welker, S., Chien, J., Attarian, M., Armstrong, T., Krasin, I., Duong, D., Wahid, A., Sindhwani, V., & Lee, J. (2022). *Transporter networks: Rearranging the visual world for robotic manipulation*. Available online: https://arxiv.org/abs/2010.14406 (accessed on).

Zhang, B., Li, J., Shen, J., Cai, Y., Zhang, Y., Chen, Y., Dai, J., Ji, J., & Yang, Y. (2025). *Vla-arena: An open-source framework for benchmarking vision-language-action models*. Available online: https://arxiv.org/abs/2512.22539 (accessed on).

Zhang, C., Hao, P., Cao, X., Hao, X., Cui, S., & Wang, S. (2025). Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation. *arXiv preprint arXiv:2505.09577*. Available online: https://arxiv.org/pdf/2505.09577 (accessed on).

Zhang, D., Sun, J., Hu, C., Wu, X., Yuan, Z., Zhou, R., Shen, F., & Zhou, Q. (2025). *Pure vision language action (vla) models: A comprehensive survey*. Available online: https://arxiv.org/abs/2509.19012 (accessed on).

Zhang, J., Chen, X., Wang, Q., Li, M., Guo, Y., Hu, Y., Zhang, J., Bai, S., Lin, J., & Chen, J. (2026). VLM4VLA: Revisiting Vision-Language-Models in Vision-Language-Action Models. *arXiv preprint arXiv:2601.03309*. Available online: https://arxiv.org/pdf/2601.03309 (accessed on).

Zhang, J., Guo, Y., Chen, X., Wang, Y.-J., Hu, Y., Shi, C., & Chen, J. (2024). Hirt: Enhancing robotic control with hierarchical robot transformers. *arXiv preprint arXiv:2410.05273*. Available online: https://arxiv.org/pdf/2410.05273 (accessed on).

Zhang, J., Wu, S., Luo, X., Wu, H., Gao, L., Shen, H. T., & Song, J. (2025). Inspire: Vision-language-action models with intrinsic spatial reasoning. *arXiv preprint arXiv:2505.13888*.

 doi:10.20944/preprints202606.0425.v1

Zhang, K., Zhang, J., Xu, R., Sun, Y., Xue, S., Wen, Y., Guo, X., Guo, M., Liufu, W., Zihou, L., et al. (2026). A1: A Fully Transparent Open-Source, Adaptive and Efficient Truncated Vision-Language-Action Model. *arXiv preprint arXiv:2604.05672*. Available online: https://arxiv.org/abs/2604.05672 (accessed on).

Zhang, L., Dong, J., Bai, K., Ni, M., Marton, Z.-C., Chen, Z., & Zhang, J. (2025). *Responsiblerobotbench: Benchmarking responsible robot manipulation using multi-modal large language models.* Available online: https://arxiv.org/abs/2512.04308 (accessed on).

Zhang, R., Dong, M., Zhang, Y., Heng, L., Chi, X., Dai, G., Du, L., Wang, D., Du, Y., & Zhang, S. (2026). Mole-vla: Dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 40, pp. 18764–18772). Available online: https://arxiv.org/pdf/2503.20384 (accessed on).

Zhang, S., Xu, Z., Liu, P., Yu, X., Li, Y., Gao, Q., Fei, Z., Yin, Z., Wu, Z., Jiang, Y.-G., & Qiu, X. (2024). *Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks.* Available online: https://arxiv.org/abs/2412.18194 (accessed on).

Zhang, T., Duan, H., Hao, H., Qiao, Y., Dai, J., & Hou, Z. (2026). Grounding actions in camera space: Observation-centric vision-language-action policy. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 40, pp. 18782–18790). Available online: https://arxiv.org/abs/2508.13103 (accessed on).

Zhang, T., Hu, Y., You, J., & Gao, Y. (2024). Leveraging locality to boost sample efficiency in robotic manipulation. *arXiv preprint arXiv:2406.10615*. Available online: https://arxiv.org/abs/2406.10615 (accessed on).

Zhang, T., Yuan, Z., Chi, D., Liu, P., Li, D., Hu, K., Zhang, L., Nie, J., Wei, Z., Chen, Z., Tang, Y., Li, J., Xiang, Z., Li, M., Luo, T., Wan, H., Li, A., Zhai, L., Zhan, Z., … Lin, L. (2026). *Joyai-ra 0.1: A foundation model for robotic autonomy.* Available online: https://arxiv.org/abs/2604.20100 (accessed on).

Zhang, W., Zhang, B., Qi, Z., Zeng, W., Jin, X., & Zhang, L. (2026). Disentangled Robot Learning via Separate Forward and Inverse Dynamics Pretraining. *arXiv preprint arXiv:2604.16391*. Available online: https://arxiv.org/abs/2604.16391 (accessed on).

Zhang, Z., Li, H., Dai, Y., Zhu, Z., Zhou, L., Liu, C., Wang, D., Tay, F. E., Chen, S., Liu, Z., et al. (2025). From spatial to actions: Grounding vision-language-action model in spatial foundation priors. *arXiv preprint arXiv:2510.17439*. Available online: https://arxiv.org/abs/2510.17439 (accessed on).

Zhang, Z., Zheng, K., Chen, Z., Jang, J., Li, Y., Han, S., Wang, C., Ding, M., Fox, D., & Yao, H. (2024). Grape: Generalizing robot policy via preference alignment. *arXiv preprint arXiv:2411.19309*. Available online: https://arxiv.org/abs/2411.19309 (accessed on).

Zhao, H., Song, W., Wang, D., Tong, X., Ding, P., Cheng, X., & Ge, Z. (2025). More: Unlocking scalability in reinforcement learning for quadruped vision-language-action models. In *2025 ieee international conference on robotics and automation (icra)* (pp. 11212–11218). Available online: https://arxiv.org/abs/2503.08007 (accessed on).

Zhao, H., Wang, J., Song, W., Chen, S., Liu, Y., Wang, Y., Li, H., & Wang, D. (2026). Frappe: Infusing world modeling into generalist policies via multiple future representation alignment. *arXiv preprint arXiv:2602.17259*. Available online: https://arxiv.org/abs/2602.17259 (accessed on).

Zhao, H., Zhang, J., Song, W., Ding, P., & Wang, D. (2025). VLA^2: Empowering Vision-Language-Action Models with an Agentic Framework for Unseen Concept Manipulation. *arXiv preprint arXiv:2510.14902*. Available online: https://arxiv.org/abs/2510.14902 (accessed on).

Zhao, J., Lu, W., Zhang, D., Liu, Y., Liang, Y., Zhang, T., Cao, Y., Xie, J., Hu, Y., Wang, S., et al. (2025). Do You Need Proprioceptive States in Visuomotor Policies? *arXiv preprint arXiv:2509.18644*.

Zhen, H., Qiu, X., Chen, P., Yang, J., Yan, X., Du, Y., Hong, Y., & Gan, C. (2024). 3D-VLA: A 3D Vision-Language-Action Generative World Model. *ArXiv*, *abs/2403.09631*.

Zhen, H., Sun, Q., Zhang, H., Li, J., Zhou, S., Du, Y., & Gan, C. (2025). Tesseract: learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*. Available online: https://arxiv.org/abs/2504.20995 (accessed on).

Zheng, J., Li, J., Liu, D., Zheng, Y., Wang, Z., Ou, Z., Liu, Y., Liu, J., Zhang, Y.-Q., & Zhan, X. (2025). Universal actions for enhanced embodied foundation models. In *Proceedings of the computer vision and pattern recognition conference* (pp. 22508–22519). Available online: https://arxiv.org/abs/2501.10105 (accessed on).

Zheng, J., Li, J., Wang, Z., Liu, D., Kang, X., Feng, Y., Zheng, Y., Zou, J., Chen, Y., Zeng, J., et al. (2025). X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. *arXiv preprint arXiv:2510.10274*. Available online: https://arxiv.org/abs/2510.10274 (accessed on).

Zheng, R., Liang, Y., Huang, S., Gao, J., Daum'e, H., Kolobov, A., Huang, F., & Yang, J. (2024). TraceVLA: Visual Trace Prompting Enhances Spatial-Temporal Awareness for Generalist Robotic Policies. *ArXiv*, *abs/2412.10345*.

 doi:10.20944/preprints202606.0425.v1

Zheng, Z., Cai, J.-F., Wu, X.-M., Wei, Y.-L., Tang, Y.-M., Wu, A., & Zheng, W.-S. (2025). imanip: Skill-incremental learning for robotic manipulation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 13890–13900). Available online: https://arxiv.org/abs/2503.07087 (accessed on).

Zhong, L., Liu, Y., Wei, Y., Xiong, Z., Yao, M., Liu, S., & Ren, G. (2026). ACoT-VLA: Action Chain-of-Thought for Vision-Language-Action Models. *arXiv preprint arXiv:2601.11404*. Available online: https://arxiv.org/abs/2601.11404 (accessed on).

Zhong, Y., Bai, F., Cai, S., Huang, X., Chen, Z., Zhang, X., Wang, Y., Guo, S., Guan, T., Lui, K. N., Qi, Z., Liang, Y., Chen, Y., & Yang, Y. (2025). *A survey on vision-language-action models: An action tokenization perspective.* Available online: https://arxiv.org/abs/2507.01925 (accessed on).

Zhong, Z., Yan, H., Li, J., Liu, X., Gong, X., Zhang, T., Song, W., Chen, J., Zheng, X., Wang, H., et al. (2025). Flowvla: Visual chain of thought-based motion reasoning for vision-language-action models. *arXiv preprint arXiv:2508.18269*. Available online: https://arxiv.org/abs/2508.18269 (accessed on).

Zhou, E., Su, Q., Chi, C., Zhang, Z., Wang, Z., Huang, T., Sheng, L., & Wang, H. (2025). Code-as-monitor: Constraint-aware visual programming for reactive and proactive robotic failure detection. In *Proceedings of the computer vision and pattern recognition conference* (pp. 6919–6929). Available online: https://arxiv.org/abs/2412.04455 (accessed on).

Zhou, H., Ma, C., & Lee, G. H. (2025). VLA-4D: Embedding 4D Awareness into Vision-Language-Action Models for SpatioTemporally Coherent Robotic Manipulation. *ArXiv*, *abs/2511.17199*.

Zhou, X., Xu, Y., Tie, G., Chen, Y., Zhang, G., Chu, D., Zhou, P., & Sun, L. (2025). LIBERO-PRO: Towards Robust and Fair Evaluation of Vision-Language-Action Models Beyond Memorization. *arXiv preprint arXiv:2510.03827*.

Zhou, Z., Atreya, P., Tan, Y. L., Pertsch, K., & Levine, S. (2025). AutoEval: Autonomous Evaluation of Generalist Robot Manipulation Policies in the Real World. *arXiv preprint arXiv:2503.24278*.

Zhu, C., Yu, R., Feng, S., Burchfiel, B., Shah, P., & Gupta, A. (2025). Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*. Available online: https://arxiv.org/abs/2504.02792 (accessed on).

Zhu, Y., Wong, J., Mandlekar, A., Martín-Martín, R., Joshi, A., Nasiriany, S., Zhu, Y., & Lin, K. (2020). robosuite: A Modular Simulation Framework and Benchmark for Robot Learning. In *arxiv preprint arxiv:2009.12293.*

Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al. (2023). Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on robot learning* (pp. 2165–2183).