A Survey of Robotic Learning for Perception and Manipulation: From Modular Pipelines to Robotic Foundation Models

Anonymous authors
Paper under double-blind review

Abstract

Over the past decade, robotic manipulation systems have undergone a fundamental paradigm shift: from carefully engineered hierarchical pipelines to data-driven foundation-model-based robotic policies. Following the 2015 DARPA Robotics Challenge, classical systems relied on decomposed perception-planning-control architectures with strong modeling assumptions and task-specific engineering. Since then, advances in machine learning, large-scale visual representation learning, and robot interaction data collection have enabled a progression toward imitation learning policies, end-to-end generative visuomotor policies, and, most recently, robotic foundation models capable of multi-task and cross-embodiment generalization.

This survey provides a structured perspective on this evolution from the viewpoint of robotic perception and manipulation. We introduce a taxonomy of manipulation systems organized along architectural transitions: *hierarchical pipelines, imitation-based policies, learning-based generative visuomotor policies, and robotic foundation models (e.g., VLAs)*, and analyze each paradigm in terms of system design, data requirements, and embodied intelligence capabilities such as compositionality, generalization, and adaptability. Beyond model architectures, we examine the scaling of data that underpins recent progress, covering developments in large-scale visual and 3D datasets, in-the-wild robot interaction corpora, and emerging multimodal sensing modalities including tactile and force feedback.

We further discuss emerging directions that integrate robotics foundation models with reinforcement learning and world models to enable online adaptation and long-horizon reasoning in physical environments. We review current benchmarks and evaluation protocols, highlighting limitations in measuring generalization, safety, and data efficiency, and conclude by outlining open challenges toward general-purpose embodied agents, including interaction-centric scaling, safety and alignment in physical deployment, multimodal perception integration, and the fusion of cognitive abstraction with physical reasoning.

By synthesizing architectural, data-centric, and systems-level trends, this survey aims to provide both a conceptual map of robotic learning's recent trajectory and a forward-looking agenda for advancing robotic manipulation toward truly general embodied intelligence.

Contents

1 Introduction

Robotic manipulation has rapidly evolved from engineered, modular pipelines to foundation-model-based robotic policies with increasingly unified perception, language grounding, and control. This survey provides a structured synthesis of that transition, with emphasis on how model scaling, data scaling, and interaction scaling jointly shape capability, robustness, and generalization in embodied systems.

1.1 A Decade of Transition: From DRC Systems to Foundation Robotics

The DARPA Robotics Challenge (DRC) Finals in 2015 are widely regarded as a watershed moment for modern field robotics Feng et al. (2015); Krotkov et al. (2018). Competing systems completed demanding tasks under severe uncertainty, including vehicle operation, door traversal, valve turning, wall cutting, and hose connection Kuindersma et al. (2016). These demonstrations established the practical value of tightly coupled perception, planning, and control.

At the same time, the DRC generation made explicit the central limitation of classical manipulation pipelines: performance was primarily bounded by engineering bandwidth. Representative systems, including Team KAIST's DRC-Hubo Lim et al. (2017), integrated hand-crafted perception modules, model-based planning, and carefully tuned controllers Johnson et al. (2015); Kim et al. (2011); Atkeson et al. (2015). Such designs were effective for known tasks and operating conditions, but transfer to new objects, scenes, or task specifications often required substantial redesign and retuning.

Over the past decade, robotic manipulation has shifted from module-centric engineering toward data-driven policy learning at scale. Influenced by foundation-model advances in language and vision Brown et al. (2020); Radford et al. (2021), robotics now increasingly adopts pre-train–adapt paradigms in which visual understanding, language grounding, and action generation are learned jointly from large and heterogeneous corpora Bommasani et al. (2021). This transition motivates a systematic survey focused not only on model families, but also on the scaling principles that govern capability gains in embodied systems.

Between these two eras, the field also experienced an important intermediate phase: the rapid rise of imitation learning and large-scale robot data collection. This phase established a practical bridge from hand-engineered pipelines to foundation-policy paradigms. On the one hand, behavior cloning and language-conditioned policies demonstrated that a single learned policy could absorb components previously implemented as separate perception and planning modules. On the other hand, emerging shared datasets and cross-lab benchmarks exposed persistent bottlenecks—distribution shift, weak long-horizon robustness, and limited cross-embodiment transfer, which could not be solved by architecture changes alone. These observations directly motivate the central position of this survey: progress in embodied manipulation should be interpreted as a co-evolution of *model capacity*, *data diversity*, and *interactive adaptation*, rather than as a linear replacement of one model family by another.

1.2 A Three-Axis Perspective on Embodied Scaling

This survey advocates a three-axis perspective to interpret progress in embodied manipulation (Figure 1). Rather than viewing recent advances as a simple sequence of model replacements, we treat capability growth as the result of coupled scaling in model capacity, data regimes, and interactive adaptation.

(1) Model Scaling. Model scaling captures the shift from task-specific modules and compact policies to integrated, high-capacity vision-language-action (VLA) architectures Zitkovich et al. (2023); Kim et al. (2024); Black et al. (2024). Recent lines of work further incorporate world modeling, including latent-dynamics world models and video-action world modeling with unified video–action pretraining Zhang et al. (2025g); Cen et al. (2025b); Zhu et al. (2025). Larger and more unified models improve semantic grounding, long-horizon action representation, and cross-task parameter sharing. However, this axis alone does not guarantee stable deployment performance when data coverage and adaptation mechanisms are insufficient.

(2) Data Scaling. Data scaling reflects the transition from small teleoperation collections Mandlekar et al. (2018) to large, heterogeneous corpora spanning institutions, embodiments, tasks, and sensing settings (e.g.,

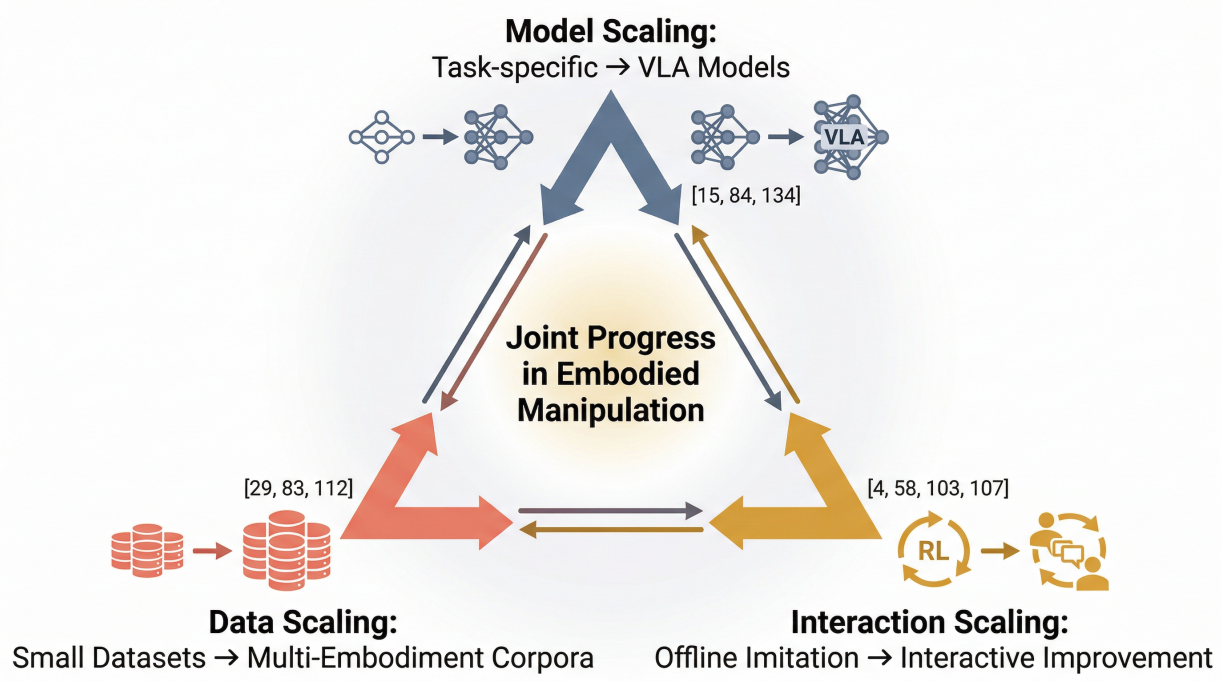

Figure 1: Three-axis view of embodied manipulation scaling used throughout this survey. The framework organizes progress along model scaling, data scaling, and interaction scaling, and emphasizes that robust general-purpose manipulation requires coordinated improvement across all three axes rather than architectural scaling alone.

Open X-Embodiment, DROID O'Neill et al. (2024); Khazatsky et al. (2024)). In manipulation, scaling is inherently multidimensional: trajectory volume, scene and object diversity, embodiment mismatch, contact complexity, temporal horizon, and modality richness jointly determine effective coverage. Consequently, "more data" is meaningful only when accompanied by diversity and consistency aligned with target deployment conditions.

(3) Interaction Scaling. Interaction scaling characterizes how policies improve beyond static supervision through online experience, including reinforcement learning, world-model-based rollouts, and continual adaptation Luo et al. (2025); Lu et al. (2025a); Zhang et al. (2025g); Cen et al. (2025a). In this survey, interaction scaling covers distinct regimes—test-time adaptation, online RL fine-tuning, and continual learning, with different stability and safety trade-offs. This axis is central to error correction, recovery behavior, and robustness under distribution shift, where purely offline policies typically underperform. In practical terms, interaction scaling determines whether a system can move from benchmark competence to persistent real-world reliability.

Taken together, the three axes provide the organizing principle of this survey: sustained progress in embodied manipulation arises from *coordinated scaling*, not from architectural innovation in isolation. This perspective guides our taxonomy, analysis of failure modes, and the research agenda presented in later sections.

1.3 Highlights of This Survey

Beyond enumerating representative methods, we emphasize unifying abstractions, cross-paradigm comparison, and deployment-relevant capability boundaries, and concretely highlighted as follows:

1. **A unified taxonomy spanning four manipulation paradigms.** We organize the field into a coherent progression—**Hierarchical → Imitation Learning → Generative Visuomotor Policies → Vision-Language-Action (VLA)**—while explicitly characterizing where these paradigms overlap

in practice rather than treating them as strictly disjoint categories. For each paradigm, we analyze architectural decomposition, supervision assumptions, data interfaces, and control abstractions, so that readers can compare methods on common axes instead of paper-specific metrics.

2. **A three-axis scaling framework for embodied manipulation.** We advocate a joint model–data–interaction scaling perspective and use it as the analytical backbone of the survey. This framework explains why architecturally similar systems can produce markedly different robustness and generalization outcomes, and why scaling parameters alone is insufficient without matching improvements in data diversity and online adaptation Kaplan et al. (2020); Hoffmann et al. (2022). We further connect this perspective to world-model-augmented policy learning, where prediction and control are increasingly co-designed.

3. **A capability-centric analysis of failure modes and bottlenecks.** Rather than evaluating progress only through headline benchmark scores, we analyze recurring failure mechanisms across paradigms: covariate shift and compounding error in behavior cloning Ross et al. (2011), underspecified contact dynamics in vision-dominant policies, weak recovery behavior under long-horizon perturbations, and safety/reliability gaps between benchmark competence and real-robot deployment. This analysis is used to clarify what current systems can do reliably, what they can do only in narrow settings, and where empirical claims remain fragile.

4. **A forward-looking agenda toward interaction-centric embodied intelligence.** We outline research directions that we view as most consequential for the next stage of progress: continual and test-time adaptation, world-model-guided planning and policy improvement, safety-aligned optimization under physical constraints, and scalable real-world data infrastructures that preserve annotation quality and embodiment coverage. A central theme is closing two persistent gaps simultaneously: *sim-to-real* transfer and *benchmark-to-deployment* reliability.

5. **A practical reading map for researchers and practitioners.** In addition to technical synthesis, we structure the survey as a decision-oriented map: which paradigm to favor under data scarcity, safety constraints, embodiment mismatch, or long-horizon task requirements; where hybrid designs are currently most effective; and which evaluation protocols are most informative for deployment planning. This practical orientation is intended to make the survey useful not only as a retrospective but also as an engineering guide for building and assessing next-generation manipulation systems.

Scope and Focus. We focus on **robotic manipulation and skill learning**, including single-arm manipulation, bimanual coordination, and dexterous/humanoid manipulation behaviors Billard & Kragic (2019); Kroemer et al. (2021). We do not attempt comprehensive coverage of navigation-only systems, social robotics, or conversational embodied agents, except when directly relevant to manipulation. Within this scope, we review the methodological trajectory from hierarchical pipelines to imitation learning, generative visuomotor policies, and VLA foundation models, together with associated developments in data infrastructure, online adaptation, and evaluation protocols.

Paper Organization. The remainder of this survey is organized as follows. Section 2 formalizes the taxonomy and comparison axes used throughout the paper. Sections 3–6 review the four paradigm families in depth: hierarchical vision-centric systems (Section 3), imitation learning and behavior cloning (Section 4), generative visuomotor policies (Section 5), and VLA foundation models (Section 6). Section 7 analyzes data scaling across perception and interaction sources. Section 8 discusses the emerging integration of VLAs with reinforcement learning and world models. Section 9 reviews benchmarks and evaluation methodology. Finally, Section 10 summarizes open challenges and outlines a roadmap toward robust, general-purpose embodied manipulation.

2 A Taxonomy of Manipulation Systems for Embodied Intelligence

The past decade of robotic manipulation research can be understood as a progression through four increasingly integrated architectural paradigms: *Hierarchical Vision-Centric Systems*, *Imitation Learning / Behavior*

Cloning, *Generative Visuomotor Policies*, and *Vision-Language-Action Foundation Models (VLA)*. Each paradigm reflects a distinct stance on how perception, language, planning, and control should be composed.

2.1 Four Paradigms: An Overview

Hierarchical vision-centric systems (Section 3). These systems follow a modular *sense–plan–act* pipeline widely used in classical robotics. A perception front-end (e.g., detection, 6D pose estimation, and scene graph construction) first produces a structured scene representation. A motion planner then generates feasible robot trajectories, and a low-level controller executes the planned trajectory while regulating contacts and dynamics. Perception modules have been dramatically upgraded by modern vision foundation models such as SAM Kirillov et al. (2023) and geometric reconstruction models like DUSt3R Wang et al. (2024b) and VGGT Wang et al. (2025b). However, the overall architecture remains largely *engineered*: each module is designed and validated independently, while task execution is coordinated through hand-designed skills or state machines.

Imitation learning and behaviour cloning (Section 4). Rather than designing each module explicitly, imitation learning (IL) formulates manipulation as a policy learning problem from demonstrations. A parametric policy $\pi_\theta(a \mid o, c)$ is trained to reproduce expert behaviour, where o denotes observations, a denotes robot actions, and c is an optional task specification such as a language instruction, goal image, or task identifier. Temporal abstraction via hierarchical policies Sutton et al. (1999); Lynch et al. (2020), as well as structured action representations such as discretised 3D voxel actions Shridhar et al. (2023) or multi-view transformer features Goyal et al. (2023), have significantly expanded the applicability of BC beyond simple single-task regression. Nevertheless, the performance of BC remains fundamentally limited by the *coverage and cost* of the demonstration distribution (Section 7).

Generative visuomotor policies (Section 5). Recent work extends imitation learning by adopting powerful generative models for policy learning. Instead of predicting a single next-step action, these approaches generate *action sequences* or trajectory segments using expressive sequence models. Representative examples include conditional diffusion policies Chi et al. (2025), transformer-based action chunking models such as ACT Zhao et al. (2023), and flow-matching policy generators Black et al. (2024). By modelling the multi-modal distribution over feasible actions, generative visuomotor policies mitigate compounding errors in long-horizon tasks and learn richer perception–action mappings. However, most existing approaches operate purely on visual observations and lack explicit language grounding or cross-embodiment generalisation.

Vision-language-action models (Section 6). Vision-language-action (VLA) models integrate visual perception, natural language understanding, and action generation within a unified foundation-model architecture. By jointly training on internet-scale vision–language corpora and large-scale robot demonstrations (e.g., Open X-Embodiment O'Neill et al. (2024)), VLAs such as RT-2 Zitkovich et al. (2023), OpenVLA Kim et al. (2024), and the π_0 family Black et al. (2024); Intelligence et al. (2025) acquire transferable semantic priors that enable multi-task and multi-embodiment generalisation from free-form language instructions. Emerging directions further combine VLAs with reinforcement learning and world models (Section 8), enabling online adaptation, planning, and constraint-aware policy refinement beyond purely offline pre-trained policies.

2.2 Three Comparison Axes

To compare paradigms on a common footing, we identify three axes—*architecture*, *data requirements*, and *embodied intelligence capabilities*—that together capture the design philosophy, practical constraints, and performance envelope of each approach. Figure 2 provides a side-by-side architectural reference for these paradigms, and we use it below to anchor the architecture axis comparison.

2.2.1 Architecture: Module Composition and End-to-End Integration

The four paradigms in Sec.2.1 span a spectrum from fully modular to fully integrated. Hierarchical systems explicitly separate perception, planning, and control, offering interpretability and the ability to inject

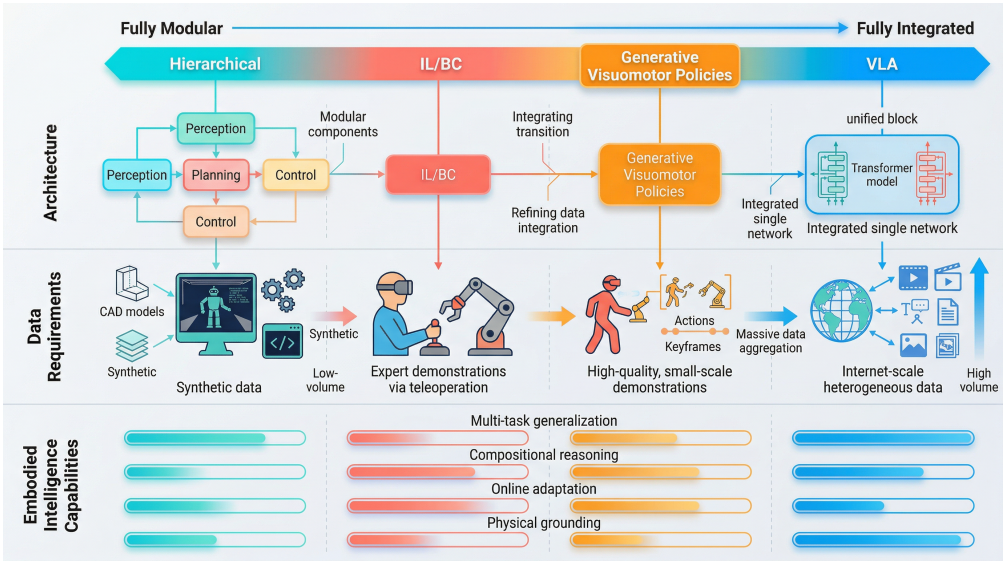

Figure 2: Illustration of the unified manipulation taxonomy used in this survey, highlighting four paradigm families and their relationships along the three comparison axes: architecture, data requirements, and embodied intelligence capabilities.

engineering priors (e.g., collision checking, impedance limits) at each interface. IL/BC collapses planning and control into a learned policy but typically still relies on a separately designed observation pipeline. Generative visuomotor policies push further toward end-to-end learning: a single network maps raw visual inputs and proprioception to action trajectories, implicitly acquiring both perceptual features and control strategies. VLA models represent the most integrated extreme, absorbing language, vision, and action into one generative model—often a large language model backbone augmented with visual encoders and action decoders.

A practical consequence of increasing integration is that the *locus of inductive bias shifts from architecture to data*. Hierarchical systems encode strong priors in module design (e.g., 6D pose as an intermediate representation); VLAs encode priors through the statistics of their pre-training corpora. This shift has profound implications for debugging, safety certification, and data efficiency, which we revisit in Section 10.

2.2.2 Data Requirements and Sources

Each paradigm is coupled to a distinct data regime, with different assumptions about data modality (structured labels vs. demonstrations vs. multimodal corpora), collection scale (task-specific to internet-scale), and transferability across tasks and embodiments; the following comparison makes these contrasts explicit.

- **Hierarchical:** CAD models, synthetic renders for pose estimation, hand-designed cost maps, and task-specific calibration data Wang et al. (2025c). The data volume per task is modest, but generalisation to new objects or scenes typically requires re-engineering.
- **IL/BC:** Expert demonstrations collected via teleoperation Mandlekar et al. (2018), kinesthetic teaching, play Lynch et al. (2020), or hardware-free handheld interfaces Chi et al. (2024). The performance ceiling scales with demonstration coverage (Section 7), and cross-embodiment transfer remains limited without careful action normalisation.
- **Generative visuomotor policies:** High-quality but typically small-scale demonstrations (tens to hundreds per task). Generative visuomotor policies compensate for limited data by leveraging powerful generative priors (diffusion, flow matching) that regularise the action space.
- **VLA:** Heterogeneous mixtures of internet-scale vision–language corpora and large-scale robot trajectories (e.g., >1M episodes in OXE O'Neill et al. (2024), >10k hours in π_0 Black et al. (2024)).

Cross-embodiment, cross-institution datasets Ebert et al. (2021); Walke et al. (2023); Khazatsky et al. (2024) and simulation-augmented data Mu et al. (2025); Chen et al. (2025c) are essential to fill coverage gaps. Here, cross-embodiment denotes transfer across robots with different morphologies, kinematics, and control interfaces.

A key trend across the paradigms is the progressive *democratisation of data sourcing*: from lab-only CAD and teleop, through crowdsourced and play-based collection, to internet-scale pre-training supplemented by in-the-wild human demonstrations Chi et al. (2024); Team (2025).

2.2.3 Embodied Intelligence Capabilities

We assess each paradigm along four capability dimensions that collectively characterise the degree of embodied intelligence a system can exhibit:

1. **Multi-task and cross-environment generalisation.** Can the system perform diverse tasks across varied scenes without per-task re-engineering or re-training?

2. **Compositional skill reasoning.** Can the system combine previously learned primitives to solve novel, long-horizon tasks specified by language or goals?

3. **Online adaptation.** Can the system improve through interaction—correcting errors, exploring, and adapting to distribution shifts encountered during deployment?

4. **Physical-world grounding.** Does the system exploit physical dynamics (contact, compliance, material properties) beyond what is captured by vision alone?

Table 1 rates each paradigm along these dimensions. In short, hierarchical systems offer strong physical grounding (through explicit dynamics models) but weak generalisation; IL/BC achieves moderate multi-task coverage but suffers from covariate shift; generative visuomotor policies improve long-horizon robustness through generative action modelling; and VLAs provide the broadest generalisation envelope, though their online adaptation and contact/dynamics grounding remain active research frontiers (Section 8).

2.3 A Timeline of Manipulation Systems: 2015 to 2025

Figure 3 places representative systems on a timeline to highlight how the field has evolved from the engineering-heavy approaches of the DRC era to today's foundation-model-driven VLAs.

Several observations emerge from this timeline. First, the four paradigms are not strictly sequential; they coexist and continue to develop in parallel. Hierarchical methods, for instance, have been substantially modernised by foundation vision models (SAM Kirillov et al. (2023), DUSt3R Wang et al. (2024b), VGGT Wang et al. (2025b)) even as VLAs have risen to prominence. Second, the transition from IL to generative visuomotor policies occurred around 2022 to 2023, marked by Diffusion Policy and ACT. This shift was driven less by architectural novelty and more by a key insight: *generative modelling of the action space* can mitigate critical failure modes of regression-based BC, especially mode averaging and compounding errors. Third, the emergence of VLAs (since 2023) coincides with, and is enabled by, the rapid scaling of both robot demonstration data (OXE, DROID, BridgeData V2) and pre-trained vision-language backbones (CLIP Radford et al. (2021), LLaVA Liu et al. (2023b), LLaMA Touvron et al. (2023)).

Table 1: Comparison of four manipulation paradigms along architecture, data, and embodied intelligence axes. Capability ratings: +++ strong, ++ moderate, + limited, – largely absent. Representative systems are illustrative, not exhaustive.

Paradigm	**Rep. Systems**	**Typical Data**	**Multi-task Gen.**	**Comp. Reason.**	**Online Adapt.**	**Ground.**
Hierarchical	MegaPose Labbé et al. (2022), FoundationPose Wen et al. (2024a), VoxPoser Huang et al. (2023), Code-as-Policies Liang et al. (2023)	CAD models, synthetic renders, task-specific calibration	+	++	–	+++
Imitation Learning / BC	Transporter, CLIPort, PerAct, RVT, BC-Z Zeng et al. (2021); Shridhar et al. (2022; 2023); Goyal et al. (2023); Jang et al. (2022)	Expert demos (teleop, play, handheld); 100s–10ks trajectories	++	+	+[†]	+
Generative Visuomotor Policies	Diffusion Policy Chi et al. (2025), ACT Zhao et al. (2023)	Small-scale high-quality demos; 10s–100s per task	+	+	–	++
Vision-Language-Action (VLA)	RT-1/RT-2 Brohan et al. (2022); Zitkovich et al. (2023), OpenVLA Kim et al. (2024), $\pi_0/\pi_{0.5}$ Black et al. (2024); Intelligence et al. (2025), GR00T N1 Bjorck et al. (2025)	Internet V-L data + large-scale robot trajectories; >1M episodes	+++	++	++[‡]	+

[†]Via interactive IL (e.g., DAgger-style interventions Ross et al. (2011), BC-Z Jang et al. (2022)). [‡]When combined with RL and/or world models (Section 8).

Beyond chronology, Figure 3 highlights a structural shift in how progress is achieved. In 2015–2021, gains are dominated by algorithmic and interface innovations under limited data regimes (modular pipelines, then IL/BC). In 2022–2025, performance improvements increasingly track *joint scaling* of model capacity and heterogeneous data (multi-lab robot trajectories plus internet-scale vision–language pre-training), enabling the emergence of generative policies and VLAs. The figure therefore supports the central thesis of this survey: modern embodied manipulation advances are driven by coordinated scaling across architecture and data, not by isolated architectural changes alone.

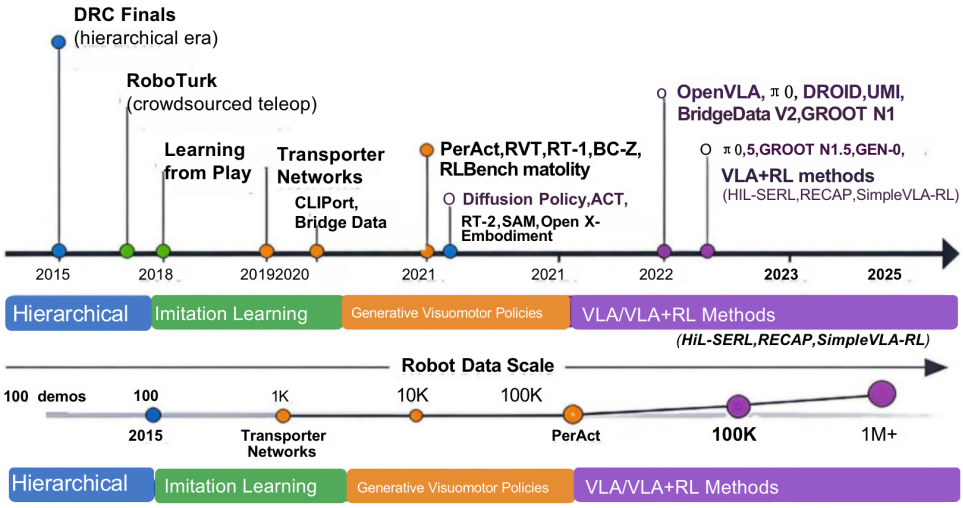

Figure 3: Timeline of representative manipulation systems from 2015 to 2025. Milestones are colour-coded by paradigm (Hierarchical, IL/BC, Generative Visuomotor Policies, VLA) and positioned by first public release/publication year, highlighting substantial temporal overlap rather than a strict sequential handoff. The lower axis reports the approximate growth of robot demonstration data scale (from hundreds to millions of trajectories), linking paradigm shifts to data scaling. Representative labels are illustrative rather than exhaustive.

3 Hierarchical Vision-Centric Manipulation Systems

Hierarchical vision-centric manipulation systems follow a modular *sense–plan–act* design in which perception estimates task-relevant geometry, planning maps that state to feasible trajectories, and control executes motion while regulating contact. Although this architecture predates foundation-model robotics, it remains a strong baseline in safety-critical and data-sparse settings because each stage is interpretable and can incorporate explicit physical constraints. A canonical perception stack comprises object discovery (detection or promptable segmentation), 6D pose estimation for rigid objects and/or part-level estimation for articulated objects, and scene-level 3D representations (e.g., scene graphs or occupancy maps) consumed by downstream planners. For rigid object pose, modern pipelines often rely on benchmarked object-pose protocols such as BOP Hodan et al. (2018) and leverage large-scale synthetic rendering plus test-time alignment.

On the other hand, recent advances in foundation models and large-scale synthetic data have reshaped hierarchical perception pipelines, shifting them from specialized instance-level training toward generalizable open-world perception; accordingly, this subsection examines how perception-side scaling progresses from promptable masking to task-driven grounding that directly supports downstream planning and manipulation.

3.1 Foundations of Rigid Object Perception

The perception front-end has been revolutionized by promptable segmentation. Segment Anything (SAM) Kirillov et al. (2023) enables zero-shot transfer across domains via promptable masks, while SAM 2 Ravi et al. (2024) extends this to video with streaming memory. Complementing segmentation, Grounded-SAM Ren et al. (2024) further bridges the gap between natural language descriptions and instance masks, effectively solving the open-set "object discovery" bottleneck.

Once isolated, object-centric manipulation relies on accurate 6D pose estimation. Modern approaches leverage large-scale synthetic corpora and *render-and-compare* strategies. MegaPose Labbé et al. (2022) aligns novel objects against CAD models via iterative refinement. FoundationPose Wen et al. (2024a) further unifies tracking and estimation for novel objects, supporting both model-based (CAD) and model-free (reference

views) onboarding. More recently, generative approaches such as PoseDiffusion Wang et al. (2023) and GenPose Zhang et al. (2023) have formulated pose estimation as a diffusion process, utilizing the probabilistic nature of diffusion models to handle pose ambiguity and distribution shifts more robustly than deterministic regression.

3.2 Open-Set Generalization and Robustness to Occlusion

In open-world scenarios, precise CAD models are often unavailable, and scenes are cluttered. This necessitates perception that handles category-level variations and severe occlusions.

Category-Level and Template-Free Understanding. To handle objects without instance-specific models, SAR-Net Lin et al. (2022a) performs category-level pose and size estimation by aligning observations against a category template. Similarly, Tax-Pose Pan et al. (2023) explores task-specific cross-category alignment, learning to associate geometric features between anchor objects and novel instances to transfer manipulation skills. Pushing further towards category-agnostic perception, Beyond Templates Zhang et al. (2025e) predicts pose, size, and dense shape from single-view RGB-D without any templates or category labels, utilizing a Transformer encoder to fuse 2D foundation features with partial point clouds.

Handling Occlusion via Amodal Perception. A critical challenge in cluttered scenes is grasping under occlusion, where visible masks are insufficient. While UOAIS Back et al. (2022) pioneered amodal instance segmentation for unseen objects, LAC-Net Zhang et al. (2024a) explicitly addresses the grasping downstream task. It employs a Linear-Fusion Attention-Guided Convolutional Network to recover the full object mask—including occluded regions—from RGB-D data. By hallucinating the complete shape, LAC-Net allows the robotic system to identify stable grasp configurations that would be missed by methods relying solely on visible surfaces.

3.3 Beyond Rigid Bodies: Articulated Parts and Deformables

Everyday manipulation extends beyond rigid pick-and-place to interacting with articulated mechanisms and deformable materials.

Part-Level State Estimation. Tasks like opening drawers or turning knobs require kinematic awareness. GAPartNet Geng et al. (2023) defines Generalizable and Actionable Parts (GAParts) and provides large-scale supervision for part segmentation and pose. CAP-Net Huang et al. (2025) complements this by jointly predicting instance segmentation and Normalized Part Coordinate Space (NPCS) to recover 6D part poses. Alternatively, flow-based approaches like FlowBot3D Eisner et al. (2022) avoid explicit part pose estimation entirely, instead predicting per-point 3D articulated flow to identify actuation directions directly from point clouds.

Perceiving Deformables. For fluids and non-rigid items, state estimation becomes more abstract. PourIt! Lin et al. (2023) demonstrates weakly-supervised liquid perception by combining Class Activation Maps (CAM) with container pose to approximate 3D liquid volume. In the domain of granular and elastoplastic materials, DiffSkill Lin et al. (2022b) and subsequent differentiable physics-based approaches highlight the importance of inferring physical parameters (e.g., stiffness, friction) alongside geometry for closed-loop control.

3.4 Scene-Level Geometry and Task Specification

Finally, the gap between raw perception and planning is being bridged by 3D foundation models and language-driven task interfaces.

3D Geometry and Scene Representation. Feed-forward models like DUSt3R Wang et al. (2024b) and VGGT Wang et al. (2025b) are replacing traditional SfM pipelines, enabling fast generation of cost maps. To support long-horizon semantic planning, ConceptGraphs Gu et al. (2024) and F3RM Shen et al. (2023) construct open-vocabulary 3D scene graphs and feature fields, allowing robots to query objects by natural

language properties (e.g., "the mug next to the laptop") rather than just geometric class IDs. To overcome limitations in open-vocabulary understanding, ReasonGrounder Liu et al. (2025c) and SpatialReasoner Liu et al. (2025d) subsequent neural representation frameworks utilize LVLM-guided hierarchical feature splatting and LLM-driven spatial reasoning to achieve precise 3D visual grounding and reasoning in complex, unstructured environments.

Language-Driven Task Grounding. Perception is increasingly used to *specify* objectives rather than just state. VoxPoser Huang et al. (2023) maps language instructions to 3D voxel value maps (affordances/constraints), serving as objective functions for motion planners. This aligns with the broader trend of LLM-based planning, as seen in Code as Policies Liang et al. (2023), which generates executable code primitives from language. Similarly, Polaris Wang et al. (2024c) integrates LLMs with a Syn2Real pipeline, using synthetic data generation to ground user intents into actionable geometric plans.

4 Imitation Learning and Behavior Cloning

This section examines the transition from module-engineered manipulation stacks to demonstration-driven policy learning. In hierarchical systems, perception, planning, and control are designed and tuned as separate components; while interpretable, this design paradigm scales poorly under open-world variation in object geometry, contact dynamics, and scene composition. **Imitation Learning (IL)**, particularly **Behavior Cloning (BC)**, addresses this bottleneck by formulating manipulation as supervised policy learning from expert trajectories. Under this view, a robot learns a closed-loop mapping from observations to actions without explicit reward design, enabling practical skill acquisition while avoiding unsafe trial-and-error exploration in early training stages.

Figure 4 serves as a roadmap for this section. We first formalize BC as conditional policy learning and discuss flat versus hierarchical control structures; we then trace how representation design evolved across Transporter, CLIPort, PerAct, RVT, BC-Z, and RT-1/RT-2. Next, we examine how data-collection interfaces determine the effective coverage and scalability of IL/BC, and finally summarize the resulting trade-offs in sample efficiency, deployment practicality, and robustness under distribution shift.

4.1 Standard BC Framework: From Action Regression to Closed-Loop Policy

Problem formulation. Let a demonstration trajectory be $\tau = \{(o_t, a_t)\}_{t=0}^{T}$, where observations o_t may include multi-view RGB(-D), proprioception, and optionally a task descriptor c (e.g., language embedding or goal image). BC fits a parametric policy $\pi_\theta(a \mid o, c)$ to approximate an expert $\pi^\star$ via maximum-likelihood:

$$\theta^\star = \arg\min_\theta \; \mathbb{E}_{(o,a,c)\sim\mathcal{D}}\big[-\log \pi_\theta(a \mid o, c)\big], \tag{1}$$

implemented as regression (continuous actions) or classification (discretized actions).

Single-task and multi-task BC (conditional policies). Single-task BC trains an individual policy $\pi_{\theta_i}(a \mid o)$ per task i, which scales poorly and wastes shared structure. Multi-task BC instead trains a unified conditional policy $\pi_\theta(a \mid o, c)$, where c specifies "what to do" (e.g., language instruction, goal image, or one-hot task ID). This better matches embodied intelligence goals such as skill reuse and composition, but raises the bar for representation learning: the policy must jointly ground semantics (what), localize geometry (where), and execute control (how), and its generalization becomes tightly coupled to **data coverage** (Section 4.3).

Flat vs. hierarchical policies (temporal abstraction). For long-horizon manipulation, a common and practically effective design is to introduce **temporal abstraction**: a high-level policy selects a subtask or skill, while a low-level policy executes motor control to accomplish it. Formally, using the options framework Sutton et al. (1999), let an option (skill) be $z \in \mathcal{Z}$ with an intra-option policy $\pi_{\text{lo}}(a \mid o, z)$ and termination $\beta(o, z)$. A high-level policy $\pi_{\text{hi}}(z \mid o, c)$ selects z at option boundaries $\{t_k\}_{k=0}^{K}$, and the option runs for

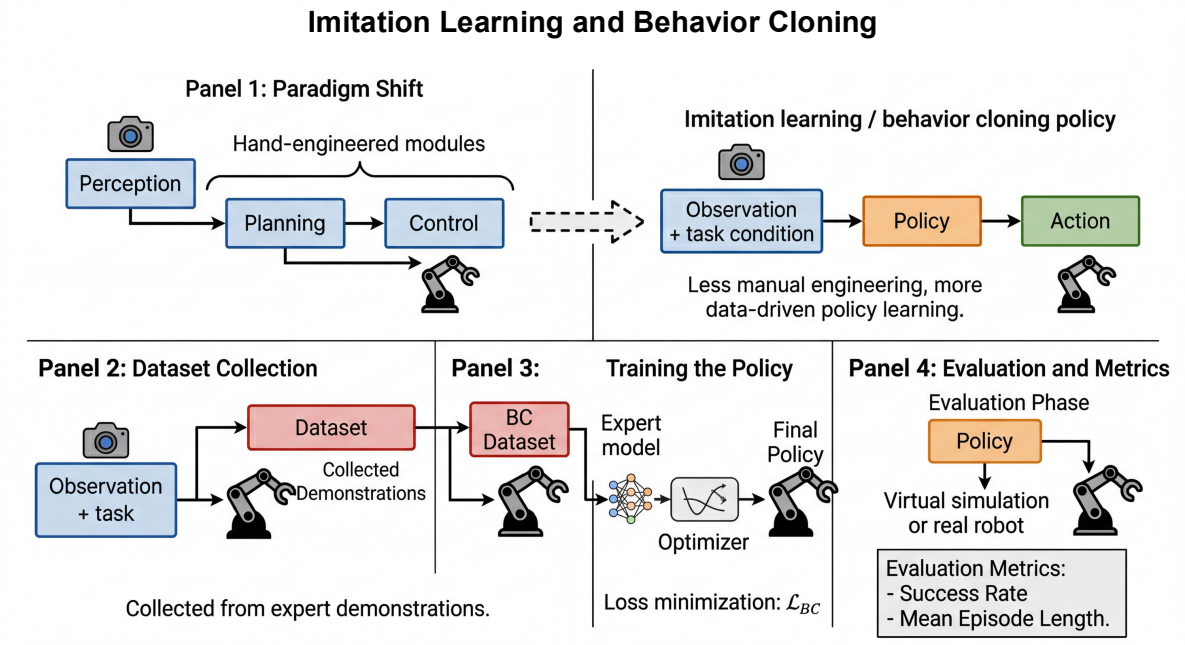

Figure 4: Overview of imitation learning and behavior cloning for robotic manipulation. The figure organizes this section into four connected layers: (i) standard BC formulation and policy structures, (ii) representative action-representation advances, (iii) data-scaling interfaces from teleoperation to in-the-wild acquisition, and (iv) practical strengths and limitations that motivate the transition to generative visuomotor policies.

$H_k = t_{k+1} - t_k$ steps until termination. The induced *flat* action distribution can be written as a mixture:

$$\pi(a_t \mid o_t, c) = \sum_{z \in \mathcal{Z}} \underbrace{\mu_t(z \mid o_t, c)}_{\text{option occupancy}} \; \pi_{\text{lo}}(a_t \mid o_t, z), \tag{2}$$

where μ_t is determined by π_{hi} and β. A standard hierarchical BC objective maximizes the (possibly latent) joint likelihood

$$\max_{\theta} \; \log p_\theta(\tau \mid c) \quad \text{with} \quad p_\theta(\tau \mid c) = \prod_{k=0}^{K-1} \pi_{\text{hi}}(z_k \mid o_{t_k}, c) \prod_{t=t_k}^{t_{k+1}-1} \pi_{\text{lo}}(a_t \mid o_t, z_k)\,, \tag{3}$$

where option boundaries t_k and assignments z_k can be (i) predefined via subtask annotations, (ii) discovered as latent variables (e.g., EM-style), or (iii) implemented implicitly by predicting discrete subgoals.

Why hierarchy helps. Compared to a flat policy $\pi_{\text{flat}}(a_t \mid o_t, c)$, hierarchy shortens the effective control horizon: the low-level controller only needs to solve H_k-step subproblems, while the high-level policy reasons over $K \approx T/\bar{H}$ decisions. This reduces compounding errors, enables modular reuse of low-level skills across tasks, and allows injecting safety and feasibility constraints at the execution layer (e.g., motion planning for a predicted waypoint). Notably, many "IL" systems in robotics are *hybrid*: learned high-level decisions and classical planners at low-level, which provides robustness in the low-data regime and motivates discretized action designs in PerAct Shridhar et al. (2023) and RVT Goyal et al. (2023).

4.2 Representative Works: Evolution of Action Representations

A key theme in modern IL is the progressive redesign of action representations to encode stronger geometric, semantic, and temporal inductive biases, thereby improving sample efficiency, robustness, and transfer across tasks. The representative works below can be read as a step-by-step evolution—from 2D spatial formulations,

to language-conditioned and 3D-aware discretized actions, to multi-view transformer policies and large-scale data-mixing paradigms—showing how action-space design has become a central lever for scaling BC beyond single-task settings.

(1) Transporter Networks: 2D spatial actions. Transporter Networks Zeng et al. (2021) formulate manipulation as predicting spatial displacements (pick-and-place) by correlating feature maps, achieving strong sample efficiency and closed-loop behavior on a range of tabletop tasks.

(2) CLIPort: language-conditioned "what & where" pathways. CLIPort Shridhar et al. (2022) couples a semantic pathway (leveraging vision-language pretraining) with a spatial pathway (transporter-style precision), providing a strong baseline for language-conditioned manipulation.

(3) PerAct: discretized 3D voxel actions. PerAct Shridhar et al. (2023) voxelizes RGB-D observations and discretizes 6-DoF actions into classification over translation, rotation, and gripper state, using a transformer (Perceiver-style) to learn multi-task language-conditioned policies from few demonstrations.

(4) RVT: multi-view transformers without explicit voxels. RVT Goyal et al. (2023) avoids voxelization cost by aggregating multi-view image features via attention and virtual re-rendering, retaining 3D reasoning while improving scalability and training/inference efficiency.

(5) BC-Z: interactive supervision (demonstrations + interventions). BC-Z Jang et al. (2022) shifts emphasis from architecture to *supervision modalities*: it learns from both demonstrations and human interventions, and can be conditioned on language or human video embeddings. Crucially, when scaling data collection to 100+ tasks, BC-Z reports non-trivial zero-shot success on unseen tasks.

(6) RT-1 / RT-2: data mixing as a bridge from IL to VLA. RT-1 Brohan et al. (2022) demonstrates that a transformer policy can *absorb heterogeneous data sources* (including simulation and different robot morphologies), often improving generalization when mixing datasets across embodiments. RT-2 Zitkovich et al. (2023) takes a further step toward VLA: it *co-fine-tunes* a vision-language model on both robot trajectories and internet-scale VLM tasks, representing robot actions as text tokens to unify training as next-token prediction (equivalent to BC loss). This establishes a clear "IL $\rightarrow$ (generative visuomotor policies / VLA)" motif: mixing **robot action data** with **web-scale semantic priors** boosts generalization beyond what robot-only BC typically achieves.

4.3 Data Scaling: From Teleoperation to "In-the-Wild" Acquisition

A practical viewpoint is:

> *For IL/BC, the performance ceiling is often set less by the learning algorithm than by the coverage radius and cost curve of demonstrations.*

This is why IL/BC is uniquely sensitive to the **data collection interface**: the interface shapes who can contribute data, how diverse the behaviors are, and whether the resulting trajectories are robot-embodiment-specific. Table 2 summarizes representative collection paradigms and their trade-offs.

(1) Remote teleoperation and crowdsourcing. RoboTurk Mandlekar et al. (2018) uses widely available mobile devices (e.g., smartphones) to enable remote 6-DoF teleoperation and crowdsourced data collection, reporting a pilot dataset of 137.5 hours and 2200+ successful demonstrations. Its key contribution is lowering the barrier from specialized hardware to commodity devices.

(2) Learning from play. Learning from Play Lynch et al. (2020) argues that unscripted play is cheaper (less reset or labeling) and covers a substantially wider interaction space, which helps learn robust behaviors including retries and recovery. This reframes "data scaling" as increasing *state-action coverage*, not just expert optimality.

Acq.	Cost	Data quality	Scale	Representative
Kinesthetic / VR teleop	High	High precision; low latency	Low	Lab teleop
Smartphone remote teleop	Low	Closed-loop; operator variance	High	RoboTurk Mandlekar et al. (2018)
Play	Low	Broad coverage; recovery-rich	Very high	Learning from Play Lynch et al. (2020)
Handheld interface	Low	Relative trajectories; cross-embodiment	High	UMI Chi et al. (2024), MV-UMI Rayyan et al. (2025)
Portable capture	Med.	Contact + dexterity signals	Med.	DexCap Wang et al. (2024a), FARM Helmut et al. (2025)

Table 2: Demonstration acquisition paradigms for IL/BC. *Note:* FARM denotes *Force-Aware Robotic Manipulation*.

(3) Hardware-free and multimodal acquisition. UMI Chi et al. (2024) introduces "robot-less" teaching: a handheld gripper with cameras records demonstrations in the wild, and uses tracking/SLAM-style estimation to recover relative 6D trajectories that can be transferred across robot platforms. MV-UMI Rayyan et al. (2025) augments egocentric wrist views with third-person perspective to mitigate limited scene context and improve performance on tasks requiring broader understanding. For dexterous manipulation, DexCap Wang et al. (2024a) provides portable wrist or finger motion capture (SLAM with additional sensing) together with 3D scene observations, supporting direct learning of dexterous skills from human capture.

(4) Vision with tactile/force: parsing contact-rich demonstrations. A key limitation of vision-only demonstrations is ambiguity around contact states (incipient slip, insertion force, compliance). Recent work demonstrates integrating tactile sensors (e.g., GelSight-like vision-based tactile) into handheld interfaces to collect demonstrations, then aligning high-frequency tactile streams with visual frames and mapping them into a force-aware action space for imitation learning Helmut et al. (2025). This suggests a concrete path for scaling *contact-rich* IL: better interfaces with synchronized multimodal parsing.

4.4 Advantages and Limitations of IL/BC

4.4.1 Advantages: Why IL/BC is effective in robotics

IL/BC is particularly attractive in real-world manipulation because it aligns with practical deployment constraints: data can be collected directly from human demonstrations, policies can be trained without delicate reward engineering, and the learned controller can be combined with classical safety-critical execution layers; the following points summarize these core strengths.

- **High sample efficiency (no reward engineering, minimal exploration).** BC avoids reward design and costly trial-and-error, which is crucial on real robots where failures are expensive and unsafe.
- **Fast deployment and hybridization with safety layers.** BC policies can be integrated with classical execution layers by predicting structured subgoals or discretized actions, combining data-driven perception with safety and feasibility constraints in execution. This is especially compatible with hierarchical designs (Section 4.1).

4.4.2 Limitations: Why covariate shift is more severe in manipulation

- **Covariate shift → compounding error.** BC minimizes one-step prediction error on expert states, but at test time the policy induces its own state distribution. In sequential decision-making, small errors can drive the robot into unseen states (contact-rich, partially observed dynamics), where errors compound. Theoretical analyses show that naive BC can incur error that grows quadratically with horizon, whereas interactive aggregation (DAgger) can reduce it to linear growth under standard assumptions Ross et al. (2011). *Robotics makes this worse* because contacts create discontinuities: a

millimeter pose error can cause a failed grasp or jammed insertion, often without an easy recovery state in the training distribution.

Typical fixes and their robotics-friendly variants. DAgger queries expert labels on learner-visited states to mitigate distribution shift Ross et al. (2011). In practice, robotics often uses *intervention-style* interactive IL: humans only correct when the robot is about to fail (lower cognitive burden), which aligns with BC-Z's demonstrations+interventions paradigm Jang et al. (2022).

- **Multi-modality and long-horizon: "averaging" hurts.** Many tasks admit multiple valid strategies (left vs. right grasp, push-then-grasp vs. grasp-then-push). Regression-based BC tends to average modes, producing unstable actions. Long horizons further amplify this issue as single-step errors accumulate. This directly motivates Section 5 (generative visuomotor policies): sequence modeling and generative policies (action chunks, diffusion) represent multi-modal trajectory distributions, reducing mode-averaging and improving robustness.

- **Generalization depends on data coverage (open-world long tail).** BC generalization is often driven by whether the dataset has seen sufficient variation. BC-Z explicitly emphasizes scaling and broadening data collection; when scaling to 100+ tasks, it reports meaningful zero-shot performance on unseen tasks Jang et al. (2022). However, open-world generalization remains limited by long-tail objects or scenes and rare failure modes, which is a key gap that motivates VLA+RL and world models later in the survey.

4.5 Summary

In this section, we reviewed IL/BC as a shift from hand-engineered perception, planning, and control pipelines to data-driven closed-loop policies learned from demonstrations. We emphasized three practical axes that shape modern BC systems. First, conditioning enables multi-task skill reuse through $\pi(a \mid o, c)$. Second, temporal abstraction supports long-horizon manipulation by decomposing behavior into reusable skills or options with hierarchical policies. Third, data collection interfaces determine the cost–coverage curve of demonstrations and often set the performance ceiling of BC. Despite these advantages, BC remains limited by distribution shift in contact-rich and partially observed manipulation, by mode averaging in multi-modal settings, and by reduced stability over long horizons in open-world environments. These limitations motivate the next stage, learning-based generative visuomotor policies. Rather than predicting a single next action, generative visuomotor policies predict action chunks or trajectories using sequence or generative modeling, which improves long-horizon robustness and better captures multi-modal behavior.

5 Learning-based Generative Visuomotor Policies

This section reviews learning-based generative visuomotor policies, which form the execution-centric layer between perception and control in modern embodied systems. Compared with classical sense–plan–act pipelines, these policies directly map high-dimensional observations to low-level action sequences, reducing dependence on manually designed state estimators and planner interfaces. The central technical challenge is to model the multimodal conditional action distribution, $P(a \mid o)$, under visual uncertainty, contact variability, and long-horizon compounding error. Recent generative visuomotor policy architectures address this challenge with autoregressive, diffusion, and flow-based action modeling, action chunking, and geometry-aware representations, yielding stronger temporal consistency and broader cross-task transfer than deterministic regression policies.

Figure 5 provides a roadmap for this section. We first define the common policy formulation and shared design principles, then trace the architectural evolution from diffusion to more efficient flow-based generation, discuss geometry-aware representations and action-space design for robust manipulation, and finally position emerging RL/world-model integrations as the path from offline imitation toward adaptive long-horizon control.

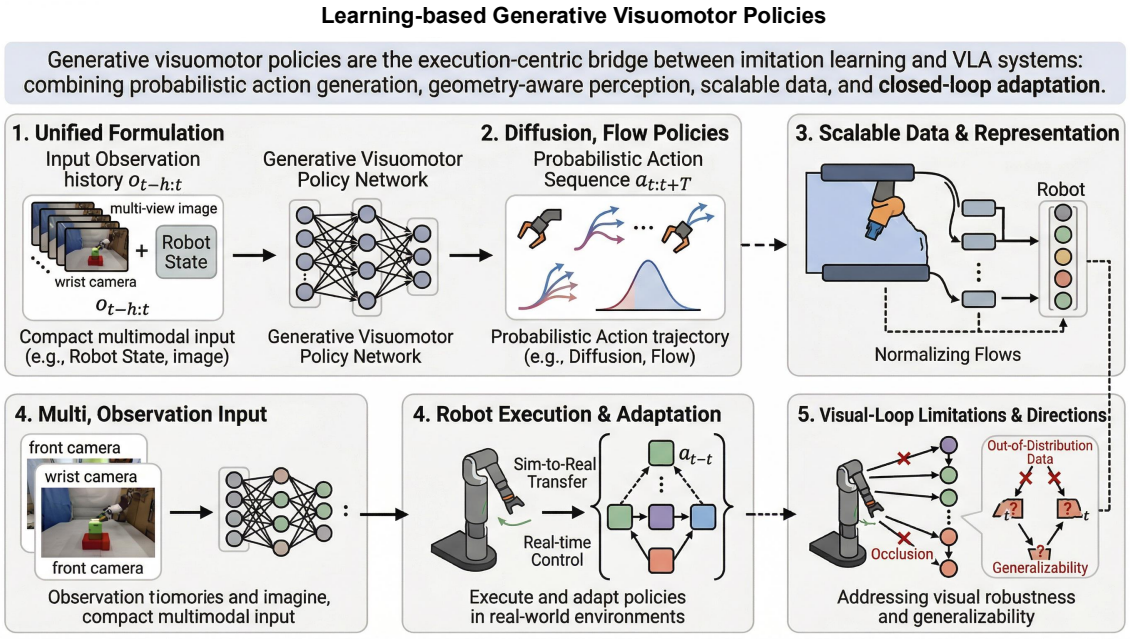

Figure 5: Overview of learning-based generative visuomotor policies. The figure summarizes the main progression covered in this section: unified visuomotor policy formulation, generative modeling evolution (autoregressive, diffusion, and flow-based methods), geometry- and representation-aware policy design, and closed-loop extensions that integrate reinforcement learning and world models for long-horizon adaptation.

5.1 Definition and Commonality

Despite architectural variations ranging from CNN-based behavior cloning to transformer-based diffusion models, modern generative visuomotor policies converge on a unified mathematical objective: approximating the optimal policy $\pi(a_{t:t+k}|o_{t-h:t})$. This function maps a history of high-dimensional visual observations o directly to a sequence of future actions a, effectively bypassing brittle explicit state estimation and kinematic solving in traditional sense–plan–act pipelines.

Multimodal Perception and State Encoding. The observation space $\mathcal{O}$ is primarily defined by visual sensory data, often augmented with proprioceptive state to handle non-Markovian dynamics. While early approaches relied on standard 2D backbones, recent architectures have shifted decisively towards geometry-aware representations to resolve spatial ambiguities and occlusion. For instance, *RoboUniView* Liu et al. (2024a) fuses multi-camera inputs into a unified representation, whereas *BridgeVLA* Li et al. (2025b) projects 3D point clouds into aligned 2D images to leverage pre-trained 2D backbones. Going further, models like *PointVLA* Li et al. (2026) and *3D Diffuser Actor* Ke et al. (2024) inject 3D point cloud data directly into the policy, decoupling geometric understanding from camera viewpoints. This is further extended by *4D-VLA* Zhang et al. (2025c), which incorporates temporal depth information to enable spatiotemporal reasoning across frames. Temporal and spatial consistency in generative modeling is further refined by DP4 Liu et al. (2025b) , which introduces a spatial-temporal aware visuomotor diffusion policy to explicitly capture the coupled dependencies within expert demonstrations. Complementing these visual inputs, compact pre-trained representations such as *R3M* Nair et al. (2022) are frequently employed to encode the robot's physical state efficiently.

Action Chunking and Universal Spaces. A critical innovation in the output space is the transition from single-step actuation to **Action Chunking**—predicting a sequence of future actions (e.g., k steps) simultaneously. As demonstrated by *Mobile ALOHA* Fu et al. (2024) and *Any-point Trajectory Modeling* Wen et al. (2023), this trajectory-level prediction significantly improves temporal consistency and mitigates the

compounding errors inherent in autoregressive execution. Beyond standard joint positions or Cartesian poses, research is expanding into generalized action spaces to support diverse embodiments. *UniAct* Zheng et al. (2025) proposes a unified latent action space to handle heterogeneous robot kinematics, while approaches like *Pixel Motion* Ranasinghe et al. (2025) and *Gripper Keypose* Yang et al. (2025b) map actions to visual flow or key-point trajectories, thereby grounding control directly in the visual domain rather than specific robot hardware.

Generative Training Regimes. The learning objective has evolved from simple Mean Squared Error (MSE) regression to sophisticated distribution matching. Traditional Behavior Cloning often fails on multimodal expert data (e.g., multiple valid ways to grasp an object), leading to averaging artifacts. To address this, *Diffusion Policy* Chi et al. (2025) formulates imitation as a conditional generative process, learning the gradient of the data distribution to represent distinct, valid modes of behavior. Furthermore, inspired by foundation models, the field is moving towards large-scale pre-training on diverse datasets like *Open X-Embodiment* O'Neill et al. (2024) and *DROID* Khazatsky et al. (2024). Approaches such as *Unified World Models* Zhu et al. (2025) and *Video2Policy* Ye et al. (2025) leverage massive internet video data to learn generalizable physical priors, effectively treating robotic control as a next-token prediction problem in a physics-grounded latent space.

5.2 Generative Evolution: From Diffusion to Flow Matching

While Diffusion Models have established a robust baseline for visuomotor control, recent research has pivoted towards optimizing inference efficiency and structural compositionality. This evolution is characterized by a shift from stochastic denoising to deterministic flow matching, and from monolithic generation to hierarchical composition.

From Denoising to Flow Matching. Although diffusion models excel in capturing multimodal distributions, their reliance on stochastic differential equations (SDEs) necessitates iterative sampling, which incurs high computational costs. To address this, **Flow Matching (FM)** has emerged as a superior alternative for continuous control. Unlike diffusion, FM learns a deterministic vector field modeled by Ordinary Differential Equations (ODEs), transforming a noise distribution to the data distribution along an optimal transport path. As demonstrated by *ManiFlow* Yan et al. (2025) and *VITA* Dong et al. (2025), enforcing straight-line trajectories in the probability path significantly reduces the Number of Function Evaluations (NFE) required for high-quality action generation. This paradigm is further advanced by the *Streaming Flow Policy* Jiang et al. (2025), which treats action sequences as continuous flow trajectories. By enabling real-time streaming execution, this approach establishes a "Flow-First" paradigm, effectively reconciling the conflict between high-frequency control loops and the computational burden of generative modeling.

Consistency and Distillation for Real-Time Control. To further mitigate the inference latency bottleneck, the community has adopted consistency models and distillation techniques. *Consistency Policy* Prasad et al. (2024) and *ManiCM* Lu et al. (2024a) fundamentally alter the generation process by learning to map any point on the trajectory directly to the origin (the clean action), thereby enabling one-step or few-step generation without sacrificing precision. Parallel to this, distillation methods such as *Falcon* Chen et al. (2025b) introduce "partial denoising" strategies to accelerate inference, while *One-Step Diffusion Policy* Wang et al. (2024d) leverages progressive distillation to synthesize policies into a single forward pass. These innovations collectively resolve the inference bottleneck, making generative policies viable for highly dynamic tasks—such as table tennis or object catching—where millisecond-level latency is critical.

Compositional and Hierarchical Generation. Beyond speed, structural compositionality has become a key research focus for handling long-horizon and multi-objective tasks. *Compose Your Policies!* Cao et al. (2025) introduces a framework for test-time distribution-level composition, allowing distinct policies (e.g., grasping and obstacle avoidance) to be mathematically combined via energy-based gradient addition during inference. Addressing temporal complexity, *Hierarchical Diffusion Policy* Ma et al. (2024) and *H3DP* Lu et al. (2025b) decompose the generation process into hierarchical layers: a high-level planner generates sub-goals or waypoints, which condition a low-level policy to generate dense action trajectories. This factorization not

only facilitates contact guidance and multi-stage reasoning but also improves the robustness of the policy against compounding errors over long horizons.

5.3 Structured Prediction: Discretization and Autoregression

While continuous generative models dominate the landscape of precise low-level control, discrete sequence prediction remains a powerful paradigm, fundamentally treating robotic manipulation as a next-token prediction problem. This approach aligns the action space with the vocabulary of Large Language Models (LLMs), enabling the direct transfer of scaling laws from NLP to robotics.

Discretization and Codebook Alignment. To leverage the Transformer architecture for continuous control, the high-dimensional action space must be mapped to a discrete latent space. Early works like RT-1 Brohan et al. (2022) utilized simple uniform binning, but recent research emphasizes learning semantic-rich discrete representations. *VQ-VLA* Wang et al. (2025e) introduces a sophisticated Vector-Quantized (VQ) tokenizer that clusters continuous actions into a learnable codebook. By minimizing the reconstruction error between continuous signals and discrete codes, it ensures that the autoregressive model captures the multimodal nature of human demonstrations. Furthermore, the granularity of tokenization is critical for temporal abstraction. *FAST* Pertsch et al. (2025) challenges the fixed-length tokenization paradigm by proposing variable-length action tokens. This method dynamically adjusts the temporal resolution based on task complexity, significantly improving precision in long-horizon tasks while reducing the sequence length required for planning.

Efficient Inference Architectures. Despite their expressivity, autoregressive Vision-Language-Action (VLA) models suffer from high computational costs during inference, primarily due to the quadratic complexity of attention mechanisms and memory bandwidth limitations. To address the deployment bottleneck on edge devices, architectural optimizations have become imperative. *BitVLA* Wang et al. (2025a) pioneers extreme model compression by employing 1-bit quantization. By binarizing weight parameters without compromising the policy's decision boundaries, it drastically reduces memory footprint and energy consumption. Complementing this, *VLA-Cache* Xu et al. (2025b) targets the redundancy in sequential decoding. It introduces an adaptive Key-Value (KV) caching mechanism that selectively retains visual-action context, eliminating redundant computations in the attention layers and enabling high-frequency control loops essential for dynamic interaction.

5.4 Spatial Intelligence: The Shift to 3D Representations

Traditional 2D image-based policies often suffer from viewpoint dependence and depth ambiguity, limiting their ability to generalize across camera poses. To overcome these limitations, the field is undergoing a paradigm shift towards *Spatial Intelligence*, moving from entangled 2D pixel spaces to explicit 3D geometric representations. This transition enables models to learn SE(3)-equivariant policies that decouple object structure ("what") from camera pose ("where").

Geometry-Aware Policy Learning. Operating directly on 3D representations allows for the injection of strong geometric inductive biases. Early voxel-based approaches like *PerAct* Shridhar et al. (2023) demonstrated the efficacy of learning actions in a discretized 3D volume. Building on this, *3D Diffuser Actor* Ke et al. (2024) integrates 3D scene encodings into a diffusion denoising process, enabling the generation of precise 6-DoF trajectories that are consistent with the physical geometry of the environment. To bridge the gap between 2D pre-training and 3D control, *OG-VLA* Singh et al. (2025) utilizes orthographic projections to construct 3D-aware visual representations. By fusing multi-view features into a unified spatial latent space, it significantly improves manipulation precision in cluttered scenes where single-view 2D policies typically fail.

Unified Representations: From Points to Gaussians. Recent research focuses on lightweight and continuous 3D representations to unify observation and action spaces. *DP3* Ze et al. (2024) establishes a strong baseline by applying diffusion models directly to point cloud data, demonstrating superior generalization to novel viewpoints compared to image-based baselines. Extending this to Vision-Language-Action models,

PointVLA Li et al. (2026) injects point cloud tokens directly into the LLM backbone, enabling the model to perform 3D spatial reasoning without the computational overhead of voxel grids. Furthermore, the emergence of 3D Gaussian Splatting has introduced a new frontier for dynamic scene representation. *ManiGaussian* Lu et al. (2024b) leverages dynamic Gaussian Splatting to model non-rigid object deformations and scene dynamics explicitly, offering a richer, texture-aware geometric prior than sparse point clouds.

5.5 Closing the Loop: Integration with RL and World Models

While Imitation Learning (IL) provides a stable initialization, it is fundamentally constrained by the quality of the demonstration data and suffers from covariate shift when deploying in novel states. To transcend the upper bound of the demonstrator and enable adaptation in dynamic environments, recent advancements integrate Vision-Language-Action models with Reinforcement Learning (RL) and predictive World Models, forming a closed-loop learning system.

Visual Reward Discovery and Policy Improvement. A critical bottleneck in scaling robotic RL is the design of dense reward functions from high-dimensional visual observations. *VIP* Ma et al. (2022) addresses this by learning a pre-trained visual representation that embeds functional distances to the goal, enabling zero-shot reward computation without manual instrumentation. Moving beyond fixed embeddings, *Eureka* Ma et al. (2023) leverages the coding capabilities of LLMs to synthesize executable reward algorithms automatically, allowing VLA policies to perform trial-and-error learning on complex manipulation tasks. These methods bridge the gap between semantic intent and low-level control, facilitating online policy improvement via standard RL algorithms like PPO or SAC.

Generative World Models for Planning. The integration of World Models endows VLA agents with "mental simulation" capabilities, enabling Model-Based Reinforcement Learning (MBRL) in latent space. Building on the success of latent dynamics models like *DreamerV3* Hafner et al. (2023), recent works utilize generative video models as predictive simulators. *Genie* Bruce et al. (2024) demonstrates that a foundation model trained on internet videos can serve as an interactive environment for training agents. Specifically for VLA architectures, *WorldVLA* Cen et al. (2025b) learns to predict future video frames conditioned on language instructions and proposed actions. This allows the agent to perform counterfactual reasoning—evaluating the outcomes of potential action sequences before execution. To ensure deployment safety, *SafeVLA* Zhang et al. (2025a) incorporates a safety critic within this imagination loop, employing constrained optimization to filter out visually plausible but physically hazardous actions.

5.6 Summary: Positioning Generative Visuomotor Policies in the Embodied Stack

The evolution of generative visuomotor policies represents a pivotal maturation in the "motor cortex" of embodied intelligence, characterized by a convergence of generative expressivity, geometric rigor, and closed-loop adaptation.

The Convergence of Efficiency and Geometry. First, the architectural landscape has shifted from deterministic regression to probabilistic generation, acknowledging that robotic action distributions are inherently multimodal. While early diffusion policies unlocked this expressivity, the transition to flow matching and consistency models has increasingly addressed the trade-off between sampling quality and inference latency, enabling high-frequency control loops essential for dynamic manipulation. Concurrently, the perceptual foundation has migrated from entangled 2D pixel spaces to explicit 3D representations. By incorporating SE(3)-equivariant structures and point-cloud encoders, modern generative visuomotor policies reduce viewpoint dependence and acquire stronger geometric inductive biases for cross-scene generalization.

From Static Imitation to Active Adaptation. Second, the learning paradigm is transcending the limits of static Behavior Cloning. Recognizing that imitation learning is upper-bounded by demonstrator quality and susceptible to covariate shift, the field is moving towards a hybrid approach integrating Reinforcement Learning and World Models. This integration endows agents with "System 2" capabilities: the ability to perform counterfactual reasoning via mental simulation (World Models) and to self-correct through online

interaction (RL). This shift transforms robots from passive mimics into active learners capable of recovering from errors and refining their policies in unseen environments.

Bridging Control with Semantic Reasoning. Finally, while generative visuomotor policies have achieved proficiency in low-level manipulation skills ("how to grasp"), they remain limited in high-level semantic intent ("what to grasp and why"). Achieving general-purpose embodied intelligence requires grounding these motor primitives within broader reasoning-capable architectures, motivating the transition to vision-language-action (VLA) models discussed in the next section.

6 Vision-Language-Action Foundation Models

The generative visuomotor policies of the preceding section resolve key failure modes of behaviour cloning—mode averaging and compounding errors—by introducing generative action modelling. Yet they remain limited in two critical dimensions: they lack natural-language understanding, and they are typically trained per-task or per-embodiment, offering little cross-platform generality.

Vision-Language-Action (VLA) models address both limitations by unifying visual perception, language comprehension, and action generation within a single foundation-model architecture. By co-training on internet-scale vision–language data and large-scale robot demonstrations, VLAs absorb transferable semantic priors that enable multi-task, multi-embodiment manipulation conditioned on free-form language instructions.

This section traces the design evolution from vision-language models (VLMs) to VLAs (§6.1), analyses generalisation properties and the role of data scale (§6.2), examines mechanisms for long-horizon reasoning (§6.3), and surveys emerging architectural trends that shape the capability boundary of next-generation VLA systems (§6.4).

Figure 6 provides a roadmap for this section. We first present the design transition from VLM to VLA, then examine how data scale and diversity drive cross-task and cross-embodiment generalization, followed by mechanisms for long-horizon reasoning and emerging design dimensions; we conclude with the key limitations that motivate integration with RL and world models in the next section.

6.1 From VLMs to VLAs: Design Principles

Semantic foundation from vision-language pre-training. Large-scale multimodal pre-training has made vision-language models a natural bridge between perception and language. Two dominant paradigms have emerged. Contrastive dual-encoder models such as CLIP Radford et al. (2021) align image and text embeddings via contrastive objectives, providing open-vocabulary visual recognition. Generative VLMs—Flamingo Alayrac et al. (2022), LLaVA Liu et al. (2023b)—integrate visual features into large language model decoders, enabling image-conditioned generation and reasoning. Crucially, generative VLMs treat language as a compositional and flexible interface rather than a fixed task label, providing a natural substrate for extending visual understanding toward action generation.

Incorporating action into the generative loop. The defining step from VLM to VLA is to bring *action tokens* into the model's output (or an attached action head), so that a single forward pass maps an observation–instruction pair to executable robot commands. Two broad strategies have been pursued.

Strategy 1: Action as language tokens. RT-2 Zitkovich et al. (2023) co-fine-tunes a pre-trained VLM on both internet data and robot trajectories, representing discretised robot actions as text tokens. This reframes robotic control as next-token prediction, allowing semantic knowledge—object recognition, spatial reasoning, even cultural common sense—to zero-shot transfer to manipulation tasks. OpenVLA Kim et al. (2024) scales this approach with a 7B-parameter model built on LLaMA 2 Touvron et al. (2023) and strong visual encoders (DINOv2 + SigLIP), trained on the Open X-Embodiment (OXE) corpus spanning ∼1M real-robot episodes across 22+ embodiments O'Neill et al. (2024).

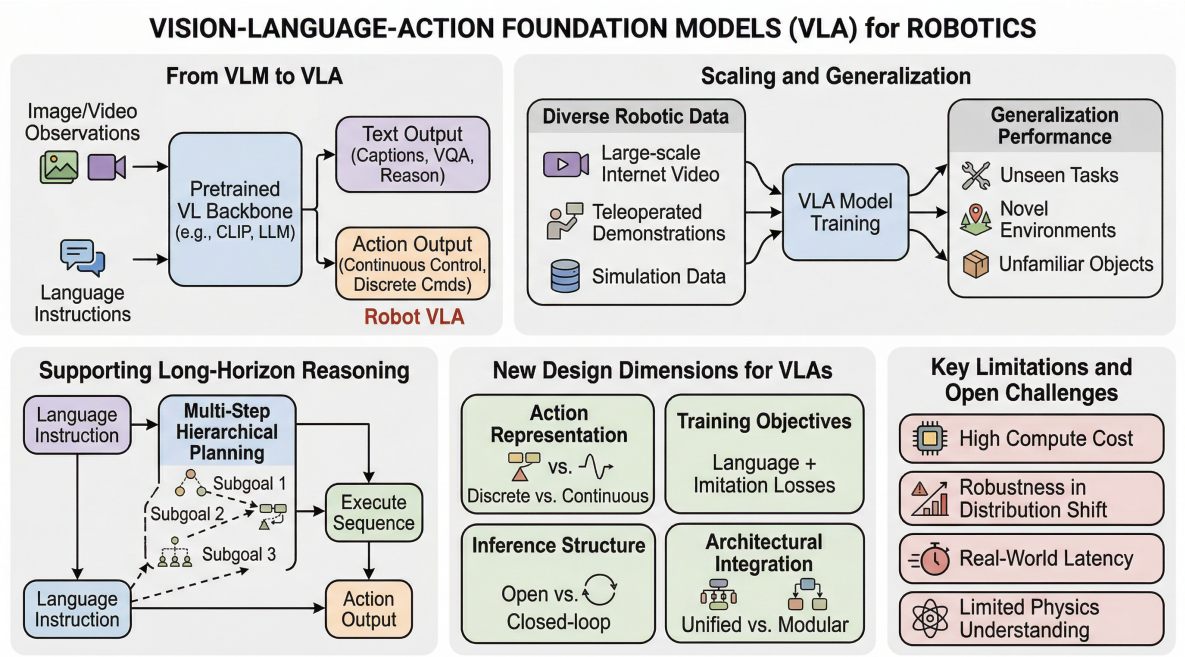

Figure 6: Overview of Vision-Language-Action (VLA) foundation models. The figure summarizes the core progression discussed in this section: extending VLMs into action-capable architectures (discrete, continuous, and hybrid action representations), scaling generalization through heterogeneous cross-embodiment data, improving long-horizon performance via language-structured and hierarchical reasoning, and advancing capability through geometry, multimodal sensing, memory, and efficient inference. It also highlights open challenges in safety, language–action alignment, and online adaptation.

Strategy 2: Continuous action generation. Rather than discretising actions, the π_0 family Black et al. (2024); Intelligence et al. (2025) employs flow matching on top of a pre-trained VLM backbone to generate continuous action distributions. This preserves the precision needed for dexterous manipulation while retaining the semantic priors of the language model. $\pi_{0.5}$ Intelligence et al. (2025) further introduces a knowledge-insulation mechanism to prevent fine-tuning from degrading the VLM's open-world reasoning capabilities. $\pi_{0.7}$ Ai et al. (2026) extends this framework by introducing a steerable multimodal prompting strategy that integrates generated subgoal images and episode metadata. This approach facilitates the ingestion of large-scale, heterogeneous datasets, allowing the model to demonstrate emergent zero-shot cross-embodiment transfer on complex and dexterous tasks.

Strategy 3: Hybrid representations. A growing family of models combines both paradigms: high-level decisions (skill selection, subgoal specification) are modelled as discrete language tokens, while low-level execution uses continuous action heads. DiffusionVLA Wen et al. (2024b) unifies autoregressive language decoding with a diffusion-based action generator; CogACT Li et al. (2024a) separates cognition and action into interacting subsystems that share a latent space. These hybrids aim to balance language alignment with control fidelity. Table 3 summarises representative VLA models and their design choices.

6.2 Generalisation and the Effects of Scale

A central promise of VLA models is that they can generalise broadly across tasks, environments, and robot platforms—a qualitative departure from prior visuomotor policies that are typically confined to narrow task distributions and single embodiments.

Dimensions of embodied generalisation. Traditional robot learning defines generalisation narrowly: robustness to unseen initial states or object instances within a fixed task distribution. VLA models target a

substantially broader envelope that spans five dimensions simultaneously: **novel tasks and compositions** (combining known primitives in unseen ways), **unseen language instructions** (interpreting paraphrases and new descriptions), **environment variation** (changes in lighting, clutter, backgrounds, and scene geometry), **cross-embodiment transfer** (deploying on robots with different kinematics, grippers, and control interfaces), and **long-horizon stability** (maintaining coherent behaviour over multi-stage tasks with delayed feedback).

Addressing these dimensions jointly exceeds the capacity of task-specific visuomotor policies and is only made possible by the combination of large-scale pre-training and unified vision–language–action representations.

Table 3: Representative VLA models and their key design choices. *Action repr.*: D = discretised tokens, C = continuous (flow/diffusion), H = hybrid. *VLM backbone*: the pre-trained vision-language model used for initialisation.

Model	Act.	VLM Backbone	Training Data	Key Contribution
RT-1 Brohan et al. (2022)	D	EfficientNet + Transformer	Large-scale real demos	Multi-task tokenised policy at scale
RT-2 Zitkovich et al. (2023)	D	PaLI-X / PaLM-E	Web V-L + robot data	Action-as-text; semantic transfer
OpenVLA Kim et al. (2024)	D	LLaMA 2 + DINOv2/SigLIP	OXE (∼1M episodes)	Open-source; cross-embodiment
π_0 Black et al. (2024)	C	Pre-trained VLM + flow head	>10k hrs multi-platform	Flow-matching action generation
$\pi_{0.5}$ Intelligence et al. (2025)	C	Pre-trained VLM + flow head	π_0 data + open-world	Knowledge insulation for generality
GR00T N1 Bjorck et al. (2025)	C	Multi-modal Transformer	Real + sim + internet video	Humanoid cross-embodiment
DiffusionVLA Wen et al. (2024b)	H	Autoregressive LM + diffusion	Multi-task robot data	Unified AR + diffusion
CogACT Li et al. (2024a)	H	VLM + separate action system	Multi-task robot data	Cognition–action separation

Data infrastructure as a scaling enabler. Generalisation in VLA models is determined less by architecture alone than by the diversity and structure of the training distribution. The Open X-Embodiment (OXE) initiative O'Neill et al. (2024) plays a foundational role here: by aggregating datasets from 34 research labs into a unified format covering 22+ robot embodiments and over one million trajectories, OXE reframes data scaling as a coordination problem rather than a single-lab collection effort. A key practical design choice is *coarse action alignment*—expressing end-effector commands through a shared, normalised interface—which proves sufficient to induce positive transfer across robots despite imperfect semantic alignment O'Neill et al. (2024).

OpenVLA Kim et al. (2024) provides empirical validation of this paradigm: trained on nearly one million OXE demonstrations, it shows clear improvements over prior baselines in multi-task and multi-embodiment settings, together with efficient adaptation to new environments via parameter-efficient fine-tuning. These results support a key empirical finding: when large-scale, heterogeneous data is combined with unified vision-language-action representations, robot policies can move beyond task-specific solutions toward general-purpose manipulation.

More recently, industrial-scale efforts have pushed data volume further. GEN-0 Team (2025) reports training on hundreds of thousands of hours of real manipulation data and observes predictable performance gains with increasing data and model capacity, echoing the scaling laws observed in NLP and computer vision, though the precise functional form for robotics remains an open question (Section 7).

6.3 Long-Horizon Reasoning

Many real-world manipulation tasks are inherently long-horizon and multi-stage: table clearing, object assembly, and garment folding all require sequential decision-making with condition-dependent execution. Traditional generative visuomotor policies struggle in this regime because small per-step errors compound over extended horizons, and flat sequence-to-sequence architectures lack explicit representations of subgoals, conditional structure, or recovery strategies. Thus, the VLA models mitigate these issues through three complementary design families.

Language as a structural prior. A distinctive advantage of VLAs for long-horizon tasks comes from the structural properties of natural language itself. Language provides a compact and compositional way to

represent task hierarchy, subgoals, and conditional logic—properties that have been widely exploited for task decomposition and skill sequencing Intelligence et al. (2025).

In practice, language serves as an intermediate abstraction between perception and action, enabling the policy to reason over high-level intent rather than relying solely on local visual feedback. SayCan Ahn et al. (2022) exemplifies this approach: a language model proposes feasible skill sequences at each step, while low-level controllers execute the chosen skill, reducing failures that arise from predicting long action sequences directly.

More recent work incorporates chain-of-thought reasoning to make intermediate decisions explicit. CoT-VLA Zhao et al. (2025) generates visual reasoning traces prior to action execution, and ECoT Zawalski et al. (2024) jointly trains reasoning and action prediction, improving long-horizon coherence and interpretability.

Hierarchical planning and execution. A second design family addresses long-horizon tasks through explicit temporal hierarchy: a high-level policy operates in a language or semantic space to decide *what to do*, while a low-level policy handles continuous control to execute the decision.

RT-H Belkhale et al. (2024) introduces *language motions* as intermediate representations, translating high-level language actions into low-level robot commands. This decouples planning from execution, reducing error accumulation over long horizons. Fast-in-Slow Chen et al. (2025a) implements a dual-system variant where a large VLM generates subgoals (slow, deliberative) and a compact policy executes them (fast, reactive), enabling complex reasoning while maintaining real-time control. HiRobot Shi et al. (2025b) extends hierarchical VLAs to open-ended instruction following, learning to decompose arbitrary language instructions into executable subgoals without predefined skill libraries.

Predictive planning with world knowledge. The third family introduces explicit or implicit world modelling to enable anticipatory planning. Instead of reacting only to current observations, these methods reason about future states or intermediate subgoals.

Representative approaches include subgoal prediction, affordance chaining, and diffusion-based future-state generation Zhang et al. (2025g); Liu et al. (2025a). The π_0 and $\pi_{0.5}$ models Black et al. (2024); Intelligence et al. (2025) combine flow-based generation with joint learning of world knowledge and control, improving stability in open-world, multi-stage tasks. GR00T N1 Bjorck et al. (2025) targets humanoid robots specifically, leveraging mixtures of real data, simulation, and internet video to support whole-body coordination and long-term temporal consistency.

By incorporating prediction into the control loop, these methods shift VLAs from reactive behaviour toward anticipatory planning—a capability that is particularly valuable for tasks with delayed consequences and irreversible actions. We discuss world models in greater depth in Section 8.

6.4 Emerging Design Dimensions

Beyond language conditioning, generalisation, and long-horizon reasoning, progress in VLA is driven by coordinated advances across several additional design dimensions. We highlight four that most directly shape the capability boundary of current systems; the integration of VLAs with reinforcement learning and world models—a fifth critical dimension—is deferred to Section 8.

Spatial and geometric reasoning. Early VLA models rely primarily on single-view RGB, limiting their spatial reasoning to what monocular appearance can support. Recent work addresses this through explicit 3D representations. 3D-VLA Zhen et al. (2024) extends world modelling to 3D scene predictions. PointVLA Li et al. (2026) injects point-cloud inputs directly into the VLA backbone, improving manipulation in cluttered scenes where 2D projections lose critical depth information. SpatialVLA Qu et al. (2025) introduces spatial token encodings to represent grasp and placement targets explicitly. Beyond static observation, ActiveVLA Liu et al. (2026) introduces the paradigm of active perception into vision-language-action architectures, enabling models to dynamically adjust sensor configurations to resolve spatial ambiguities during high-precision 3D manipulation tasks.

These advances improve robustness to occlusion, viewpoint variation, and fine-grained alignment errors that are common in real-world deployment.

Multimodal perception beyond vision. Vision alone is often insufficient for contact-rich manipulation. Recent VLA systems therefore integrate tactile and force sensing alongside visual and language inputs. ForceVLA Yu et al. (2025) uses mixture-of-experts to handle different contact regimes (free space, light contact, heavy contact). VLA-Touch Bi et al. (2025) fuses dual-level tactile feedback with visual observations for precision assembly. Audio-VLA Wei et al. (2025) augments the VLA with contact audio, targeting failure modes where vision is insufficient (e.g., slip onset, sustained wiping contact). This trend reflects a broader shift from purely vision-driven policies toward closed-loop agents that actively sense and respond to physical interaction (Section 7).

Memory and context management. Long-horizon tasks require memory that extends beyond the current observation. MemoryVLA Shi et al. (2025a) incorporates external memory modules that store and retrieve task-relevant representations across time. VLA-Cache Xu et al. (2025b) caches stable tokens (static background, task descriptions) across timesteps, simultaneously improving inference efficiency and temporal context retention. These mechanisms enable VLA systems to maintain long-term context and move beyond purely reactive, single-frame policies.

Inference efficiency. Deploying VLAs on real robots requires inference rates above $\sim$10 Hz for reactive control, yet foundation-scale models are computationally expensive. Several training-free acceleration strategies have emerged: token pruning removes uninformative visual tokens (EfficientVLA Yang et al. (2025a)); dynamic early exit allows easy inputs to bypass deeper layers (DeeR-VLA Yue et al. (2024)); parallel decoding generates action chunks simultaneously rather than autoregressively (PD-VLA Song et al. (2025)); and learned action tokenisers reduce sequence length (FAST Pertsch et al. (2025)). These methods achieve 3–5$\times$ speedups with minimal accuracy loss, though the trade-off between inference speed, action precision, and model capacity remains an active research frontier.

6.5 Open Challenges for VLA Models

Despite rapid progress, several fundamental challenges remain before VLAs can serve as reliable, deployable embodied agents.

Data distribution gaps. Current VLA training corpora are heavily biased toward tabletop pick-and-place in Western household settings. Complex assembly, industrial manipulation, outdoor environments, and culturally diverse domestic scenes remain severely under-represented, creating predictable failure modes on out-of-distribution deployments Wang et al. (2025d).

Language–action alignment. Language instructions in robot datasets are typically short, low-entropy commands ("pick up the red cup") that provide far weaker supervision than the rich, diverse text corpora used in NLP. This shallow language–action coupling limits the depth of semantic reasoning VLAs can perform and may partly explain why chain-of-thought and hierarchical approaches (Section 6.3) yield disproportionate gains.

Safety and reliability. End-to-end VLA models lack an explicit safety layer: there is no analogue of the motion-planning collision checker or impedance controller that hierarchical systems provide. Recent work on safety-constrained VLAs (SafeVLA Zhang et al. (2025a)) and refusal mechanisms is promising but nascent. The absence of formal safety guarantees remains a major barrier to deployment in human-shared environments.

Static deployment. Pre-trained VLAs follow a "train–freeze–deploy" paradigm: once deployed, they cannot self-improve through interaction. When robots encounter novel objects, unexpected dynamics, or environmental variations absent from training data, frozen models can only wait for human re-training.

Addressing this core limitation through online reinforcement learning, world-model-based imagination, and continual learning is the central focus of Section 8.

6.6 Critical Synthesis

VLA foundation models represent a paradigm shift in robotic manipulation: from task-specific, single-embodiment policies toward unified models that absorb language, vision, and action into a common representational framework. Before turning to data scaling and online adaptation, we distil three structural observations that cut across the design families reviewed above.

The semantic–contact/dynamics grounding gap. A central tension in current VLA systems is the asymmetry between *semantic grounding*, strongly supported by internet-scale vision-language pretraining; and *contact/dynamics grounding* limited by relatively small-scale robotic interaction data. In practice, this manifests as a characteristic failure signature: VLAs frequently identify the correct object, decompose the task into a valid sequence, and produce linguistically plausible reasoning, yet execute unstable grasps, imprecise insertions, or physically infeasible motions. The gap suggests that semantic scaling has substantially outpaced physical interaction scaling, and that closing it requires not merely more robot data but richer modalities (force, tactile, audio) and online adaptation mechanisms that ground language-level plans in contact-level reality.

VLAs as partial world models. An underexplored interpretation is that VLAs implicitly approximate *structured world models*. Because they integrate semantic reasoning and action generation, VLAs often exhibit capabilities associated with world models, such as task decomposition, intermediate-state inference, error explanation, yet these capabilities remain predominantly *symbolic and perceptual* rather than *dynamically predictive*. Unlike explicit world models used in model-based RL (Section 8), current VLAs rarely simulate future physical states or reason about contact dynamics. In this sense, today's VLA systems are *semantically rich but dynamically shallow*. Bridging this gap—integrating predictive physical modelling with language-conditioned planning—is a key motivation for the VLA+RL+world-model frontier discussed in Section 8.

Positioning within the three-scale framework. Within the framework introduced in Section 1, VLA models represent the first paradigm to strongly scale along two axes simultaneously: very high model scaling (billion-parameter architectures) and high data scaling (internet-scale vision-language corpora combined with million-trajectory robot datasets). However, their interaction scaling remains nascent—the dominant "train–freeze–deploy" paradigm means that VLAs cannot self-correct, discover strategies beyond their demonstrations, or adapt to distribution shifts encountered in deployment. This positioning clarifies both the transformative contribution of VLAs (unprecedented generalisation breadth) and their fundamental limitation (brittleness without online adaptation), and motivates the integration with reinforcement learning and world models examined in the following sections.

The following sections examine the data scaling revolution (Section 7) and the emerging VLA+RL+world-model frontier (Section 8) that together define the path forward.

7 Scaling Robot Learning Data

Data scaling is a primary driver of recent progress in robotic manipulation, but its role is qualitatively different from scaling in language-only or vision-only settings. In embodied learning, performance depends not only on sample count but also on coverage over embodiments, environments, contact regimes, temporal horizons, and sensing modalities. This section synthesizes how scaling-law intuition transfers to robotics, why naive "more data" strategies often underperform, and what dataset design principles are most predictive of downstream manipulation generalization.

Figure 7 provides a roadmap for this section. We first revisit scaling-law principles and explain why robotic data scaling is fundamentally multi-dimensional, then review representative dataset and infrastructure efforts,

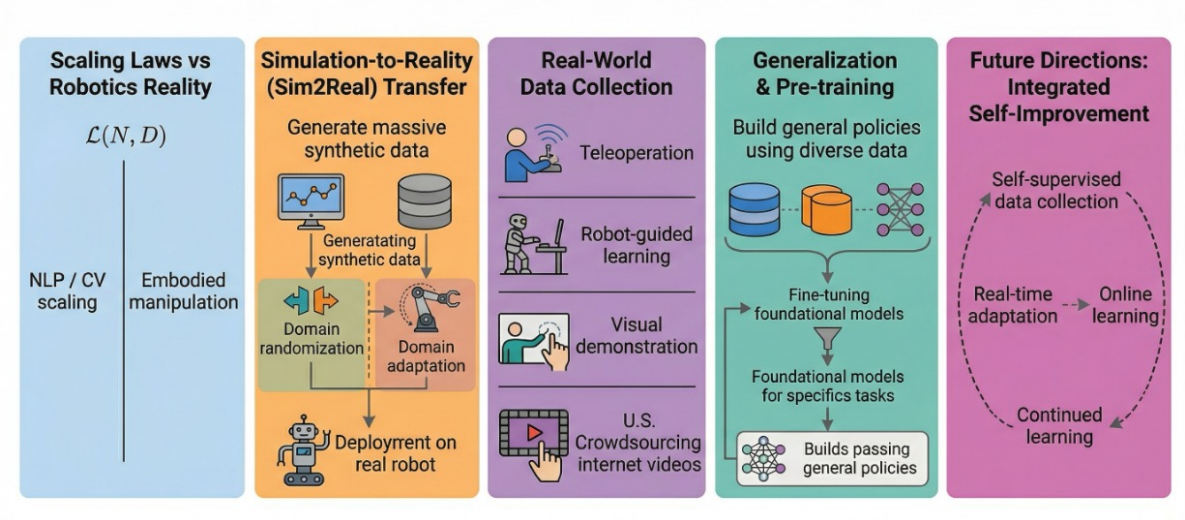

Figure 7: Overview of scaling robot learning data. The figure summarizes the central themes of this section: adapting scaling-law intuition from NLP/CV to embodied learning, building reusable large-scale datasets and cross-embodiment data infrastructure, designing effective heterogeneous pretraining mixtures for VLA models, expanding modality coverage beyond vision (audio, force/torque, tactile), and outlining future scaling directions centered on distribution quality, standardization, and hybrid real–synthetic pipelines.

followed by practical mixture-design considerations for VLA pretraining and multimodal sensing expansion; we conclude with forward-looking scaling directions and their implications for downstream generalization.

7.1 From scaling laws in NLP/CV to multimodal foundation models

Large-scale learning in NLP, computer vision, and multimodal modeling has been guided by empirical *scaling laws*: as model size, dataset size, and training compute increase, performance (or training loss) often improves following smooth power-law trends Kaplan et al. (2020); Hestness et al. (2017). A common abstraction is to model the irreducible loss floor plus decaying terms governed by model and data scale:

$$\mathcal{L}(N, D) \approx \mathcal{L}_\infty + aN^{-\alpha} + bD^{-\beta}, \tag{4}$$

where N denotes parameter count and D denotes the amount of training data (e.g., tokens, images, or pairs). While such formulas are not mechanistic theories, they have proven practically useful for forecasting returns on additional compute/data, comparing architectures at fixed scale, and informing data-collection investment.

A second key insight from large language modeling is that *compute-optimal training* depends on the balance between model size and dataset size: for a given compute budget, overly large models trained on too little data underperform smaller models trained on more data Hoffmann et al. (2022). This observation directly motivates the data-first mindset in foundation-model development: scaling D is not merely helpful, but can be *necessary* to unlock the gains of scaling N.

In multimodal representation learning, similar scaling regularities appear once training pipelines become stable and data becomes sufficiently large. For example, contrastive image–text pretraining exhibits reproducible scaling behavior with public datasets and open-source training stacks Cherti et al. (2023). These results also highlight that the *training distribution* matters: scaling trends can shift when the mixture of sources changes, even if the architecture and objectives remain similar Cherti et al. (2023). This point will reappear in robotics, where distribution shift across environments, embodiments, and contact dynamics is the norm rather than the exception.

7.2 Why scaling data is uniquely hard in robot manipulation

Although robotics increasingly borrows the *foundation model* playbook (pretrain broadly, then adapt), robot manipulation data differs qualitatively from web-scale text and images. Several properties make data scaling both expensive and scientifically subtle:

Embodiment and action-space heterogeneity. Robots differ in kinematics, actuation, control frequency, grippers, calibration, and safety envelopes. Unlike classification, the output space is continuous and platform-specific. Bridging embodiments demands careful action representations, normalization, and interface design (e.g., relative actions, chunked trajectories, latency matching).

Contact-rich partial observability. Manipulation is governed by frictional contacts, compliance, and unobserved state (e.g., grasp stability, micro-slip, deformation). Pure vision often underdetermines the contact state, motivating additional sensing (audio, force/torque, tactile).

Long-horizon composition. Real tasks require multi-stage composition (search, grasp, regrasp, tool use, recovery). Scaling must therefore include *behavioral diversity*: failure cases, retries, recovery maneuvers, and long-horizon interaction patterns.

Safety, logistics, and annotation. Collecting diverse interactions in diverse homes/labs is hardware- and labor-intensive, and safety constraints limit aggressive exploration. Labeling language instructions, goals, or subtask boundaries can also be costly, which makes learning from weak supervision and self-supervision attractive.

These considerations suggest that "more data" in robotics is multi-dimensional. Important scaling axes include: number of trajectories, environment diversity, object diversity, task diversity, embodiment diversity, temporal horizon, and modality richness.

7.3 Early large-scale datasets for manipulation

Early efforts toward reusable datasets for end-to-end manipulation emphasized *multi-task* and *multi-domain* coverage, aiming to make robotic data more like ImageNet-style reusable corpora.

Bridge Data and BridgeData V2. Bridge Data explicitly asked what it would take to enable practical data reuse for end-to-end skill learning and introduced a multi-task, multi-domain dataset with 7,200 demonstrations spanning 71 tasks across 10 environments Ebert et al. (2021). A central finding is that even modest amounts of target-domain data can benefit strongly from being combined with a diverse, reusable multi-domain corpus, improving cross-task and cross-domain generalization Ebert et al. (2021).

BridgeData V2 pushed this philosophy toward scale on a low-cost, reproducible robot platform and emphasized open-vocabulary conditioning (goal images or language) to support broader task sets and institutions Walke et al. (2023). Besides the headline scale (tens of thousands of trajectories across dozens of environments), BridgeData V2 is noteworthy as an *infrastructure* contribution: it helped normalize the expectation that scalable robot learning should be enabled by public datasets and reproducible loaders/formatting conventions Walke et al. (2023).

DROID: distributed, in-the-wild collection. DROID (Distributed Robot Interaction Dataset) significantly expanded both scale and geographic diversity by coordinating distributed data collection. It reports 76k demonstration trajectories (about 350 hours) across 564 scenes and 84 tasks, collected by 50 data collectors spanning North America, Asia, and Europe Khazatsky et al. (2024). This design targets a core bottleneck in manipulation generalization: robustness to *real* scene variability rather than lab-controlled diversity.

Open X-Embodiment: unifying many datasets across embodiments. A complementary scaling direction is to aggregate many pre-existing datasets and unify them under a shared schema. Open X-Embodiment (OXE) pools open-sourced robot datasets into a standardized format and couples them with

Table 4: Representative manipulation datasets and scaling strategies. Reported scales are taken from the respective papers.

Dataset	Collection mode	Scaling highlights
Bridge Data Ebert et al. (2021)	real robot, multi-domain	7,200 demos; 71 tasks; 10 environments; emphasizes cross-domain data reuse.
BridgeData V2 Walke et al. (2023)	real robot, reproducible setup	Tens of thousands of trajectories across 24 environments; open-vocabulary conditioning for scalable multi-task learning.
DROID Khazatsky et al. (2024)	real robot, distributed "in-the-wild"	76k trajectories / 350h; 564 scenes; 84 tasks; 50 collectors across continents.
Open X-Embodiment O'Neill et al. (2024)	aggregation + standardization	1M+ trajectories; 22 embodiments; pooled from 60 datasets across 34 labs; enables cross-robot training.
RoboTwin / RoboTwin 2.0 Mu et al. (2025); Chen et al. (2025c)	simulation + digital twins	Synthetic scaling with strong domain randomization; object library (731 objects / 147 categories) and 50 dual-arm tasks in RoboTwin 2.0.

cross-robot model training recipes (RT-X models) O'Neill et al. (2024). The OXE dataset is reported to contain over one million real-robot trajectories spanning 22 robot embodiments, constructed by pooling 60 datasets from 34 research labs O'Neill et al. (2024). Conceptually, OXE reframes scaling as a *coordination* problem: the community can scale data faster by standardizing storage, metadata, and action/observation representations than by any single lab collecting everything alone.

RoboTwin and RoboTwin 2.0: scaling via simulation and digital twins. While real-world data is the gold standard, simulation offers a powerful scaling lever: synthetic expert trajectories can be generated at high volume, with controllable coverage over object categories, lighting, clutter, and language prompts. RoboTwin introduced a generative digital-twin approach to produce simulated data aligned to real scenes and provide standardized benchmarks for dual-arm tasks Mu et al. (2025). RoboTwin 2.0 further emphasizes scalable task generation and robust domain randomization, anchored by RoboTwin-OD, an object library of 731 instances across 147 categories, instantiated across 50 dual-arm tasks and five robot embodiments Chen et al. (2025c). From a scaling perspective, these works highlight a pragmatic view: when real-world collection is rate-limiting, simulation can provide wide coverage, and a small amount of real data can be used to calibrate, adapt, or validate downstream policies Mu et al. (2025); Chen et al. (2025c).

A compact comparison. Table 4 summarizes representative dataset scaling strategies discussed above.

7.4 From datasets to VLA foundation models: scaling the pretraining mixture

As manipulation datasets grew, the research focus shifted from "collect a dataset for one policy" to "train a generalist policy that can absorb heterogeneous corpora". This motivates *Vision–Language–Action* (VLA) foundation models: policies that (i) ingest rich visual observations, (ii) condition on language goals/instructions, and (iii) output continuous control actions.

A representative example is π_0, which proposes a flow-matching architecture on top of a pretrained vision–language model and evaluates large-scale pretraining followed by downstream fine-tuning Black et al. (2024). Crucially for the theme of this chapter, π_0 reports pretraining on *over 10,000 hours* of robot data and emphasizes a *data mixture* spanning multiple robot platforms and a broad set of tasks Black et al. (2024). This mirrors the NLP lesson from compute-optimal training Hoffmann et al. (2022): architecture innovations matter, but the scaling regime is determined by the ability to build and continuously refine a sufficiently large and sufficiently diverse training distribution. In addition to scaling data volume, $\pi_{0.7}$ Ai et al. (2026) presents a strategy for managing heterogeneous datasets of varying quality. By conditioning the policy on metadata specific to each episode, the model distills fundamental physical priors from noisy or suboptimal

trajectories. This approach mitigates the challenges associated with data imbalance and distribution shifts that typically constrain the training of large-scale VLA models.

Practical implications for data scaling in VLA training. Beyond raw volume, successful VLA pretraining depends on a stable interface layer (consistent definitions of observations, action parameterizations, and control frequencies across platforms), conditioning richness (language instructions, goal images, or task descriptors that turn heterogeneous behavior into a compositional training signal), and coverage of recovery (datasets biased toward "clean" expert demos often underrepresent failure and recovery). In addition, mixture design is critical: scaling can be bottlenecked by data imbalance, such as too many easy pick-and-place clips and too few contact-rich long-horizon trajectories.

7.5 Scaling beyond robots: UMI-style interfaces and in-the-wild human demonstrations

A fundamental limitation of robot-centric scaling is throughput: even with distributed collection, robots are costly and slow. An alternative is to scale *human* manipulation demonstrations in real environments and transfer them to robots through carefully designed interfaces.

Universal Manipulation Interface (UMI). UMI proposes a portable, low-cost, information-rich pipeline for collecting bimanual and dynamic manipulation demonstrations using hand-held grippers and an explicit policy interface that facilitates transfer to robots Chi et al. (2024). The key scaling idea is decoupling data acquisition from robot deployment: one can collect large, diverse demonstrations in the wild and later map them into deployable robot policies, reducing the dependence on in-situ robot data Chi et al. (2024).

GEN-0 and the prospect of "internet-scale" physical interaction. Recent efforts push this direction toward foundation-scale corpora of physical interaction. For example, GEN-0 reports scaling embodied interaction to hundreds of thousands of hours of data with continued growth Team (2025). Regardless of the precise implementation details, the broader trend is clear: *robot learning data is beginning to inherit the growth curves of web data*, but only by exploiting new collection channels (wearables, interfaces, large human pipelines) rather than relying solely on autonomous robot rollouts.

7.6 Scaling modalities: toward multimodal contact understanding

Scaling the *quantity* of trajectories is only one axis. Scaling the *modality richness* of each trajectory can be equally important, especially for contact-rich and safety-critical manipulation. We highlight several modality trends that are increasingly central to object property estimation, contact safety, and fine-grained dexterity.

Audio as a contact sensor. Contact produces structure in sound: impacts, scraping, frictional slip, and resonant vibrations correlate with contact events and material properties. **Hearing Touch** demonstrates that contact microphones can be treated as an "audio tactile" channel and that audio–visual pretraining can improve representations for contact-rich manipulation Mejia et al. (2024). **ManiWAV** further explores learning manipulation-relevant cues from in-the-wild audio–visual data Liu et al. (2024b). More recently, **Audio-VLA** explicitly augments VLA models with contact audio, targeting failure modes where vision alone is insufficient (e.g., slip onset, sustained wiping contact) Wei et al. (2025). From a scaling standpoint, audio is attractive because it is cheap, ubiquitous, and can be collected at high frequency with minimal additional hardware complexity.

Force/torque sensing and force-controlled manipulation. Although not always highlighted in dataset papers, force/torque (F/T) sensing is a practical cornerstone for safe interaction: it provides direct feedback for compliance, contact stabilization, and limiting peak forces. As policies become more general and operate in unstructured settings, scaling F/T signals in datasets (or learning to infer them) becomes key for contact safety, robust insertion and assembly, and sustained-contact tasks such as wiping, scrubbing, and polishing. A useful viewpoint is that force signals define *constraints* and *stability margins* that are invisible in RGB.

High-resolution tactile and visuo-tactile sensors. Vision-based tactile sensors (e.g., GelSight-like designs) provide high spatial resolution of contact geometry and shear patterns. The improved GelSight design demonstrates compact sensing for geometry and slip, enabling direct access to phenomena that typically only appear indirectly in kinematics Dong et al. (2017). DIGIT provides a low-cost, compact, high-resolution tactile sensor design intended to make tactile sensing more broadly deployable Lambeta et al. (2020). A broader perspective on generating and using visuo-tactile data at scale, including applications such as hardness/texture estimation and slip detection, is discussed in a recent review Sun et al. (2025). These sensors are important because they can turn contact into a rich supervisory signal: contact patches provide self-supervision for grasp stability, material classification, and precise alignment.

Wearable tactile gloves and tactile skins. Wearables and skins offer a route to scale tactile data collection *outside* the robot lab. OSMO proposes an open-source wearable tactile glove designed for human-to-robot skill transfer, explicitly targeting the embodiment gap by using comparable tactile channels on both human and robot Yin et al. (2025). DOGlove explores dense tactile gloves that support skill transfer with rich contact measurements Zhang et al. (2025b). On the robot side, tactile skins aim for scalable, replaceable, and reusable sensing. ReSkin proposes replaceable tactile skins using magnetic sensing and learning-based calibration Bhirangi et al. (2021), and AnySkin pushes toward plug-and-play skin sensing with cross-instance generalization Bhirangi et al. (2025).

7.7 Outlook: what "scaling laws" might mean for robot data

Compared with NLP/CV, robotics still lacks universally agreed-upon scaling laws that reliably predict downstream performance from dataset/model/compute knobs. Nevertheless, the trajectory of the field suggests several likely directions:

Scaling will be distribution-limited, not just volume-limited. As in multimodal scaling studies Cherti et al. (2023), the *mixture*—contact richness, long-horizon composition, and recovery behavior—may dominate returns once basic coverage is achieved.

Standardization is a first-class scaling primitive. Unifying storage formats and interfaces (as in OXE O'Neill et al. (2024)) directly increases the effective dataset size by enabling reuse and cross-institution training.

Modality scaling will unlock new capability regimes. Audio, F/T, and tactile channels are not optional add-ons for dexterity; they may be prerequisites for the next stage of progress in (i) object physical property estimation, (ii) contact-safe interaction, and (iii) fine manipulation under occlusion.

Hybrid real + synthetic pipelines will likely dominate. Simulation-based generators such as RoboTwin Mu et al. (2025); Chen et al. (2025c) can cheaply expand coverage, while targeted real data anchors transfer, calibrates sensors, and defines safety constraints.

8 Beyond Supervision: VLA + RL and World Models

The preceding sections have shown how VLA foundation models achieve broad cross-task generalisation through large-scale offline pre-training. Yet offline training alone imposes a fundamental ceiling: when a robot encounters novel objects, unexpected dynamics, or environmental variation absent from the training distribution, a frozen VLA model cannot self-correct—it can only wait for humans to collect more data and retrain. This *train–freeze–deploy* paradigm is inherently static, whereas genuine embodied intelligence should be dynamic, capable of learning and adapting through continuous physical interaction.

This section reviews three complementary pathways that break this bottleneck: reinforcement learning for online policy improvement (§8.1), world models for sample-efficient imagination and planning (§8.2), and learning from human videos to scale training data beyond robot demonstrations (§8.3). We close with a discussion of how these components interact and the key open challenges (§8.4).

8.1 VLA Meets Reinforcement Learning

Question: Why VLAs Need RL? VLA foundation models and reinforcement learning are complementary rather than competing. VLAs provide the *prior*: broad visual semantics, language grounding, and action regularities learned from large-scale offline data, enabling a single policy to transfer across tasks, objects, and scenes. RL supplies the missing mechanism for *task-specific adaptation*: through trial-and-error interaction, the policy can correct its own compounding errors, specialise to local dynamics (friction, compliance, occlusions), and discover behaviour that surpasses the demonstration envelope. Equally important, RL can serve as a *constraint-aware policy refinement layer*—by optimising under constraints or incorporating human feedback as reward signals, RL reshapes a powerful but imperfect VLA prior into behaviour that is risk-aware in the real world. However, applying RL directly to large VLAs faces severe practical challenges: low sample efficiency on real hardware, gradient instability in billion-parameter models, and safety risks from unconstrained exploration. Recent work addresses these challenges through three distinct technical strategies.

Trajectory-Level Policy Optimisation. A natural mismatch exists between VLAs, which generate action sequences autoregressively, and classical RL, which optimises step-level Bellman equations. Trajectory-level methods resolve this by treating the entire generated sequence as the unit of optimisation.

VLA-RL Lu et al. (2025a) reformulates policy gradients at the trajectory level, introducing trajectory-level value functions $V(\tau)$ and *process reward models* (PRMs) learned from human-annotated success/failure pairs. Importance-sampling-based gradient estimation sidesteps the credit-assignment difficulties of long-horizon tasks. TGRPO Chen et al. (2025e) extends group relative policy optimisation to trajectory-wise settings, comparing relative performance across trajectory groups rather than individual actions. SimpleVLA-RL Li et al. (2025a) demonstrates that straightforward GRPO-style updates with careful hyperparameter tuning can scale VLA training competitively, suggesting that algorithmic simplicity is viable when the VLA prior is strong.

Advantage-Conditioned and Human-in-the-Loop Methods. A second family initialises RL from the VLA prior and refines it through a mixture of autonomous practice and human feedback.

Physical Intelligence's RECAP framework Physical Intelligence et al. (2025) introduces *advantage-conditioned policies* $\pi(a \mid s, \hat{A}(s,a))$ that predict action advantages rather than raw actions, trained jointly on offline demonstrations, human corrections, and autonomous exploration. Experiments report success-rate improvements from 60% (pure BC) to 85% through 100–200 hours of online interaction. HIL-SERL Luo et al. (2025) targets stringent real-robot constraints: demonstration-guided exploration from 10–50 human demos, safety-boundary constraints, and off-policy SAC for sample efficiency. On a Franka platform, HIL-SERL learns fine manipulation tasks (USB insertion, screw tightening) with $\sim$3 hours of total interaction—an order of magnitude more sample-efficient than human-free RL.

Preference-based methods offer an alternative to scalar reward design. GRAPE Zhang et al. (2024b) adapts RLHF techniques from language modelling to robotic manipulation, learning from human preference comparisons rather than hand-designed rewards. RIPT-VLA Tan et al. (2025) enables real-time human corrections during deployment that are immediately incorporated into policy updates, supporting rapid adaptation without extensive offline data collection.

Reasoning-Enhanced RL. An emerging direction integrates chain-of-thought reasoning with RL optimisation, allowing the model to learn *when* deliberation improves task success. Embodied-R1 Yuan et al. (2025) trains VLAs to generate explicit reasoning traces before action prediction, jointly optimising reasoning quality and action outcomes. VLAC Zhai et al. (2025) introduces a critic network that shares visual representations with the policy to evaluate both reasoning and action quality, improving sample efficiency for real-world RL. Table 5 provides a consolidated comparison of representative VLA-RL methods.

8.2 World Models for Manipulation

World models learn environment dynamics $p(s_{t+1}, r_t \mid s_t, a_t)$, providing three capabilities that directly address VLA limitations: *imagined rollouts* for sample-efficient policy optimisation without real interaction, *safety*

rehearsal by simulating dangerous scenarios virtually, and *long-horizon planning* that compensates for the end-to-end nature of VLAs which sacrifices explicit lookahead.

Table 5: Representative VLA+RL methods. *Optim. level*: T = trajectory, S = step. *Human role*: D = demos only, C = corrections/interventions, P = preferences.

Method	Optim.	Human	Key Mechanism	Reported Gain
VLA-RL Lu et al. (2025a)	T	D	Trajectory value + PRM	Improved long-horizon SR
TGRPO Chen et al. (2025e)	T	D	Group relative policy optim.	Competitive with complex methods
SimpleVLA-RL Li et al. (2025a)	S	D	GRPO with tuned hyperparams	Scales with minimal complexity
RECAP (π_0) Physical Intelligence et al. (2025)	S	D+C	Advantage-conditioned policy	60%→85% SR (100–200h)
HIL-SERL Luo et al. (2025)	S	D+C	Demo-guided SAC + safety	Fine manip. in ∼3h
GRAPE Zhang et al. (2024b)	T	P	RLHF for manipulation	No hand-designed rewards
RIPT-VLA Tan et al. (2025)	S	C	Real-time interactive updates	Rapid online adaptation
Embodied-R1 Yuan et al. (2025)	S	D	Joint reasoning + action RL	Improved deliberation
VLAC Zhai et al. (2025)	S	D	Shared-repr. critic	Better sample efficiency

From Dreamer to Robotic World Models. The Dreamer series Hafner et al. (2020; 2023) established the paradigm of learning dynamics in a compact latent space rather than raw pixels: an encoder compresses observations, a recurrent transition model propagates latent states forward, and a decoder reconstructs observations for training. DreamerV3 Hafner et al. (2023) achieves strong results across Atari, DMControl, and Minecraft through mixed discrete–continuous latent representations.

Adapting this paradigm to robotic manipulation introduces three key challenges that exceed the complexity of game and locomotion domains: (i) *contact-rich dynamics*—friction, collision, and elastic deformation are difficult to capture in learned latent models; (ii) *partial observability*—monocular RGB misses occlusions, internal joint states, and contact forces; and (iii) *long-tail failure modes*—rare but critical events (object drops, jams) are undersampled in training data yet dominate real-world risk.

VLA-Integrated World Models. Recent work addresses these challenges by designing world models as *VLA-specific modules* that predict richer representations than visual futures alone.

DreamVLA Zhang et al. (2025g) adds multi-dimensional world-knowledge prediction heads to a shared VLA visual encoder, jointly predicting dynamics, spatial relationships (3D object poses, contact states), and semantic states (task progress such as "door handle rotated 45°"). On RLBench long-horizon tasks, adding world-knowledge prediction improves success rates by 12% over baselines without world models. RynnVLA-002 Cen et al. (2025a) unifies policy and world model with shared bottom-layer representations and bidirectional enhancement: predicted future states serve as additional policy context, while action-selection history improves dynamics prediction. On the CALVIN benchmark, this unified architecture improves single-task success from 62% to 73% compared with separated models.

An alternative strategy uses world models for *offline data augmentation* rather than online planning. DreamGen Jang et al. (2025) trains an initial world model from few real demonstrations, samples diverse trajectories in imagination, filters low-quality rollouts, and jointly trains the VLA on real and synthetic data. On BridgeData V2, 1000 real plus 5000 synthetic demonstrations match the performance of 3000 pure real demonstrations—a particularly effective strategy for long-tail tasks where real data is scarce.

Autoregressive and Video-Based Approaches. A parallel line of work explores autoregressive architectures that jointly model visual futures and actions within a unified token sequence. WorldVLA Cen et al. (2025b) treats both visual and action tokens as parts of a single autoregressive sequence, enabling mutual reinforcement between action generation and visual prediction. FlowVLA Zhong et al. (2025) uses optical flow prediction as an intermediate representation, achieving more robust dynamics modelling that generalises across visual variations. Meta's V-JEPA 2 Assran et al. (2025) learns predictive representations through joint-embedding prediction in latent space, avoiding expensive pixel-level reconstruction while supporting understanding, prediction, and planning for embodied tasks. Unlike frameworks based on explicit dynamics modeling, $\pi_{0.7}$ Ai et al. (2026) adopts a steerable visual-guidance paradigm. By conditioning the action expert on subgoal images that serve as intermediate visual targets, the model reconciles high-level semantic

intent with low-level reactive control. This approach provides a scalable mechanism for the coordination of long-horizon tasks without the necessity for explicit latent dynamics simulation.

8.3 Learning from Human Videos

A critical bottleneck in VLA training is the scarcity of robot demonstration data. Human video—abundant on platforms like YouTube, Ego4D, and EPIC-KITCHENS—captures rich manipulation knowledge but lacks action labels and exhibits a substantial domain gap from robot observations. Recent methods bridge this gap through two main strategies.

Latent Action Learning. The core insight is that consecutive video frames implicitly encode the "action" that caused the visual transition. LAPA Ye et al. (2024) trains an inverse-dynamics model to predict latent actions from frame pairs, then uses these latent representations to pre-train VLA policies; during fine-tuning on robot data, the latent space is aligned to real action spaces. UniVLA Bu et al. (2025) extends this with task-centric clustering, grouping latent actions by task semantics for better transfer to specific manipulation categories.

Domain Alignment and Video-to-Action Generation. HumanRobotAlign Zhou et al. (2025a) addresses the visual domain gap directly by training domain-invariant encoders that map both human hands and robot grippers to shared representations, enabling direct policy transfer. For humanoid robots, whose morphology resembles the human body, Humanoid-VLA Ding et al. (2025) exploits keypoint correspondences for action retargeting without explicit action labels.

A more ambitious approach uses video generation as an intermediate step. Gen2Act Bharadhwaj et al. (2024) synthesises robot-execution videos from human demonstrations using video generation models, then extracts actions from the generated robot videos. RigVid Patel et al. (2025) pushes further toward zero-shot execution: given only a task description, it generates a video of successful execution and applies inverse dynamics to extract actions, bypassing the need for any physical demonstration.

These methods are complementary to robot-centric data scaling (Section 7): human videos provide breadth of manipulation knowledge, while robot data provides the embodiment-specific grounding needed for precise control.

8.4 Synergy and Open Challenges

Discussion: the Complementarity Argument. The components reviewed in this section are not competing alternatives but complementary layers of a unified system. VLA foundation models provide the *generalisation prior*, broading cross-task, cross-embodiment knowledge learned from diverse offline data. Reinforcement learning provides *task specialisation* which is environment-specific optimisation through online interaction that corrects compounding errors and discovers behaviour beyond the demonstration envelope. World models provide *sample efficiency and safety* by reducing the need for costly real interaction through imagined rollouts and enabling safety verification before physical execution. Human video learning provides *data scaling* on leveraging abundant non-robot data to expand the knowledge base without proportional robot-hours.

Early evidence of synergistic integration is emerging. NORA-1.5 Hung et al. (2025) trains world models to predict action preferences ("action A more likely to succeed than action B") rather than exact pixel futures, using preference signals as reward for policy RL—a strategy well-suited to contact-rich tasks where pixel prediction is unreliable. RoboScape-R Tang et al. (2025) jointly predicts observations and rewards in a unified world model, providing dense training signals for RL even in sparse-reward manipulation environments.

Key Open Challenges. Despite promising results, several fundamental challenges must be addressed before VLA+RL systems can achieve genuine continual learning in the real world.

World model fidelity for contact. Learned world models still struggle with contact-rich dynamics: friction, collision, and deformation involve discontinuities and high-dimensional state that are poorly captured

by current latent-space architectures. Integrating explicit physics priors—differentiable physics engines Heiden et al. (2021b), tactile sensing (Section 6)—may be necessary to bridge this gap.

Real-interaction cost. Even with world-model augmentation, hundreds of real-robot trajectories are typically needed for RL fine-tuning. Multi-robot parallelism, improved sim-to-real transfer, and better human-in-the-loop protocols remain essential for making VLA+RL practical outside well-resourced labs.

Reward design. Hand-designing manipulation reward functions is notoriously difficult. Learning rewards from human preferences Zhang et al. (2024b) or generating them via language models Zhang et al. (2025d) are promising directions, but their robustness and scalability to diverse open-world tasks remain unproven.

Training stability. RL gradients applied to billion-parameter VLAs can cause optimisation instability. Consistency regularisation Chen et al. (2025d), trajectory-level optimisation Lu et al. (2025a), and progressive stage-aware training Xu et al. (2025a) are partial solutions, but principled methods for stable large-model RL remain an open problem.

From episodic adaptation to lifelong learning. Current VLA+RL methods focus on episodic adaptation—fine-tuning on a specific task for hours, then freezing. True embodied intelligence requires lifelong learning: accumulating skills across tasks, actively identifying knowledge gaps, and balancing retention of old capabilities with acquisition of new ones. Modular memory systems, meta-reinforcement learning, and multi-robot collaborative learning are promising but largely unexplored directions in the VLA context.

In summary, the combination of VLA foundation models with reinforcement learning, world models, and human-video learning defines the frontier of embodied manipulation research. VLAs provide the broad prior; RL provides task-specific refinement; world models provide sample efficiency and safety; and human videos provide scalable knowledge. The transition from "large-scale pre-trained static VLAs" to "continuously evolving embodied agents" will require progress on all four fronts simultaneously—a challenge that motivates the benchmarking and evaluation frameworks discussed in the next section.

9 Datasets & Benchmarks

The rise of Vision-Language-Action (VLA) systems has made embodied evaluation substantially more demanding. The core question is no longer whether a policy can solve a fixed benchmark layout, but whether benchmark performance reliably predicts deployment behavior on real robots. Simple success-rate leaderboards are increasingly insufficient: models can exploit protocol regularities (e.g., static layouts, narrow perturbations, clean sensing) while remaining brittle under contact variation, latency, or scene shift. This section therefore reviews datasets and benchmarks through a reliability-oriented lens, emphasizing generalization, robustness, efficiency, and safety in addition to nominal task completion.

9.1 Simulation benchmarks

Simulation remains the dominant method for scalable evaluation due to its convenience and reproducibility. However, the VLA era highlights a central failure mode: under overly clean and weakly randomized simulation environments, policies may behave as if they "understand" instructions, while they are replaying memorized action patterns in essence. Recent trends on benchmark design can be summarized as four directions: (i) scaling up task diversity; (ii) constructing VLA-oriented benchmarks; (iii) introducing systematic robustness and generalization tests; and (iv) building real-relevant simulation proxies with rankings correlated with real-world outcomes.

9.1.1 Large-scale multi-task simulation benchmarks

RLBench(James et al., 2020). RLBench provides various hand-designed manipulation tasks with standardized success criteria and unified proprioceptive observations, making it a widely used benchmark for multi-task IL/RL and vision-guided manipulation. Its major strength lies in standardization, while a typical limitation

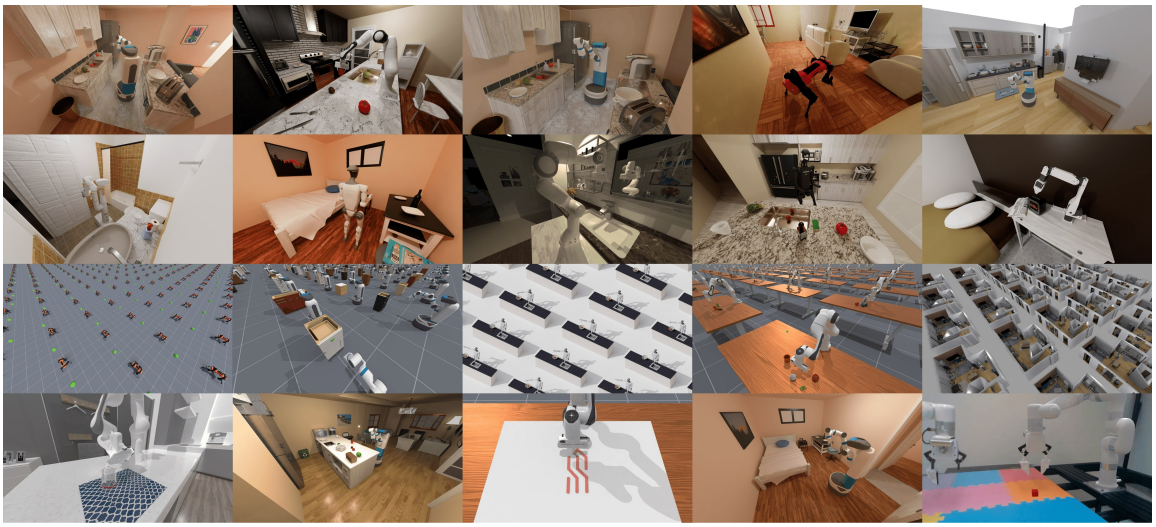

Figure 8: Large-scale embodied tasks covered in ManiSkill3

is that task/scene structure can be relatively fixed, allowing strong policies to exploit layout/geometric regularities.

Robosuite(Zhu et al., 2020). Robosuite is a modular MuJoCo-based simulation framework and benchmark suite designed for reproducible robot-learning research, supporting multiple robots and extensible task composition.

ManiSkill2(Gu et al., 2023). ManiSkill2 targets generalizable manipulation evaluation with broader task families and object diversities, and provides unified interfaces to compare policies more fairly under a larger range of tasks.

ManiSkill3(Tao et al., 2025). As shown in Figure 8, ManiSkill3 pushes scalability via GPU-parallel simulation and rendering, enabling evaluation with many random seeds and perturbation grids at feasible cost. This supports not only estimates of success rates, but also the stability and sensitivity of models with failure-mode statistics.

Meta-World(Yu et al., 2020) & **Meta-World+**(McLean et al., 2025). Meta-World is a classic benchmark for multi-task and meta-RL over many manipulation tasks, emphasizing standardized APIs and protocols; Meta-World+ revisits reproducibility issues and releases a more standardized and feasible version.

RoboCasa(Nasiriany et al., 2024). RoboCasa focuses on more realistic household environments and everyday task structures, advocating scalable simulation as a route to broaden environment and task coverage for generalist robotics.

Scaling task coverage helps, but once a benchmark becomes mainstream, overfitting and shortcut strategies tend to emerge, making it less significant on real-world robot performance.

9.1.2 VLA-oriented simulation benchmarks

VLA benchmarks introduce language and richer perceptual inputs, but also increase the risk of "answer memorization": a policy may replay layout-specific trajectories while appearing to follow instructions. Modern VLA-oriented benchmarks therefore emphasize long-horizon inference, stronger randomization, instruction sensitivity and reasoning ability.

CALVIN(Mees et al., 2022). CALVIN is a long-horizon language-conditioned manipulation benchmark that evaluates sequential instruction following and composition under environment or object variations, making it a common testbed for language-conditioned policy learning beyond single-step tasks.

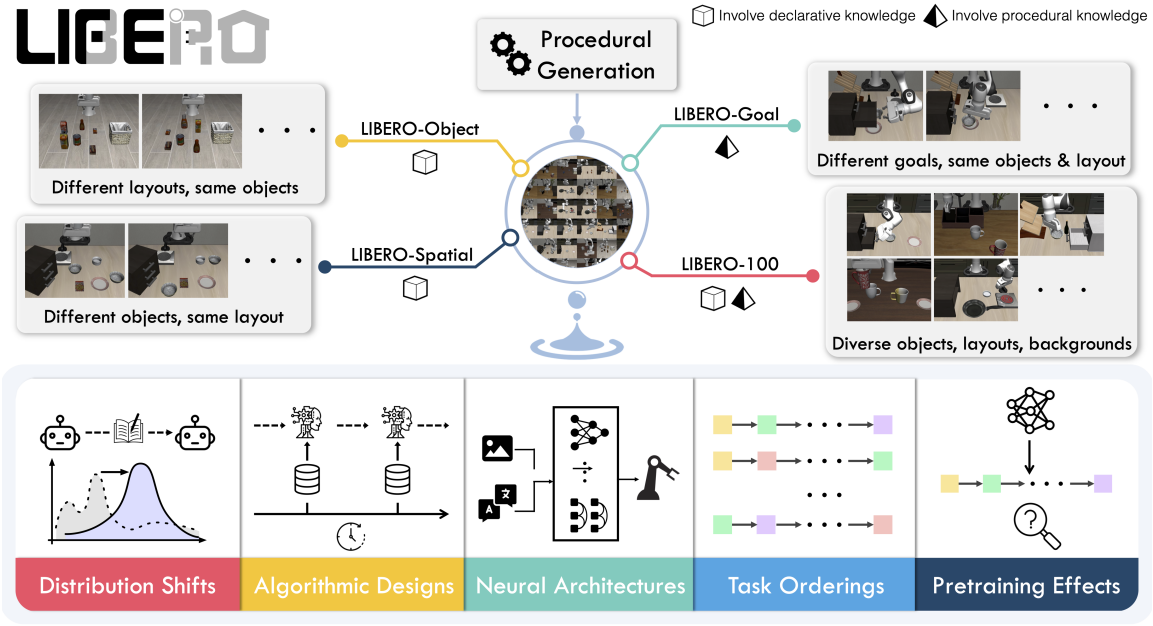

Figure 9: Overview of classic LIBERO benchmark

LIBERO(Liu et al., 2023a). LIBERO is a classic benchmark suite for knowledge-transfer lifelong robot manipulation, generating language-conditioned tasks organized to probe distribution shifts and transfer across tasks. Figure 9 gives an overview of this benchmark.

LIBERO-Plus(Fei et al., 2025). LIBERO-Plus performs systematic robustness analysis for VLA models via controlled perturbations across multiple dimensions (layout, viewpoint, robot initial states, instruction variation, lighting, background and sensor noise, etc.), revealing brittleness hidden by standard evaluation.

LIBERO-PRO(Zhou et al., 2025b). LIBERO-PRO argues that standard protocols can lead to inflated scores and proposes more robust generalization settings. Under these settings, performance of high-rank models on other benchmarks can collapse dramatically, exposing their memorization-driven behaviors.

SafeLIBERO(Hu et al., 2025). SafeLIBERO extends language-conditioned tasks with safety-critical obstacle or interference scenarios and evaluates both task completion and unsafe outcomes (e.g., collisions), shifting emphasis toward deployment-relevant safety metrics.

VLABench(Zhang et al., 2025f). VLABench emphasizes long-horizon language-conditioned manipulation together with knowledge transfer, aiming to jointly evaluate action policy and instruction understanding of models, which aligns more closely with what VLAs claim to provide.

9.1.3 Robustness and generalization tests

A key trend of recent benchmarks on embodied manipulation is to make generalization evaluation reproducible by defining perturbations and reporting sensitivity profiles, and validate which simulated perturbations correlate with real-world disturbances.

THE COLOSSEUM(Pumacay et al., 2024). THE COLOSSEUM defines perturbations across multiple axes (appearance, lighting, distractors, camera pose and physical properties, etc.) and reports correlations between simulation findings and analogous real-world perturbations.

RoboEval(Wang et al., 2025f). RoboEval argues that simple success rate metric hides behavior-quality issues (e.g., slipping, unstable contact and poor bimanual coordination) ,which shows that successful executions may not have high quality. It proposes structured, scalable evaluation to overcome such weaknesses.

9.1.4 Real-relevant simulation evaluation

Because real-robot evaluation is expensive and hard to reproduce, a promising direction is to make simulation a validated proxy by aligning sim-to-real gaps and measuring correlation with real-world outcomes.

SIMPLER(Li et al., 2024b). SIMPLER builds simulated evaluation environments aligned with common real-robot setups and demonstrates strong correlation between SIMPLER scores and real-world performance in paired sim-and-real evaluations.

9.2 Real-robot datasets and benchmarks

Real-world datasets address aspects that simulation cannot fully capture, including uncontrolled visual variation, contact dynamics, latency, sensor noise, and human or scene diversity. Beyond the problem of scale, a major practical challenge is evaluation consistency (e.g., demonstration style, action-space/control-rate conventions, semantic labeling). Without careful standardization, models may learn shortcuts rather than transferable abilities.

9.2.1 Large-scale real-world datasets

BridgeData V2(Walke et al., 2023). BridgeData V2 provides a large, diverse set of real manipulation trajectories collected across many environments on a low-cost robot platform, supporting open-vocabulary conditioning like goal images or language.

DROID(Khazatsky et al., 2024). DROID targets "in-the-wild" diversity by distributing real-world collection across many scenes, tasks, and collectors, explicitly capturing variation in viewpoints, clutter, lighting, and layouts.

RT-1(Brohan et al., 2022). RT-1 introduces the Robotics Transformer and trains it on large-scale real multi-task demonstrations, highlighting the role of scaling task-agnostic data for generalist embodied manipulation.

RH20T(Fang et al., 2024). RH20T focuses on contact-rich manipulation across diverse skills or modalities, and supports few-shot imitation style evaluation on real robots.

AgiBot World(AgiBot, 2025). AgiBot World emphasizes broad scenario coverage and large-scale trajectory collection with standardized pipelines, aiming to support scalable generalist manipulation learning.

9.2.2 Cross-embodiment evaluation

Open X-Embodiment(O'Neill et al., 2024). Open X-Embodiment aggregates data across many real-world embodiments and tasks, enabling systematic evaluation of cross-embodiment transfer under unified dataset organization and model training pipelines.

9.2.3 Failure data and reliability

RoboMIND(Wu et al., 2024). RoboMIND provides a unified multi-embodiment dataset and explicitly includes failure demonstrations with annotated causes, supporting evaluation of failure reflection, correction, and reliability beyond average success rates. Figure 10 shows different criteria concerned by RoboMIND to recover from failure data and improve reliability in embodied evaluation.

9.2.4 A reproducibility–realism compromise

RoboTwin(Mu et al., 2025) & **RoboTwin 2.0**(Chen et al., 2025c). RoboTwin combines generative digital twins with dual-arm task evaluation, mixing simulation expert data and real teleoperation data under unified task and asset settings to improve both reproducibility and real-world alignment.

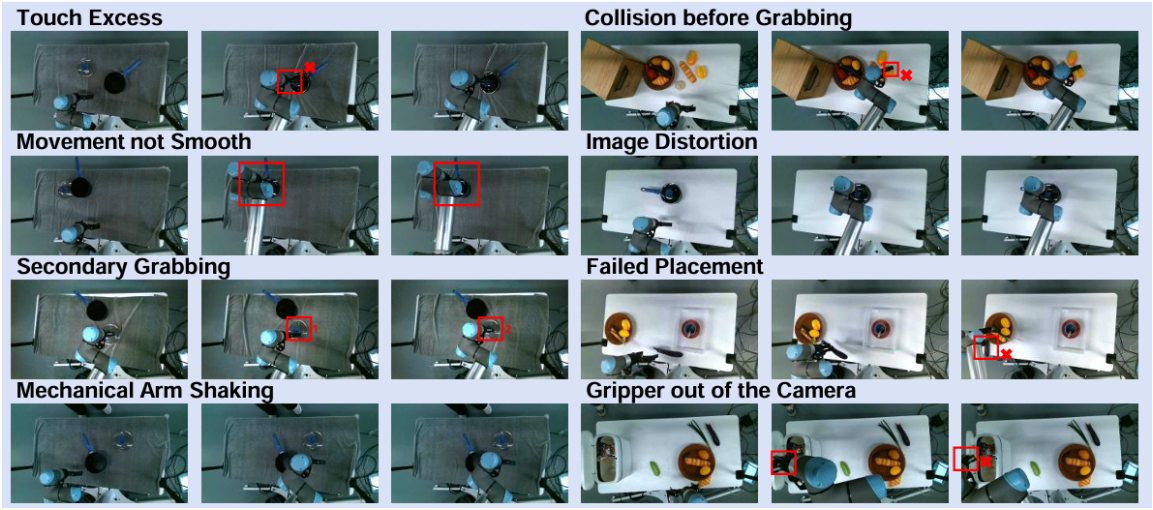

Figure 10: Quality assurance criteria defined in RoboMIND

9.3 Conclusion: balance between simulated reproducibility and real-world reliability

A central tension on current embodied benchmarks is that simulation is scalable and reproducible but may be unreliable, while real-robot evaluation is faithful but expensive and hard to reproduce. Two recurring failure patterns follow: policies can score highly in narrow simulation protocols yet fail on real robots due to contact, latency or visual gaps, and simple success rates can obscure behavior quality where "successful" rollouts are still unsafe, unstable, or inefficient. Benchmarks thus increasingly push three complementary directions: validated sim proxies, quantifiable behavior quality, and failure-aware evaluation with safety or recovery considerations.

10 Challenges, Open Problems, and Outlook

This survey has traced how robotic manipulation systems evolved from the engineering-heavy pipelines of the DRC era to today's VLA foundation models, driven by the intertwined scaling of models, visual data, and physical interaction data. Yet the distance between impressive laboratory demonstrations and reliably deployed embodied intelligence remains substantial. In this closing section we identify four directional pillars that, in our view, define the research agenda for the next stage of progress.

10.1 From Policy Scaling to Interaction Scaling

The dominant paradigm of the past three years has been *policy scaling*: training ever-larger models on ever-larger offline datasets. GEN-0 Team (2025) demonstrates that real-world manipulation performance improves predictably with data and model capacity, and Open X-Embodiment O'Neill et al. (2024) shows that cross-institution data aggregation can accelerate this process. Yet offline scaling has inherent limits.

First, the training distribution is *coverage-bounded*: however large the dataset, it cannot anticipate every contact event, material property, or failure mode a robot will encounter in deployment. Second, the data generation rate is *throughput-bounded*: even with distributed collection Khazatsky et al. (2024) and UMI-style hardware-free interfaces Chi et al. (2024), acquiring diverse physical interaction data remains orders of magnitude slower than scraping web text or images. Third, offline data is inherently *improvement-bounded*: a policy trained on fixed demonstrations cannot, by construction, discover strategies that surpass its teachers.

True embodied intelligence therefore requires a shift from policy scaling to *interaction scaling*—systems that learn continually through trial-and-error in the real world, actively seek out informative experiences, and accumulate skills over their operational lifetime. The VLA+RL methods reviewed in Section 8—

RECAP Physical Intelligence et al. (2025), HIL-SERL Luo et al. (2025), trajectory-level optimisation Lu et al. (2025a)—are early steps in this direction, but they remain episodic: fine-tune on one task for hours, then freeze. Scaling from episodic adaptation to lifelong, autonomous learning requires solving several intertwined sub-problems:

Active data collection. Robots should identify knowledge gaps in their current policy and actively design interactions—choosing objects, scenes, and strategies—that maximise information gain, rather than passively replaying pre-scripted data-collection protocols.

Catastrophic forgetting. As robots acquire new skills through online interaction, they must retain previously learned capabilities. Modular memory systems Ajay et al. (2023), experience replay, and progressive network expansion are promising but largely unexplored in the VLA context.

Multi-robot collaborative learning. A fleet of robots deployed across diverse environments collectively encounters far more distributional variety than any single platform. Federated or collaborative learning protocols that share experience without centralising raw data could dramatically accelerate interaction scaling while respecting privacy and bandwidth constraints.

10.2 Safety, Alignment, and Human Oversight

The failure modes of embodied AI are physical: a misaligned grasp shatters a glass, a wayward arm collides with a human, a mis-planned insertion jams an assembly line. Unlike language-model hallucinations, which produce incorrect text, embodied errors produce forces, impacts, and irreversible state changes. Recent policy analysis has identified physical, informational, economic, and social risk categories specific to embodied AI systems Perlo et al. (2025), underscoring that existing frameworks for industrial robotics and autonomous vehicles are insufficient for the generality of foundation-model-driven robots.

Safety layers for end-to-end models. Hierarchical manipulation systems benefit from explicit safety primitives: collision checkers, impedance controllers, and workspace boundaries. End-to-end VLAs sacrifice these modular safeguards. Restoring safety without sacrificing the generality of VLAs requires new architectures that combine learned policies with verifiable safety constraints—through constrained RL, learned barrier functions Xiao et al. (2023); Hewing et al. (2020), or runtime monitoring layers that override unsafe actions. SafeVLA Zhang et al. (2025a) and the SafeLIBERO benchmark Hu et al. (2025) represent initial steps, but formal safety guarantees for foundation-model policies remain an open problem.

Alignment with human intent. As VLAs become more autonomous, ensuring that their behaviour aligns with human preferences—not just task success, but also style, risk tolerance, and social norms—becomes critical. Preference learning Zhang et al. (2024b) and language-guided reward shaping Zhang et al. (2025d) adapt techniques from LLM alignment to robotics, but the embodied setting introduces unique challenges: reward signals are sparse, delayed, and entangled with physical dynamics, and the cost of misalignment is physical rather than reputational.

Human oversight at scale. Current human-in-the-loop approaches (HIL-SERL Luo et al. (2025), RIPT-VLA Tan et al. (2025)) require a human operator per robot. Deploying embodied agents at fleet scale demands new oversight paradigms: exception-based monitoring where humans intervene only on flagged uncertainty, shared-autonomy interfaces that allow variable levels of control delegation, and transparent reasoning traces (CoT-VLA Zhao et al. (2025), ECoT Zawalski et al. (2024)) that enable human auditing of robot decisions.

10.3 Multi-Modal, Multi-Body, Multi-Agent

Current VLA models overwhelmingly consume RGB images and output end-effector commands for single-arm desktop manipulation. Expanding along three "multi" axes is essential for the next generation of embodied intelligence.

Multi-modal sensing. As discussed in Section 7, tactile, force/torque, and audio modalities are not optional add-ons for dexterous manipulation—they are prerequisites for object-property estimation, contact-safe interaction, and fine manipulation under occlusion. Yet these modalities remain severely under-represented in large-scale robot datasets. Wearable capture systems (OSMO Yin et al. (2025), DexCap Wang et al. (2024a)) and plug-and-play tactile skins (AnySkin Bhirangi et al. (2025)) offer paths to scalable multimodal data collection, but integrating these signals into VLA architectures at training time—rather than as post-hoc additions—remains a significant engineering and modelling challenge.

Multi-body platforms. The field is rapidly expanding beyond single-arm tabletop setups. Humanoid robots are transitioning from laboratory curiosities to early industrial deployment, with companies like Figure, Unitree, and NVIDIA (GR00T Bjorck et al. (2025)) investing heavily in whole-body loco-manipulation. Dual-arm coordination is now a mainstream research topic (RoboTwin Mu et al. (2025); Chen et al. (2025c)). Dexterous hands add an order of magnitude in action-space dimensionality. Each of these embodiments requires rethinking action representations, training data, and safety constraints. Cross-embodiment transfer—training a single VLA that generalises from a Franka arm to a humanoid torso—is an ambitious goal that current architectures can support in principle O'Neill et al. (2024) but not yet reliably in practice.

Multi-agent collaboration. Real-world tasks frequently require coordination among multiple robots (e.g., collaborative assembly, warehouse logistics) or between robots and humans. Multi-agent embodied intelligence introduces challenges in communication protocols, task allocation, shared world models, and safety in shared workspaces. These issues are largely orthogonal to the single-agent VLA paradigm reviewed in this survey and represent a significant open frontier.

10.4 Bridging Cognitive Models and Robot Learning

Current VLA models are, at their core, statistical pattern matchers: they learn correlations between visual-linguistic inputs and action outputs from massive data. This approach achieves impressive interpolation within the training distribution but offers limited guarantees on out-of-distribution behaviour, physical consistency, or compositional reasoning.

A complementary research direction seeks to introduce explicit cognitive and physical structure into the learning pipeline—not to replace data-driven methods, but to provide inductive biases that improve sample efficiency, interpretability, and robustness.

Physics-informed world models. Current learned world models (Section 8) operate in latent spaces that encode dynamics implicitly. Integrating explicit physics priors—differentiable rigid-body and contact simulators Heiden et al. (2021a), material-property estimators, kinematic constraints—could dramatically improve fidelity for contact-rich manipulation without requiring orders of magnitude more training data. The challenge is to make such integration modular and differentiable, so that physics priors can be "plugged into" VLA training loops without breaking end-to-end gradient flow.

Symbolic abstraction and compositional planning. Language provides a natural interface for compositional task specification, but current VLAs do not reason symbolically over task structure. Code-as-Policies Liang et al. (2023) and VoxPoser Huang et al. (2023) demonstrate that LLMs can generate executable plans, but these rely on pre-defined primitive libraries. Learning to discover, name, and compose new abstractions from interaction—a form of program synthesis grounded in physical experience—could enable the combinatorial generalisation that statistical pattern matching alone struggles to achieve.

Causal and counterfactual reasoning. Manipulation is fundamentally causal: pushing an object *causes* it to move; tightening a screw *causes* the joint to lock. Yet current VLA training optimises for statistical association, not causal structure. Incorporating causal inference frameworks—interventional training, counterfactual data augmentation via world models Jang et al. (2025)—could improve robustness to spurious correlations and enable more reliable transfer to novel environments.

10.5 Closing Remarks

The trajectory of the past decade, from the DRC's engineered pipelines to today's VLA foundation models, can be understood as a sustained expansion along two scaling axes: model capacity and offline data volume (Figure 1). The four challenge pillars identified above collectively point toward a third axis—*interaction scaling*—that we believe will define the next era of embodied intelligence. Progress along this axis demands not only algorithmic innovation (lifelong RL, physics-informed world models, causal reasoning) but also infrastructure innovation (multi-robot fleets, standardised safety protocols, multimodal data pipelines) and institutional innovation (shared benchmarks, open-source platforms, cross-disciplinary collaboration between robotics, cognitive science, and AI safety).

Ten years ago, the DRC teams controlled barely stable humanoids with painstakingly engineered perception–planning–control pipelines. Today, general-purpose VLA models begin to control diverse robots through natural language. The embodied intelligence we seek will likely emerge not from either extreme alone, but from the productive tension between data-driven generality and structured physical understanding—between scaling and reasoning, between learning and engineering. Navigating this tension, with safety and human oversight as non-negotiable constraints, is the central challenge and the central opportunity of the decade ahead.

References

AgiBot. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Bo Ai, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Greg Balke, Kevin Black, George Bokinsky, Shihao Cao, Thomas Charbonnier, Vedant Choudhary, James Darpinian, Danny Driess, Chelsea Finn, Karol Hausman, Brian Ichter, Sergey Levine, et al. $\pi_{0.7}$: A steerable generalist robotic foundation model with emergent capabilities. *preprint*, 2026. URL `https://pi.website/pi07`.

Anurag Ajay, Seungwook Han, Yilun Du, Shuang Li, Abhi Gupta, Tommi Jaakkola, Josh Tenenbaum, Leslie Kaelbling, Akash Srivastava, and Pulkit Agrawal. Compositional foundation models for hierarchical planning. *Advances in Neural Information Processing Systems*, 36:22304–22325, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Christopher G Atkeson, Benzun P Wisely Babu, Nandan Banerjee, Dmitry Berenson, Christoper P Bove, Xiongyi Cui, Mathew DeDonato, Ruixiang Du, Siyuan Feng, Perry Franklin, et al. No falls, no resets: Reliable humanoid behavior in the darpa robotics challenge. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pp. 623–630. IEEE, 2015.

Seunghyeok Back, Joosoon Lee, Taewon Kim, Sangjun Noh, Raeyoung Kang, Seongho Bak, and Kyoobin Lee. Unseen object amodal instance segmentation via hierarchical occlusion modeling. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 5085–5092. IEEE, 2022.

Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language. *ArXiv*, abs/2403.01823, 2024. URL `https://api.semanticscholar.org/CorpusID:268249108`.

Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.

Raunaq Bhirangi, Tess Hellebrekers, Carmel Majidi, and Abhinav Gupta. Reskin: versatile, replaceable, lasting tactile skins. *arXiv preprint arXiv:2111.00071*, 2021.

Raunaq Bhirangi, Venkatesh Pattabiraman, Enes Erciyes, Yifeng Cao, Tess Hellebrekers, and Lerrel Pinto. Anyskin: Plug-and-play skin sensing for robotic touch. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16563–16570. IEEE, 2025.

Jianxin Bi, Kevin Yuchen Ma, Ce Hao, Mike Zheng Shou, and Harold Soh. Vla-touch: Enhancing vision-language-action models with dual-level tactile feedback. *ArXiv*, abs/2507.17294, 2025. URL `https://api.semanticscholar.org/CorpusID:280228347`.

Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446), 2019.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Sanjeev Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. doi: 10.48550/arXiv.2212.06817. URL `https://arxiv.org/abs/2212.06817`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

Jiahang Cao, Yize Huang, Hanzhong Guo, Rui Zhang, Mu Nan, Weijian Mai, Jiaxu Wang, Hao Cheng, Jingkai Sun, Gang Han, et al. Compose your policies! improving diffusion-based or flow-based robot policies via test-time distribution-level composition. *arXiv preprint arXiv:2510.01068*, 2025.

Jun Cen, Siteng Huang, Yuqian Yuan, Kehan Li, Hangjie Yuan, Chaohui Yu, Yuming Jiang, Jiayan Guo, Xin Li, Hao Luo, et al. Rynnvla-002: A unified vision-language-action and world model. *arXiv preprint arXiv:2511.17502*, 2025a.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025b.

Hao Chen, Jiaming Liu, Chenyang Gu, Zhuoyang Liu, Renrui Zhang, Xiaoqi Li, Xiao He, Yandong Guo, Chi-Wing Fu, Shanghang Zhang, et al. Fast-in-slow: A dual-system foundation model unifying fast manipulation within slow reasoning. *arXiv preprint arXiv:2506.01953*, 2025a.

Haojun Chen, Minghao Liu, Chengdong Ma, Xiaojian Ma, Zailin Ma, Huimin Wu, Yuanpei Chen, Yifan Zhong, Mingzhi Wang, Qing Li, et al. Falcon: Fast visuomotor policies via partial denoising. *arXiv preprint arXiv:2503.00339*, 2025b.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025c.

Yuhui Chen, Shuai Tian, Shugao Liu, Yingting Zhou, Haoran Li, and Dongbin Zhao. Conrft: A reinforced fine-tuning method for vla models via consistency policy. *arXiv preprint arXiv:2502.05450*, 2025d.

Zengjue Chen, Runliang Niu, He Kong, Qi Wang, Qianli Xing, and Zipei Fan. Tgrpo: Fine-tuning vision-language-action model via trajectory-wise group relative policy optimization. *arXiv preprint arXiv:2506.08440*, 2025e.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2818–2829, 2023.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Pengxiang Ding, Jianfei Ma, Xinyang Tong, Binghong Zou, Xinxin Luo, Yiguo Fan, Ting Wang, Hongchao Lu, Panzhong Mo, Jinxin Liu, et al. Humanoid-vla: Towards universal humanoid control with visual integration. *arXiv preprint arXiv:2502.14795*, 2025.

Shaoqi Dong, Chaoyou Fu, Haihan Gao, Yi-Fan Zhang, Chi Yan, Chu Wu, Xiaoyu Liu, Yunhang Shen, Jing Huo, Deqiang Jiang, et al. Vita-vla: Efficiently teaching vision-language models to act via action expert distillation. *arXiv preprint arXiv:2510.09607*, 2025.

Siyuan Dong, Wenzhen Yuan, and Edward H Adelson. Improved gelsight tactile sensor for measuring geometry and slip. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 137–144. IEEE, 2017.

Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021.

Ben Eisner, Harry Zhang, and David Held. Flowbot3d: Learning 3d articulation flow to manipulate articulated objects. *arXiv preprint arXiv:2205.04382*, 2022.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 653–660. IEEE, 2024.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Siyuan Feng, Eric Whitman, X Xinjilefu, and Christopher G Atkeson. Optimization-based full body control for the darpa robotics challenge. *Journal of field robotics*, 32(2):293–312, 2015.

Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7081–7091, 2023.

Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*, pp. 694–710. PMLR, 2023.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. Maniskill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations*, 2023.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028. IEEE, 2024.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Eric Heiden, Miles Macklin, Yashraj Narang, Dieter Fox, Animesh Garg, and Fabio Ramos. Disect: A differentiable simulation engine for autonomous robotic cutting. *arXiv preprint arXiv:2105.12244*, 2021a.

Eric Heiden, David Millard, Erwin Coumans, Yizhou Sheng, and Gaurav S. Sukhatme. Interactive differentiable simulation. *arXiv preprint arXiv:2108.10073*, 2021b.

Erik Helmut, Niklas Funk, Tim Schneider, Cristiana de Farias, and Jan Peters. Tactile-conditioned diffusion policy for force-aware robotic manipulation. *arXiv preprint arXiv:2510.13324*, 2025.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Lukas Hewing, Kim P Wabersich, Marcel Menner, and Melanie N Zeilinger. Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, 3 (1):269–296, 2020.

Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 19–34, 2018.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Songqiao Hu, Zeyi Liu, Shuang Liu, Jun Cen, Zihan Meng, and Xiao He. Vlsa: Vision-language-action models with plug-and-play safety constraint layer. *arXiv preprint arXiv:2512.11891*, 2025.

Jingshun Huang, Haitao Lin, Tianyu Wang, Yanwei Fu, Xiangyang Xue, and Yi Zhu. Cap-net: A unified network for 6d pose and size estimation of categorical articulated parts from a single rgb-d image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11654–11664, 2025.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

Chia-Yu Hung, Navonil Majumder, Haoyuan Deng, Liu Renhang, Yankang Ang, Amir Zadeh, Chuan Li, Dorien Herremans, Ziwei Wang, and Soujanya Poria. Nora-1.5: A vision-language-action model trained using world model-and action-based preference rewards. *arXiv preprint arXiv:2511.14659*, 2025.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 2020.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pp. 991–1002. PMLR, 2022.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

Sunshine Jiang, Xiaolin Fang, Nicholas Roy, Tomás Lozano-Pérez, Leslie Pack Kaelbling, and Siddharth Ancha. Streaming flow policy: Simplifying diffusion / flow-matching policies by treating action trajectories as flow trajectories. *arXiv preprint arXiv:2505.21851*, 2025.

Matt Johnson, Bertrand Shrewsbury, Sylvain Bertrand, Tingfan Wu, Daniel Duran, Marshall Floyd, Peter Abeles, Daniel Stephen, Nathan Mertins, Alex Lesman, et al. Team ihmc's lessons learned from the darpa robotics challenge trials. *Journal of Field Robotics*, 32(2):192–208, 2015.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*, 2024.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Jung-Hoon Kim, Jong Hyun Choi, and Baek-Kyu Cho. Walking pattern generation for a biped walking robot using convolution sum. *Advanced Robotics*, 25(9-10):1115–1137, 2011.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Oliver Kroemer, Scott Niekum, and George Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *Journal of machine learning research*, 22(30):1–82, 2021.

Eric Krotkov, Douglas Hackett, Larry Jackel, Michael Perschbacher, James Pippine, Jesse Strauss, Gill Pratt, and Christopher Orlowski. The darpa robotics challenge finals: Results and perspectives. In *The DARPA robotics challenge finals: Humanoid robots to the rescue*, pp. 1–26. Springer, 2018.

Scott Kuindersma, Robin Deits, Maurice Fallon, Andrés Valenzuela, Hongkai Dai, Frank Permenter, Twan Koolen, Pat Marion, and Russ Tedrake. Optimization-based locomotion planning, estimation, and control design for the atlas humanoid robot. *Autonomous robots*, 40(3):429–455, 2016.

Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022.

Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020.

Chengmeng Li, Junjie Wen, Yaxin Peng, Yan Peng, and Yichen Zhu. Pointvla: Injecting the 3d world into vision-language-action models. *IEEE Robotics and Automation Letters*, 11(3):2506–2513, 2026.

Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Zhaohui Yang, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, et al. Simplevla-rl: Scaling vla training via reinforcement learning. *arXiv preprint arXiv:2509.09674*, 2025a.

Peiyan Li, Yixiang Chen, Hongtao Wu, Xiao Ma, Xiangnan Wu, Yan Huang, Liang Wang, Tao Kong, and Tieniu Tan. Bridgevla: Input-output alignment for efficient 3d manipulation learning with vision-language models. *arXiv preprint arXiv:2506.07961*, 2025b.

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024a.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024b.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International conference on robotics and automation (ICRA)*, pp. 9493–9500. IEEE, 2023.

Jeongsoo Lim, Inho Lee, Inwook Shim, Hyobin Jung, Hyun Min Joe, Hyoin Bae, Okkee Sim, Jaesung Oh, Taejin Jung, Seunghak Shin, et al. Robot system of drc-hubo+ and control strategy of team kaist in darpa robotics challenge finals. *Journal of Field Robotics*, 34(4):802–829, 2017.

Haitao Lin, Zichang Liu, Chilam Cheang, Yanwei Fu, Guodong Guo, and Xiangyang Xue. Sar-net: Shape alignment and recovery network for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6707–6717, 2022a.

Haitao Lin, Yanwei Fu, and Xiangyang Xue. Pourit!: Weakly-supervised liquid perception from a single image for visual closed-loop robotic pouring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 241–251, 2023.

Xingyu Lin, Zhiao Huang, Yunzhu Li, Joshua B Tenenbaum, David Held, and Chuang Gan. Diffskill: Skill abstraction from differentiable physics for deformable object manipulations with tools. *arXiv preprint arXiv:2203.17275*, 2022b.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36: 44776–44791, 2023a.

Fanfan Liu, Feng Yan, Liming Zheng, Chengjian Feng, Yiyang Huang, and Lin Ma. Robouniview: Visual-language model with unified view representation for robotic manipulation. *arXiv preprint arXiv:2406.18977*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Zeyi Liu, Cheng Chi, Eric Cousineau, Naveen Kuppuswamy, Benjamin Burchfiel, and Shuran Song. Maniwav: Learning robot manipulation from in-the-wild audio-visual data. *arXiv preprint arXiv:2406.19464*, 2024b.

Zhenyang Liu, Yongchong Gu, Sixiao Zheng, Xiangyang Xue, and Yanwei Fu. Trivla: A triple-system-based unified vision-language-action model for general robot control. *arXiv e-prints*, pp. arXiv–2507, 2025a.

Zhenyang Liu, Yikai Wang, Kuanning Wang, Longfei Liang, Xiangyang Xue, and Yanwei Fu. Spatial-temporal aware visuomotor diffusion policy learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7122–7131, 2025b.

Zhenyang Liu, Yikai Wang, Sixiao Zheng, Tongying Pan, Longfei Liang, Yanwei Fu, and Xiangyang Xue. Reasongrounder: Lvlm-guided hierarchical feature splatting for open-vocabulary 3d visual grounding and reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3718–3727, 2025c.

Zhenyang Liu, Sixiao Zheng, Siyu Chen, Cairong Zhao, Longfei Liang, Xiangyang Xue, and Yanwei Fu. A neural representation framework with llm-driven spatial reasoning for open-vocabulary 3d visual grounding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1042–1051, 2025d.

Zhenyang Liu, Yongchong Gu, Yikai Wang, Xiangyang Xue, and Yanwei Fu. Activevla: Injecting active perception into vision-language-action models for precise 3d robotic manipulation. *arXiv preprint arXiv:2601.08325*, 2026.

Guanxing Lu, Zifeng Gao, Tianxing Chen, Wenxun Dai, Ziwei Wang, Wenbo Ding, and Yansong Tang. Manicm: Real-time 3d diffusion policy via consistency model for robotic manipulation. *arXiv preprint arXiv:2406.01586*, 2024a.

Guanxing Lu, Shiyi Zhang, Ziwei Wang, Changliu Liu, Jiwen Lu, and Yansong Tang. Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation. In *European Conference on Computer Vision*, pp. 349–366. Springer, 2024b.

Guanxing Lu, Wenkai Guo, Chubin Zhang, Yuheng Zhou, Haonan Jiang, Zifeng Gao, Yansong Tang, and Ziwei Wang. Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning. *arXiv preprint arXiv:2505.18719*, 2025a.

Yiyang Lu, Yufeng Tian, Zhecheng Yuan, Xianbang Wang, Pu Hua, Zhengrong Xue, and Huazhe Xu. H3dp: Triply-hierarchical diffusion policy for visuomotor learning. *arXiv preprint arXiv:2505.07819*, 2025b.

Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. *Science Robotics*, 10(105):eads5033, 2025.

Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. Learning latent plans from play. *arXiv preprint arXiv:1903.01973*, 2020.

Xiao Ma, Sumit Patidar, Iain Haughton, and Stephen James. Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18081–18090, 2024.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv:2310.12931*, 2023.

Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, pp. 879–893. PMLR, 2018.

Reginald McLean, Evangelos Chatzaroulas, Luc McCutcheon, Frank Röder, Tianhe Yu, Zhanpeng He, KR Zentner, Ryan Julian, Jordan K Terry, Isaac Woungang, et al. Meta-world+: An improved, standardized, rl benchmark. *arXiv preprint arXiv:2505.11289*, 2025.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

Jared Mejia, Victoria Dean, Tess Hellebrekers, and Abhinav Gupta. Hearing touch: Audio-visual pretraining for contact-rich manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6912–6919. IEEE, 2024.

Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, et al. Robotwin: Dual-arm robot benchmark with generative digital twins. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27649–27660, 2025.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems*, 2024.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Chuer Pan, Brian Okorn, Harry Zhang, Ben Eisner, and David Held. Tax-pose: Task-specific cross-pose estimation for robot manipulation. In *Conference on Robot Learning*, pp. 1783–1792. PMLR, 2023.

Shivansh Patel, Shraddhaa Mohan, Hanlin Mai, Unnat Jain, Svetlana Lazebnik, and Yunzhu Li. Robotic manipulation by imitating generated videos without physical demonstrations. *arXiv preprint arXiv:2507.00990*, 2025.

Jared Perlo, Alexander Robey, Fazl Barez, Luciano Floridi, and Jakob MĂśkander. Embodied ai: emerging risks and opportunities for policy action. *arXiv preprint arXiv:2509.00117*, 2025.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models, 2025. URL `https://arxiv.org/abs/2501.09747`.

Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, et al. $\pi^*_{0.6}$: A VLA That Learns From Experience. *arXiv preprint arXiv:2511.14759*, 2025.

Aaditya Prasad, Kevin Lin, Jimmy Wu, Linqi Zhou, and Jeannette Bohg. Consistency policy: Accelerated visuomotor policies via consistency distillation. *arXiv preprint arXiv:2405.07503*, 2024.

Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation, 2024. URL `https://arxiv.org/abs/2402.08191`.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Kanchana Ranasinghe, Xiang Li, E-Ro Nguyen, Cristina Mata, Jongwoo Park, and Michael S Ryoo. Pixel motion as universal representation for robot control. *arXiv preprint arXiv:2505.07817*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Omar Rayyan, John Abanes, Mahmoud Hafez, Anthony Tzes, and Fares Abu-Dakka. Mv-umi: A scalable multi-view interface for cross-embodiment learning. *arXiv preprint arXiv:2509.18757*, 2025.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. *arXiv preprint arXiv:2308.07931*, 2023.

Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang, Erjin Zhou, Haoqiang Fan, Xiangyu Zhang, and Gao Huang. Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. *ArXiv*, abs/2508.19236, 2025a. URL `https://api.semanticscholar.org/CorpusID:280869931`.

Lucy Xiaoyang Shi, Brian Ichter, Michael Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025b.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. *Conference on robot learning*, pp. 894–906, 2022.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pp. 785–799. PMLR, 2023.

Ishika Singh, Ankit Goyal, Stan Birchfield, Dieter Fox, Animesh Garg, and Valts Blukis. Og-vla: 3d-aware vision language action model via orthographic image generation. *arXiv preprint arXiv:2506.01196*, 2025.

Wenxuan Song, Jiayi Chen, Pengxiang Ding, Han Zhao, Wei Zhao, Zhide Zhong, Zongyuan Ge, Zhijun Li, Donglin Wang, Jun Ma, et al. Pd-vla: Accelerating vision-language-action model integrated with action chunking via parallel decoding. *arXiv preprint arXiv:2503.02310*, 2025.

Yuhao Sun, Ning Cheng, Shixin Zhang, Wenzhuang Li, Lingyue Yang, Shaowei Cui, Huaping Liu, Fuchun Sun, Jianwei Zhang, Di Guo, et al. Tactile data generation and applications based on visuo-tactile sensors: A review. *Information Fusion*, 121:103162, 2025.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive post-training for vision-language-action models. *arXiv preprint arXiv:2505.17016*, 2025.

Yinzhou Tang, Yu Shang, Yinuo Chen, Bingwen Wei, Xin Zhang, Shu'ang Yu, Liangzhi Shi, Chao Yu, Chen Gao, Wei Wu, et al. Roboscape-r: Unified reward-observation world models for generalizable robotics training via rl. *arXiv preprint arXiv:2512.03556*, 2025.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Viswesh Nagaswamy Rajesh, Yong Woo Choi, Yen-Ru Chen, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *Robotics: Science and Systems*, 2025.

Generalist AI Team. Gen-0: Embodied foundation models that scale with physical interaction. *Generalist AI Blog*, 2025. https://generalistai.com/blog/nov-04-2025-GEN-0.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023.

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. In *RSS 2024 Workshop: Data Generation for Robotics*, 2024a.

Hongyu Wang, Chuyan Xiong, Ruiping Wang, and Xilin Chen. Bitvla: 1-bit vision-language-action models for robotics manipulation. *arXiv preprint arXiv:2506.07530*, 2025a.

Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9773–9783, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025b.

Kuanning Wang, Yuqian Fu, Tianyu Wang, Yanwei Fu, Longfei Liang, Yu-Gang Jiang, and Xiangyang Xue. Rag-6dpose: Retrieval-augmented 6d pose estimation via leveraging cad as knowledge base. *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 15375–15382, 2025c. URL `https://api.semanticscholar.org/CorpusID:279999947`.

Kuanning Wang, Yongchong Gu, Yu Fu, Zeyu Shangguan, Sicheng He, Xiangyang Xue, Yanwei Fu, and Daniel Seita. Scoop'd: Learning mixed-liquid-solid scooping via sim2real generative policy. *ArXiv*, abs/2510.11566, 2025d. URL `https://api.semanticscholar.org/CorpusID:282058020`.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024b.

Tianyu Wang, Haitao Lin, Junqiu Yu, and Yanwei Fu. Polaris: Open-ended interactive robotic manipulation via syn2real visual grounding and large language models. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9676–9683, 2024c. URL `https://api.semanticscholar.org/CorpusID:271874650`.

Yating Wang, Haoyi Zhu, Mingyu Liu, Jiange Yang, Hao-Shu Fang, and Tong He. Vq-vla: Improving vision-language-action models via scaling vector-quantized action tokenizers. *arXiv preprint arXiv:2507.01016*, 2025e.

Yi Ru Wang, Carter Ung, Grant Tannert, Jiafei Duan, Josephine Li, Amy Le, Rishabh Oswal, Markus Grotz, Wilbert Pumacay, Yuquan Deng, Ranjay Krishna, Dieter Fox, and Siddhartha Srinivasa. Roboeval: Where robotic manipulation meets structured and scalable evaluation, 2025f. URL `https://arxiv.org/abs/2507.00435`.

Zhendong Wang, Zhaoshuo Li, Ajay Mandlekar, Zhenjia Xu, Jiaojiao Fan, Yashraj Narang, Linxi Fan, Yuke Zhu, Yogesh Balaji, Mingyuan Zhou, et al. One-step diffusion policy: Fast visuomotor policies via diffusion distillation. *arXiv preprint arXiv:2410.21257*, 2024d.

Xiangyi Wei, Haotian Zhang, Xinyi Cao, Siyu Xie, Weifeng Ge, Yang Li, and Changbo Wang. Audio-vla: Adding contact audio perception to vision-language-action model for robotic manipulation. *arXiv preprint arXiv:2511.09958*, 2025.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17868–17879, 2024a.

Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023.

Junjie Wen, Minjie Zhu, Yichen Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Chengmeng Li, Xiaoyu Liu, Yaxin Peng, Chaomin Shen, and Feifei Feng. Diffusion-vla: Scaling robot foundation models via unified diffusion and autoregression. *ArXiv*, abs/2412.03293, 2024b. URL `https://api.semanticscholar.org/CorpusID:278535637`.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024.

Wei Xiao, Tsun-Hsuan Wang, Chuang Gan, Ramin Hasani, Mathias Lechner, and Daniela Rus. Safediffuser: Safe planning with diffusion probabilistic models. In *The thirteenth international conference on learning representations*, 2023.

Feng Xu, Guangyao Zhai, Xin Kong, Tingzhong Fu, Daniel FN Gordon, Xueli An, and Benjamin Busam. Stare-vla: Progressive stage-aware reinforcement for fine-tuning vision-language-action models. *arXiv preprint arXiv:2512.05107*, 2025a.

Siyu Xu, Yunke Wang, Chenghao Xia, Dihao Zhu, Tao Huang, and Chang Xu. Vla-cache: Efficient vision-language-action manipulation via adaptive token caching, 2025b. URL `https://arxiv.org/abs/2502.02175`.

Ge Yan, Jiyue Zhu, Yuquan Deng, Shiqi Yang, Ri-Zhao Qiu, Xuxin Cheng, Marius Memmel, Ranjay Krishna, Ankit Goyal, Xiaolong Wang, et al. Maniflow: A general robot manipulation policy via consistency flow training. *arXiv preprint arXiv:2509.01819*, 2025.

Yantai Yang, Yuhao Wang, Zichen Wen, Luo Zhongwei, Chang Zou, Zhipeng Zhang, Chuan Wen, and Linfeng Zhang. Efficientvla: Training-free acceleration and compression for vision-language-action models. *arXiv preprint arXiv:2506.10100*, 2025a.

Yuyin Yang, Zetao Cai, Yang Tian, Jia Zeng, and Jiangmiao Pang. Gripper keypose and object pointflow as interfaces for bimanual robotic manipulation. *arXiv preprint arXiv:2504.17784*, 2025b.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.

Weirui Ye, Fangchen Liu, Zheng Ding, Yang Gao, Oleh Rybkin, and Pieter Abbeel. Video2policy: Scaling up manipulation tasks in simulation through internet videos. *arXiv preprint arXiv:2502.09886*, 2025.

Jessica Yin, Haozhi Qi, Youngsun Wi, Sayantan Kundu, Mike Lambeta, William Yang, Changhao Wang, Tingfan Wu, Jitendra Malik, and Tess Hellebrekers. Osmo: Open-source tactile glove for human-to-robot skill transfer. *arXiv preprint arXiv:2512.08920*, 2025.

Jiawen Yu, Hairuo Liu, Qiaojun Yu, Jieji Ren, Ce Hao, Haitong Ding, Guangyu Huang, Guofan Huang, Yan Song, Panpan Cai, et al. Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation. *arXiv preprint arXiv:2505.22159*, 2025.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pp. 1094–1100. PMLR, 2020.

Yifu Yuan, Haiqin Cui, Yaoting Huang, Yibin Chen, Fei Ni, Zibin Dong, Pengyi Li, Yan Zheng, and Jianye Hao. Embodied-r1: Reinforced embodied reasoning for general robotic manipulation. *arXiv preprint arXiv:2508.13998*, 2025.

Yang Yue, Yulin Wang, Bingyi Kang, Yizeng Han, Shenzhi Wang, Shiji Song, Jiashi Feng, and Gao Huang. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. *Advances in Neural Information Processing Systems*, 37:56619–56643, 2024.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024.

Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, et al. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, pp. 726–747. PMLR, 2021.

Shaopeng Zhai, Qi Zhang, Tianyi Zhang, Fuxian Huang, Haoran Zhang, Ming Zhou, Shengzhe Zhang, Litao Liu, Sixu Lin, and Jiangmiao Pang. A vision-language-action-critic model for robotic real-world reinforcement learning. *arXiv preprint arXiv:2509.15937*, 2025.

Borong Zhang, Yuhao Zhang, Jiaming Ji, Yingshan Lei, Josef Dai, Yuanpei Chen, and Yaodong Yang. Safevla: Towards safety alignment of vision-language-action model via safe reinforcement learning. *ArXiv*, abs/2503.03480, 2025a. URL `https://api.semanticscholar.org/CorpusID:276782535`.

Han Zhang, Songbo Hu, Zhecheng Yuan, and Huazhe Xu. Doglove: Dexterous manipulation with a low-cost open-source haptic force feedback glove. *arXiv preprint arXiv:2502.07730*, 2025b.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, et al. 4d-vla: Spatiotemporal vision-language-action pretraining with cross-scene calibration. *arXiv preprint arXiv:2506.22242*, 2025c.

Jiahui Zhang, Yusen Luo, Abrar Anwar, Sumedh Anand Sontakke, Joseph J Lim, Jesse Thomason, Erdem Biyik, and Jesse Zhang. Rewind: Language-guided rewards teach robot policies without new demonstrations. *arXiv preprint arXiv:2505.10911*, 2025d.

Jinyu Zhang, Yongchong Gu, Jianxiong Gao, Haitao Lin, Qiang Sun, Xinwei Sun, Xiangyang Xue, and Yanwei Fu. Lac-net: Linear-fusion attention-guided convolutional network for accurate robotic grasping under the occlusion. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10059–10065. Ieee, 2024a.

Jinyu Zhang, Haitao Lin, Jiashu Hou, Xiangyang Xue, and Yanwei Fu. Beyond'templates': Category-agnostic object pose, size, and shape estimation from a single view. *arXiv preprint arXiv:2510.11687*, 2025e.

Jiyao Zhang, Mingdong Wu, and Hao Dong. Genpose: Generative category-level object pose estimation via diffusion models. *arXiv preprint arXiv:2306.10531*, 2023.

Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11142–11152, 2025f.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, Fan Lu, He Wang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *arXiv preprint arXiv:2507.04447*, 2025g.

Zijian Zhang, Kaiyuan Zheng, Zhaorun Chen, Joel Jang, Yi Li, Siwei Han, Chaoqi Wang, Mingyu Ding, Dieter Fox, and Huaxiu Yao. Grape: Generalizing robot policy via preference alignment. *arXiv preprint arXiv:2411.19309*, 2024b.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Ming-Yu Liu, Donglai Xiang, Gordon Wetzstein, and Tsung-Yi Lin. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1702–1713, 2025. URL https://api.semanticscholar.org/CorpusID:277435005.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *ArXiv*, abs/2403.09631, 2024. URL https://api.semanticscholar.org/CorpusID:268385444.

Jinliang Zheng, Jianxiong Li, Dongxiu Liu, Yinan Zheng, Zhihao Wang, Zhonghong Ou, Yu Liu, Jingjing Liu, Ya-Qin Zhang, and Xianyuan Zhan. Universal actions for enhanced embodied foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22508–22519, 2025.

Zhide Zhong, Haodong Yan, Junfeng Li, Xiangchen Liu, Xin Gong, Wenxuan Song, Jiayi Chen, and Haoang Li. Flowvla: Thinking in motion with a visual chain of thought. *arXiv e-prints*, pp. arXiv–2508, 2025.

Jiaming Zhou, Teli Ma, Kun-Yu Lin, Zifan Wang, Ronghe Qiu, and Junwei Liang. Mitigating the human-robot domain discrepancy in visual pre-training for robotic manipulation. In *Proceedings of the computer vision and pattern recognition conference*, pp. 22551–22561, 2025a.

Xueyang Zhou, Yangming Xu, Guiyao Tie, Yongchao Chen, Guowen Zhang, Duanfeng Chu, Pan Zhou, and Lichao Sun. Libero-pro: Towards robust and fair evaluation of vision-language-action models beyond memorization. *[arXiv preprint arXiv:2510.03827]*, 2025b.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Kevin Lin, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. In *arXiv preprint arXiv:2009.12293*, 2020.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.