The Embodiment Gap in Robot Foundation Models

Anonymous authors
Paper under double-blind review

Abstract

Robot foundation models (RFMs), including vision-language-action (VLA) policies, are often read through a familiar scaling story: more data, larger models, and broader benchmarks. Robotics adds a practical follow-up: when a shared model reaches a new body, what work lets it act there? This survey asks what travels across robot bodies and what has to be realized on the target robot. We call the mismatch between reusable structure and target-specific execution the *embodiment gap*. The gap identifies which structures become reusable, where body-specific work remains, and what evidence should accompany cross-embodiment success claims. We organize this lens around three scaling directions–semantic meaning and perception, physical robot data and interfaces, and embodiment correspondence–and use it to define a reporting agenda for target-body residuals. The goal is to make cross-embodiment progress easier to compare, reproduce, and build on, while encouraging systems that leave new robots with less target-specific work, clearer failure attribution, and safer recovery.

1 Introduction

Robot learning has begun to borrow the shape of recent progress in language and vision: train large models on heterogeneous data, and expect useful generalization to emerge. Robot foundation models, especially VLA policies and generalist robot policies, are the most visible expression of this borrowing. Language, perception, trajectories, teleoperation, and pretraining now act as reusable resources for embodied control; RT-X, Octo, OpenVLA, RoboCat, and the π_0 family are representative examples (Octo Model Team et al., 2024; Kim et al., 2025c; Open X-Embodiment Collaboration et al., 2023; Khazatsky et al., 2024; Zhao et al., 2023; Brohan et al., 2023b;a; Driess et al., 2023; Bousmalis et al., 2024; Black et al., 2024; Physical Intelligence et al., 2025). Robots, however, must act through bodies: a plan, affordance, object-motion cue, latent action, or policy backbone may travel across robots, but it becomes behavior only when it meets a particular morphology, sensor suite, controller, contact regime, and safety envelope. We call this distance between transferable structure and physical realization the *embodiment gap*.

The reporting problem is simple but consequential. Two systems can report the same final success rate while requiring very different work on the target robot: one may need a small number of demonstrations and a calibrated controller, while another may need extensive fine-tuning, repeated real-robot trials, manual resets, or safety interventions. The question is therefore not only whether a robot eventually succeeds, but what had to happen on that robot before success became possible. We call that remaining work *residual adaptation burden*.

The recent debate over scaling robot data and models motivates the same diagnostic question. Goldberg argues that robotics lacks the naturally available observation-action pairs that enabled internet-scale learning in language and vision (Goldberg, 2025); a related *Science Robotics* debate frames the role of data, models, and engineering as open rather than settled (Amato et al., 2025). For this survey, the central question is: *what transfers across bodies, and what remains on the target robot?*

We answer with a lens and a reporting view. The lens separates *shareability level*, or what kind of structure transfers, from *dominant residual locus*, or where target-side work mainly remains. It also organizes three

The authors used GPT-5.5 for language polishing and for drafting Figure 1. The authors verified all outputs and take full responsibility for the ideas, claims, citations, figures, and final manuscript.

overlapping scaling directions: semantic scale, physical-data scale, and embodiment correspondence scale. The reporting view asks papers to expose the target-body pathway behind success: target data, updates, setup, robot operation, safety events, recovery, and failure attribution.

Recent surveys have mapped foundation models in robotics (Firoozi et al., 2025; Hu et al., 2023), VLA and action-tokenization work (Zhong et al., 2025; Kawaharazuka et al., 2025; Motoda et al., 2025), video-to-control learning (McCarthy et al., 2025; Zheng et al., 2026b), data and benchmark resources (Wang et al., 2026), evaluation taxonomies (Gao et al., 2025), and embodied-AI mapping (Raychaudhuri and Chang, 2025). These surveys map the landscape; this survey asks a different question: not which capability, module, or benchmark is present, but what shared structure transfers across bodies and what target-body burden remains. Methodologically, this paper is a lens-driven scoping survey rather than an exhaustive systematic review. We focus on manipulation-oriented RFMs, especially VLA policies and generalist robot policies, and on mechanisms that connect shared structures to concrete target robots: model backbones, data, interfaces, correspondence mechanisms, morphology-aware models, contact-rich grounding methods, and evaluation practices. Appendix A details the search, inclusion protocol, boundary cases, and status policy for recent preprints and OpenReview submissions.

This survey makes three contributions. First, it defines the embodiment gap as the distance between reusable structure and target-body realization. Second, it organizes recent robot-foundation-model work by shareability level, residual locus, and three scaling directions without treating those directions as disjoint taxonomies. Third, it turns the lens into a reporting agenda for target-body residuals, including burden profiles, adaptation curves, and failure attribution.

The rest of the paper is organized as follows. Section 2 defines the embodiment gap and residual adaptation burden. Section 3 introduces the two-axis lens and the three scaling directions. Sections 4-6 examine semantic scale, physical-data scale, and embodiment correspondence scale. Section 7 develops target-body residual reporting. Section 8 discusses embodiment-gap-aware scaling, and Section 9 concludes. Appendix A gives scope notes, Appendix B summarizes adjacent concepts, Appendix C reports placement rationales, and Appendix D gives extended reporting checklists.

2 The Embodiment Gap in Robot Foundation Models

Robot foundation models, including VLA policies and generalist robot policies, are built on the promise that structure can be reused across tasks, environments, datasets, and robot bodies. Task knowledge, perceptual cues, object-motion structure, action interfaces, and policy backbones can all make a new robot less alone. The last step, however, still happens through a particular machine: reusable structure has to become executable under the physical and operational conditions of the target robot.

We use the term *embodiment gap* for the distance between *shared structure* and *target-body realization*. The concept does not name every distribution shift in robot learning. It refers to the gap that appears when structure transferred across bodies must be re-grounded in a target robot's morphology, sensing, control, contact, and safety conditions. We use embodiment in an extended robot-learning sense: the body-sensor-controller-contact stack through which a policy becomes executable.

2.1 Shared structure is not directly executable

We use *shared structure* as an umbrella term for the part of a robot-foundation-model system intended to transfer across robot bodies. It may be task-level, such as a goal or plan; perceptual, such as an object or affordance cue; interaction-centric, such as object flow or point tracks; or closer to execution, such as a latent action, skill, action token, or morphology-aware policy component. The distance to execution changes across these representations: a task plan travels easily but says little about contact, while an end-effector trajectory is closer to execution but depends on kinematics and control.

The other side of the gap is *target-body realization*: the point at which shared structure becomes executable under the robot's concrete conditions. A language model may choose the right subtask, a vision model may identify the right object, and a policy backbone may output a plausible action representation; the cup is

still picked up by a particular gripper, with a particular controller, under particular contact conditions. The instruction "pick up the cup" may remain semantically stable across suction, parallel-jaw, dexterous-hand, and mobile-manipulator setups, while grasps, collision risks, stability margins, and recovery options change with the body. Figure 1 gives the schematic version of this distinction.

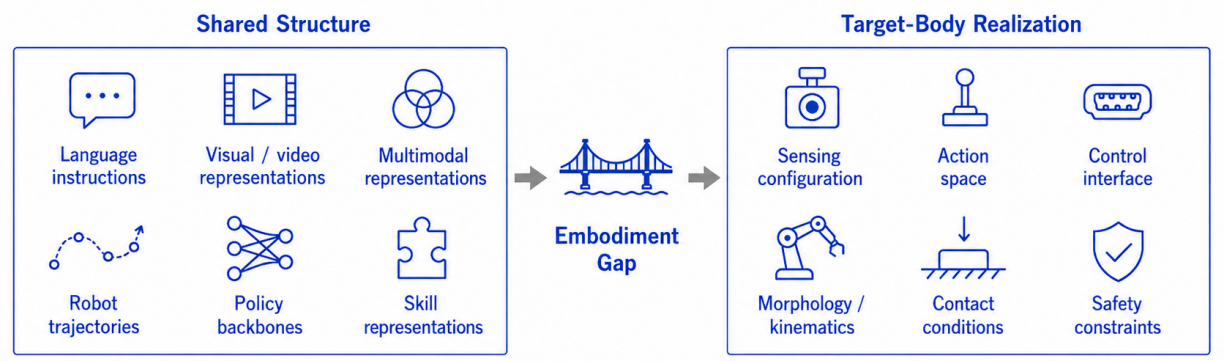

Figure 1: The embodiment gap. Shared structure can travel across bodies, but execution requires target-body realization through the selected robot's sensing, action, control, morphology, contact, and safety conditions.

2.2 Relation to adjacent gaps

The embodiment gap overlaps with familiar notions in robot learning, but its focus is narrower. *Domain shift* concerns changes in visual statistics, environments, object appearances, tasks, or observation conditions (Csurka, 2017); these shifts may occur without changing the robot body. A *sim-to-real gap* concerns differences between simulated and real physical processes, including contact, dynamics, noise, latency, and unmodeled effects (Zhao et al., 2020; Tobin et al., 2017). Sim-to-real and embodiment shifts can meet in human-to-robot transfer (Lum et al., 2025), but sim-to-real transfer can also occur for a fixed body.

General *transfer learning* asks how knowledge acquired in one domain, task, or setting can be reused in another (Taylor and Stone, 2009; Pan and Yang, 2010). Cross-embodiment robot learning is one form of transfer; the embodiment-gap lens focuses on how shared structure maps to the target robot's action possibilities, physical constraints, and operational risks. Likewise, *action-space mismatch* is an important manifestation of the gap, but not the whole gap: even a nominally shared action interface can imply different reachability, contact stability, sensing uncertainty, and recovery requirements.

A shared Cartesian end-effector command illustrates the boundary: the nominal action interface may match while a suction cup, parallel-jaw gripper, dexterous hand, or mobile manipulator realizes the same displacement through different reachability, compliance, tactile sensing, contact stability, and recovery options. Body differences define action possibilities, so the design question is which differences should be shared, standardized, transformed, or left for target-body adaptation.

2.3 Residual adaptation burden

Even after shared structure has been transferred, some work often remains on the target robot. We call this *residual adaptation burden*: the target-side data, model changes, setup, real-robot operation, and recovery needed before the transferred structure can be used successfully. Many robot systems deliberately leave work to a controller, action decoder, adapter, calibration step, or safety layer. This can be a good design choice when the residual is small, stable, and visible; it becomes difficult to compare when the residual is hidden. In this survey, we use five burden-profile items: target-data burden, model-update burden, calibration/setup burden, real-robot operation burden, and safety/intervention/recovery burden. These items form a reporting profile, not a universal scalar metric: success rate tells us whether the target robot eventually completed the task, while residual adaptation burden tells us what had to be paid on that robot before success became

possible. The embodiment gap is visible in the relation between the two: what was shared, how it was realized, and what remained.

3 Shareability Levels, Residual Loci, and Scaling Directions

We use a two-axis lens to compare how shared structure returns to the body. The first axis, *shareability level*, asks what kind of structure is intended to transfer across bodies. The second, *dominant residual locus*, asks where target-side work most visibly remains. The full burden is multi-dimensional and is reported later through the report card and case-study profiles; the two-axis lens shows where a system places the boundary between shared structure and target-body realization.

3.1 A two-axis lens: what transfers and where residuals remain

The *shareability level* describes how far the shared structure is from target-body execution. Highly abstract structures are easier to share but leave more grounding to the target robot; execution-proximal structures can reduce some residuals but are more sensitive to morphology, sensing, control, and contact. We use five working levels, drawing on foundation-model surveys, VLA/action-tokenization work, video-to-control studies, and evaluation taxonomies: semantic/task, perceptual/affordance, object-interaction, action/skill, and morphology-conditioned or execution-proximal sharing (Firoozi et al., 2025; Zhong et al., 2025; McCarthy et al., 2025; Gao et al., 2025).

At the semantic/task level, systems share goals, instructions, plans, or commonsense knowledge. At the perceptual/affordance level, they share visual or spatial cues such as object representations, affordance maps, or video-derived subgoals. At the object-interaction level, the shared structure describes how the manipulated object or scene should change, for example through object flow, point tracks, tool-centered motion, or object pose trajectories. At the action or skill level, systems share latent actions, skill embeddings, action tokens, action decoders, or manipulation primitives. At the morphology-conditioned level, systems condition policies or action generation on body structure itself, as in body graphs, kinematic tokens, morphology embeddings, or body-specific prompts (Sferrazza et al., 2024; Patel and Song, 2025; Zheng et al., 2026a; Suzuki et al., 2026; Zhang et al., 2026).

Boundary cases are assigned by the representation's primary commitment. Affordance maps remain near perceptual/affordance when they localize possible interaction; object flow and point tracks move toward object-interaction when they specify scene change (Xu et al., 2025; Bharadhwaj et al., 2024); action motifs sit near the object-interaction–action boundary when they encode action-oriented abstractions (Zhi et al., 2026). A single system may combine several levels, so the level in Figure 2 indicates the dominant shared structure for the cross-embodiment claim.

The second axis is a *dominant residual locus*, not a residual amount. Target-side burden includes target data, model updates, calibration, controller alignment, contact grounding, real-robot operation, safety, and recovery; these cannot be collapsed into one reliable score. The vertical axis instead records where the remaining work mainly appears: skill/API grounding, calibration and action-interface alignment, contact-, force-, and tactile execution, or recovery, safety, and closed-loop robustness. Section 7 and Tables 1–2 provide the multi-dimensional reporting view.

3.2 From the emphasis map to three scaling directions

Figure 2 summarizes representative systems by dominant shareability level and residual locus. The map is a qualitative design-space view, not a taxonomy score or a burden ranking: moving rightward brings the shared structure closer to execution, while moving downward marks a different locus of target-body realization. Placements follow two coding rules–primary shared representation and most visible target-body residual–with row-level rationale in Appendix C; Section 7 reports burden profiles separately.

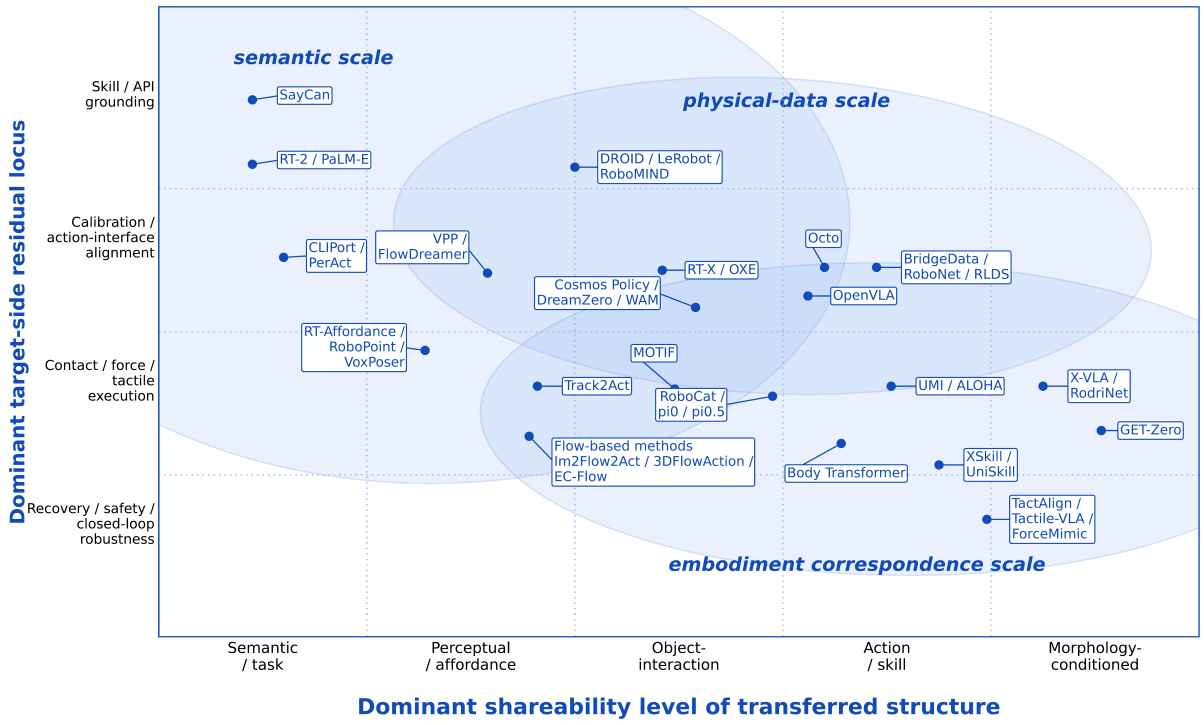

Figure 2: A two-axis emphasis map under the embodiment-gap lens. The horizontal axis indicates the dominant shareability level of the structure transferred across bodies; the vertical axis indicates the dominant locus at which target-side residual work remains. Coordinates indicate qualitative emphases rather than scalar residual amounts. The translucent overlays denote overlapping scaling directions; Appendix C gives placement rationales.

The map is most useful when read through contrasts. SayCan, RT-2/PaLM-E, CLIPort/PerAct, and RT-Affordance/RoboPoint/VoxPoser, for example, share goals, visual cues, or affordances, but their target-body residuals remain in skill availability, API grounding, calibration, and feasible execution. RT-X/OXE, Octo, OpenVLA, BridgeData/RoboNet/RLDS, and DROID/LeRobot/RoboMIND occupy the middle of the map: shared data and action schemas make trajectories and interfaces trainable across bodies, but do not remove controller binding, target data, setup, or evaluation burden. Track2Act, flow-based methods, MOTIF, and latent-action methods move the shared structure closer to object motion or action abstraction. The lower-right methods—Body Transformer, GET-Zero, X-VLA/RodriNet, TactAlign, Tactile-VLA, and ForceMimic—make morphology, contact, or force part of the representation itself. The lesson is that sharing more structure changes where the residual sits, and these contrasts motivate the three scaling directions used in Sections 4–6. By cross-embodiment scaling, we mean making more structure travel across bodies while the remaining target-body work becomes smaller, safer, better localized, or easier to report. *Level* names the granularity of shared structure; *scale* names a research direction that expands what can be shared and connected to target bodies. Figure 3 summarizes the three directions as overlapping design-and-reporting axes: semantic scale routes reusable meaning and perception through foundation models; physical-data scale makes robot trajectories and interfaces reusable and traceable; embodiment correspondence scale learns cross-body relations among robot-task realizations.

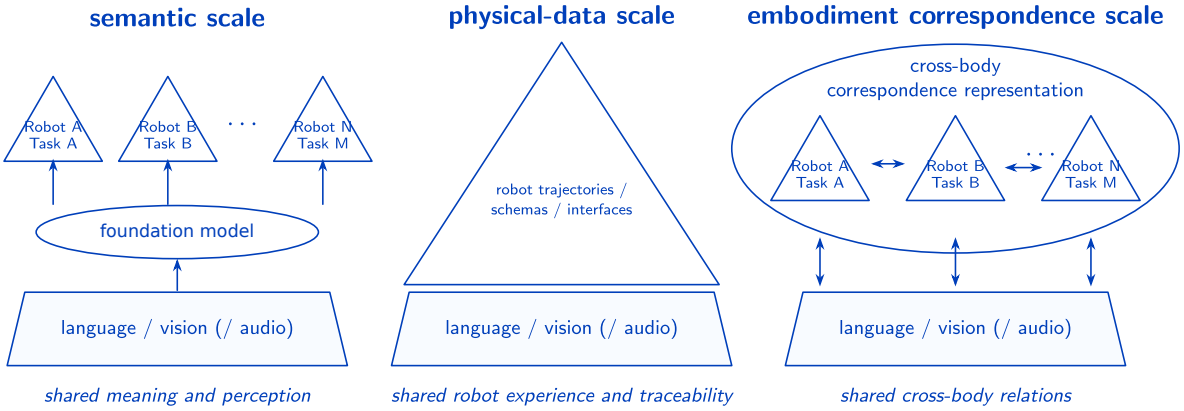

Figure 3: Three scaling directions under the embodiment-gap lens. Semantic scale expands reusable meaning and perception; physical-data scale makes robot trajectories and interfaces reusable and traceable; embodiment correspondence scale learns relations across bodies through object motion, latent actions, body conditioning, and contact cues. The directions are overlapping design-and-reporting axes.

4 Semantic Scale: High-Level Shared Structure and Execution Residuals

Semantic-scale methods reuse task meaning, language, perception, and video-derived cues, but leave feasibility, contact, and recovery to the target robot. At the semantic/task level, what is shared is not a low-level action but a high-level description of what should happen. SayCan is a representative example (Ahn et al., 2023). Other LLM planning interfaces, including Code as Policies, ProgPrompt, and LLM+P, expose the same boundary between semantic structure and executable robot behavior (Liang et al., 2023; Singh et al., 2023; Liu et al., 2023a). A language model may propose actions, programs, or plans, but target feasibility still depends on available skills, planners, execution, and safety.

Perceptual and affordance methods move semantic scale closer to execution. CLIPort combines semantic what information from language-image pretraining with spatial where information for language-conditioned manipulation (Shridhar et al., 2022); PerAct predicts actions over a 3D voxel representation (Shridhar et al., 2023). R3M, RT-Affordance, RoboPoint, and VoxPoser provide reusable visual or spatial cues that constrain robot behavior (Nair et al., 2023; Nasiriany et al., 2024a; Yuan et al., 2025; Huang et al., 2023). Perception Stitching makes a related separation between transferable visual encoders and downstream visuomotor execution (Jian et al., 2024), while ActionEQA shows failures shifting from high-level perception toward geometric and physical reasoning as semantic instructions become lower-level actions (Bao et al., 2026).

Video and world-model cues add temporal structure. Visual foresight, video prediction, affordance/value-map methods, and human or egocentric video imitation can narrow what the target robot should attempt (Nasiriany et al., 2024a; Ebert et al., 2018; Hu et al., 2024; Wu et al., 2023; Yuan et al., 2025; Guo et al., 2025; Huang et al., 2023; Wang et al., 2023; Kareer et al., 2024; Hoque et al., 2025). They still do not settle the target-body questions of reachability, contact, and recovery.

Recent robot foundation models, including VLA policies and generalist robot policies, combine semantic scale with physical-data scale. RT-2, PaLM-E, RT-X, Octo, OpenVLA, RoboCat, and the π_0 family train large policy backbones on robot data, often using pretrained vision-language models and multi-robot datasets (Octo Model Team et al., 2024; Kim et al., 2025c; Open X-Embodiment Collaboration et al., 2023; Khazatsky et al., 2024; Zhao et al., 2023; Brohan et al., 2023b;a; Driess et al., 2023; Bousmalis et al., 2024; Black et al., 2024; Physical Intelligence et al., 2025). From the embodiment-gap perspective, the key question is where the shared backbone ends and target-body realization begins; tuning and VLA fine-tuning studies make that residual especially visible (Zhang et al., 2025b; Kim et al., 2025b).

5 Physical-Data Scale: Shared Robot Experience and Traceable Residuals

Physical-data scale shifts the focus from reusable meaning to the robot experience through which meaning is grounded: how that experience is collected, formatted, normalized, and made traceable across embodiments. A robot trajectory is produced by a particular body, sensor setup, teleoperation interface, controller, workspace, and task protocol, so the central tension is simple: common format is not common physical meaning.

Large robot datasets and data infrastructures have changed how generalist robot policies are trained. RoboNet, BridgeData V2, Open X-Embodiment / RT-X, DROID, RLDS, LeRobot, RoboMIND, OXE-AugE, and RoboWheel treat robot experience as a shared learning resource (Open X-Embodiment Collaboration et al., 2023; Khazatsky et al., 2024; Ramos et al., 2021; Cadène et al., 2026; Walke et al., 2023; Dasari et al., 2020; Wu et al., 2025; Ji et al., 2025; Zhang et al., 2025c). From the embodiment-gap perspective, their key contribution is traceability: useful cross-embodiment data record not only actions, but also the robot, sensors, action space, controller assumptions, calibration, and success/failure criteria behind those actions. Collection interfaces shape what human operation becomes as robot-learning data: ALOHA-style systems provide low-cost bimanual teleoperation for fine-grained manipulation (Zhao et al., 2023; Fu et al., 2025; ALOHA 2 Team et al., 2024), DROID pushes toward large-scale in-the-wild manipulation data with standardized procedures and metadata (Khazatsky et al., 2024), and UMI, OPEN TEACH, and AnyTeleop show how interface design affects the resulting embodiment residuals (Chi et al., 2024; Iyer et al., 2025; Qin et al., 2023).

Schemas and normalizations make heterogeneous data trainable, but not automatically executable. RLDS represents datasets as episodes and steps (Ramos et al., 2021); Open X-Embodiment integrates datasets under common conventions (Open X-Embodiment Collaboration et al., 2023); LeRobot connects dataset formats with training and deployment tools (Cadène et al., 2026). Object-centric coordinates, end-effector poses, SE(3) transformations, task frames, and equivariant representations can make behavior more comparable across robots (Wang et al., 2019; Pan et al., 2023; Liu et al., 2023c; Zhang et al., 2025d; Yang et al., 2023a), but a geometrically valid displacement may still be physically difficult for a particular target robot.

Interfaces and benchmarks allocate residuals. Frameworks such as PyRobot, Robot Control Stack, robosuite, RoboHive, and LeRobot define reusable control, simulation, benchmarking, or deployment boundaries (Cadène et al., 2026; Murali et al., 2019; Jülg et al., 2026; Zhu et al., 2020; Kumar et al., 2023). Benchmarks such as LIBERO, RLBench, CALVIN, RoboCasa, FurnitureBench, AnyBody, AutoEval, and trustworthy manipulation evaluation make task, embodiment, and evaluation assumptions more explicit (Liu et al., 2023b; James et al., 2020; Mees et al., 2022; Nasiriany et al., 2024b; Heo et al., 2023; Parakh et al., 2025; Zhou et al., 2025; Liu et al., 2026). They become most informative when they state the bodies involved, adaptation rules, setup assumptions, evaluation rollouts, and any reset or safety burden.

6 Embodiment Correspondence Scale: Reducing Residuals Through Cross-Body Correspondence

Embodiment correspondence scale revisits some of the systems above from a third angle: what relation can be carried from one body to another, and where that relation still has to be grounded. Task stages, affordances, object motion, latent actions, body graphs, paired demonstrations, force, and touch can each give a model a relation that survives some body changes while leaving other work to the target robot.

High-level correspondence matches task stages, intentions, affordances, or visual alignment rather than low-level actions. XSkill and UniSkill learn skill representations that are meaningful across embodiments (Xu et al., 2023; Kim et al., 2025a). Human-video affordance and intention-alignment methods share where interaction should happen rather than how the demonstrator moved (Nasiriany et al., 2024a; Bahl et al., 2023; Srirama et al., 2024; Chen et al., 2026). Mirage, SHADOW, and RoVi-Aug reduce visual embodiment mismatch through augmentation and masking (Chen et al., 2024; Lepert et al., 2025; Chen et al., 2025b). These methods narrow the realization problem, but grasping, force, control, and recovery remain target-body dependent.

Intermediate correspondence uses scene change as a bridge from video or source-domain demonstrations to robot manipulation. Object-flow, object-centric motion, pose-trajectory, and tool-centered methods share

how the object or scene should move (Xu et al., 2025; Zhi et al., 2025; Chen et al., 2025c; Yin et al., 2025; Hsu et al., 2025; Chen et al., 2025a). Track2Act predicts point tracks, turns them into object and end-effector targets, and then learns a target-body residual policy (Bharadhwaj et al., 2024). In this reading, the shared structure is object motion; the residual is closed-loop contact, grasp recovery, and real-robot operation.

Video/world-model methods can also act as correspondence channels when scene-evolution prediction is coupled to action generation. VPP and FlowDreamer use predictive visual or flow-based world-model representations (Hu et al., 2024; Guo et al., 2025); Cosmos Policy adapts a pretrained video-to-world model into a robot policy (Kim et al., 2026); DreamZero, introduced by Ye et al. as a World Action Model (WAM), reports a video-diffusion approach that models future world states together with actions and studies cross-embodiment transfer from video-only demonstrations with few-shot adaptation (Ye et al., 2026). These methods move video/world prediction toward executable policies, while residuals remain in action-interface alignment, contact, force, safety, and recovery.

Latent-action and action-motif methods address correspondence from the action side. Rather than directly sharing low-level robot commands, they learn intermediate action spaces that capture task progress or action intent (Bauer et al., 2025; Ye et al., 2025; Bu et al., 2025; Zha et al., 2026; Huang et al., 2026a). MOTIF learns action motifs from heterogeneous robot data for few-shot cross-embodiment transfer (Zhi et al., 2026). These methods move closer to execution without assuming that all bodies share one action space.

A more execution-proximal route makes body structure an explicit model input. Body Transformer, GET-Zero, X-VLA, morphology-aware Transformers, and RodriNet use sensor-actuator graphs, body prompts, topology-aware attention, joint attributes, or kinematic priors to condition policy computation on the target body (Sferrazza et al., 2024; Patel and Song, 2025; Zheng et al., 2026a; Suzuki et al., 2026; Zhang et al., 2026). Related hardware- and morphology-conditioned policy work shows the same design logic across control settings (Chen et al., 2018; Wang et al., 2018; Gupta et al., 2022; Hu et al., 2022; Xiong et al., 2023; Wei et al., 2024; Xiong et al., 2024; Przystupa et al., 2025; Wang et al., 2024b; Wu et al., 2026). Kinematic priors can reduce part of the target-body residual, while geometry, touch, force, contact-rich execution, and closed-loop recovery remain the next realization questions.

Contact-rich manipulation also depends on force and tactile signals, which change the kind of structure that can be shared. TactAlign and UniTacHand align tactile observations across embodiments, while Feel the Force and ForceMimic transfer force- or contact-relevant evidence from human or robot-free demonstrations (Wi et al., 2026; Zhang et al., 2025a; Adeniji et al., 2025; Liu et al., 2024). Tactile-VLA and TaF-VLA suggest that touch and force can become policy inputs or correspondence signals (Huang et al., 2025; 2026b), and tactile-perception work shows that touch can compensate for cross-embodiment capability differences (van den Bogert et al., 2024). The shared structure can include contact evidence alongside visual goals, trajectories, and action abstractions, while the remaining residual shifts toward sensor calibration, contact stability, safety, and recovery.

Correspondence data provide another route. UMI uses a handheld gripper and policy interface to collect in-the-wild human demonstrations that can transfer to robot policies (Chi et al., 2024). DexCap connects human hand motion to dexterous robot-hand learning through a portable motion-capture system and DexIL, its imitation-learning pipeline for training robot-hand skills from that mocap data (Wang et al., 2024a). Human2Robot, Data Analogies, Polybot, and Scaling Cross-Embodied Learning use paired or multi-embodiment data to help models learn what is preserved and what changes across bodies (Xie et al., 2025; Yang et al., 2026; 2023b; Doshi et al., 2024).

7 Reporting Target-Body Residuals: From Success Rates to Adaptation Burden

After organizing mechanisms, the next step is to report the target-body evidence behind cross-embodiment claims. This section proceeds through four layers–report card, burden profile, EACs, and failure attribution–that turn success rate into a traceable target-body pathway. Table 1 gives the minimum report card and separates adaptation operations from evaluation rollouts: the former are target-body burden, while the latter are performance evidence.

Table 1: Minimum report card for cross-embodiment reporting.

Item	**Report**	**Purpose**
Source embodiment	Robots; sensors; controllers	Origin of transfer
Target embodiment	Body; sensors; workspace	Realization context
Shared structure	Plans; representations; skills	What transfers
Updated components	Heads; adapters; decoders; controllers	What changes
Target-data burden	Demos; adaptation rollouts; recovery data	Target experience
Model-update burden	Frozen / adapter / LoRA / full FT	Update cost
Calibration/setup burden	Frames; cameras; gripper; control rate	Deployment prep
Real-robot operation burden	Adaptation trials; resets; robot-hours; re-executions	Pre-eval physical cost
Safety/intervention/recovery burden	Stops; unsafe contacts; interventions; recovery	Physical risk and recovery cost
Evaluation rollouts	Held-out episodes; tasks; seeds; success criteria	Performance evidence
Failure attribution	Semantic / data / correspondence / execution	Where residual arises

The report card structures the five burden-profile items from Section 2.3 together with context, evaluation evidence, and failure attribution. It complements success rate by making the pathway behind that success inspectable. Appendix D gives a worked OpenVLA example showing what can be reconstructed from public reporting and which target-body burdens remain N/R, while Table 2 applies the same profile as an illustrative reporting audit for representative cross-embodiment claims; its final column records which parts of the target-body pathway are exposed. In these profiles, N/R is a reporting-visibility marker.

Table 2: Compact case study: applying the burden profile to representative cross-embodiment claims (Open X-Embodiment Collaboration et al., 2023; Octo Model Team et al., 2024; Kim et al., 2025c;b; Zhang et al., 2025b; Bousmalis et al., 2024; Black et al., 2024; Physical Intelligence et al., 2025; Bharadhwaj et al., 2024; Zhi et al., 2026; Sferrazza et al., 2024; Patel and Song, 2025; Zheng et al., 2026a; Zhang et al., 2026; Wi et al., 2026; Huang et al., 2025). Contact-related observations are reported under safety/intervention/recovery when they affect physical risk, intervention, or recovery. N/R denotes information that we could not identify in the main paper and associated public supplementary material at the time of writing; it is used as a reporting-visibility marker.

System	**Shared structure and target-body connection**	**Target data / model update**	**Setup / operation**	**Safety / intervention / recovery**	**Target-body pathway exposed**
RT-X / OXE	Multi-robot data → robot-specific actions	Large mixed data; shared policy	Data/control conventions	Reset/safety often N/R	Robot count and adaptation cost differ
Octo	Generalist policy → target interface	Target data/adaptation	Controller alignment	Intervention/ recovery N/R	Target binding remains
OpenVLA	7B VLA + action tokens → FT/LoRA	10-150 demos; full FT or LoRA	Controller freq. partly reported	Recovery qualitative; counts N/R	Decoder/control-rate residuals
RoboCat / π_0	Generalist backbone → robot/task adaptation	RoboCat: 100-1000 examples; π_0: FT data	Setup N/R	Safety/reset N/R	Zero-shot/FT regimes mixed
Track2Act	Point tracks/object motion → residual policy	400 Spot teleop; residual BC	Depth/ transform fitting	Failure modes reported; counts N/R	Residual policy still needed
MOTIF	Action motifs → few-shot transfer	1-50 shots; motif-conditioned	Setup mostly N/R	Operation/safety N/R	Operation cost to report
Body Transformer / GET-Zero	Body graph → body-conditioned policy	Low/zero target data; frozen/generated	Morphology specification	Contact/recovery outside model	Kinematics and contact separate
X-VLA / RodriNet	Body prompts/kinematic priors → action computation	Target adaptation varies	Body metadata; calibration N/R	Touch/force/recovery remain	Body-aware residuals remain
TactAlign / Tactile-VLA	Touch/force grounding → tactile/VLA channel	Sensor/task data; tactile modules	Sensor placement/calibration	Contact explicit; safety counts N/R	Contact evidence adds sensor burden

Across the nine rows in Table 2, seven include an N/R marker in the safety/intervention/recovery column, and three explicitly mark setup or calibration metadata as N/R or mostly N/R. More broadly, operation cost is usually reported indirectly rather than as robot-hours, resets, or intervention counts. The N/R entries are useful markers of the next reporting frontier: they show which parts of the target-body pathway are already easy to reconstruct and which parts future papers can make routine, including calibration time, adaptation trials, intervention counts, reset burden, recovery evidence, and safety events.

7.1 Burden profiles and Embodiment Adaptation Curves

A burden profile records what the target robot still had to do before evaluation, preserving differences among adaptation forms: a method with little target data may need careful calibration, and a method with high success may still require frequent intervention. Appendix D, Table 7 gives the extended format. Embodiment Adaptation Curves (EACs) then plot task capability against a fixed target-body burden, such as demonstrations, adaptation rollouts, robot-hours, interventions, resets, or safety events. When a study reports performance across several budgets, the result can be reconstructed as a dense EAC; MOTIF is a suitable example because it reports few-shot cross-embodiment transfer across several demonstration budgets (Zhi et al., 2026). When a study reports only terminal conditions, the appropriate representation is an endpoint or arrow; Track2Act illustrates this endpoint evidence through open-loop and residual-policy results (Bharadhwaj et al., 2024).

Figure 4 illustrates the difference. Panel (a) is a conceptual monotonicity hypothesis, panel (b) reconstructs MOTIF's few-shot results with and without motif guidance, and panel (c) shows Track2Act endpoint evidence under Mild (MG), Standard (G), Combinatorial (CG), and Type (TG) Generalization categories.

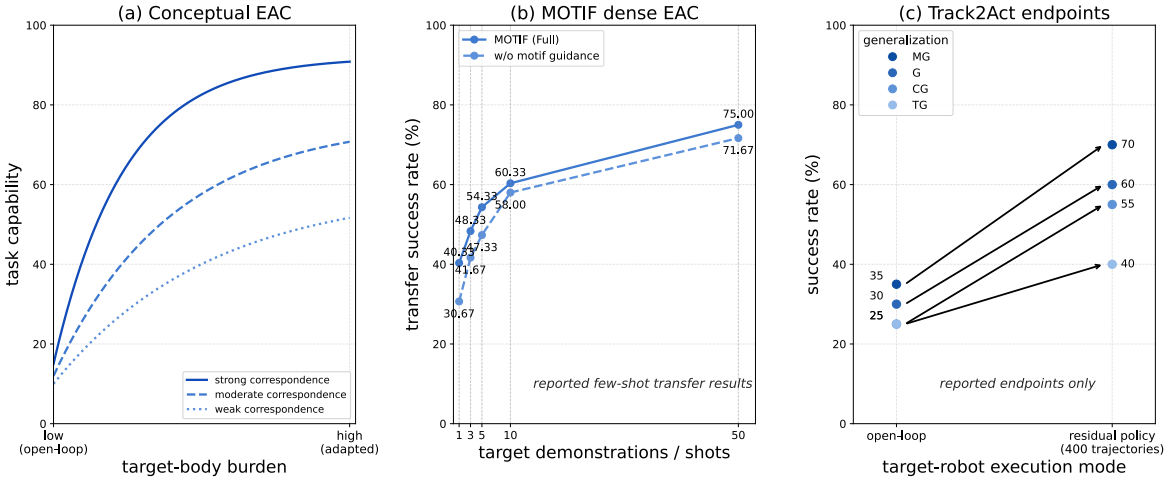

Figure 4: Embodiment Adaptation Curves. EACs plot task capability or safety-related performance against a reported target-body burden, such as demonstrations, robot trials, or interventions. Panel (a) is schematic; panels (b) and (c) visualize reported values or endpoints from the cited works. Dense EACs, illustrated by MOTIF (Zhi et al., 2026), show performance across several budgets. Endpoint evidence, illustrated by Track2Act (Bharadhwaj et al., 2024), is shown as before/after arrows; in panel (c), each arrow compares reported endpoint conditions within the same Track2Act generalization category. MG/G/CG/TG denote Track2Act's Mild (MG), Standard (G), Combinatorial (CG), and Type (TG) Generalization categories, and the horizontal axis denotes the reported target-robot execution mode.

7.2 Failure attribution

Failure attribution closes the reporting loop by linking observed breakdowns back to the target-body pathway. We use four attribution layers: semantic-scale failures, physical-data-scale failures, embodiment-correspondence failures, and target-embodiment execution failures. A useful attribution record names the observed failure,

the system output immediately before failure, and the first interface at which that output no longer supported feasible target-body execution; unobserved causes are marked N/R.

For example, if a model predicts the correct object flow but the target gripper slips during execution, the primary attribution may be target-embodiment execution and the secondary attribution may be embodiment correspondence. If an affordance point lies outside the target robot's reachable workspace, the primary attribution may be embodiment correspondence and the secondary attribution may be target execution. If a VLA policy fails because the action representation does not match the controller frequency or action decoder, the primary attribution may be physical-data scale.

Morphology-aware architectures illustrate the distinction. A system such as RodriNet can reduce residuals associated with articulated kinematics, but failures may still arise from geometry, touch, force, contact-rich execution, or recovery. Tactile- or force-aligned methods try to move contact residuals from hidden execution failures into explicit correspondence signals, making them more available for adaptation-aware reporting.

8 Toward Embodiment-Gap-Aware Scaling

Once reporting tools make the target-body pathway visible, embodiment-gap-aware scaling asks whether new bodies require less residual work, safer adaptation, and clearer evidence. It turns the lens into a three-part agenda: robot data should make embodied experience traceable; models and interfaces should make correspondence executable; and evaluation should make adaptation, safety, and recovery visible. Existing datasets, interfaces, correspondence mechanisms, benchmarks, and evaluation protocols already expose parts of this pathway (Open X-Embodiment Collaboration et al., 2023; Khazatsky et al., 2024; Cadène et al., 2026; Bharadhwaj et al., 2024; Zhi et al., 2026; Sferrazza et al., 2024; Liu et al., 2023b; Parakh et al., 2025; Zhou et al., 2025). The next step is to ask, whenever a new body is added, whether the remaining work becomes smaller, safer, better localized, and easier to explain.

Data and traceability. The next data opportunity is traceability as much as scale. A robot trajectory should record the embodiment context that produced its observations and actions: body, sensors, action interface, controller, setup, failures, and recovery. With that information, multi-robot data can support not only training, but also interpretation of why target-body transfer succeeds or fails. For manipulation, correspondence-rich data should record robot motion together with object-state change, contact, task stage, failure cause, and recovery; dataset design should be guided by the residuals a model is expected to reduce, and data standardization should make embodiment differences visible rather than hide them.

Executable correspondence. Models and interfaces should make cross-body relations executable rather than implicit. Object-centered motion, point tracks, latent actions, skill motifs, and language-action units can reduce the gap between high-level task structure and robot-specific commands. Body-aware architectures can condition policies on the structure of the target robot, and tactile or force signals can become correspondence signals rather than merely low-level feedback. Embodiment-gap-aware model design should state what is shared, what is body-conditioned, what is target-specific, and what is safety-critical; residual adaptation can be a design feature when it is small, stable, well isolated, and easy to report, but making that placement explicit helps others reproduce the pathway by which the robot learned to act.

Evaluation, deployment, and safety. Evaluation should specify the adaptation contract: the bodies involved, the components that may change, the target data and setup allowed before evaluation, and the safety or recovery evidence to be reported. These details define the conditions under which the generalization claim holds. Deployment makes safety central because target-body adaptation happens in the physical world: a policy that reaches high success with fewer unsafe contacts, manual resets, or human interventions teaches a stronger lesson about reusable control, and safety/intervention/recovery burden is therefore one of the residuals that determines whether a cross-embodiment system is practically useful.

Open problems. Once target-body residuals are visible, three practical research problems become easier to organize. First, *standardizing residuals without flattening them.* Target demonstrations and fine-tuning steps are relatively easy to count, while calibration difficulty, reset burden, safety risk, and recovery complexity are richer to describe. RLDS, Open X-Embodiment, DROID, and LeRobot already standardize parts of the

data and training pipeline; the next step is to extend reporting standards toward target-side calibration, intervention, recovery, and safety burden.

Second, *comparing residuals across bodies and protocols.* A small residual on a simple gripper may not be comparable to a small residual on a dexterous hand or mobile manipulator. Benchmarks such as LIBERO, RLBench, CALVIN, RoboCasa, FurnitureBench, and AnyBody provide task and embodiment contexts, and future benchmark reporting can make target-body adaptation cost more directly comparable by recording body differences and adaptation contracts alongside task success.

Third, *tracking residuals over time, including safety and recovery.* Residual adaptation burden can shrink through continual learning, fleet learning, better calibration tools, improved interfaces, or repeated deployments. EACs become more informative when they record not only success against demonstrations, but also unsafe exploration, frequent resets, interventions, execution quality, and recovery cost. Together, these problems turn residual adaptation burden from an implicit deployment cost into an explicit object of study: which residuals are shrinking, which are becoming safer, and which are being deliberately localized in reusable parts of the system.

9 Conclusion

This survey asked what transfers across robot bodies and what has to be realized on the target robot. The embodiment-gap lens gives that question a vocabulary: shareability levels name the structure that travels, dominant residual loci name where target-side work appears, and the three scaling directions show how reusable meaning, traceable robot data, and cross-body correspondence move toward behavior.

The concrete implication is reporting. Success rate becomes most useful when paired with the pathway that produced it: target data, model updates, calibration and setup, real-robot operation, safety events, recovery, and failure attribution. Report cards, burden profiles, and EACs make that pathway visible, so two systems with the same final success rate can still teach different lessons about target-body adaptation.

Embodiment-gap-aware scaling asks whether larger models, broader data, and stronger correspondence mechanisms leave new robots with less target-specific work, clearer failure attribution, and safer recovery. Vision can be reused. Language can be reused. Policy backbones can be reused. The gripper, the controller, the moment of contact, and the recovery after a failed trial show what robotics adds to the scaling conversation. The embodiment gap is the part of robot learning that keeps the robot in the story.

References

Ademi Adeniji, Zhuoran Chen, Vincent Liu, Venkatesh Pattabiraman, Raunaq Bhirangi, Siddhant Haldar, Pieter Abbeel, and Lerrel Pinto. Feel the force: Contact-driven learning from humans, 2025. URL `https://arxiv.org/abs/2506.01944`. arXiv preprint. arXiv:2506.01944.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, et al. Do as i can, not as i say: Grounding language in robotic affordances, 2023. URL `https://arxiv.org/abs/2204.01691`. In Proceedings of the 6th Conference on Robot Learning, PMLR 205, 2023.

ALOHA 2 Team, Jorge Aldaco, Travis Armstrong, et al. Aloha 2: An enhanced low-cost hardware for bimanual teleoperation, 2024. URL `https://arxiv.org/abs/2405.02292`. arXiv preprint. arXiv:2405.02292.

Nancy M. Amato, Seth Hutchinson, Animesh Garg, Aude Billard, Daniela Rus, Russ Tedrake, Frank Park, and Ken Goldberg. "data will solve robotics and automation: True or false?": A debate, 2025. Science Robotics, 2025. doi: 10.1126/scirobotics.aea7897.

Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics, 2023. URL `https://arxiv.org/abs/2304.08488`. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

Tianwei Bao, Qineng Wang, Kangrui Wang, Mingkai Deng, Guangyi Liu, Jiayuan Mao, Lawrence Birnbaum, Zhiting Hu, Eric P. Xing, Zhaoran Wang, and Manling Li. ActionEQA: Action Interface for Embodied Question Answering, 2026. URL https://openreview.net/forum?id=HY2ruqdMt4. Transactions on Machine Learning Research, 2026. Accepted by TMLR.

Erik Bauer, Elvis Nava, and Robert K. Katzschmann. Latent action diffusion for cross-embodiment manipulation, 2025. URL https://arxiv.org/abs/2506.14608. arXiv preprint. arXiv:2506.14608.

Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation, 2024. URL https://arxiv.org/abs/2405.01527. In European Conference on Computer Vision (ECCV), 2024.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi 0$: A Vision-Language-Action Flow Model for General Robot Control, 2024. URL https://arxiv.org/abs/2410.24164. arXiv preprint. arXiv:2410.24164.

Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Manon Devin, Alex X. Lee, Maria Bauza Villalonga, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, et al. Robocat: A self-improving generalist agent for robotic manipulation, 2024. URL https://openreview.net/forum?id=vsCpILiWHu. Transactions on Machine Learning Research, 2024. Accepted by TMLR; J2C certification.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023a. URL https://arxiv.org/abs/2307.15818. In Proceedings of The 7th Conference on Robot Learning, PMLR 229:2165-2183, 2023a.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale, 2023b. URL https://arxiv.org/abs/2212.06817. In Robotics: Science and Systems, 2023b.

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, 2025. URL https://arxiv.org/abs/2505.06111. In Robotics: Science and Systems, 2025. doi: 10.48550/arXiv.2505.06111.

Rémi Cadène, Simon Aliberts, et al. LeRobot: An open-source library for end-to-end robot learning, 2026. URL https://arxiv.org/abs/2602.22818. arXiv preprint. arXiv:2602.22818.

Anthony Chemero. An outline of a theory of affordances, 2003. Ecological Psychology, 2003.

Haonan Chen, Cheng Zhu, Shuijing Liu, Yunzhu Li, and Katherine Driggs-Campbell. Tool-as-interface: Learning robot policies from observing human tool use, 2025a. URL https://arxiv.org/abs/2504.04612. In Conference on Robot Learning (CoRL), 2025a. doi: 10.48550/arXiv.2504.04612.

Lawrence Yunliang Chen, Karthik Dharmarajan, Kush Hari, Chenfeng Xu, Quan Vuong, and Ken Goldberg. Mirage: Cross-embodiment zero-shot policy transfer with cross-painting, 2024. URL https://arxiv.org/abs/2402.19249. In Robotics: Science and Systems, 2024. doi: 10.15607/RSS.2024.XX.069.

Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning, 2025b. URL https://arxiv.org/abs/2409.03403. In Proceedings of The 8th Conference on Robot Learning, PMLR 270:209-233, 2025b.

Tao Chen, Adithyavairavan Murali, and Abhinav Gupta. Hardware conditioned policies for multi-robot transfer learning, 2018. URL https://arxiv.org/abs/1811.09864. In Advances in Neural Information Processing Systems, 2018.

Xi Chen, Yuan Gao, Hangxin Liu, Fangkai Yang, Ali Ghadirzadeh, Jun Yang, Bin Liang, Chongjie Zhang, Tin Lun Lam, and Song-Chun Zhu. Cross-robot behavior adaptation through intention alignment, 2026. Science Robotics, 2026. doi: 10.1126/scirobotics.adv2250.

Yixiang Chen, Peiyan Li, Yan Huang, Jiabing Yang, Kehan Chen, and Liang Wang. Ec-flow: Enabling versatile robotic manipulation from action-unlabeled videos via embodiment-centric flow, 2025c. URL https://arxiv.org/abs/2507.06224. In IEEE/CVF International Conference on Computer Vision (ICCV), 2025c. doi: 10.48550/arXiv.2507.06224.

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, 2024. URL https://arxiv.org/abs/2402.10329. In Robotics: Science and Systems, 2024.

Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey, 2017. URL https://arxiv.org/abs/1702.05374. Domain Adaptation in Computer Vision Applications. arXiv:1702.05374.

Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning, 2020. URL https://arxiv.org/abs/1910.11215. In Proceedings of the Conference on Robot Learning, PMLR 100:885-897, 2020.

Ria Doshi, Homer Rich Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation, 2024. URL https://arxiv.org/abs/2408.11812. arXiv preprint. arXiv:2408.11812.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model, 2023. URL https://arxiv.org/abs/2303.03378. arXiv preprint. arXiv:2303.03378.

Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex X. Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control, 2018. URL https://arxiv.org/abs/1812.00568. arXiv preprint. arXiv:1812.00568.

Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, Brian Ichter, Danny Driess, et al. Foundation models in robotics: Applications, challenges, and the future, 2025. The International Journal of Robotics Research, 2025. doi: 10.1177/02783649241281508.

Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation using low-cost whole-body teleoperation, 2025. URL https://arxiv.org/abs/2401.02117. In Proceedings of The 8th Conference on Robot Learning, PMLR 270:4066-4083, 2025.

Jensen Gao, Suneel Belkhale, Sudeep Dasari, Ashwin Balakrishna, Dhruv Shah, and Dorsa Sadigh. A taxonomy for evaluating generalist robot policies, 2025. URL https://arxiv.org/abs/2503.01238. arXiv preprint. arXiv:2503.01238.

James J. Gibson. The Ecological Approach to Visual Perception, 1979. Houghton Mifflin, 1979.

Ken Goldberg. Good old-fashioned engineering can close the 100,000-year "data gap" in robotics, 2025. Science Robotics, 2025. doi: 10.1126/scirobotics.aea7390.

Jun Guo, Xiaojian Ma, Yikai Wang, Min Yang, Huaping Liu, and Qing Li. Flowdreamer: A rgb-d world model with flow-based motion representations for robot manipulation, 2025. URL https://arxiv.org/abs/2505.10075. arXiv preprint. arXiv:2505.10075.

Agrim Gupta, Linxi Fan, Surya Ganguli, and Li Fei-Fei. Metamorph: Learning universal controllers with transformers, 2022. URL https://arxiv.org/abs/2203.11931. In International Conference on Learning Representations (ICLR), 2022.

Minho Heo, Youngwoon Lee, Doohyun Lee, and Joseph J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation, 2023. URL https://arxiv.org/abs/2305.12821. In Robotics: Science and Systems, 2023. doi: 10.15607/RSS.2023.XIX.041.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video, 2025. URL https://arxiv.org/abs/2505.11709. arXiv preprint. arXiv:2505.11709.

Cheng-Chun Hsu, Bowen Wen, Jie Xu, Yashraj Narang, Xiaolong Wang, Yuke Zhu, Joydeep Biswas, and Stan Birchfield. Spot: Se(3) pose trajectory diffusion for object-centric manipulation, 2025. URL https://arxiv.org/abs/2411.00965. In IEEE International Conference on Robotics and Automation (ICRA):4853-4860, 2025.

Edward S. Hu, Kun Huang, Oleh Rybkin, and Dinesh Jayaraman. Know thyself: Transferable visual control policies through robot-awareness, 2022. URL https://arxiv.org/abs/2107.09047. In International Conference on Learning Representations, 2022.

Yafei Hu, Quanting Xie, Vidhi Jain, Jonathan Francis, Jay Patrikar, Nikhil Keetha, Seungchan Kim, Yaqi Xie, Tianyi Zhang, Hao-Shu Fang, Shibo Zhao, Shayegan Omidshafiei, et al. Toward general-purpose robots via foundation models: A survey and meta-analysis, 2023. URL https://arxiv.org/abs/2312.08782. arXiv preprint. arXiv:2312.08782.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations, 2024. URL https://arxiv.org/abs/2412.14803. arXiv preprint. arXiv:2412.14803.

Huang Huang et al. Latent action robot foundation world models for cross-embodiment adaptation, 2026a. URL https://openreview.net/forum?id=vEZgPr1deb. OpenReview submission to ICLR 2026.

Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-VLA: Unlocking Vision-Language-Action Model's Physical Knowledge for Tactile Generalization, 2025. URL https://arxiv.org/abs/2507.09160. arXiv preprint. arXiv:2507.09160.

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, 2023. URL https://arxiv.org/abs/2307.05973. In Proceedings of The 7th Conference on Robot Learning, PMLR 229, 2023.

Yuzhe Huang, Pei Lin, Wanlin Li, Daohan Li, Jiajun Li, Jiaming Jiang, Chenxi Xiao, and Ziyuan Jiao. TaF-VLA: Tactile-Force Alignment in Vision-Language-Action Models for Force-Aware Manipulation, 2026b. URL https://arxiv.org/abs/2601.20321. arXiv preprint. arXiv:2601.20321.

Aadhithya Iyer, Zhuoran Peng, Yinlong Dai, Irmak Guzey, Siddhant Haldar, Soumith Chintala, and Lerrel Pinto. Open teach: A versatile teleoperation system for robotic manipulation, 2025. URL https://arxiv.org/abs/2403.07870. In Proceedings of the 8th Conference on Robot Learning, 2025.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. Rlbench: The robot learning benchmark and learning environment, 2020. URL https://arxiv.org/abs/1909.12271. IEEE Robotics and Automation Letters, 2020. doi: 10.1109/LRA.2020.2974707.

Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and José Santos-Victor. Affordances in psychology, neuroscience, and robotics: A survey, 2018. IEEE Transactions on Cognitive and Developmental Systems, 2018. doi: 10.1109/TCDS.2016.2594134.

Guanhua Ji et al. OXE-AugE: A large-scale robot augmentation of OXE for scaling cross-embodiment policy learning, 2025. URL https://arxiv.org/abs/2512.13100. arXiv preprint. arXiv:2512.13100.

Pingcheng Jian, Easop Lee, Zachary I. Bell, Michael M. Zavlanos, and Boyuan Chen. Perception stitching: Zero-shot perception encoder transfer for visuomotor robot policies, 2024. URL https://openreview.net/forum?id=tYxRyNT0TC. Transactions on Machine Learning Research, 2024. Accepted by TMLR.

Tobias Jülg, Pierre Krack, Seongjin Bien, Yannik Blei, Khaled Gamal, Ken Nakahara, Johannes Hechtl, Roberto Calandra, Wolfram Burgard, and Florian Walter. Robot control stack: A lean ecosystem for robot learning at scale, 2026. URL https://arxiv.org/abs/2509.14932. arXiv preprint. arXiv:2509.14932.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video, 2024. URL https://arxiv.org/abs/2410.24221. arXiv preprint. arXiv:2410.24221.

Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. Vision-language-action models for robotics: A review towards real-world applications, 2025. URL https://arxiv.org/abs/2510.07077. IEEE Access, 2025. doi: 10.1109/ACCESS.2025.3609980.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, et al. Droid: A large-scale in-the-wild robot manipulation dataset, 2024. URL https://arxiv.org/abs/2403.12945. arXiv preprint. arXiv:2403.12945.

Hanjung Kim, Jaehyun Kang, Hyolim Kang, Meedeum Cho, Seon Joo Kim, and Youngwoon Lee. Uniskill: Imitating human videos via cross-embodiment skill representations, 2025a. URL https://arxiv.org/abs/2505.08787. arXiv preprint. arXiv:2505.08787.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success, 2025b. URL https://arxiv.org/abs/2502.19645. In Robotics: Science and Systems, 2025b. doi: 10.48550/arXiv.2502.19645.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, et al. OpenVLA: An open-source vision-language-action model, 2025c. URL https://arxiv.org/abs/2406.09246. In Proceedings of The 8th Conference on Robot Learning, PMLR 270:2679-2713, 2025c.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos Policy: Fine-Tuning Video Models for Visuomotor Control and Planning, 2026. URL https://arxiv.org/abs/2601.16163. arXiv preprint. arXiv:2601.16163.

Vikash Kumar, Rutav Shah, Gaoyue Zhou, Vincent Moens, Vittorio Caggiano, Jay Vakil, Abhishek Gupta, and Aravind Rajeswaran. Robohive: A unified framework for robot learning, 2023. URL https://arxiv.org/abs/2310.06828. In Advances in Neural Information Processing Systems, 2023.

Marion Lepert, Ria Doshi, and Jeannette Bohg. Shadow: Leveraging segmentation masks for cross-embodiment policy transfer, 2025. In Proceedings of The 8th Conference on Robot Learning, PMLR 270:3536-3550, 2025.

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, 2023. URL https://arxiv.org/abs/2209.07753. In IEEE International Conference on Robotics and Automation (ICRA), 2023.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. Llm+p: Empowering large language models with optimal planning proficiency, 2023a. URL https://arxiv.org/abs/2304.11477. arXiv preprint. arXiv:2304.11477.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning, 2023b. URL https://arxiv.org/abs/2306.03310. In Advances in Neural Information Processing Systems, 2023b.

Mengyuan Liu, Juyi Sheng, Peiming Li, Ziyi Wang, Tianming Xu, Tiantian Xu, and Hong Liu. Trustworthy Evaluation of Robotic Manipulation: A New Benchmark and AutoEval Methods, 2026. URL https://arxiv.org/abs/2601.18723. arXiv preprint. arXiv:2601.18723.

Minghua Liu, Xuanlin Li, Zhan Ling, Yangyan Li, and Hao Su. Frame mining: A free lunch for learning robotic manipulation from 3d point clouds, 2023c. In Proceedings of the 6th Conference on Robot Learning, 2023c.

Wenhai Liu, Junbo Wang, Yiming Wang, Weiming Wang, and Cewu Lu. Forcemimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation, 2024. URL https://arxiv.org/abs/2410.07554. arXiv preprint. arXiv:2410.07554.

Tyler Ga Wei Lum, Olivia Y. Lee, C. Karen Liu, and Jeannette Bohg. Crossing the human-robot embodiment gap with sim-to-real rl using one human demonstration, 2025. URL https://arxiv.org/abs/2504.12609. arXiv preprint. arXiv:2504.12609.

Max Lungarella, Giorgio Metta, Rolf Pfeifer, and Giulio Sandini. Developmental robotics: A survey, 2003. Science, 2003. doi: 10.1080/09540090310001655110.

Robert McCarthy, Daniel C. H. Tan, Dominik Schmidt, Fernando Acero, Nathan Herr, Yilun Du, Thomas G. Thuruthel, and Zhibin Li. Towards generalist robot learning from internet video: A survey, 2025. Journal of Artificial Intelligence Research, 2025. doi: 10.1613/jair.1.17400.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks, 2022. URL https://arxiv.org/abs/2112.03227. IEEE Robotics and Automation Letters, 2022. doi: 10.1109/LRA.2022.3180108.

Tomohiro Motoda, Koshi Makihara, Ryoichi Nakajo, Hanbit Oh, Keisuke Shirai, Ryo Hanai, Masaki Murooka, Yuma Suzuki, Hiroki Nishihara, Mitsuru Takeda, Takumi Takada, Takayuki Hori, and Yukiyasu Domae. Recipe for vision-language-action models in robotic manipulation: A survey, 2025. TechRxiv preprint. doi:10.36227/techrxiv.175624610.06665789/v1.

Vincent C. Müller and Matej Hoffmann. What is morphological computation? on how the body contributes to cognition and control, 2017. Artificial Life, 2017.

Adithyavairavan Murali, Tao Chen, Kalyan Vasudev Alwala, Dhiraj Gandhi, Lerrel Pinto, Saurabh Gupta, and Abhinav Gupta. Pyrobot: An open-source robotics framework for research and benchmarking, 2019. URL https://arxiv.org/abs/1906.08236. arXiv preprint. arXiv:1906.08236.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation, 2023. URL https://arxiv.org/abs/2203.12601. In Proceedings of the 6th Conference on Robot Learning, 2023.

Soroush Nasiriany, Sean Kirmani, Tianli Ding, Laura Smith, Yuke Zhu, Danny Driess, Dorsa Sadigh, and Ted Xiao. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation, 2024a. URL https://arxiv.org/abs/2411.02704. arXiv preprint. arXiv:2411.02704.

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, 2024b. URL https://arxiv.org/abs/2406.02523. In Robotics: Science and Systems, 2024b.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, et al. Octo: An open-source generalist robot policy, 2024. In Robotics: Science and Systems, 2024.

Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, et al. Open x-embodiment: Robotic learning datasets and rt-x models, 2023. URL https://arxiv.org/abs/2310.08864. arXiv preprint. arXiv:2310.08864.

Chuer Pan, Brian Okorn, Harry Zhang, Ben Eisner, and David Held. Tax-pose: Task-specific cross-pose estimation for robot manipulation, 2023. In Proceedings of the 6th Conference on Robot Learning, 2023.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning, 2010. IEEE Transactions on Knowledge and Data Engineering, 2010. doi: 10.1109/TKDE.2009.191.

Meenal Parakh, Alexandre Kirchmeyer, Beining Han, and Jia Deng. Anybody: A benchmark suite for cross-embodiment manipulation, 2025. URL https://arxiv.org/abs/2505.14986. arXiv preprint. arXiv:2505.14986.

Austin Patel and Shuran Song. Get-zero: Graph embodiment transformer for zero-shot embodiment generalization, 2025. URL https://arxiv.org/abs/2407.15002. In IEEE International Conference on Robotics and Automation (ICRA), 2025. doi: 10.1109/ICRA55743.2025.11127922.

Rolf Pfeifer and Josh Bongard. How the Body Shapes the Way We Think: A New View of Intelligence, 2007. MIT Press, 2007.

Rolf Pfeifer, Max Lungarella, and Fumiya Iida. Self-organization, embodiment, and biologically inspired robotics, 2007. Science, 2007. doi: 10.1126/science.1145803.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, et al. π0.5: A Vision-Language-Action Model with Open-World Generalization, 2025. URL https://arxiv.org/abs/2504.16054. arXiv preprint. arXiv:2504.16054.

Michael Przystupa, Hongyao Tang, Martin Jagersand, Santiago Miret, Mariano Phielipp, Matthew E. Taylor, and Glen Berseth. Efficient morphology-aware policy transfer to new embodiments, 2025. URL https://arxiv.org/abs/2508.03660. In Reinforcement Learning Conference (RLC), 2025. doi: 10.48550/arXiv.2508.03660.

Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system, 2023. URL https://arxiv.org/abs/2307.04577. In Robotics: Science and Systems, 2023. doi: 10.15607/RSS.2023.XIX.015.

Sabela Ramos, Sertan Girgin, Léonard Hussenot, Damien Vincent, Hanna Yakubovich, Daniel Toyama, Anita Gergely, Piotr Stanczyk, Raphael Marinier, Jeremiah Harmsen, Olivier Pietquin, and Nikola Momchev. Rlds: an ecosystem to generate, share and use datasets in reinforcement learning, 2021. URL https://arxiv.org/abs/2111.02767. arXiv preprint. arXiv:2111.02767.

Sonia Raychaudhuri and Angel X. Chang. Semantic Mapping in Indoor Embodied AI: A Survey on Advances, Challenges, and Future Directions, 2025. URL https://openreview.net/forum?id=USgQ38RG6G. Transactions on Machine Learning Research, 2025. Accepted by TMLR.

Carmelo Sferrazza, Dun-Ming Huang, Fangchen Liu, Jongmin Lee, and Pieter Abbeel. Body transformer: Leveraging robot embodiment for policy learning, 2024. URL https://arxiv.org/abs/2408.06316. In Proceedings of The 8th Conference on Robot Learning, PMLR 270:3407-3424, 2024.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation, 2022. URL https://arxiv.org/abs/2109.12098. In Proceedings of the Conference on Robot Learning, PMLR 164:894-906, 2022.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation, 2023. URL https://arxiv.org/abs/2209.05451. In Proceedings of the Conference on Robot Learning, PMLR 205:785-799, 2023.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models, 2023. URL https://arxiv.org/abs/2209.11302. In IEEE International Conference on Robotics and Automation (ICRA), 2023. doi: 10.1109/ICRA48891.2023.10161317.

Mohan Kumar Srirama, Sudeep Dasari, Shikhar Bahl, and Abhinav Gupta. Hrp: Human affordances for robotic pre-training, 2024. URL https://arxiv.org/abs/2407.18911. In Robotics: Science and Systems, 2024. doi: 10.15607/RSS.2024.XX.068.

Kei Suzuki, Jing Liu, Ye Wang, Chiori Hori, Matthew Brand, Diego Romeres, and Toshiaki Koike-Akino. Embedding morphology into transformers for cross-robot policy learning, 2026. URL https://arxiv.org/abs/2603.00182. arXiv preprint. arXiv:2603.00182.

Matthew E. Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey, 2009. URL https://www.jmlr.org/papers/v10/taylor09a.html. Journal of Machine Learning Research 10(1):1633-1685. URL: https://www.jmlr.org/papers/v10/taylor09a.html.

Joshua Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world, 2017. URL https://arxiv.org/abs/1703.06907. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017.

William van den Bogert, Madhavan Iyengar, and Nima Fazeli. Built different: Tactile perception to overcome cross-embodiment capability differences in collaborative manipulation, 2024. URL https://arxiv.org/abs/2409.14896. arXiv preprint. arXiv:2409.14896.

Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, et al. Bridgedata v2: A dataset for robot learning at scale, 2023. URL https://arxiv.org/abs/2308.12952. In Proceedings of The 7th Conference on Robot Learning, PMLR 229:1723-1736, 2023.

Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play, 2023. URL https://arxiv.org/abs/2302.12422. In Proceedings of The 7th Conference on Robot Learning, PMLR 229, 2023.

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C. Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation, 2024a. URL https://arxiv.org/abs/2403.07788. In Robotics: Science and Systems, 2024a.

He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation, 2019. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers, 2024b. In Advances in Neural Information Processing Systems 37, 2024b. doi: 10.52202/079017-3952.

Tingwu Wang, Renjie Liao, Jimmy Ba, and Sanja Fidler. Nervenet: Learning structured policy with graph neural networks, 2018. In International Conference on Learning Representations (ICLR), 2018.

Ziyao Wang, Bingying Wang, Hanrong Zhang, Tingting Du, Tianyang Chen, Guoheng Sun, Yexiao He, Zheyu Shen, Wanghao Ye, and Ang Li. Vision-Language-Action in Robotics: A Survey of Datasets, Benchmarks, and Data Engines, 2026. URL https://arxiv.org/abs/2604.23001. arXiv preprint. arXiv:2604.23001.

Yunze Wei, Maria Attarian, and Igor Gilitschenski. Geomatch++: Morphology conditioned geometry matching for multi-embodiment grasping, 2024. URL https://arxiv.org/abs/2412.18998. arXiv preprint. arXiv:2412.18998.

Youngsun Wi, Jessica Yin, Elvis Xiang, Akash Sharma, Jitendra Malik, Mustafa Mukadam, Nima Fazeli, and Tess Hellebrekers. TactAlign: Human-to-Robot Policy Transfer via Tactile Alignment, 2026. URL https://arxiv.org/abs/2602.13579. arXiv preprint. arXiv:2602.13579.

Margaret Wilson. Six views of embodied cognition, 2002. Psychonomic Bulletin and Review, 2002. doi: 10.3758/BF03196322.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, et al. RoboMIND: Benchmark on multi-embodiment intelligence normative data for robot manipulation, 2025. URL https://arxiv.org/abs/2412.13877. In Robotics: Science and Systems, 2025. doi: 10.15607/RSS.2025.XXI.152.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning, 2023. URL https://arxiv.org/abs/2206.14176. In Proceedings of the 6th Conference on Robot Learning, 2023.

Yuliang Wu, Yanhan Lin, WengKit Lao, Yuhao Lin, Yi-Lin Wei, Wei-Shi Zheng, and Ancong Wu. DexGrasp-Zero: A Morphology-Aligned Policy for Zero-Shot Cross-Embodiment Dexterous Grasping, 2026. URL https://arxiv.org/abs/2603.16806. arXiv preprint. arXiv:2603.16806.

Sicheng Xie, Haidong Cao, Zejia Weng, Zhen Xing, Haoran Chen, Shiwei Shen, Jiaqi Leng, Zuxuan Wu, and Yu-Gang Jiang. Human2robot: Learning robot actions from paired human-robot videos, 2025. URL https://arxiv.org/abs/2502.16587. arXiv preprint. arXiv:2502.16587.

Zheng Xiong, Jacob Beck, and Shimon Whiteson. Universal morphology control via contextual modulation, 2023. In Proceedings of the 40th International Conference on Machine Learning, 2023.

Zheng Xiong, Risto Vuorio, Jacob Beck, Matthieu Zimmer, Kun Shao, and Shimon Whiteson. Distilling morphology-conditioned hypernetworks for efficient universal morphology control, 2024. URL https://arxiv.org/abs/2402.06570. In International Conference on Machine Learning (ICML), 2024. doi: 10.48550/arXiv.2402.06570.

Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery, 2023. URL https://arxiv.org/abs/2307.09955. In Proceedings of The 7th Conference on Robot Learning, PMLR 229:3536-3555, 2023.

Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface, 2025. URL https://arxiv.org/abs/2407.15208. In Proceedings of The 8th Conference on Robot Learning, 2025.

Jingyun Yang, Congyue Deng, Jimmy Wu, Rika Antonova, Leonidas Guibas, and Jeannette Bohg. Equivact: Sim(3)-equivariant visuomotor policies beyond rigid object manipulation, 2023a. URL https://arxiv.org/abs/2310.16050. arXiv preprint. arXiv:2310.16050.

Jonathan Yang, Chelsea Finn, and Dorsa Sadigh. Data analogies enable efficient cross-embodiment transfer, 2026. URL https://arxiv.org/abs/2603.06450. arXiv preprint. arXiv:2603.06450.

Jonathan Heewon Yang, Dorsa Sadigh, and Chelsea Finn. Polybot: Training one policy across robots while embracing variability, 2023b. URL https://arxiv.org/abs/2307.03719. In Proceedings of the 7th Conference on Robot Learning, 2023b.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, et al. Latent action pretraining from videos, 2025. URL https://arxiv.org/abs/2410.11758. In International Conference on Learning Representations (ICLR), 2025.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi Fan, and Joel Jang. World Action Models are Zero-shot Policies, 2026. URL https://arxiv.org/abs/2602.15922. arXiv preprint. arXiv:2602.15922.

Zhao-Heng Yin, Sherry Yang, and Pieter Abbeel. Object-centric 3d motion field for robot learning from human videos, 2025. URL https://arxiv.org/abs/2506.04227. arXiv preprint. arXiv:2506.04227.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. RoboPoint: A vision-language model for spatial affordance prediction in robotics, 2025. URL https://arxiv.org/abs/2406.10721. In Proceedings of The 8th Conference on Robot Learning, 2025.

Philipp Zech, Simon Haller, Safoura Rezapour Lakani, Barry Ridge, Emre Ugur, and Justus Piater. Computational models of affordance in robotics: A taxonomy and systematic classification, 2017. Adaptive Behavior, 2017. doi: 10.1177/1059712317726357.

Lihan Zha, Asher J. Hancock, Mingtong Zhang, Tenny Yin, Yixuan Huang, Dhruv Shah, Allen Z. Ren, and Anirudha Majumdar. LAP: Language-Action Pre-Training Enables Zero-Shot Cross-Embodiment Transfer, 2026. URL https://arxiv.org/abs/2602.10556. arXiv preprint. arXiv:2602.10556.

Chi Zhang, Penglin Cai, Haoqi Yuan, Chaoyi Xu, and Zongqing Lu. UniTacHand: Unified Spatio-Tactile Representation for Human to Robotic Hand Skill Transfer, 2025a. URL https://arxiv.org/abs/2512.21233. arXiv preprint. arXiv:2512.21233.

Jialiang Zhang, Haoran Geng, Yang You, Congyue Deng, Pieter Abbeel, Jitendra Malik, and Leonidas Guibas. Rodrigues Network for Learning Robot Actions, 2026. URL https://arxiv.org/abs/2506.02618. International Conference on Learning Representations (ICLR), Oral, 2026. arXiv:2506.02618.

Wenbo Zhang, Yang Li, Yanyuan Qiao, Siyuan Huang, Jiajun Liu, Feras Dayoub, Xiao Ma, and Lingqiao Liu. Effective tuning strategies for generalist robot manipulation policies, 2025b. URL https://arxiv.org/abs/2410.01220. In IEEE International Conference on Robotics and Automation (ICRA):7255-7262, 2025b. doi: 10.48550/arXiv.2410.01220.

Yuhong Zhang, Zihan Gao, Shengpeng Li, Ling-Hao Chen, Kaisheng Liu, Runqing Cheng, Xiao Lin, Junjia Liu, Zhuoheng Li, Jingyi Feng, Ziyan He, Jintian Lin, Zheyan Huang, Zhifang Liu, and Haoqian Wang. RoboWheel: A data engine from real-world human demonstrations for cross-embodiment robotic learning, 2025c. URL https://arxiv.org/abs/2512.02729. arXiv preprint. arXiv:2512.02729.

Zhiyuan Zhang, Zhengtong Xu, Jai Nanda Lakamsani, and Yu She. Canonical policy: Learning canonical 3d representation for se(3)-equivariant policy, 2025d. URL https://arxiv.org/abs/2505.18474. arXiv preprint. arXiv:2505.18474.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware, 2023. URL https://arxiv.org/abs/2304.13705. In Robotics: Science and Systems, 2023. doi: 10.15607/RSS.2023.XIX.016.

Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: A survey, 2020. URL https://arxiv.org/abs/2009.13303. IEEE Symposium Series on Computational Intelligence (SSCI). arXiv:2009.13303. doi:10.1109/SSCI47803.2020.9308468.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, Ya-Qin Zhang, Jiangmiao Pang, Jingjing Liu, Tai Wang, and Xianyuan Zhan. X-VLA: Soft-Prompted Transformer as a Scalable Cross-Embodiment Vision-Language-Action Model, 2026a. URL https://arxiv.org/abs/2510.10274. International Conference on Learning Representations (ICLR), Poster, 2026. arXiv:2510.10274.

Linfang Zheng, Zikai Ouyang, Chen Wang, Jia Pan, and Wei Zhang. From Video to Control: A Survey of Learning Manipulation Interfaces from Temporal Visual Data, 2026b. URL https://arxiv.org/abs/2604.04974. arXiv preprint. arXiv:2604.04974.

Heng Zhi, Wentao Tan, Lei Zhu, Fengling Li, Jingjing Li, Guoli Yang, and Heng Tao Shen. MOTIF: Learning Action Motifs for Few-Shot Cross-Embodiment Transfer, 2026. URL https://arxiv.org/abs/2602.13764. arXiv preprint. arXiv:2602.13764.

Hongyan Zhi, Peihao Chen, Siyuan Zhou, Yubo Dong, Quanxi Wu, Lei Han, and Mingkui Tan. 3DFlowAction: Learning cross-embodiment manipulation from 3D flow world model, 2025. URL https://arxiv.org/abs/2506.06199. arXiv preprint. arXiv:2506.06199.

Yifan Zhong, Fengshuo Bai, Shaofei Cai, Xuchuan Huang, Zhang Chen, Xiaowei Zhang, Yuanfei Wang, Shaoyang Guo, Tianrui Guan, Ka Nam Lui, Zhiquan Qi, Yitao Liang, et al. A Survey on Vision-Language-Action Models: An Action Tokenization Perspective, 2025. URL https://arxiv.org/abs/2507.01925. arXiv preprint. arXiv:2507.01925.

Zhiyuan Zhou, Pranav Atreya, You Liang Tan, Karl Pertsch, and Sergey Levine. Autoeval: Autonomous evaluation of generalist robot manipulation policies in the real world, 2025. URL https://arxiv.org/abs/2503.24278. In Proceedings of The 9th Conference on Robot Learning, PMLR 305:1997-2017, 2025.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Kevin Lin, Abhiram Maddukuri, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning, 2020. URL https://arxiv.org/abs/2009.12293. arXiv preprint. arXiv:2009.12293.

A Scope, Literature Selection, and Boundary Cases

A.1 Scope of the survey

This survey focuses primarily on manipulation-oriented RFMs, especially VLA policies and generalist robot policies, because manipulation makes the embodiment gap especially visible: reaching, grasping, contact, end-effectors, sensors, controllers, and resets all shape whether a shared representation becomes executable. The lens can extend to locomotion or navigation, but the residuals there often emphasize terrain, dynamics, localization, and long-horizon autonomy rather than object contact and grasp realization. We therefore use manipulation as the main setting and discuss adjacent embodied-AI work only where it clarifies reporting or evaluation.

We include VLA backbones and generalist policies, together with the data, interface, correspondence, morphology-aware, tactile/force, and evaluation mechanisms that clarify target-body realization. Dataset, benchmark, and interface papers are included when they determine what cross-body information is visible: robot metadata, action schema, controller assumptions, adaptation operations, evaluation protocol, or safety/recovery evidence.

A.2 Literature selection strategy

The literature set is constructed as a lens-driven scoping survey. We assembled the corpus from recent surveys, venue proceedings, OpenReview records, arXiv records, dataset and benchmark documentation, and targeted searches around the following query families: robot foundation models; vision-language-action policies; generalist robot policies; multi-embodiment or cross-embodiment robot learning; robot data engines and schemas; teleoperation datasets; retargeting and cross-body correspondence; object flow, point tracks, and latent actions; morphology-conditioned control; tactile/force grounding; and robot-policy evaluation. The main coverage window is work available up to early 2026, with earlier papers included when they define an important mechanism used by recent systems.

Works were included when they satisfied at least one of four roles: they introduce a widely used robot foundation model or generalist robot policy; define cross-embodiment data, interfaces, schemas, or benchmarks; propose a distinct correspondence mechanism between bodies, objects, actions, or sensors; or expose evaluation evidence about target-body residuals. We excluded papers whose main contribution was outside embodied control, papers that did not affect transfer or target-body realization, and works used only as generic background unless they clarified an adjacent concept. Recent surveys are used as coverage checks to avoid turning this paper into a general robot-foundation-model catalog. The main text emphasizes representative systems and mechanisms rather than every paper in the corpus; Figure 2 uses 21 grouped placements to keep the map readable, and Appendix C gives the coding rule and placement rationale for each group. Recent preprints and OpenReview submissions are retained when they clarify emerging mechanisms or fast-moving empirical practice, but we distinguish their status from accepted or published work and do not use them as the sole evidence for the main conceptual claims. Core claims are anchored whenever possible in published, accepted, or widely used systems, datasets, and benchmarks; emerging work is used primarily to show where the field is moving. Table 2 applies the burden profile to nine compact case-study rows.

A.3 Boundary cases and inclusion decisions

Table 3 summarizes boundary cases and inclusion decisions used to keep the survey scope explicit.

Table 3: Boundary cases and inclusion decisions.

Boundary case	Why boundary	Treatment
Body Transformer	Body-structured policy, not VLA	Action/skill sharing with body conditioning
GET-Zero	Body graph-conditioned policy generation	Morphology-conditioned mechanism
RodriNet	Kinematic prior, not RFM	Architecture-level mechanism
Tactile/force methods	Contact signals, not always VLA	Execution-proximal grounding
Common APIs	Planning or control interface	Treated by role

Boundary case	Why boundary	Treatment
Object flow / point tracks	Visual cue or correspondence	Treated by target connection
Latent actions	Action abstraction	Action/skill correspondence
Evaluation frameworks	Not models	Reveal residuals

B Adjacent Concepts and Terminology

Table 4 summarizes adjacent concepts and how they relate to the embodiment gap.

Table 4: Adjacent concepts and relation to the embodiment gap.

Concept	Main concern	Relation to embodiment gap
Domain shift	Observation/environment shift	May occur without body change
Sim-to-real gap	Simulation-real mismatch	Overlaps under real-body deployment
Transfer learning	Reuse across settings	Broader than embodiment realization
Action-space mismatch	Different action interfaces	Important but narrower
Embodiment gap	Shared structure $\rightarrow$ target realization	Focus of this survey

Embodied cognition, ecological psychology, affordance theory, morphological computation, and developmental robotics ground action possibilities in body-environment relations (Wilson, 2002; Gibson, 1979; Pfeifer and Bongard, 2007; Jamone et al., 2018; Müller and Hoffmann, 2017; Chemero, 2003; Lungarella et al., 2003; Zech et al., 2017; Pfeifer et al., 2007). We use these traditions only as conceptual grounding for an engineering reading of action possibilities, not as a comprehensive review of embodied cognition.

C Extended Shareability and Residual-Locus Lens

Table 5 gives the extended definitions used for the shareability levels in Section 3.

Table 5: Shareability levels.

Shareability level	Shared structure	Target-body link	Typical residual
Semantic / task	Goals; plans	Skills; APIs	Feasibility
Perceptual / affordance	Features; affordances	Grasp/motion generation	Calibration; reachability
Object-interaction	Flow; tracks	Retargeting; contact planning	Timing; force
Action / skill	Latent actions; tokens	Decoders; adapters	Controller alignment
Morphology-conditioned	Body graphs; kinematic tokens	Body-conditioned policy	Tactile; recovery

The boundary between levels is decided by the representation's primary commitment. Affordance maps primarily localize possible interaction and therefore sit near perceptual/affordance. Object flow and point tracks primarily specify scene change and therefore sit near object-interaction. If the same method also learns a residual policy or action decoder, that second component is treated as the target-body connection rather than as the shared structure itself.

The two-axis map follows a fixed qualitative coding rule. The primary shareability level is the structure most central to the cross-embodiment claim. The dominant residual locus is the target-side work most visibly left after that structure is transferred. When a system combines multiple levels, the placement follows its distinctive contribution rather than all of its components; in such cases, the table notes the body-conditioned or residual component in the rationale. For dataset and infrastructure groups, the shareability entry names the reusable data or interface resource and the plotted coordinate approximates the representation level that resource makes reusable. The translucent scale overlays are therefore overlapping emphasis regions, not exclusive classes, and the map should not be read as a scalar ranking. Table 6 gives the corresponding placement rationale for Figure 2.

Table 6: Canonical placement rationale for the two-axis emphasis map.

System / mechanism	Primary shareability	Dominant residual locus	Scale overlay	Placement rationale
SayCan	Semantic / task	Skill/API grounding	Semantic	Language-level task structure is shared; feasibility and available skills remain target-specific.
RT-2 / PaLM-E	Semantic / task	Skill/API grounding	Semantic	Web-scale semantic and visual-language priors are central; executable action still depends on the robot stack.
CLIPort / PerAct	Perceptual / affordance	Calibration / action-interface alignment	Semantic	Visual and spatial affordance structure is shared; motion generation and calibration remain.
RT-Affordance / RoboPoint / VoxPoser	Perceptual / affordance	Calibration / action-interface alignment	Semantic–physical-data overlap	Spatial or affordance cues constrain action, but target-body grounding remains necessary.
VPP / FlowDreamer	Perceptual–object interaction	Calibration–contact	Semantic–physical-data or correspondence overlap	Video prediction and RGB-D world models provide predictive visual or flow-based representations of scene evolution; target-body action grounding, timing, contact, and control remain.
DROID / LeRobot / RoboMIND	Physical-data infrastructure / robot-experience schema	Skill/API grounding–calibration / action-interface alignment	Physical-data	Reusable robot experience, schemas, metadata, and tooling are the main shared resource; physical meaning and embodiment interfaces remain target-specific.
RT-X / OXE	Object-interaction–action	Calibration / action-interface alignment	Physical-data	Multi-robot data and common action conventions support transfer, while embodiment interfaces remain target-specific.
Octo	Action / skill	Calibration / action-interface alignment	Physical-data	A generalist policy is reused across tasks and robots, but target interfaces and adaptation remain central.
OpenVLA	Action / skill	Calibration / action-interface alignment	Physical-data	The VLA backbone and action generation are shared; target data, action decoding, and controller alignment remain.
BridgeData / RoboNet / RLDS	Physical-data infrastructure / trajectory-action schema	Calibration / action-interface alignment	Physical-data	Dataset infrastructure and episode schemas support reusable training and action-schema sharing, but do not by themselves fix embodiment-specific physical meaning.
Track2Act	Object-interaction	Contact / force / tactile execution	Correspondence	Point tracks act as object-motion correspondence; residual policy learning and contact execution remain on the target body.
MOTIF	Object-interaction–action	Contact / force / tactile execution	Correspondence	Action motifs carry cross-body structure, while few-shot target execution still bears residual work.
Cosmos Policy / DreamZero (WAM)	Object-interaction–action	Contact / force / tactile execution	Physical-data–correspondence overlap	World-action models connect video/world dynamics to action generation, future-state prediction, and planning. DreamZero, the WAM introduced in (Ye et al., 2026), reports cross-embodiment transfer from video-only demonstrations and few-shot adaptation, but target-body residuals remain in action-interface alignment, contact, force, safety, and recovery.
Flow-based methods	Object-interaction	Contact / force / tactile execution	Correspondence	Object or scene flow specifies desired change, while timing, contact, force, and target-body closed-loop realization remain.
RoboCat / π_0 / $\pi_{0.5}$	Action / skill	Contact / force / tactile execution	Physical-data–correspondence overlap	Generalist policies share action-level structure but still rely on target adaptation and execution robustness.
UMI / ALOHA	Action / skill	Contact / force / tactile execution	Correspondence	UMI emphasizes correspondence-oriented demonstration interfaces, whereas ALOHA is primarily a physical-data collection platform; both expose skill-data pathways while contact-rich execution remains body-specific.
XSkill / UniSkill	Action / skill	Contact / force / tactile execution	Correspondence	Shared skill abstractions are central, but target execution details remain.
Body Transformer	Action/skill–morphology-conditioned	Contact / force / tactile execution	Correspondence	Sensor-actuator graph computation makes body structure explicit within policy learning, while contact, force, tactile execution, safety, and recovery remain target-body residuals.
GET-Zero	Morphology-conditioned	Contact / force / tactile execution	Correspondence	A robot-body graph conditions policy generation directly, so the plotted point sits farther toward morphology-conditioned sharing; contact, force, tactile execution, safety, and recovery remain outside the shared representation.
X-VLA / RodriNet	Morphology-conditioned	Contact / force / tactile execution	Correspondence	Body-specific prompting or kinematic priors make embodiment explicit, while contact and deployment residuals remain.
TactAlign / Tactile-VLA / ForceMimic	Action/skill + tactile/force grounding	Contact / force / tactile execution	Correspondence	Tactile and force signals make execution residuals observable and partially transferable, but the primary shared structure is not morphology itself; robust deployment, safety, and recovery remain target-body residuals.

D Extended Reporting Checklists

Table 7 expands the report-card items into a checklist for future papers.

Table 7: Extended burden checklist.

Burden item	**Suggested details**
Target-data burden	Demos; rollouts; failure/recovery data
Model-update burden	Frozen parts; adapters/LoRA; decoder/full FT
Calibration/setup burden	Cameras; frames; gripper; control rate; workspace
Real-robot operation burden	Robot hours; resets; re-executions
Safety/intervention/recovery burden	Stops; unsafe contacts; manual/autonomous recovery
Evaluation evidence	Rollouts; tasks; seeds; success criteria
Failure attribution	Primary and secondary layer

Adaptation operations and evaluation rollouts should be separated. Adaptation operations are target-robot trials, resets, or data used to improve or configure the system. Evaluation rollouts are performance evidence. Evaluation rollouts should not be inferred as adaptation burden, although they may contribute to real-robot operation burden.

Worked report-card example. Table 8 applies the report-card format to OpenVLA as a partial worked example, and Table 9 gives example failure-attribution readings for common target-body breakdowns. The OpenVLA example does not re-evaluate the system; it shows what can be reconstructed from public reporting and which reporting fields future work can make routine under the embodiment-gap lens. N/R denotes information that we could not identify in the main paper and associated public supplementary material at the time of writing.

Table 8: Worked report-card example for OpenVLA.

Item	**Reconstructed report**	**Reading under the embodiment-gap lens**
Source embodiment	Open X-Embodiment pretraining and reported out-of-box settings on source robots	Origin of the shared VLA backbone and action-token interface
Target embodiment	Franka target fine-tuning setting	Target-body realization context
Shared structure	7B VLA backbone; tokenized actions	What transfers across settings
Updated components	Full fine-tuning or LoRA adaptation	What changes for the target body
Target-data burden	10–150 demonstrations per Franka task; seven tasks	Target experience required for adaptation
Model-update burden	Full FT or LoRA; LoRA rank 32 trains a small fraction of parameters	Update cost is visible
Calibration/setup burden	Controller frequencies partly reported; calibration metadata N/R	Setup residual partly visible
Real-robot operation burden	Reported evaluation rollouts; adaptation trials, resets, and robot-hours N/R	Physical operation cost is a reporting frontier
Safety/intervention/ recovery burden	Qualitative recovery examples; stops and interventions N/R	Safety and recovery burden are reporting frontiers
Evaluation evidence	Reported task rollouts and success rates	Performance evidence is visible
Failure attribution	Not systematically reported	Residual source is ready for systematic reporting

Table 9: Example failure-attribution readings.

Observed failure	**Primary attribution**	**Secondary attribution**
Correct object flow, but the gripper slips	Target-embodiment execution	Embodiment correspondence
Affordance point outside the reachable workspace	Embodiment correspondence	Target-embodiment execution
Action token mismatches controller frequency	Physical-data scale	Target-embodiment execution
Correct task plan, unavailable skill	Semantic scale	Skill/API grounding