MM-ABC: TOWARDS GENERALIST MOBILE MANIPULATION VIA SEEING, COORDINATING AND IMAGINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Mobile manipulation lets a robot control the extent of its own workspace, but it couples two problems that fixed-base policies can avoid: perception must stay spatially grounded under continuous ego-motion, and heterogeneous arm and base actions must be coordinated. We present MM-ABC, a vision–language–action model built around *seeing*, *coordinating*, and *imagining* arm–base collaboration. MM-ABC conditions its action expert on sparse multilevel VLM features, supervises the shared perceptual representation with a training-only branch that predicts the future features of a frozen geometric teacher, and generates manipulation and body actions with MM-APT, a two-stream transformer that couples the streams through masked near–far joint attention and predicts clean action chunks. A toy experiment motivates clean-action prediction for coupled streams, and in the full policy, replacing it with velocity prediction lowers success on RoboCasa365 composite-seen tasks from 32.8% to 29.2%. We pretrain MM-ABC on 5,000+ hours of robot data spanning 400K+ episodes, 12 datasets, and 17 embodiments, including MM-30, a new multi-view dual-arm mobile manipulation dataset. MM-ABC achieves 44.71% success on EBench, 61.2% on RoboCasa365, the highest mean success on three ManiSkill-HAB suites, 99.1% on LIBERO, 82.8% on LIBERO-Plus without perturbation training, and 83% on five long-horizon real-world tasks, 12 points above $\pi_{0.5}$.

1 INTRODUCTION

Robot manipulation ultimately requires bringing end effectors into configurations that interact with objects. A fixed-base manipulator can do so only inside its kinematic workspace, and a target beyond that region is out of reach for even a capable policy. A mobile base turns the workspace from a fixed property into a variable that the policy controls. Hence, *mobile manipulation is not navigation followed by manipulation, but manipulation over a workspace that the policy itself reconfigures*.

This view exposes two coupled challenges. First, **spatial perception becomes dynamic**: ego-motion continuously changes viewpoints and reference frames, while interaction still requires precise relative geometry among the robot, the target, and the scene. Second, **arm and base actions must be coordinated**: the base repositions the workspace coarsely and the arm performs fine interaction, yet base motion determines what the arm can reach and the intended interaction determines where the base should move. Existing work strengthens perception with point clouds, depth, geometric tokens, or auxiliary geometric objectives (Liu et al., 2026a; Zhu et al., 2026; Liang et al., 2026b; Tu et al., 2026) and learns future visual or geometric representations (Zheng et al., 2025; Chen et al., 2026b; Li et al., 2026e), but explicit reconstruction or dense future generation adds modeling and inference cost. For coordination, recent methods decouple mobility and manipulation into separate branches, subsystem-specific perception, or structured action spaces (Chen et al., 2026c; Liu et al., 2026a; Zhu et al., 2026; Chen et al., 2026a). Decoupling alone does not guarantee collaboration: once arm and base occupy different streams, the central question is how the streams should exchange information while preserving their distinct control semantics.

We present **MM-ABC**, a vision–language–action (VLA) model for mobile manipulation organized around **seeing, coordinating, and imagining** (Figure 1). For *seeing*, a sparse DeepStack interface

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

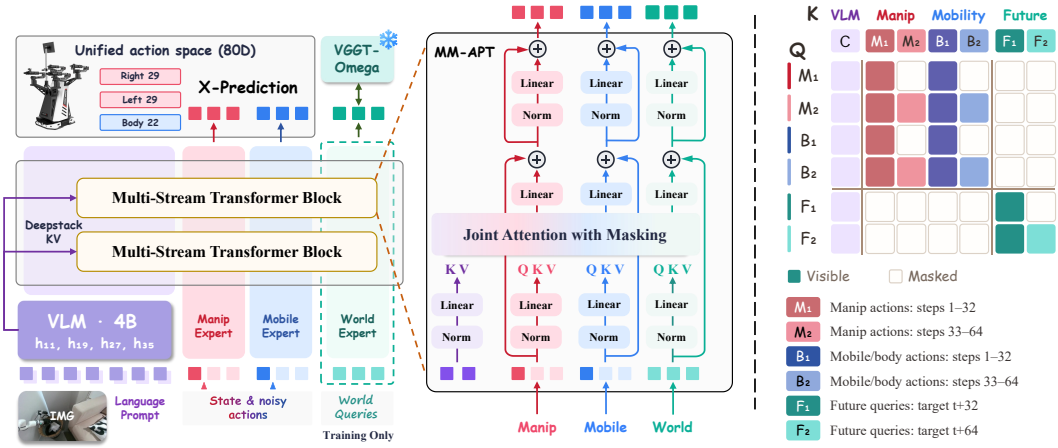


Figure 1: **MM-ABC architecture.** *Left:* sparse VLM features condition MM-APT, and a training-only future branch is supervised by frozen VGGT-Ω features. *Middle:* joint blocks (MM-JiT in the diagram) combine stream-specific transformations with masked attention over read-only context. *Right:* action streams communicate within each segment, far tokens also read near tokens, and future queries stay isolated from actions.

injects intermediate VLM features into the action expert as read-only context, so that fine visual evidence reaches the policy alongside the final semantic representation. For *imagining*, learnable future queries predict the features of a frozen geometric teacher at future horizons; the branch supervises the shared representation during training and is removed at deployment, so world imagination serves purely as extra supervision of geometric intent. For *coordinating*, the Mobile Manipulation Action Prediction Transformer (**MM-APT**) keeps separate manipulation and body streams, couples them through masked joint attention with an asymmetric near–far pattern, and predicts clean action chunks. This last choice rests on a representational argument: under rectified flow, clean-action (x) and velocity (v) prediction are algebraically interchangeable, yet a velocity head must carry and cancel the full-dimensional flow noise. A toy two-stream experiment illustrates the point: x-prediction keeps far less injected noise in its endpoint estimates and realizes a prescribed arm–body task allocation more accurately under few-step sampling (Section 3.5).

To train MM-ABC across embodiments, we build a 5,000+-hour corpus from 12 datasets and 17 embodiments with a data engine that maps every source into a masked 80-dimensional state/action interface and keeps only temporally contiguous valid segments. It includes **MM-30**, a new dataset of 30+ hours of multi-view teleoperated demonstrations over 40+ tasks on a dual-arm mobile manipulator with an omnidirectional base (Section 4). Empirically, MM-ABC exceeds the strongest baselines on EBench, RoboCasa365, and all three ManiSkill-HAB suites, reaches 99.1% on LIBERO and 82.8% on LIBERO-Plus without perturbation training, and averages 83% success on five long-horizon real-world tasks, 12 points above $\pi_{0.5}$ (Intelligence et al., 2025). In summary, this paper makes the following three contributions:

- **MM-ABC**, a mobile manipulation VLA combining sparse multilevel perceptual context, training-only future-geometry supervision, and a two-stream action transformer with near–far joint attention that predicts clean action chunks (Section 3).
- **A multi-embodiment data engine and corpus** that includes **MM-30**, a new multi-view dual-arm mobile manipulation dataset (Section 4).
- **Extensive evaluation** on five simulation benchmarks and five real-world tasks, with clear leads on mobile manipulation and strong fixed-base performance (Section 5).

2 Related Work

Mobile manipulation and generalist policies. Model-based whole-body control (Khatib, 1999) was followed by learned mobile manipulation (Hu et al., 2023; Ehsani et al., 2024; Fu et al., 2024; Jiang et al., 2025b; Liang et al., 2026a), and generalist VLA policies based on discrete tokens, dif-

fusion, or flow matching (Kim et al., 2024; Chi et al., 2025; Black et al., 2024; Intelligence et al., 2025) now extend to mobile and whole-body control (Wu et al., 2025b; Chen et al., 2026b; Team et al., 2026b; Wu et al., 2026). For coordination, AC-DiT conditions manipulation on a mobility prior (Chen et al., 2026c), InCoM and GeoHAT structure arm–base coordination (Liu et al., 2026a; Zhu et al., 2026), MoPA aligns perception per subsystem (Chen et al., 2026a), and VITRA, Qwen-VLA, and RLDX-1 couple VLM features with action streams through attention (Li et al., 2025a; Wang et al., 2026c; Kim et al., 2026a). MM-ABC applies joint attention (Esser et al., 2024) to specialized manipulation and body streams with asymmetric temporal visibility and read-only perceptual context.

Perceptual interfaces and predictive supervision. Spatially grounded policies use 3D representations, point clouds, and geometry-aware designs (Shridhar et al., 2023; Ke et al., 2024; Li et al., 2026b; Qu et al., 2025; Liang et al., 2026b) or geometric feature teachers (Li et al., 2026c; Guo et al., 2025), and action experts read VLM features through cross-attention (Bjorck et al., 2025), the final layer (Yang et al., 2026b), or several depths (Meng et al., 2024; Luo et al., 2026). Predictive objectives target future images, video embeddings, or latent world states (Wu et al., 2024; Assran et al., 2025; Zheng et al., 2025), and removable prediction branches keep their training benefit without deployment cost (Soleymanzadeh et al., 2026; Li et al., 2026e); MM-ABC predicts future VGGT-Ω features (Wang et al., 2026b) in a branch masked from the action streams.

Prediction targets. The prediction target is a longstanding design choice in diffusion and flow models (Salimans & Ho, 2022; Karras et al., 2022; Lipman et al., 2022); DP3, RDT-1B, and ManiCM predict clean actions (Ze et al., 2024; Liu et al., 2025; Gao et al., 2024), ABot-M0 pairs clean outputs with a velocity loss (Yang et al., 2026b), and JiT motivates clean-data prediction through manifold structure (Li & He, 2026). We revisit this choice for coupled arm–body streams with a toy experiment and a full-policy ablation; Section A extends this discussion.

3 MM-ABC

Overview. At environment step n, MM-ABC receives multi-view images I_n, a language instruction with embodiment metadata, and a normalized state $s_n \in \mathbb{R}^{80}$ with validity mask m_n^s (Section 4). It generates a 64-step chunk $X \in \mathbb{R}^{64\times 80}$, split into manipulation $X_M \in \mathbb{R}^{64\times 58}$ and body motion $X_B \in \mathbb{R}^{64\times 22}$. A Qwen3-VL-4B backbone (Bai et al., 2025) encodes images and instruction, MM-APT jointly denoises the two action streams, and a training-only future stream predicts later geometric features (Figure 1).

3.1 Seeing: Sparse Multilevel Perceptual Context

Conditioning only on the final VLM layer places both semantic recognition and local geometry on one representation, while attending to every layer gives the action expert many interfaces to reconcile. Following DeepStack (Meng et al., 2024), we keep the final representation as a common context and add a few intermediate features. With hidden states $H^l \in \mathbb{R}^{S\times 2560}$ and $C_{-1} = H^{35}$, we inject intermediate features into the first three of the 16 expert blocks as gated residual updates:

$$C_i = C_{i-1} + g_i \odot P_i\big(\mathrm{LN}(H^{l_i})\big), \qquad (l_0, l_1, l_2) = (11, 19, 27), \tag{1}$$

where P_i is a tokenwise linear map and g_i a zero-initialized gate, so the initial interface equals final-layer conditioning. Action tokens read the resulting keys and values without updating them, which holds perceptual evidence fixed throughout a denoising trajectory.

3.2 Coordinating: Structured Two-Stream Action Generation

The body changes reachability and viewpoint while the arm controls local interaction, so MM-APT gives each a separate stream and lets them adjust to each other through joint attention. MM-APT has 16 blocks of width 1024; each stream encodes its noisy actions with a time embedding and prepends a state token from $[s_n; m_n^s]$, and segment embeddings split each chunk into a *near* segment (state token and first 32 actions) and a *far* segment (last 32 actions). The streams keep separate normalization, attention projections, and SwiGLU layers, but share one masked attention

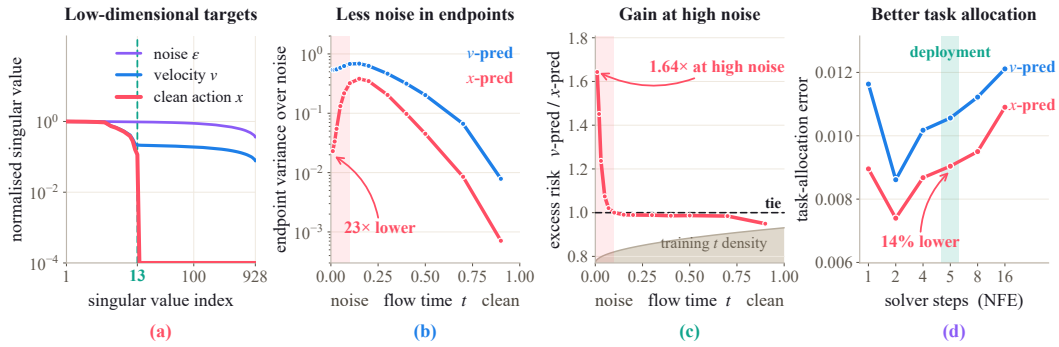


Figure 2: **Toy experiment: clean-action versus velocity prediction under a common loss.** **(a)** Clean chunks occupy a 13-dimensional subspace, whereas velocity and noise span the ambient space. **(b)** Variance of the endpoint estimate across noise draws with the clean chunk fixed. **(c)** Excess risk over the Bayes-optimal denoiser as a v/x ratio. **(d)** Task-allocation error of sampled chunks against the number of Euler steps.

over manipulation, body, and read-only context tokens. Manipulation and body tokens communicate bidirectionally within each segment; far queries also read near keys, whereas near queries cannot read far keys. The executable prefix is thus grounded in the current observation and unaffected by speculative far-future actions, while the tail builds on the immediate plan; separate decoders then predict the clean chunks $\hat{X}_M$ and $\hat{X}_B$ without a noisy-input skip.

3.3 Imagining: Future Geometry as Auxiliary Supervision

A robot approaching a handle needs features that anticipate how the handle and the end effector will align, not only the handle's present appearance. A training-only world expert therefore predicts geometric features at steps $n + 32$ and $n + 64$ from the current context, using 192 learned queries per horizon (8×8 per view for up to three views) that form a separate stream and never see action tokens or future images. A frozen VGGT-Ω teacher (Wang et al., 2026b) encodes the recorded future images into pooled targets F^*_{brj}, and a linear decoder maps query outputs to predictions $\hat{F}_{brj}$:

$$\mathcal{L}_{\text{future}} = \frac{\sum_{b,r,j} q_{brj} \left[1 - \cos_\delta(\hat{F}_{brj}, F^*_{brj})\right]}{\max\left(1, \sum_{b,r,j} q_{brj}\right)}, \tag{2}$$

where b, r, and j index examples, horizons, and view–spatial tokens, q_{brj} masks unavailable targets, and $\cos_\delta$ is a stabilized cosine. Because the future stream reads the same VLM features as the action streams, this loss trains the backbone to expose predictive geometric information to the policy; the branch and the teacher are removed at inference.

3.4 Clean-Action Prediction and Training Objective

With Gaussian noise E and $t \sim \text{Beta}(1.5, 1)$, flow time defines $Z_t = (1 - t)E + tX$ and velocity $V = X - E$. An x-head predicts the endpoint directly, $\hat{X} = f_\theta(Z_t, t, c)$, whereas a v-head predicts $g_\theta \approx V$ with endpoint $\hat{X} = Z_t + d(t)\, g_\theta$ and $d(t) = \max(1 - t, 0.05)$; MM-APT uses the x-head. With normalized weights $\bar{w}_b = d(t_b)^{-2} / [B^{-1} \sum_{b'} d(t_{b'})^{-2}]$, masks $M_{b,s}$, and active streams $\mathcal{S}_b$,

$$\mathcal{L}_{\text{action}} = \frac{1}{B} \sum_{b=1}^{B} \frac{\bar{w}_b}{|\mathcal{S}_b|} \sum_{s \in \mathcal{S}_b} \frac{\|M_{b,s} \odot (\hat{X}_{b,s} - X_{b,s})\|_F^2}{\|M_{b,s}\|_1}, \qquad \mathcal{L} = \mathcal{L}_{\text{action}} + 0.05\, \mathcal{L}_{\text{future}}, \tag{3}$$

which equals velocity regression for a v-head wherever $d(t) = 1 - t$, so the two parameterizations differ only in what the network emits. $\mathcal{L}_{\text{action}}$ is averaged over four noise/time draws that share one context computation. At inference, the streams integrate from noise with five Euler steps using $(\hat{X} - Z_t)/d(t)$, and the robot executes a prefix before replanning.

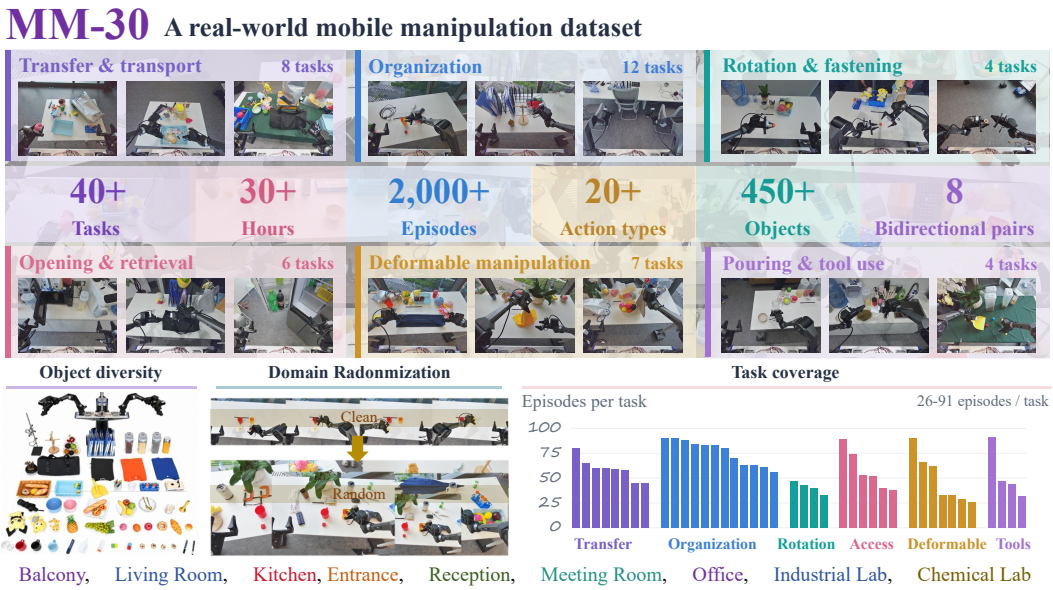


Figure 3: **MM-30: a multi-view dual-arm mobile manipulation dataset.** Representative tasks across six task families, object and domain diversity, and episode counts per task.

3.5 Why Predict Clean Actions? A Toy Experiment

As motivation for this choice, we run a small toy experiment rather than a full analysis. An x-head and a v-head share architecture, conditioning, data, noise, and the loss of Equation (3), and differ only in what the network emits. They denoise synthetic paired arm and body chunks that lie in a known 13-dimensional subspace and split a task vector between the two streams, so the Bayes-optimal denoiser is available in closed form; the generator, training protocol, and metrics are given in Section C. As Figure 2 illustrates, a v-head must reproduce noise directions that are absent from the action subspace (a). Accordingly, x-prediction keeps far less injected noise in its endpoint estimates, with $23\times$ lower variance at $t = 0.01$ (b), stays closer to the Bayes-optimal denoiser at high noise (c), and realizes the prescribed arm–body shares more accurately under few-step sampling (d). We use this toy setting only to motivate the design, and Section 5.4 evaluates clean-action prediction in the full policy, where the allocation is inferred from observations.

4 Pretraining Data and the MM-30 Dataset

Multi-embodiment corpus and data engine. Our corpus contains 5,166.2 hours and 400K+ episodes from 12 datasets, 51 subsets, and 17 embodiments, with 65K+ unique instructions. It combines eleven public sources, including ABC-130k, RoboCOIN, BEHAVIOR-1K, AgiBot World, InternData-A1, Galaxea Open-World, DROID, and HIW-500 (Allshire et al., 2026; Wu et al., 2025a; Li et al., 2023; Bu et al., 2025a; Tian et al., 2026; Jiang et al., 2025a; Khazatsky et al., 2024; BitRobot et al., 2026), with MM-30; real and simulated data account for 80% and 20% of the duration, and mobile-capable platforms for about 61%. Every source is mapped into a fixed 80-dimensional interface $\mathbf{a}_t = \mathbf{a}_t^L \oplus \mathbf{a}_t^R \oplus \mathbf{a}_t^B \in \mathbb{R}^{29+29+22}$ with a validity mask, whose arm blocks hold joint, end-effector, rotation, and gripper slots and whose body block holds base, torso, lift, and head controls. A per-source *frame contract* fixes units, action type, parent frame, tool center, rotation convention, and temporal semantics before conversion, and staged checks keep only the longest contiguous valid segment of each trajectory, so no unrecorded time jump enters an action chunk. Each source profile is sampled in proportion to $h_i^{0.4}$ for h_i cleaned hours, capped at 0.2 (Sections D and E).

MM-30. Public data rarely contain coordinated base and bimanual motion, so we collect **MM-30** on a platform with a HexFellow Trigger-A3 omnidirectional base and two AgileX PiPER-X 6-DoF arms (Figure 3). The base repositions the robot and the arms interact with objects, including bi-

manual handovers; these channels map to the base and the left and right arm slots of the interface. MM-30 contains 30+ hours and 2,000+ teleoperated episodes over 40+ tasks, 450+ objects, and nine indoor domains, spanning transport, organization, opening and retrieval, deformable manipulation, rotation and fastening, and pouring and tool use. Each task contains 26 to 91 episodes, recorded in domains that range from living rooms and kitchens to offices and industrial and chemistry laboratories. Because multi-view demonstrations benefit manipulation learning beyond viewpoint generalization (Cai et al., 2026), every task is recorded from multiple viewpoints under randomized layouts, distractors, and scene appearance.

5 EXPERIMENTS

We ask whether MM-ABC improves long-horizon mobile manipulation over generalist VLAs, world-action models, and mobile manipulation specialists (Section 5.2), whether it remains strong in fixed-base manipulation and under distribution shift (Section 5.3), which components drive these results (Section 5.4), and whether it transfers to a physical dual-arm mobile manipulator (Section 5.5).

5.1 SETUP

Benchmarks. **EBench** (Gao et al., 2026) evaluates one generalist checkpoint on 26 mobile pick-and-place, long-horizon, and dexterous tasks with success rate (SR) and a stage-wise progress score. **RoboCasa365** (Nasiriany et al., 2026) evaluates 50 kitchen tasks in held-out scenes, comprising 18 atomic-seen, 16 composite-seen, and 16 composite-unseen tasks. **ManiSkill-HAB** (Shukla et al., 2025) evaluates low-level whole-body control of a Fetch robot on the SetTable, TidyHouse, and PrepareGroceries suites. **LIBERO** (Liu et al., 2023) covers four fixed-base suites with 50 trials per task, and **LIBERO-Plus** (Fei et al., 2025) applies seven perturbation types over 10,030 episodes to the LIBERO-trained policy. We follow the official splits, success criteria, and evaluation protocol of every benchmark; with five real-world tasks and 300 physical trials, the evaluation spans mobile and fixed-base control, seen and unseen tasks, and distribution shifts.

Baselines and post-training. We compare with generalist VLAs (Kim et al., 2024; 2025; Liu et al., 2025; Black et al., 2024; Pertsch et al., 2025; Intelligence et al., 2025; Bu et al., 2025b; Bjorck et al., 2025; Zheng et al., 2026; Ye et al., 2026; Ma et al., 2026; Team et al., 2026a; Yang et al., 2026b), world-action models (WAMs) (Yuan et al., 2026; Kim et al., 2026b; Li et al., 2026d; Chen et al., 2026b), imitation policies (Ze et al., 2024; Zhao et al., 2023; Chi et al., 2025), and mobile manipulation specialists (Chen et al., 2026c; Lim et al., 2026; Fan et al., 2026; Zhu et al., 2026; Liu et al., 2026a), using results reported by the benchmark leaderboards or the cited papers under each benchmark's official evaluation protocol. Starting from the pretrained weights, MM-ABC is post-trained as one checkpoint for all 26 EBench tasks (100k steps, batch size 512), on the full RoboCasa365 target demonstrations (120k steps, batch size 512), as one policy per ManiSkill-HAB suite from head- and wrist-camera RGB (100k steps, batch size 64), and on the standard LIBERO demonstrations; implementation details are given in Section B.

5.2 MOBILE MANIPULATION

EBench and RoboCasa365. Table 1 compares MM-ABC with generalist VLAs and WAMs. On EBench, a single MM-ABC checkpoint achieves 44.71% SR and a progress score of 59, exceeding the strongest baseline, $\pi_{0.5}$, by 3.30 points in SR and 5 points in score, so the gains appear in intermediate progress as well as in full task completion; it also exceeds the world-action model Fast-WAM by 19.07 points in SR. On RoboCasa365, MM-ABC improves every split over ABot-M0.5 and raises the average by 7.0 points to 61.2%; the largest gain, 8.4 points, occurs on composite-seen tasks, which chain several interactions with navigation and require the base and arms to act in concert as reachability and viewpoint change. On composite-unseen tasks, which lie outside the benchmark's pretraining task set, MM-ABC exceeds GR00T-N1.5, Fast-WAM, and LingBot-VA by 17.0 to 18.2 points, so the gains are not limited to seen task compositions.

ManiSkill-HAB. Table 2 compares MM-ABC with imitation policies, VLAs, and mobile manipulation specialists, as reported by MoPA (Chen et al., 2026a). MM-ABC achieves the highest mean

Table 1: **Mobile manipulation on EBench and RoboCasa365.** Left: success rate (%) and stage-wise progress score of one generalist checkpoint on the 26 EBench tasks. Right: success rate (%) on the 50 RoboCasa365 target tasks, with the average weighted by split size (18/16/16). Best: shaded; second best: bold; gains are over the strongest baseline.

EBench	Success rate	Progress score
X-VLA	23.72	35
Fast-WAM	25.64	37
GigaBrain-0.7	33.27	46
InternVLA-A1.5	34.17	46
$\pi_{0.5}$	**41.41**	**54**
MM-ABC	**44.71**	**59**
gain	+3.30	+5

RoboCasa365	Atomic seen	Composite seen	Composite unseen	Avg.
GR00T-N1.5	60.6	35.0	33.3	43.7
Fast-WAM	59.1	36.4	33.2	43.5
LingBot-VA	63.5	37.3	32.1	45.1
ABot-M0.5	**70.6**	**44.3**	**45.6**	**54.2**
MM-ABC	**78.4**	**52.7**	**50.3**	**61.2**
gain	+7.8	+8.4	+4.7	+7.0

Table 2: **ManiSkill-HAB** success rate (%). Left: SetTable, where Pick, Place, and Artic. average the two picking, two placing, and three articulated-object skills (per-skill results in Table 6). Right: TidyHouse and PrepareGroceries, where pick and place rates are averaged over nine object categories. *Depth input. Best: shaded; second best: bold.

SetTable	Pick	Place	Artic.	Mean
RGB-D input				
DP3*	10.0	31.5	22.7	21.6
AC-DiT*	34.7	25.3	89.8	55.6
GeoHAT*	**75.8**	69.0	88.3	79.2
InCoM*	71.8	**83.3**	92.1	**83.8**
RGB input				
ACT	28.0	10.9	29.2	23.6
DP	21.0	48.7	20.8	28.8
RDT-1B	11.4	25.4	75.6	42.9
π_0	26.6	52.9	84.9	59.1
AnchorVLA	33.6	64.1	94.5	64.0
MobileWAM	46.0	64.2	**96.8**	73.0
MM-ABC	**84.0**	**76.2**	**94.8**	**86.4**

TidyHouse	Pick	Place	Mean
DP3*	0.0	61.0	30.5
ACT	2.2	31.6	16.9
DP	0.0	30.3	15.2
GeoHAT*	**30.3**	**73.3**	**51.8**
InCoM*	16.7	**78.9**	47.8
MM-ABC	**71.5**	68.8	**70.2**
PrepareGroceries	Pick	Place	Mean
DP3*	0.0	31.3	15.7
ACT	2.0	27.5	14.8
DP	0.4	17.7	9.1
GeoHAT*	**19.7**	60.7	40.2
InCoM*	15.0	**65.9**	**40.5**
MM-ABC	**64.4**	**66.8**	**65.6**

success among the compared methods on all three suites from RGB input alone, whereas DP3, AC-DiT, GeoHAT, and InCoM additionally use depth. The gains concentrate in picking: MM-ABC reaches 84.0% pick success on SetTable against 75.8% for GeoHAT, and on TidyHouse and PrepareGroceries it raises picking from at most 30.3% and 19.7% to 71.5% and 64.4%. Picking after base motion requires acquiring objects of varied geometry, where multilevel perception and coupled manipulation–body prediction matter most; it lifts the TidyHouse and PrepareGroceries means by 18.4 and 25.1 points. Averaged over the three suites, MM-ABC reaches 74.1%, 16.7 points above InCoM, the strongest baseline evaluated on all three.

5.3 Fixed-Base Manipulation and Robustness

With the inactive body stream masked, the same architecture transfers to fixed-base control, where we compare with VLAs and WAMs (Table 3). MM-ABC achieves 99.1% mean success on LIBERO, 0.5 points above the strongest baseline, ABot-M0, with the highest success on Spatial, a tie for the highest on Object, and 98.4% on the long-horizon suite, 1.8 points above ABot-M0. Evaluated on LIBERO-Plus without further training, the same policy reaches an 82.8% total, above Cosmos-Policy (82.2%) and ABot-M0 (80.5%), and it also exceeds OpenVLA-OFT, GR00T-N1.6, and π_0 fine-tuned on LIBERO-Plus perturbed demonstrations (79.6%, 79.4%, and 67.4%; Table 8). MM-ABC attains the highest language (88.9%) and background (96.1%) scores, consistent with conditioning actions on shared multilevel visual–language features. High LIBERO success does not

Table 3: **Fixed-base manipulation on LIBERO and robustness on LIBERO-Plus.** Success rate (%). LIBERO uses 50 trials per task; LIBERO-Plus evaluates LIBERO-trained policies on all 10,030 perturbed episodes, with the total weighted by episode count. Best: shaded; second best: bold.

	LIBERO					LIBERO-Plus (no perturbation training)							
Method	Spa.	Obj.	Goal	Long	Avg.	Cam.	Robot	Lang.	Light	Back.	Noise	Layout	Total
OpenVLA	84.7	88.4	79.2	53.7	76.5	0.8	3.5	23.0	8.1	34.8	15.2	28.5	15.6
π_0-FAST	96.4	96.8	88.6	60.2	85.5	65.1	21.6	61.0	73.2	73.2	74.4	68.8	61.6
π_0	96.8	98.8	95.8	85.2	94.2	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
UniVLA	96.5	96.8	95.6	92.0	95.2	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
OpenVLA-OFT	97.6	98.4	97.9	94.5	97.1	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
GR00T-N1.6	**99.3**	99.2	98.4	92.9	97.5	20.9	40.2	35.0	65.4	76.3	27.8	51.3	42.8
Fast-WAM	98.2	**100.0**	97.0	95.2	97.6	16.4	44.5	68.9	78.2	53.7	37.7	60.7	50.0
X-VLA	98.2	98.6	97.8	97.6	98.1	23.4	**89.7**	75.7	88.2	**96.0**	62.7	71.8	70.5
LingBot-VA	98.5	99.6	97.2	**98.5**	98.5	40.9	**83.0**	**86.4**	82.3	53.1	64.4	76.2	69.5
Cosmos-Policy	98.1	**100.0**	98.2	97.6	98.5	**75.8**	63.3	81.7	**96.5**	88.9	**92.7**	**82.2**	**82.2**
ABot-M0	98.8	**99.8**	**99.0**	96.6	**98.6**	60.4	67.9	**86.4**	**96.2**	91.6	**86.4**	**82.6**	80.5
MM-ABC	**99.4**	**100.0**	**98.6**	**98.4**	**99.1**	**73.6**	66.5	**88.9**	94.9	**96.1**	**86.4**	80.5	**82.8**

Table 4: **Component ablation on the 16 RoboCasa365 composite-seen tasks.** All variants are trained from scratch without robot-data pretraining under a common budget; the top block lists the design of each variant. Success rate (%); best per task: shaded.

Variant	Interleaved	Last layer	No future	v-pred	**MM-ABC**
Prediction target	x	x	x	v	x
Future supervision	✓	✓	✗	✓	✓
VLM conditioning	interleaved	final layer	multilevel	multilevel	multilevel
DeliverStraw	0.0	**0.5**	**0.5**	**0.5**	0.0
GetToastedBread	0.0	**1.0**	0.0	**1.0**	0.5
KettleBoiling	33.0	29.5	35.5	41.5	**49.0**
LoadDishwasher	18.5	28.0	22.0	27.0	**33.0**
PackIdentical	0.5	1.5	1.5	**7.0**	6.0
PreSoakPan	29.5	50.5	48.5	52.0	**67.5**
PrepareCoffee	13.0	17.5	15.0	**22.0**	17.5
RinseSinkBasin	53.5	54.5	**62.0**	59.0	59.0
ScrubCuttingBoard	24.5	**47.0**	19.5	20.0	28.5
SearingMeat	10.5	3.5	8.5	11.5	**23.0**
SetUpCuttingStation	21.5	14.0	18.0	28.0	**31.0**
StackBowlsCabinet	57.5	69.0	63.5	67.5	**69.5**
SteamInMicrowave	24.5	12.5	10.0	27.0	**29.5**
StirVegetables	19.0	5.5	12.5	15.5	**24.5**
StoreLeftovers	11.5	28.5	28.5	34.5	**48.0**
WashLettuce	35.5	46.0	**53.0**	**53.0**	38.0
Average	22.0	25.6	24.9	**29.2**	**32.8**

by itself imply robustness: GR00T-N1.6 reaches 97.5% on LIBERO but 42.8% on LIBERO-Plus, whereas MM-ABC ranks first on both benchmarks.

5.4 Ablation Study

To isolate the components, we train five variants from scratch on the 16 RoboCasa365 composite-seen tasks, without robot-data pretraining and with a common budget of 120k steps at batch size 64 (Table 4). The full model achieves the highest average success, 32.8%, and is best or tied for best on 9 of the 16 tasks. Replacing clean-action prediction with velocity prediction, while keeping future supervision, lowers the average to 29.2%, in line with the toy experiment of Section 3.5; velocity prediction is the strongest alternative, yet the full model exceeds it on 10 of the 16 tasks. Removing future supervision while keeping x-prediction causes a larger drop, to 24.9%, showing that geometric anticipation improves the shared representation even though the future branch never reaches the action tokens. Conditioning only on the final VLM layer or on interleaved VLM layers yields 25.6% and 22.0%, favoring sparse injection of intermediate features. The margins are largest



Figure 4: **Real-world tasks.** Representative stages and step-level instructions of Birthday Party Setup (top), Office Folder Arrangement and Kitchen Work (middle), and Industrial Parts Organization and Chemistry Lab Operation (bottom); frames progress from left to right.

Table 5: **Real-world success rate (%)** over 20 trials per task and method; a trial succeeds only when all five stages are completed. Best: shaded; second best: bold; gains are over $\pi_{0.5}$.

Method	Birthday Party	Office Folder	Kitchen Work	Industrial Parts	Chemistry Lab	Mean
StarVLA-GR00T	50	55	55	60	50	54
$\pi_{0.5}$	**65**	**80**	**70**	**75**	**65**	**71**
MM-ABC	**75**	**90**	**90**	**75**	**85**	**83**
gain	+10	+10	+20	0	+20	+12

on PreSoakPan, StoreLeftovers, and SearingMeat, where the full model exceeds every variant by 11.5 to 15.5 points.

5.5 Real-World Mobile Manipulation

We deploy MM-ABC on the MM-30 platform for five long-horizon tasks in household, office, workcell, and laboratory scenes (Figure 4). Each task has five stages specified by a single instruction and requires base repositioning between interactions (Section H). All methods are fine-tuned on the same 200 teleoperated demonstrations per task and evaluated over 20 trials per task. As shown in Table 5, MM-ABC achieves 83% mean success, 12 points above $\pi_{0.5}$ and 29 points above StarVLA-GR00T, and it is best or tied for best on every task. The largest margins over $\pi_{0.5}$, 20 points each, occur on Kitchen Work and Chemistry Lab Operation, which combine base repositioning with pouring, lid placement, cabinet-door closing, and stirring.

6 Conclusion

We presented MM-ABC, a vision–language–action model that treats mobile manipulation as manipulation over a workspace the policy itself reconfigures. MM-APT couples manipulation and body streams through near–far joint attention; in the ablation, sparse multilevel conditioning, future-geometry supervision, and clean-action prediction each contribute, the last also motivated by a toy experiment. Pretrained on a 5,000+-hour corpus that includes the new MM-30 dataset, MM-ABC leads the compared methods on EBench, RoboCasa365, and ManiSkill-HAB, performs strongly on LIBERO and LIBERO-Plus, and reaches 83% mean success on long-horizon real-world tasks. Because the future branch is removed at deployment, these gains come without additional inference cost from world imagination. Extending the same interface to dexterous hands and multi-room tasks is a natural next step, and we will release the code, model weights, and MM-30 to support research on generalist mobile manipulation across embodiments.

AI USE STATEMENT

In this work, we used generative AI tools (large language model assistants) to aid and polish the writing and to draft parts of the paper, including turning the authors' notes and results into English text, restructuring sections, and tightening wording. We also used them for retrieval and discovery: searching for related work, collecting the reported numbers of published baselines from papers and public leaderboards, and checking bibliographic metadata against arXiv and publisher records. In addition, they helped format references and assemble figures and tables from our recorded results, for example by re-plotting logged metrics of the toy experiment in Section 3.5 and recomputing table averages and episode-weighted totals. We have not used generative AI tools to generate synthetic datasets, to formulate or prove mathematical claims, to design the model, to collect data, or to run the training and evaluation experiments; these were carried out by the authors. We have reviewed all AI-assisted work: the authors checked every reported number against the experiment records, verified every citation against its source, and edited all generated text. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

MM-30 was collected by trained operators teleoperating our robot in laboratory and staged indoor environments; the recordings contain robot observations and actions and are not intended to capture identifiable people. All public datasets are used under their respective licenses, several of which restrict use to non-commercial research, and we follow these terms for training and release. Real-world experiments were supervised by an operator with emergency-stop access. We do not foresee direct harmful applications beyond the general risks of deploying autonomous mobile manipulators, which call for supervised operation and safety validation before use around people.

REPRODUCIBILITY STATEMENT

Section 3 and Section B specify the architecture, training objective, and hyperparameters of MM-ABC. Section C gives the generator, architecture, training protocol, and metric definitions of the toy experiment. Sections D and E list all pretraining sources and describe the canonical interface, cleaning pipeline, and sampling mixture, and Section F describes the MM-30 platform and collection protocol. Section G reports per-benchmark post-training budgets, protocols, and full per-task results, and Section H describes the real-world tasks and evaluation. We will release the code, model weights, and the MM-30 dataset upon publication.

REFERENCES

Arthur Allshire, Himanshu Gaurav Singh, Ritvik Singh, Adam Rashid, Hongsuk Choi, David McAllister, Justin Yu, Yiyuan Chen, Huang Huang, Pieter Abbeel, et al. Scalable behavior cloning with open data, training, and evaluation. *arXiv preprint arXiv:2606.27375*, 2026.

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

BitRobot, Unitree, and Hugging Face. Hiw-500: Humanoids in-the-wild dataset for robot learning. https://bitrobot-foundation.github.io/humanoids-in-the-wild-500-hours/, 2026.

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

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

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. *arXiv preprint arXiv:2307.15818*, 2023.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025a.

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

Boyang Cai, Qiwei Liang, Jiawei Li, Shihang Weng, Zhaoxin Zhang, Tao Lin, Xiangyu Chen, Wenjie Zhang, Jiaqi Mao, Weisheng Xu, et al. Beyond viewpoint generalization: What multi-view demonstrations offer and how to synthesize them for robot manipulation? *arXiv preprint arXiv:2603.26757*, 2026.

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

Guangyu Chen, Qiwei Liang, Shaolong Zhu, Tianxing Chen, Zikuan Xiao, Yifan Xie, Lingfeng Zhang, Ping Luo, Renjing Xu, and Wenbo Ding. Mopa: Coordinated mobile manipulation via subsystem-specific perception alignment. *arXiv preprint arXiv:2609.12081*, 2026a.

Ronghan Chen, Yandan Yang, Zuojin Tang, Dongjie Huo, Tong Lin, Haoning Wu, Haoyun Liu, Yuzhi Chen, Lulu Zheng, Botai Yuan, et al. Abot-m0. 5: Unified mobility-and-manipulation world action model. *arXiv preprint arXiv:2607.00678*, 2026b.

Sixiang Chen, Jiaming Liu, Siyuan Qian, Han Jiang, Zhuoyang Liu, Chenyang Gu, Xiaoqi Li, Chengkai Hou, Pengwei Wang, Zhongyuan Wang, et al. Ac-dit: Adaptive coordination diffusion transformer for mobile manipulation. *Advances in Neural Information Processing Systems*, 38:64008–64036, 2026c.

Yizhi Chen, Zhanxiang Cao, Xinyi Peng, Yixiao Zheng, Xiaxi Si, Yiheng Li, Liyun Yan, Keqi Zhu, Xueyun Chen, Shengcheng Fu, et al. Geoalign: Beyond semantics with state-guided spatial alignment in vla models. *arXiv preprint arXiv:2606.03240*, 2026d.

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

Xingyu Ding, Yuzhong Zhao, Yang Wu, Chaoyang Zhao, Chunhai Zhao, Yifan Zhang, and Jian Cheng. Mind-vla: Instruction-aware spatial representation alignment for vision-language-action models. *arXiv preprint arXiv:2608.04633*, 2026a.

Xingyu Ding, Yuzhong Zhao, Chunhai Zhao, Yinghuan Shi, Chaoyang Zhao, and Yifan Zhang. Temporal forcing: 4d representation alignment for vision-language-action models. *arXiv preprint arXiv:2608.30643*, 2026b.

Kiana Ehsani, Tanmay Gupta, Rose Hendrix, Jordi Salvador, Luca Weihs, Kuo-Hao Zeng, Kunal Pratap Singh, Yejin Kim, Winson Han, Alvaro Herrasti, et al. Spoc: Imitating shortest paths in simulation enables effective navigation and manipulation in the real world. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16238–16250. IEEE, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Zehua Fan, Junjie He, Wenxuan Song, Xi Wang, Wenqi Lyu, Linge Zhao, Fuhao Li, Zihan You, Yifei Yang, Kaiming Xu, et al. Mobilewam: Bridging world action models to mobile manipulation with chain-of-foresight. *arXiv preprint arXiv:2608.04657*, 2026.

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

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

Ning Gao, Jinliang Zheng, Xing Gao, Haoxiang Ma, Hanqing Wang, Yukai Wang, Jiantong Chen, Zanxin Chen, Shujie Zhang, Mingda Jia, et al. Ebench: Elemental diagnosis of generalist mobile manipulation policies. *arXiv preprint arXiv:2606.18239*, 2026.

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

Minghao Guo, Meng Cao, Jiachen Tao, Rongtao Xu, Yan Yan, Xiaodan Liang, Ivan Laptev, and Xiaojun Chang. Glad: Geometric latent distillation for vision-language-action models. *arXiv preprint arXiv:2512.09619*, 2025.

Jisang Han, Seonghu Jeon, Jaewoo Jung, René Zurbrügg, Honggyu An, Tifanny Portela, Marco Hutter, Marc Pollefeys, Seungryong Kim, and Sunghwan Hong. Geometric action model for robot policy learning. *arXiv preprint arXiv:2606.17046*, 2026.

Jiaheng Hu, Peter Stone, and Roberto Martín-Martín. Causal policy gradient for whole-body mobile manipulation. *arXiv preprint arXiv:2305.04866*, 2023.

Chia-Yu Hung, Qi Sun, Pengfei Hong, Amir Zadeh, Chuan Li, U Tan, Navonil Majumder, Soujanya Poria, et al. Nora: A small open-sourced generalist vision language action model for embodied tasks. *arXiv preprint arXiv:2504.19854*, 2025.

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

Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea open-world dataset and g0 dual-system vla model. *arXiv preprint arXiv:2509.00576*, 2025a.

Yunfan Jiang, Ruohan Zhang, Josiah Wong, Chen Wang, Yanjie Ze, Hang Yin, Cem Gokmen, Shuran Song, Jiajun Wu, and Li Fei-Fei. Behavior robot suite: Streamlining real-world whole-body manipulation for everyday household activities. *arXiv preprint arXiv:2503.05652*, 2025b.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

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

Oussama Khatib. Mobile manipulation: The robotic assistant. *Robotics and Autonomous Systems*, 26(2-3):175–183, 1999.

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

Dongyoung Kim, Huiwon Jang, Myungkyu Koo, Suhyeok Jang, Taeyoung Kim, Beomjun Kim, Byungjun Yoon, Changsung Jang, Daewon Choi, Dongsu Han, et al. Rldx-1 technical report. *arXiv preprint arXiv:2605.03269*, 2026a.

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

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026b.

Prajwal Koirala and Mark Campbell. Vgfm: Expressive robot policies via dense value guidance in flow matching. *arXiv preprint arXiv:2609.14261*, 2026.

Iok Tong Lei, Ying Jie Yap, Wei Huang, Qingchen Xie, Qianzhi Li, Yujie Zhang, Xiaolong Liu, and Zhidong Deng. Teaching tiny vla models where to look and how to move. *arXiv e-prints*, pp. arXiv–2607, 2026.

Andrew Ting Yan Li, Zhuo Li, Zhelin Yang, Zhipeng Dong, Quentin Rouxel, and Fei Chen. Reasoning without inference cost: Latent semantic scaffolding for robot vla policies. *arXiv preprint arXiv:2609.04893*, 2026a.

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

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pp. 80–93. PMLR, 2023.

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. In *International Conference on Learning Representations*, volume 2026, pp. 132324–132345, 2026c.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026d.

Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025a.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Xiaoqi Li, Liang Heng, Jiaming Liu, Yan Shen, Chenyang Gu, Zhuoyang Liu, Hao Chen, Nuowei Han, Renrui Zhang, Hao Tang, et al. 3ds-vla: A 3d spatial-aware vision language action model for robust multi-task manipulation. In *9th Annual Conference on Robot Learning*, 2025b.

Ying Li, Xiaobao Wei, Jiajun Cao, Hao Wang, Xiaowei Chi, Chengyu Bai, Qianpu Sun, Jiajun Li, Xiaojie Zhang, Peidong Jia, et al. Wam4d: Fast 4d world action model via spatial register tokens. *arXiv preprint arXiv:2606.14048*, 2026e.

Yixuan Li, Yuhui Chen, Mingcai Zhou, Haoran Li, Zhengtao Zhang, and Dongbin Zhao. Qdepth-vla: Quantized depth prediction as auxiliary supervision for vision-language-action models. *arXiv preprint arXiv:2510.14836*, 2025c.

Zhe Li, Zhenzhe Zhang, Yangyang Wei, Wenjie Zhang, Xichen Yuan, Peiyuan Zhi, Gen Li, Xinying Guo, Fengjie Gao, Jianfei Yang, et al. ω-0: A latent predictive world action model for concurrent humanoid loco-manipulation. *arXiv preprint arXiv:2608.06375*, 2026f.

Qiwei Liang, Boyang Cai, Rongyi He, Hui Li, Tao Teng, Haihan Duan, Changxin Huang, and Runhao Zeng. Whole-body coordination for dynamic object grasping with legged manipulators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18434–18442, 2026a.

Qiwei Liang, Boyang Cai, Minghao Lai, Sitong Zhuang, Tao Lin, Yan Qin, Yixuan Ye, Jiaming Liang, and Renjing Xu. Bootstrap dynamic-aware 3d visual representation for scalable robot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13419–13429, 2026b.

Jia Syuen Lim, Zhizhen Zhang, Peter Bohm, Brendan Tidd, Zi Huang, and Yadan Luo. Anchorvla: Anchored diffusion for efficient end-to-end mobile manipulation. *arXiv preprint arXiv:2604.01567*, 2026.

Tao Lin, Gen Li, Yilei Zhong, Yanwen Zou, Yuxin Du, Jiting Liu, Encheng Gu, and Bo Zhao. Evo-0: Vision-language-action model with implicit spatial understanding. *arXiv preprint arXiv:2507.00416*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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

Jiahao Liu, Cui Wenbo, Zhongpu Xia, Yongliang Wang, Haoran Li, and Dongbin Zhao. Incom: Intent-driven perception and structured coordination for mobile manipulation. *arXiv preprint arXiv:2602.23024*, 2026a.

Jiaming Liu, Qingpo Wuwu, Nuowei Han, Hao Chen, Zhuoyang Liu, Fan Fei, Yueru Jia, Chenyang Gu, Yandong Guo, Boxin Shi, et al. Lift3d-vla: Lifting vla models to 3d geometry and dynamics-aware manipulation. *arXiv preprint arXiv:2607.06564*, 2026b.

Mengya Liu, Baoxiong Jia, Jiangyong Huang, Jingze Zhang, and Siyuan Huang. Lara: Latent action representation alignment for vision-language-action models. *arXiv preprint arXiv:2606.07100*, 2026c.

Mohan Liu, Zhihao Gu, Xuanyu Chen, Haitian Zhang, Kaimin Mao, Yan Wu, Wei-Yun Yau, and Lin Wang. Vistavla: Geometry-and semantic-aware 3d gaussian-grounded vla for robotic manipulation. *arXiv preprint arXiv:2607.12356*, 2026d.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 29982–30009, 2025.

Zonghe Liu, Shanyuan Jie, Xiaoquan Sun, Chen Cao, Zetian Xu, Zongsheng Liu, and Jiayu Chen. Sam3d-guided object-centric representation alignment for vision-language-action models. *arXiv preprint arXiv:2607.25912*, 2026e.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026.

Yulin Luo, Hao Chen, Zhuangzhe Wu, Bowen Sui, Jiaming Liu, Chenyang Gu, Zhuoyang Liu, Qiuxuan Feng, Jiale Yu, Shuo Gu, et al. Look before acting: Enhancing vision foundation representations for vision-language-action models. *arXiv preprint arXiv:2603.15618*, 2026.

Haoxiang Ma, Junhao Cai, Xiaoxu Xu, Hao Li, Yuyin Yang, Yang Tian, Jiafei Cao, Hongrui Zhu, Zherui Qiu, Yuqiang Yang, et al. Internvla-a1. 5: Unifying understanding, latent foresight, and action for compositional generalization. *arXiv preprint arXiv:2607.04988*, 2026.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems*, 37:23464–23487, 2024.

Shangchen Miao, Ningya Feng, Jialong Wu, Ye Lin, Xu He, Dong Li, and Mingsheng Long. Jepa-vla: Video predictive embedding is needed for vla models. *arXiv preprint arXiv:2602.11832*, 2026.

Soroush Nasiriany, Sep Nasiriany, Abhiram Maddukuri, and Yuke Zhu. Robocasa365: A large-scale simulation framework for training and benchmarking generalist robots. In *International Conference on Learning Representations*, volume 2026, pp. 98643–98667, 2026.

Chaoyi Pan, Giridharan Anantharaman, Nai-Chieh Huang, Claire Jin, Daniel Pfrommer, Chenyang Yuan, Frank Permenter, Guannan Qu, Nicholas Boffi, Guanya Shi, et al. Much ado about noising: Dispelling the myths of generative robotic control. In *International Conference on Learning Representations*, volume 2026, pp. 90575–90614, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

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

RealSource. Realsource world: A large-scale real-world dual-arm manipulation dataset. https://huggingface.co/datasets/RealSourceData/RealSource-World, 2025.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Jinghuan Shang, Karl Schmeckpeper, Brandon B May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. *arXiv preprint arXiv:2407.20179*, 2024.

Jiaxin Shi, Xidong Zhang, Fucai Zhu, Zhe Li, Siyu Zhu, and Weihao Yuan. 3dthinkvla: Endowing vision-language-action models with latent 3d priors via 3d-thinking-guided co-training. *arXiv preprint arXiv:2606.04436*, 2026.

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

Arth Shukla, Stone Tao, and Hao Su. Maniskill-hab: A benchmark for low-level manipulation in home rearrangement tasks. In *International Conference on Learning Representations*, volume 2025, pp. 15288–15317, 2025.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

Davood Soleymanzadeh, Kaidi Zhang, Zhiyuan Zhang, Bihao Zhang, Xiao Liang, Yu She, and Minghui Zheng. Phr-vla: Planning horizon reasoning for vision-language-action models. *arXiv preprint arXiv:2608.27609*, 2026.

Wenxuan Song, Ziyang Zhou, Han Zhao, Jiayi Chen, Pengxiang Ding, Haodong Yan, Yuxin Huang, Feilong Tang, Donglin Wang, and Haoang Li. Reconvla: Reconstructive vision-language-action model as effective robot perceiver. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18549–18557, 2026.

Guoheng Sun, Tingting Du, Kaixi Feng, Chenxiang Luo, Xingguo Ding, Zheyu Shen, Ziyao Wang, Yexiao He, and Ang Li. Rocket: Residual-oriented multi-layer alignment for spatially-aware vision-language-action models. *arXiv preprint arXiv:2602.17951*, 2026a.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. Vla-jepa: Enhancing vision-language-action model with latent world model. In *European Conference on Computer Vision*, pp. 478–497. Springer, 2026b.

Lin Sun, Bin Xie, Yingfei Liu, Hao Shi, Tiancai Wang, and Jiale Cao. Geovla: Empowering 3d representations in vision-language-action models. *arXiv preprint arXiv:2508.09071*, 2025.

AgiBot World Team. Agibot world 2026. https://huggingface.co/datasets/agibot-world/AgiBotWorld2026, 2026.

GigaBrain Team, Angen Ye, Axiang Sun, Can Jin, Chenxi Cheng, Chong Shi, Dengke Shang, Dingqian Zhang, Guan Huang, Guangqiang Wang, et al. Gigabrain-0.7: Scaling embodied foundation models to emergent capabilities with a three-system architecture. *arXiv preprint arXiv:2608.15875*, 2026a.

Xiaomi Robotics Team, Jun Guo, Piaopiao Jin, Jason Li, Peiyan Li, Yingyan Li, Futeng Liu, Wanli Peng, Optimus Qin, Yifei Su, et al. Xiaomi-robotics-1: Scaling vision-language-action models with over 100k hours of real-world trajectories. *arXiv preprint arXiv:2607.15330*, 2026b.

Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, et al. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 976–985, 2026.

Ruisen Tu, Arth Shukla, Sohyun Yoo, Xuanlin Li, Junxi Li, Jianwen Xie, Hao Su, and Zhuowen Tu. Sg-vla: Learning spatially-grounded vision-language-action models for mobile manipulation. *arXiv preprint arXiv:2603.22760*, 2026.

Hao Wang, Xiaobao Wei, Jingyang He, Chengyu Bai, Chun-Kai Fan, Jiajun Cao, Jintao Chen, Ying Li, Shanyu Rong, Ming Lu, et al. Vega: Visual encoder grounding alignment for spatially-aware vision-language-action models. *arXiv preprint arXiv:2605.10485*, 2026a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-ω. *arXiv preprint arXiv:2605.15195*, 2026b.

Qiuyue Wang, Mingsheng Li, Jian Guan, Jinhui Ye, Sicheng Xie, Yitao Liu, Junhao Chen, Zhixuan Liang, Jie Zhang, Xintong Hu, et al. Qwen-vla: Unifying vision-language-action modeling across tasks, environments, and robot embodiments. *arXiv preprint arXiv:2605.30280*, 2026c.

Xianbang Wang, Hanhong Zhao, Yiyang Lu, Kangyang Zhou, Linrui Ma, and Kaiming He. Minit2i: A minimalist baseline for text-to-image generation, 2026d. URL https://peppaking8.github.io/#/post/minit2i.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, volume 2024, pp. 10641–10662, 2024.

Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, Bowen Yang, Zhe Li, Kai Zhu, Hongyu Wu, Yiheng Liu, et al. Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025a.

Wei Wu, Fangjing Wang, Fan Lu, He Sun, Shi Liu, Yunnan Wang, Yibin Yan, Yong Wang, Shuailei Ma, Xinyang Wang, et al. From foundation to application: Improving vla models in practice. *arXiv preprint arXiv:2607.06403*, 2026.

Zhenyu Wu, Yuheng Zhou, Xiuwei Xu, Ziwei Wang, and Haibin Yan. Momanipvla: Transferring vision-language-action models for general mobile manipulation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1714–1723. IEEE, 2025b.

Lishan Yang, Wenxuan Song, Xi Wang, Pingyue Sheng, Zheng Fang, Ziyang Zhou, Junjie He, Haodong Yan, Jiayi Chen, Nan Sun, et al. 4d-wam: Infusing spatiotemporal awareness into world action models through trajectory fields. *arXiv preprint arXiv:2608.08023*, 2026a.

Yandan Yang, Shuang Zeng, Tong Lin, Xinyuan Chang, Dekang Qi, Junjin Xiao, Haoyun Liu, Ronghan Chen, Yuzhi Chen, Dongjie Huo, et al. Abot-m0: Vla foundation model for robotic manipulation with action manifold learning. *arXiv preprint arXiv:2602.11236*, 2026b.

Zheng Yang, Wenjie Zhang, Xiangyu Chen, Wenxuan Song, Xianpeng Wang, Yihang Kang, Jiawen Wen, Wen Chen, Lujia Wang, Renjing Xu, et al. Dreamtrajectory: Trajectory-guided action generation with world model alignment for mobile manipulation. *arXiv preprint arXiv:2608.01381*, 2026c.

Jinhui Ye, Ning Gao, Senqiao Yang, Jinliang Zheng, Zixuan Wang, Yuxin Chen, Pengguang Chen, Yilun Chen, Shu Liu, and Jiaya Jia. Starvla-α: Reducing complexity in vision-language-action systems. *arXiv preprint arXiv:2604.11757*, 2026.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

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

He Zhang, Lingzhu Xiang, Haitao Lin, Zeyu Huang, Minghui Wang, Dingyan Zhong, Yubo Dong, Yihao Wu, Yongming Rao, Dongsheng Zhang, et al. Hy-embodied-0.5-vla: From vision-language-action models to a real-world robot learning stack. *arXiv preprint arXiv:2606.14409*, 2026a.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, et al. 4d-vla: Spatiotemporal vision-language-action pretraining with cross-scene calibration. *Advances in Neural Information Processing Systems*, 38: 33914–33937, 2026b.

Jianjun Zhang, Jian Zhu, Taiyi Su, Chong Ma, Zitai Huang, Yi Xu, and Hanli Wang. Learning 4d geometric priors for inference-efficient world action models. *arXiv preprint arXiv:2607.05468*, 2026c.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *Advances in Neural Information Processing Systems*, 38: 24195–24228, 2026d.

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

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *International Conference on Learning Representations*, volume 2026, pp. 60580–60606, 2026.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, et al. Flare: Robot learning with implicit world modeling. *arXiv preprint arXiv:2505.15659*, 2025.

Xiangyu Zhu, Renjun Wu, Luzhou Ge, Jinyan Liu, and Xuesong Li. Geohat: Geometry-adaptive hybrid action transformer for mobile manipulation. *arXiv preprint arXiv:2606.13394*, 2026.

Lifeng Zhuo, Wendi Chen, Han Xue, Shirun Tang, Jun Lv, Cewu Lu, and Chuan Wen. Fa-rdp: A frequency-adaptive reactive diffusion policy for contact-rich manipulation. *arXiv preprint arXiv:2607.28596*, 2026.

A Extended Related Work

Mobile manipulation. Mobile manipulation couples base motion with object interaction, extending robot control beyond a fixed workspace. Early model-based whole-body controllers (Khatib, 1999) were followed by reinforcement and imitation learning approaches (Hu et al., 2023; Ehsani et al., 2024; Fu et al., 2024; Jiang et al., 2025b), including whole-body coordination for dynamic grasping with legged manipulators (Liang et al., 2026a). Generalist policies support discrete action decoding (Brohan et al., 2023; Kim et al., 2024), diffusion (Chi et al., 2025), and flow matching (Black et al., 2024), with mobile extensions addressing unfamiliar environments (Intelligence et al., 2025), trajectory optimization (Wu et al., 2025b), and joint mobility–manipulation modeling (Chen et al., 2026b). Xiaomi-Robotics-1 scales UMI pretraining beyond 100K hours before aligning the policy with robot embodiments and instructions (Team et al., 2026b), and LingBot-VLA 2.0 combines broader pretraining with whole-body action interfaces and predictive semantic and geometric supervision (Wu et al., 2026). Beyond scale, architecture determines how perceptual features and action streams interact. VITRA concatenates an extracted VLM cognition feature with state and noisy action tokens in a DiT (Li et al., 2025a), Qwen-VLA jointly processes VLM hidden states and noisy actions through self-attention (Wang et al., 2026c), and RLDX-1 extends this approach to modality-specific cognition, action, and physical-signal streams (Kim et al., 2026a). For coordination across body subsystems, AC-DiT conditions manipulation on a mobility prior (Chen et al., 2026c), InCoM and GeoHAT introduce structured arm–base coordination (Liu et al., 2026a; Zhu et al., 2026), and MoPA aligns perception separately for each subsystem (Chen et al., 2026a). DreamTrajectory guides whole-body actions with end-effector trajectories and refines candidates through a trajectory world model at test time (Yang et al., 2026c), and ω-0 combines future observation embeddings with controller-compatible action latents for humanoid loco-manipulation (Li et al., 2026f).

Spatial representations and perceptual interfaces. Spatially grounded policies incorporate three-dimensional scene representations (Shridhar et al., 2023; Ke et al., 2024), point-cloud conditioning (Li et al., 2026b; Sun et al., 2025), and spatially informed VLA architectures (Qu et al., 2025; Li et al., 2025b). Further work models temporal structure (Zhang et al., 2026b; Liu et al., 2026b), learns dynamic-aware 3D representations for scalable robot learning (Liang et al., 2026b), or integrates geometry with semantic features (Lin et al., 2025; Chen et al., 2026d; Liu et al., 2026d; Tu et al., 2026). The interface to the action expert also matters: policies use cross-attention (Bjorck et al., 2025; Shukor et al., 2025), final-layer features (Yang et al., 2026b), or predictive embeddings (Miao et al., 2026), and DeepStack, Qwen3-VL, and DeepVision-VLA motivate conditioning across network depth (Meng et al., 2024; Bai et al., 2025; Luo et al., 2026).

Representation alignment and predictive supervision. Teacher supervision builds on feature alignment and distillation (Yu et al., 2024; Shang et al., 2024). Spatial Forcing (Li et al., 2026c) and GLaD (Guo et al., 2025) supervise internal policy features with geometry, ROCKET extends alignment across layers (Sun et al., 2026a), and VEGA targets the visual encoder (Wang et al., 2026a). Supervision ranges from object geometry (Ding et al., 2026a; Liu et al., 2026e), depth (Li et al., 2025c), and attended-region reconstruction (Song et al., 2026) to spatial and semantic knowledge (Shi et al., 2026; Lei et al., 2026; Li et al., 2026a), latent actions (Liu et al., 2026c), and temporal geometry (Ding et al., 2026b). Predictive supervision extends these ideas to future images (Wu et al., 2024), video embeddings (Assran et al., 2025), and latent world representations (Zheng et al., 2025; Sun et al., 2026b; Zhang et al., 2026d), including geometric evolution (Han et al., 2026; Yang et al., 2026a). PHR-VLA, WAM4D, and MECo-WAM use removable prediction branches to retain training benefits without deploying the auxiliary predictor (Soleymanzadeh et al., 2026; Li et al., 2026e; Zhang et al., 2026c). MM-ABC predicts future VGGT-Ω features during training (Wang et al., 2026b); its future and action streams remain mutually masked, so geometric supervision shapes shared perceptual features without making future tokens policy inputs.

Clean-action prediction. The prediction target is a longstanding design choice in diffusion and flow models (Salimans & Ho, 2022; Karras et al., 2022; Lipman et al., 2022), including Transformer-based generators (Peebles & Xie, 2023). DP3 uses sample prediction for high-dimensional action generation (Ze et al., 2024), RDT-1B regresses clean action chunks for bimanual manipulation (Liu et al., 2025), ManiCM adopts action-sample prediction within consistency distillation for one-step

control (Gao et al., 2024), and FA-RDP uses action-space reparameterization in distillation for reactive contact-rich manipulation (Zhuo et al., 2026). ABot-M0 combines clean-action outputs with a velocity-space objective (Yang et al., 2026b), and VGFM uses action-space predictions for value guidance in offline reinforcement learning (Koirala & Campbell, 2026). In image generation, JiT motivates clean-data prediction through manifold structure and finite capacity (Li & He, 2026), Pixel MeanFlow separates clean outputs from velocity-based objectives (Lu et al., 2026), and MiniT2I explores a minimalist generation framework (Wang et al., 2026d). Network outputs, training objectives, and sampling procedures are separate design choices, and controlled robot-policy studies emphasize separating their effects from architecture and iterative prediction (Pan et al., 2026).

B Implementation Details

Architecture. The backbone is Qwen3-VL-4B (Bai et al., 2025). Hidden states $H^l \in \mathbb{R}^{S\times 2560}$ are read from zero-indexed layers $\{11, 19, 27, 35\}$, where S is the image–text sequence length; layer 35 provides the base context, and layers 11, 19, and 27 enter the first three expert blocks through the gated updates of Equation (1). Each block projects its context into width-1024 keys and values, which action tokens read without updating the context sequence; gradients still reach the VLM, and the projections are reused across noise draws and sampling steps. MM-APT has 16 transformer blocks of width 1024, 16 attention heads, and SwiGLU feed-forward width 4096. Separate stream encoders project the noisy actions, concatenate a broadcast sinusoidal time embedding, and fuse them through a multilayer perceptron; each stream prepends a state token encoded from $[s_n; m_n^s] \in \mathbb{R}^{160}$. Learned position and segment embeddings define the near segment (state token and first 32 actions) and the far segment (remaining 32 actions). Padding and inactive body streams are masked, and separate normalized multilayer decoders predict the clean chunks after discarding the state tokens. The world expert uses 192 learned width-1024 queries per horizon at environment steps $n + 32$ and $n + 64$, a frozen VGGT-Ω teacher whose features are pooled to 8×8 grids per view, and a layer-normalized linear decoder to 2048-dimensional targets; the stabilized cosine is $\cos_\delta(u, v) = u^\top v/[\sqrt{\|u\|_2^2 + \delta}\sqrt{\|v\|_2^2 + \delta}]$ for a small $\delta > 0$.

Training and inference. Flow times follow $t \sim \mathrm{Beta}(1.5, 1)$, clipped at 0.999. The action loss of Equation (3) gives each active stream equal weight regardless of its number of supervised dimensions and is averaged over four independent noise/time draws that share one context computation; the future loss is evaluated once per example. At inference, the action streams integrate from Gaussian noise with five Euler steps, and the robot executes a prefix of the chunk before replanning.

Post-training budgets. EBench: 100k steps at batch size 512, with one generalist checkpoint for all 26 tasks. RoboCasa365: 120k steps at batch size 512 on the full target demonstrations. ManiSkill-HAB: one policy per suite, 100k steps at batch size 64, using head- and wrist-camera RGB images, proprioception, and language at every control step. LIBERO: the standard LIBERO demonstrations, with 50 evaluation trials per task; the same checkpoint is evaluated on LIBERO-Plus. Real world: 200 teleoperated demonstrations per task for every method. Component ablations: 120k steps at batch size 64 from scratch, without robot-data pretraining.

C Toy Experiment on the Prediction Target

This appendix details the toy experiment of Section 3.5, which asks whether a two-stream action network should predict clean trajectories or flow velocities. Following JiT's distinction between prediction and loss spaces (Li & He, 2026), we compare both parameterizations under a common objective on a synthetic task whose action geometry and Bayes-optimal denoiser are known in closed form. The two variants share architecture, conditioning, data, noise, and loss, and differ only in what the network emits, so any difference between them stems from the prediction target.

Parameterizations. For a clean chunk x and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, the flow time $t \in [0, 1]$ defines the interpolant and the velocity

$$z_t = (1 - t)\epsilon + tx, \qquad v = x - \epsilon. \tag{4}$$

An x-head estimates x directly, whereas a v-head estimates v. With shared conditioning c, the two networks produce the following endpoint estimates of the clean chunk:

$$\hat{x}^{(x)} = f_\theta(z_t, t, c), \qquad \hat{x}^{(v)} = z_t + d(t)\, g_\theta(z_t, t, c), \tag{5}$$

where $d(t) = \max(1-t, 0.05)$ stabilizes conversion near the clean endpoint. Both minimize masked endpoint MSE with inverse-square weights $d(t)^{-2}$, normalized to unit minibatch mean and with equal weighting of active streams, which equals velocity regression for the v-head wherever $d(t) = 1 - t$. Both variants recover the sampling velocity as $(\hat{x} - z_t)/d(t)$.

Synthetic task. We generate paired manipulation and body chunks of shapes 16×58 and 16×22 online with a fixed generator and independent training and evaluation draws. A Gaussian condition $c_0 \in \mathbb{R}^{12}$ maps to an eight-dimensional task vector τ through a fixed normalized nonlinear projection, and an equiprobable mode $m \in \{0, 1\}$ sets the arm share $\alpha_m \in \{0.35, 0.65\}$:

$$\mathrm{vec}(x_s) = a_{e,s} Q_{e,s} \left[\rho_s(m)\tau;\, b_s(2m-1);\, u_s\right], \qquad s \in \{M, B\}, \tag{6}$$

where $\rho_M = \alpha_m$, $\rho_B = 1 - \alpha_m$, $b_M = 0.12$, $b_B = 1.60$, $u_s \sim \mathcal{N}(0, I_4)$ is private variation, and $a_{e,s}$ is a variance scale. Each $Q_{e,s}$ has 13 orthonormal, temporally smooth columns on the valid channels of embodiment profile e, one of six profiles, two of which have no body supervision. Both networks receive $(z_{t,M}, z_{t,B}, t, c_0, e, m)$ with channel masks, so the arm–body allocation is given and the comparison isolates the prediction target.

Architecture and training. A standalone four-block, width-512, eight-head transformer uses separate stream transformations, shared conditions, masked cross-stream attention, and no raw-input skip inside either decoder. Paired runs share initialization, batches, noise, and flow times, and train for 12,000 updates with batch size 256, AdamW, and $t \sim \mathrm{Beta}(1.5, 1)$ clipped at 0.999. We report eight paired seeds, or sixteen seeds for the time-resolved metrics in Figure 2(b) and (c).

(a) Target geometry. For one embodiment, we stack flattened 928-dimensional manipulation chunks into matrices for x, v, and ϵ, center them, and plot normalized singular values. The clean spectrum terminates at rank 13, as prescribed by the generator, while velocity and noise retain variation across the ambient space. A v-head must therefore reproduce noise directions that are absent from the clean-action subspace.

(b) Noise in the endpoint estimate. We compare both heads through their endpoint estimates $\hat{x}$, which live in the same space. Holding a clean chunk fixed, we draw 24 independent noises, compute $\hat{x}$ for each, and report its variance normalized by the clean-chunk variance. x-prediction yields lower variance at every tested flow time: 0.023 versus 0.54 at $t = 0.01$, a $23\times$ reduction, and 0.35 versus 0.63 at $t = 0.2$. Ridge probes on the final manipulation-token features show the same pattern internally. The probes regress each token's injected noise from its features, are fit on sampled token rows, and are evaluated on held-out rows, where R^2 is one minus the squared error divided by the target variation around the training mean, pooled over coordinates. The injected noise is recoverable with held-out R^2 of 0.62 for x-prediction and 0.94 for v-prediction, against -0.09 for both with permuted rows. Because neither decoder has a raw-input skip, a v-head must carry this noise through its layers to form its endpoint estimate $z_t + d(t)g_\theta$ from the noisy input.

(c) Error beyond the Bayes-optimal denoiser. At fixed t, we compute endpoint MSE over valid entries, averaging active examples within each stream and then the two streams, and subtract the MSE of the Bayes-optimal conditional mean given the same inputs, which is available analytically by Gaussian posterior estimation in the known basis. This excess risk measures approximation error beyond the uncertainty inherent in denoising. The v/x ratio reaches 1.64 at $t = 0.01$, so x-prediction is markedly more accurate in the high-noise regime where every sampling trajectory begins.

(d) Sampled trajectories. We sample from Gaussian noise with uniform Euler steps and read each generated chunk out in the known latent coordinates. The task-allocation error is the distance between each stream's realized share of τ and the nearest valid share, averaged over the two streams; it counts coordination errors even when the chunk stays inside the action subspace. x-prediction lowers this error at every tested budget from 1 to 16 steps, from 0.0106 to 0.0090 at five steps, the budget used by our policy. It also yields lower off-subspace mass $\|(I - QQ^\mathsf{T})\hat{x}\|/\|\hat{x}\|$ on valid

entries at every tested budget (0.095 versus 0.104 at five steps). Its sliced Wasserstein distance to the true latent distribution, which also reflects the private variation u_s, is lower at 8 and 16 steps (0.075 versus 0.076 and 0.046 versus 0.050).

Relation to the full policy. In this task, the allocation is supplied through m. In MM-ABC, it is inferred from observations, the instruction, and robot state, and the two action streams exchange it through masked joint attention (Section 3.2). The component ablation in Section 5.4 evaluates clean-action prediction in this learned setting, where it outperforms velocity prediction by 3.6 points on RoboCasa365 composite-seen tasks.

D Pretraining Corpus

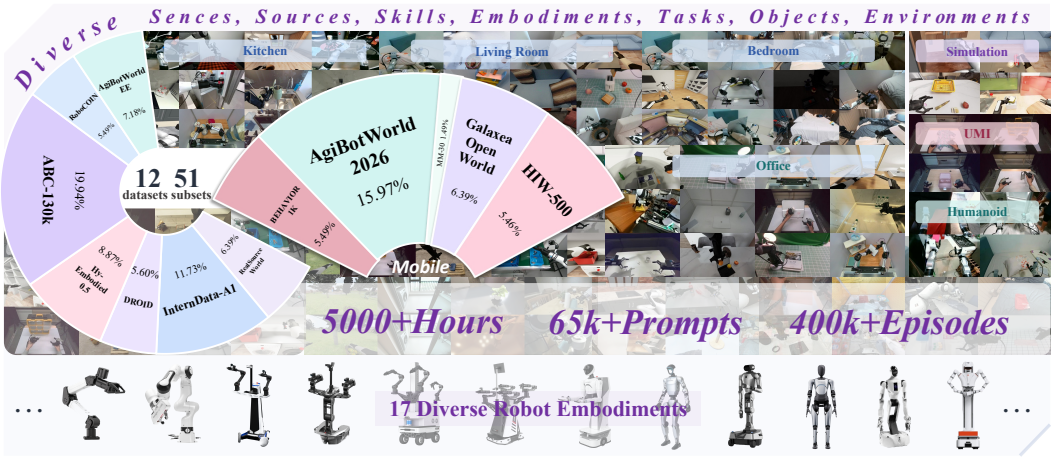
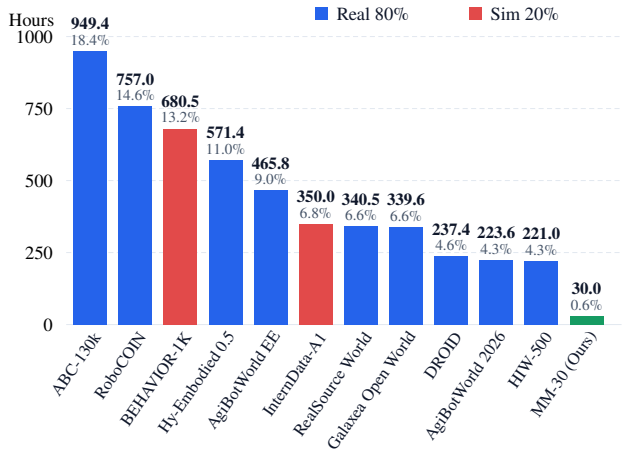


Figure 5: **Overview of the pretraining corpus.** 5,166.2 cleaned hours from 12 datasets and 51 subsets, spanning 17 embodiments. Sector areas show training sampling probabilities aggregated by dataset; the surrounding panels illustrate environments and embodiments.

(a) Pretraining data



(b) Embodiment profiles

Embodiment	Arms	Mobility	Frequency (Hz)
YAM	Dual	Fixed	30
Galaxea R1-Pro	Dual	Mobile	30
AgiBot A2-D	Dual	Mobile	30
Galaxea R1-Lite	Dual	Mobile	15/30
UMI	Paired	Handheld	30
RealSource RS-02	Dual	Mobile	30
AgileX Cobot Magic	Dual	Mobile	30
Franka Panda	Single	Fixed	15/30
AgiBot G2/A2	Dual	Mobile	30
Unitree G1	Dual	Legged	30
Realman RMC-AIDA-L	Dual	Mobile	30
Galbot G1	Dual	Mobile	30
AgileX Split Aloha	Dual	Fixed	30
ARX Lift-2	Dual	Lift	30
Yinhe Dual-Arm	Dual	Mobile	30
Trigger-A3 + PiPER-X	Dual	Mobile	30
AgiBot Genie-1	Dual	Mobile	30

Figure 6: **Pretraining data and embodiment composition.** **(a)** Dataset hours and duration shares. **(b)** Arm configuration, mobility, and control frequency.

The corpus combines eleven public datasets with MM-30 (Figures 5 and 6). BEHAVIOR-1K and InternData-A1 are simulated; the other sources are recorded on physical robots or, for Hy-Embodied 0.5, with a handheld UMI device.

- **ABC-130k** (Allshire et al., 2026): a bimanual teleoperation dataset of 134,806 episodes and 3,553 hours, collected on low-cost stations with two 6-DoF YAM arms; its 195 tasks cover pick-and-place, folding, handover, insertion, tool use, and assembly.
- **RoboCOIN** (Wu et al., 2025a): over 180K teleoperated bimanual demonstrations from 15 robotic platforms, spanning 421 tasks in 16 residential, commercial, and working scenarios, with hierarchical annotations from trajectory-level concepts to frame-level kinematics.
- **BEHAVIOR-1K** (Li et al., 2023): a simulation benchmark of 1,000 everyday household activities built on OmniGibson; its 2026 challenge release provides 20,000 teleoperated demonstrations of 100 long-horizon tasks on the wheeled bimanual Galaxea R1-Pro.
- **Hy-Embodied 0.5** (Zhang et al., 2026a): bimanual demonstrations collected with a fingertip UMI device tracked by optical motion capture and recorded from head and wrist views; the full corpus exceeds 10,000 hours, and the public release provides 2,163 hours over 70+ tasks.
- **AgiBotWorld EE** (Bu et al., 2025a): AgiBot World contains 1,001,552 trajectories and 2,976.4 hours over 217 tasks, 87 skills, and 106 scenes, collected by more than 100 AgiBot robots in five deployment domains, spanning gripper, dexterous-hand, and visuo-tactile data.
- **InternData-A1** (Tian et al., 2026): synthetic data from a compositional simulation pipeline, with 630K+ trajectories and 7,433 hours over 70 tasks and 227 scenes, covering rigid, articulated, deformable, and fluid objects on Franka Panda, AgileX Split Aloha, ARX Lift-2, and AgiBot Genie-1; a π_0 model pretrained on it alone matches the official π_0.
- **RealSource World** (RealSource, 2025): 11,428 episodes of long-horizon manipulation over 35 tasks on the RS-02 dual-arm humanoid in kitchens, conference rooms, convenience stores, homes, and industrial settings, with atomic-skill segments and per-episode quality assessments.
- **Galaxea Open-World** (Jiang et al., 2025a): 500+ hours of real-world mobile manipulation over 150+ tasks in 50 scenes on a single embodiment, the Galaxea R1-Lite with two 6-DoF arms, a 3-DoF torso, and an omnidirectional base, with bilingual subtask labels.
- **DROID** (Khazatsky et al., 2024): in-the-wild single-arm Franka Panda demonstrations comprising 76K trajectories and 350 hours across 564 scenes and 86 tasks.
- **AgiBotWorld 2026** (Team, 2026): real-world data collected on the AgiBot G2 platform in a free-form mode, covering commercial spaces such as retail stores and home scenarios, with 1:1 digital-twin scenes released for simulation.
- **HIW-500** (BitRobot et al., 2026): 500+ hours and 23K+ episodes of whole-body teleoperation of Unitree G1 humanoids in 12 real homes in Southeast Asia, covering 10+ household tasks with 161 subtask labels and 148K+ subtask annotations.
- **MM-30 (ours)**: 30+ hours of coordinated base and dual-arm demonstrations over 40+ tasks (Section F).

Mobile content. Platforms capable of mobile manipulation contribute approximately 61% of the cleaned hours; within this group, 1,371.7 hours (26.6% of the corpus) both retain base-control channels and exhibit nontrivial base motion.

Canonical interface. Each embodiment is mapped into

$$\mathbf{a}_t = \mathbf{a}_t^L \oplus \mathbf{a}_t^R \oplus \mathbf{a}_t^B \in \mathbb{R}^{29+29+22}, \qquad \mathbf{m}_t \in \{0,1\}^{80}, \tag{7}$$

where $\oplus$ denotes concatenation and $\mathbf{m}_t$ is an embodiment-specific validity mask. The 29-dimensional arm blocks reserve slots for arm joints, 3D end-effector position, 6D rotation, gripper state, and optional hand joints; the 22-dimensional body block reserves slots for base motion, torso, lift, head, and auxiliary controls. Only channels supported by the verified source schema are populated, and the number of valid stored dimensions ranges from 10 to 43, with control frequencies from 15 to 30 Hz. When a source stores both joint and end-effector representations, both occupy the vector, and the source's control mode determines which representation provides action supervision.

Balanced sampling. For a source profile i with h_i cleaned hours, the unnormalized weight is $\tilde{p}_i = h_i^{0.4}$, and $p_i = \mathrm{CapNorm}(\tilde{p}_i; 0.2)$ normalizes the weights, caps each profile at 0.2, and redistributes the excess among uncapped profiles. Figure 5 aggregates these probabilities by dataset.

E DATA ENGINE

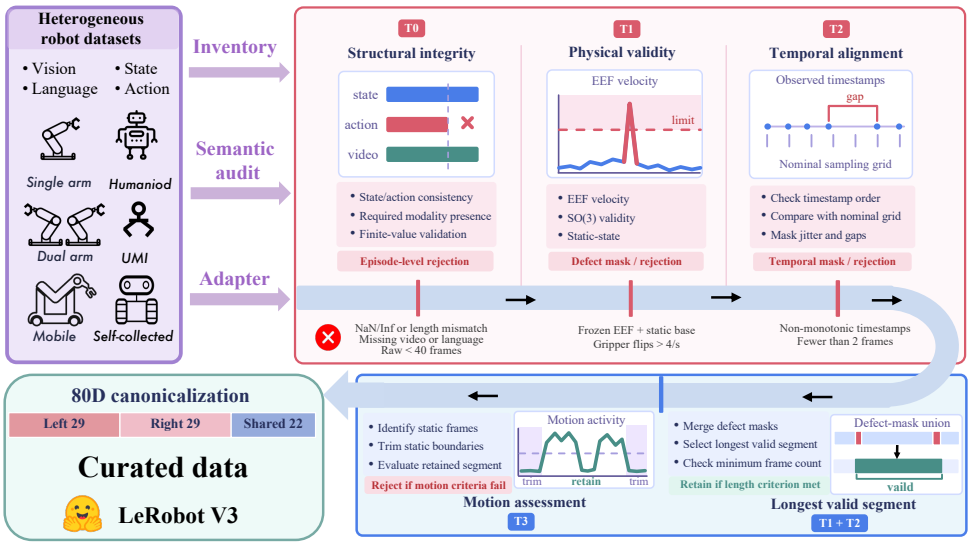


Figure 7: **Data engine.** Sources undergo semantic auditing, conversion with dataset-specific adapters, and four stages of quality filtering; retained contiguous segments are encoded in the masked 80D interface, independently validated, and incorporated into the training mixture.

State and action tensors can encode joint targets, Cartesian poses, deltas, velocities, or delayed controller commands, and sources differ in parent frames, tool center points (TCPs), rotation conventions, and gripper conventions (Figure 7).

Source audit and field semantics. For each source we fix its revision and record its schema, camera streams, control frequency, and robot metadata. A dataset-specific frame contract specifies the physical meaning of each state and action field: units, action type, parent frame, TCP, rotation convention, gripper direction, and temporal semantics. Adapters then convert units, unwrap Euler angles, reconstruct rotations in $\mathrm{SO}(3)$, and align coordinate frames and TCPs, distinguishing velocity commands from pose targets.

Staged quality filtering. **Tier 0** checks structural integrity, rejecting trajectories with insufficient length, state/action length mismatches, non-finite values, or missing videos or language annotations. **Tier 1** checks physical validity, including implausible end-effector speed, workspace violations, invalid rotations, frozen signals, and excessive gripper switching. **Tier 2** checks timestamp monotonicity and flags temporal jitter and large gaps. We merge the Tier 1–2 defect masks and retain the longest contiguous valid segment. **Tier 3** trims stationary prefixes and suffixes and rejects trajectories dominated by stationary behavior or negligible displacement of the end effector and base. Interior frames are never removed from a retained segment, which avoids unrecorded time jumps inside action chunks. A two-pass implementation first records all retention, rejection, and trimming decisions in a manifest and then writes the retained segments in the canonical format.

Independent validation. A separate validator checks schema consistency, finite values, timestamp continuity, video seeking and decoding, rotation orthogonality and round-trip consistency, compliance with the frame contracts, validity masks, and normalization statistics; corpus statistics and normalization parameters are recomputed after serialization.

F MM-30 COLLECTION

The platform comprises a HexFellow Trigger-A3 omnidirectional base and two AgileX PiPER-X 6-DoF arms with parallel grippers. Operators teleoperate all channels, so demonstrations contain

coordinated base and bimanual motion, including handovers between the two arms. The 40+ tasks form six families (transfer and transport, organization, rotation and fastening, opening and retrieval, deformable manipulation, and pouring and tool use) and are recorded in nine indoor domains, from living rooms and kitchens to offices and industrial and chemistry laboratories. Each task is recorded from multiple viewpoints under randomized initial configurations, object layouts, distractors, and scene appearance, with 26 to 91 episodes per task. After the semantic audit, coordinate alignment, and quality filtering of Section E, all cleaned demonstrations are included in pretraining, contributing 30.0 cleaned hours to the corpus.

G Benchmark Details and Full Results

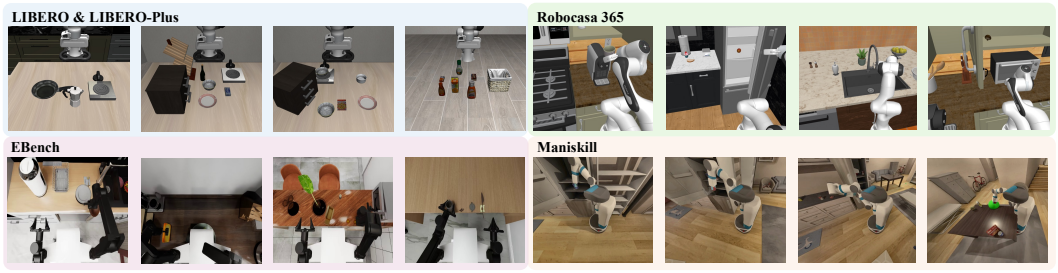


Figure 8: **Simulation benchmarks.** Representative scenes from LIBERO/LIBERO-Plus, RoboCasa365, EBench, and ManiSkill-HAB, covering fixed-base and mobile manipulation.

EBench. EBench comprises 26 indoor tasks spanning mobile pick-and-place, long-horizon mobile manipulation, and dexterous tabletop manipulation (Gao et al., 2026), which vary in scene, skill, horizon, precision, and operating mode. A single checkpoint is evaluated on the held-out test split following the official protocol, using task success and a stage-wise progress score (Figure 9).

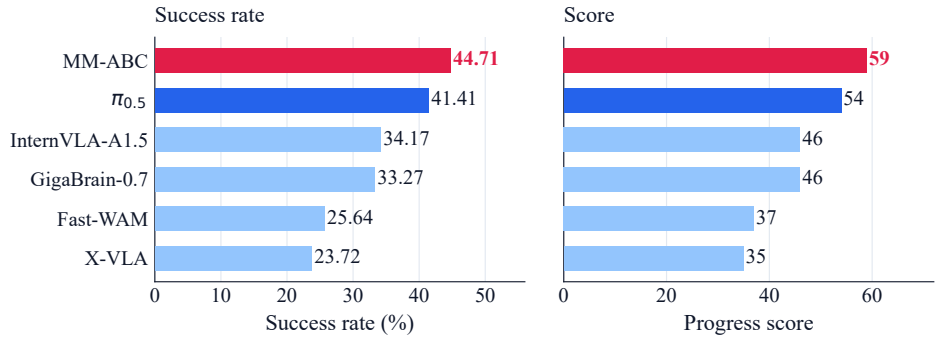


Figure 9: **EBench.** Success rate (left) and stage-wise progress score (right) for MM-ABC and baselines (Intelligence et al., 2025; Ma et al., 2026; Team et al., 2026a; Yuan et al., 2026; Zheng et al., 2026). Higher is better.

RoboCasa365. The target evaluation comprises 50 household tasks in held-out kitchens (Nasiriany et al., 2026); atomic tasks emphasize individual interactions with objects and articulated fixtures, whereas composite tasks combine successive interactions with navigation, and seen/unseen labels refer to inclusion in the benchmark's predefined pretraining set. Across all compared methods, success rates are lower on composite than on atomic tasks (Table 1).

ManiSkill-HAB. ManiSkill-HAB evaluates low-level mobile manipulation with a Fetch robot, including physically simulated grasping and interaction with articulated objects (Shukla et al., 2025). SetTable requires retrieving a bowl from a drawer and an apple from a refrigerator and placing them on a table, evaluated on seven skills; TidyHouse rearranges objects among open receptacles, and

PrepareGroceries transfers objects between a refrigerator and a counter, both over nine object categories. Baseline results are reported by MoPA (Chen et al., 2026a); Table 6 lists per-skill SetTable results, and TidyHouse and PrepareGroceries results are given in Table 2.

Table 6: **ManiSkill-HAB SetTable.** Skill and mean success rates (%). *Depth input; – unavailable, so AnchorVLA's mean covers six skills. Baselines: DP3 (Ze et al., 2024), ACT (Zhao et al., 2023), DP (Chi et al., 2025), RDT-1B (Liu et al., 2025), AC-DiT (Chen et al., 2026c), π_0 (Black et al., 2024), AnchorVLA (Lim et al., 2026), MobileWAM (Fan et al., 2026), GeoHAT (Zhu et al., 2026), InCoM (Liu et al., 2026a). Best: shaded; second best: bold.

Method	Pick Apple	Pick Bowl	Place Apple	Place Bowl	Open Fridge	Open Drawer	Close Drawer	Mean
DP3*	0.0	20.0	31.0	32.0	0.0	0.0	68.0	21.6
ACT	28.0	28.0	8.7	13.0	2.0	0.0	85.7	23.6
DP	21.3	20.7	28.0	69.3	7.3	0.0	55.0	28.8
RDT-1B	12.0	10.7	32.0	18.7	82.7	44.0	**100.0**	42.9
AC-DiT*	33.3	36.0	33.3	17.3	90.7	81.3	**97.3**	55.6
π_0	26.6	26.6	48.9	56.8	90.5	75.5	88.7	59.1
AnchorVLA	22.7	44.5	64.3	63.8	88.9	–	**100.0**	64.0
MobileWAM	46.0	46.0	63.7	64.7	**99.3**	**91.0**	**100.0**	73.0
GeoHAT*	**82.3**	69.3	60.0	78.0	83.7	86.0	95.3	79.2
InCoM*	59.4	**84.1**	**84.1**	**82.5**	87.3	88.9	**100.0**	**83.8**
MM-ABC	**81.3**	**86.7**	**71.0**	**81.3**	**96.7**	**92.0**	95.7	**86.4**

LIBERO and LIBERO-Plus. LIBERO evaluates four fixed-base suites—Spatial, Object, Goal, and Long—with ten tasks each and 50 trials per task (Liu et al., 2023). LIBERO-Plus perturbs camera configuration, robot initial state, language, lighting, background, sensor noise, and scene layout over 10,030 episodes (Fei et al., 2025); its total is the success rate over all episodes, and MM-ABC uses only the original LIBERO demonstrations (Tables 7 and 8). MM-ABC attains the highest total, language, and background scores among all compared methods, including those trained on perturbed demonstrations, while training only on the original LIBERO demonstrations.

Table 7: **LIBERO** success rate (%) with 50 trials per task, including suite-specific and shared-policy configurations. Baselines: OpenVLA (Kim et al., 2024), WorldVLA (Cen et al., 2025), π_0-FAST (Pertsch et al., 2025), NORA (Hung et al., 2025), π_0 (Black et al., 2024), UniVLA (Bu et al., 2025b), $\pi_{0.5}$ (Intelligence et al., 2025), OpenVLA-OFT (Kim et al., 2025), GR00T-N1.6 (Bjorck et al., 2025), Fast-WAM (Yuan et al., 2026), X-VLA (Zheng et al., 2026), Cosmos-Policy (Kim et al., 2026b), LingBot-VA (Li et al., 2026d), ABot-M0 (Yang et al., 2026b). Best: shaded; second best: bold.

Method	Spatial	Object	Goal	Long	Average
OpenVLA	84.7	88.4	79.2	53.7	76.5
WorldVLA	87.6	96.2	83.4	60.0	81.8
π_0-FAST	96.4	96.8	88.6	60.2	85.5
NORA	92.2	95.4	89.4	74.6	87.9
π_0	96.8	98.8	95.8	85.2	94.2
UniVLA	96.5	96.8	95.6	92.0	95.2
$\pi_{0.5}$	98.8	98.2	98.0	92.4	96.9
OpenVLA-OFT	97.6	98.4	97.9	94.5	97.1
GR00T-N1.6	**99.3**	99.2	98.4	92.9	97.5
Fast-WAM	98.2	**100.0**	97.0	95.2	97.6
X-VLA	98.2	98.6	97.8	97.6	98.1
Cosmos-Policy	98.1	**100.0**	98.2	97.6	98.5
LingBot-VA	98.5	99.6	97.2	**98.5**	98.5
ABot-M0	98.8	**99.8**	**99.0**	96.6	**98.6**
MM-ABC	**99.4**	**100.0**	**98.6**	**98.4**	**99.1**

Metrics and aggregation. Every summary score follows the official definition of its benchmark, and all values are percentages except the EBench progress score. EBench reports the success rate

Table 8: **LIBERO-Plus** success rate (%) across seven perturbation categories (Fei et al., 2025). Total is the success rate over all 10,030 evaluation episodes, following the official protocol. $^{+}$Additional training on LIBERO-Plus perturbed demonstrations. Best: shaded; second best: bold.

Method	Camera	Robot	Language	Light	Backgr.	Noise	Layout	Total
Additional training on perturbed demonstrations								
π_0^+	79.6	21.1	72.5	84.7	86.2	68.3	69.4	67.4
GR00T-N1.6$^+$	**92.6**	33.5	80.1	93.6	95.4	**93.6**	75.0	79.4
OpenVLA-OFT$^+$	**92.8**	30.3	85.8	94.9	93.9	89.3	77.6	79.6
Trained on original LIBERO only								
OpenVLA	0.8	3.5	23.0	8.1	34.8	15.2	28.5	15.6
GR00T-N1.6	20.9	40.2	35.0	65.4	76.3	27.8	51.3	42.8
UniVLA	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
Fast-WAM	16.4	44.5	68.9	78.2	53.7	37.7	60.7	50.0
π_0	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
π_0-FAST	65.1	21.6	61.0	73.2	73.2	74.4	68.8	61.6
LingBot-VA	40.9	**83.0**	**86.4**	82.3	53.1	64.4	76.2	69.5
OpenVLA-OFT	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
X-VLA	23.4	**89.7**	75.7	88.2	**96.0**	62.7	71.8	70.5
ABot-M0	60.4	67.9	**86.4**	**96.2**	91.6	86.4	**82.6**	80.5
Cosmos-Policy	75.8	63.3	81.7	**96.5**	88.9	**92.7**	**82.2**	**82.2**
MM-ABC	73.6	66.5	**88.9**	94.9	**96.1**	86.4	80.5	**82.8**

over the test episodes of all 26 tasks, together with the stage-wise progress score defined by the benchmark. The RoboCasa365 average weights the atomic-seen, composite-seen, and composite-unseen splits by their numbers of tasks (18, 16, and 16). For ManiSkill-HAB, the SetTable mean averages its seven skills, whereas the TidyHouse and PrepareGroceries means average the pick and place rates, each of which is first averaged over nine object categories. The LIBERO average is the mean of the four suites, each evaluated with ten tasks and 50 trials per task. The LIBERO-Plus total weights each category by its number of episodes: 1,599 for camera, 1,550 for robot initial state, 1,537 for language, 1,142 for lighting, 1,076 for background, 1,601 for sensor noise, and 1,525 for layout. The total is therefore not the arithmetic mean of the seven category scores. In every table, best and second-best entries are ranked per column among all listed methods, including methods trained with additional perturbed demonstrations.

H Real-World Experiments

Tasks. Each task is specified by a single natural-language instruction and requires base repositioning between object interactions.

- **Birthday Party Setup.** The robot picks up a cake from the table ahead, carries it to the decorated table on the left, and sets it down; it then passes a birthday candle between its hands, inserts it into the cake, moves beside the table, and turns on the speaker to play music.
- **Office Folder Arrangement.** The robot picks up a folder from the office desk in front of it, passes it between both hands and holds it steady, carries it to the desk on its right, aligns it with the file organizer, and inserts it.
- **Kitchen Work.** The robot picks up a food container from the table and pours the food into a pot, puts the lid on the container, places it in the cabinet, and closes the cabinet door.
- **Industrial Parts Organization.** The robot places scattered industrial parts neatly into a parts bin, carries the filled bin to the table behind it, and stacks it neatly with the other bins.
- **Chemistry Lab Operation.** The robot picks up a test tube, moves to the laboratory bench behind it, pours the reagent into a beaker holding another reagent, and stirs until they are fully mixed.

Protocol. All methods are fine-tuned on the same 200 teleoperated demonstrations per task and evaluated over 20 trials per task; a trial counts as successful only when all steps of the task are com-

pleted. Figure 10 visualizes the per-task success rates of Table 5. On Industrial Parts Organization, MM-ABC and $\pi_{0.5}$ both succeed in 75% of trials.

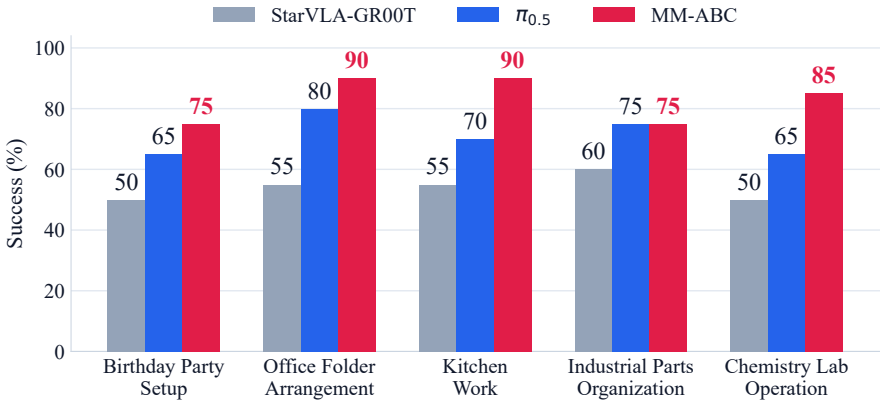


Figure 10: **Real-world success rate** over 20 trials per task.