EGOFORESIGHT: EFFECTIVELY SCALING POLICY LEARNING WITH NOISY EGOCENTRIC ACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling policy learning requires broad experience and reliable supervision. Robot demonstrations provide high-quality trajectory supervision but are costly to collect. Egocentric videos broaden task and environment coverage, yet motion reconstruction and retargeting introduce noisy action labels. To address this challenge, we introduce EgoForesight, a vision-language-action framework that learns jointly from both sources by reducing the influence of unreliable action labels and exploiting demonstrated outcomes as complementary supervision. A joint vision-action flow couples future visual representation prediction with action generation, unifying reliability-aware action learning with outcome-informed representation transfer. Specifically, *reliability-conditioned corruption* allocates action supervision across noise levels according to source reliability, while reliability-weighted regression regulates each source's contribution. Complementing this treatment of noisy labels, *privileged-future representation distillation* draws supervision from demonstrated outcomes. An exponential-moving-average teacher observes the current scene and clean future to encode action-relevant structure from the resulting scene change. Its deeper representations supervise an earlier student layer receiving a corrupted future, providing outcome-informed structure beyond noisy action annotations. Together, these mechanisms support effective scaling with noisy egocentric actions. Over the tested range, EgoForesight exhibits steeper scaling slopes than the baseline, reducing real-world open-loop action MSE by 20.5% and improving RoboCasa few-shot success by 8.1 percentage points at 30,000 hours of pretraining data. After post-training, it also achieves strong absolute performance in both simulated and real-world dexterous manipulation.

1 INTRODUCTION

Vision-language-action (VLA) models connect broad visual and linguistic knowledge to robot control (Zitkovich et al., 2023; Black et al., 2025b). We organize their supervision into an embodied data pyramid: vision–language data offer semantic coverage, egocentric demonstrations capture diverse interactions, and robot trajectories provide executable controls but require dedicated hardware and collection effort (Hoque et al., 2026; Open X-Embodiment Collaboration, 2024; Khazatsky et al., 2024). EgoMimic (Kareer et al., 2025) and EgoScale (Zheng et al., 2026b) demonstrate the value of egocentric data through human–robot co-training and large-scale policy pretraining.

However, recovering actions from egocentric videos introduces annotation noise. Camera tracking and hand-pose reconstruction can produce errors, particularly under occlusion and rapid motion (Zhang et al., 2025; Hoque et al., 2026). Retargeting adds a geometric mismatch: differences in palm size, finger proportions, and joint structure mean that accurately tracked human poses need not admit an equivalent robot configuration (Wang et al., 2024; Xin et al., 2026). Appearance and dynamics further complicate transfer. Effective learning from egocentric data therefore requires more than increasing its scale: *policies must limit the influence of unreliable action labels while drawing on supervision beyond those labels, enabling additional ego-data to provide effective supervision.*

To meet these requirements, we introduce EgoForesight, a VLA framework that reduces the influence of unreliable actions while learning from demonstrated outcomes. Inspired by prior unified video–action and dynamics–action models (Zhu et al., 2025; Lyu et al., 2026), our design couples experts for future visual representations and action chunks and trains them with modality-specific

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.

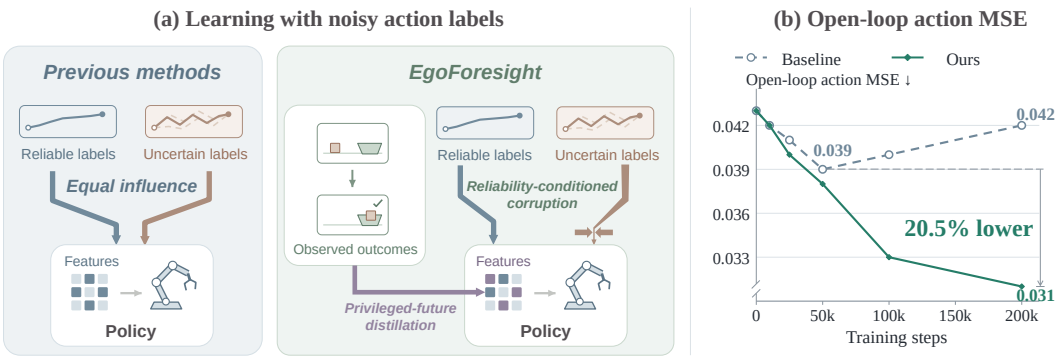


Figure 1: **Learning with noisy action labels.** **(a)** Unlike methods that treat action labels equally, EgoForesight accounts for label reliability and enriches action representations through privileged-future distillation. **(b)** Expansion training on 20 kh of egocentric data. At each checkpoint, both methods are fine-tuned on real-robot data and evaluated by test-set open-loop action MSE. See Figure 5 for a more detailed scaling analysis.

flow-matching paths. Separate corruption times control the information available in the vision and action streams, allowing us to regulate noisy action supervision and distill outcome-informed representations within the same model.

First, *reliability-conditioned corruption* allocates action supervision across noise levels according to source reliability. Motivated by ADP's finding that stronger diffusion noise can contract target and suboptimal distributions (Wei et al., 2026), our continuous source-reliability schedule shifts less reliable actions toward stronger corruption. Reliability-weighted action regression complements this schedule by controlling how strongly each source contributes to learning. Together, they enable learning from diverse egocentric data without assuming uniformly reliable labels.

Second, *privileged-future representation distillation* draws on demonstrated outcomes for action-relevant evidence beyond noisy annotations. An exponential-moving-average (EMA) teacher receives the current observation and observed future at the clean endpoint, acting as a conditional inverse dynamics model (IDM). Initialized from higher-quality demonstrations, the teacher provides action representations informed by the observed scene change. Adapting Self-Flow's cleaner, noise-shifted EMA teacher and cross-layer alignment (Chefer et al., 2026), we apply this mechanism to robot action tokens: a deeper teacher feature supervises an earlier student layer, which enriches early policy features with outcome-informed structure. In our setting, observed-future conditioning gives the teacher its inverse-dynamics role, and the student learns to anticipate the resulting outcome-informed representation while receiving a corrupted future.

For model training, we first establish visual–action correspondences on higher-quality robot demonstrations and curated egocentric trajectories. This checkpoint initializes the student and teacher before training expands to a larger action-annotated egocentric-data mixture with reliability-conditioned corruption and privileged-future distillation enabled. Our experiments examine scaling as a test of how effectively policies use noisy egocentric supervision. Under matched data and optimization budgets, EgoForesight translates added pretraining data into stronger gains in action prediction and closed-loop control. Ablations support complementary roles for reliability conditioning and future distillation, and show the value of establishing action correspondences on higher-quality data before expanding coverage.

2 Related Work

Embodied Foundation Models. Levine et al. (2016) demonstrated that robot policies could be learned end to end from images by jointly training perception and control. Subsequent work expanded this scope toward sharing knowledge across tasks and embodiments, with RT-1 and Octo scaling transformer policies into generalist robot models (Brohan et al., 2023; Ghosh et al., 2024). At the action interface, tokenized prediction (Zitkovich et al., 2023; Kim et al., 2025; Pertsch et al., 2025) and continuous diffusion or flow policies (Chi et al., 2023; Black et al., 2025b; Wu et al., 2026) offer different ways to connect perception to control, with flow-based variants grounded

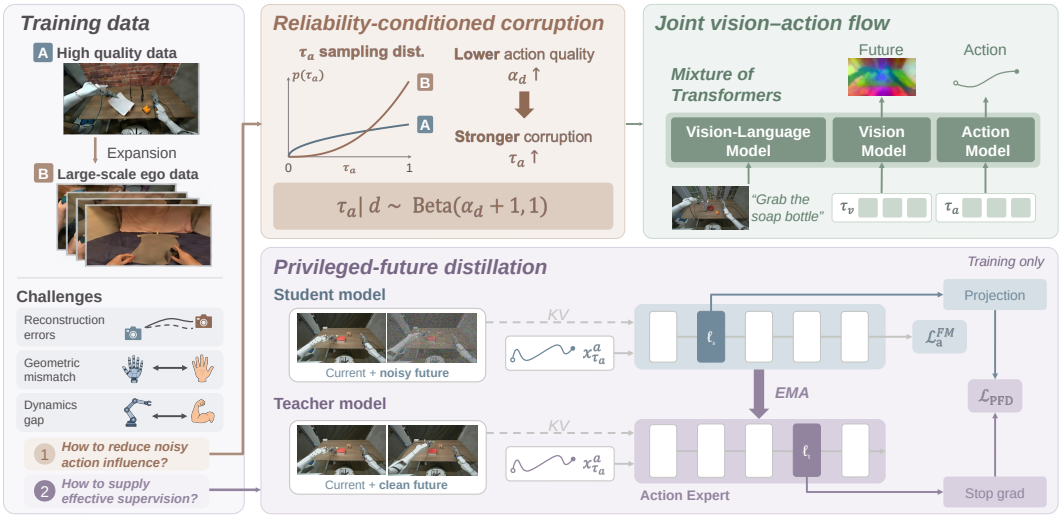


Figure 2: **Overview of EgoForesight.** Large-scale human videos expand training data but have less reliable action labels. Reliability-conditioned corruption samples stronger action corruption as reliability falls. The vision–action flow jointly predicts future visual representations and action chunks; during training, an observed-future teacher distills features into the student.

in flow matching and rectified flow (Lipman et al., 2023; Liu et al., 2023). For models built on pretrained VLMs, effective transfer benefits from retaining pretrained semantic competence while grounding spatial and task knowledge in robot actions (Black et al., 2025a; Driess et al., 2025; Chen et al., 2025). This emphasis extends to the information used to condition a policy: $\pi_{0.7}$ (Physical Intelligence et al., 2026) combines language, visual subgoals, and execution metadata to exploit heterogeneous data and recombine skills across tasks. Future prediction offers another source of guidance by relating actions to their anticipated effects. This idea developed from video-based planning (Ebert et al., 2018) into joint future/action prediction (Wu et al., 2024) and a broader family of unified future/action models, many of which use diffusion (Zhu et al., 2025; Bi et al., 2026; Ye et al., 2026; Cai et al., 2026; Li et al., 2026). Within joint diffusion, separate modality noise times support flexible conditioning (Bao et al., 2023) and allow generation schedules such as X-WAM's action-before-video denoising (Guo et al., 2026). The choice of prediction target matters as well: compact future representations can guide or be jointly coupled to action generation (Su et al., 2026; Tang et al., 2026; Ma et al., 2026). LDA-1B (Lyu et al., 2026) uses future DINO features and allocates policy, dynamics, and forecasting objectives according to the quality of the available data. Appendix G discusses the data pyramid, representation learning for generation, and learning from noisy data.

3 EgoForesight

EgoForesight addresses two complementary questions: how to reduce the influence of noisy action labels in egocentric data, and how to supply effective supervision beyond those annotations. Both designs operate on the flow timesteps of a joint future–action model. *Reliability-conditioned corruption* assigns less reliable actions greater corruption while keeping future vision comparatively informative. *Privileged-future representation distillation* places a teacher at the observed-future endpoint and transfers its action representations to a student with a corrupted future. Figure 2 summarizes how these two uses of flow time connect data reliability to policy learning.

3.1 Problem Formulation

We consider datasets $\mathcal{D} = \bigcup_d \mathcal{D}_d$ of robot and human manipulation demonstrations, all paired with action labels. At step t, the context c comprises the current observation o_t, a language instruction, available current state measurements, and embodiment information. For a horizon of K action steps, a training example supplies

$$y_v = D(o_{t+K}) \in \mathbb{R}^{N_v \times d_v}, \qquad y_a = (\widetilde{a}_t, \ldots, \widetilde{a}_{t+K-1}) \in \mathbb{R}^{K \times d_a}, \tag{1}$$

where D is a frozen visual encoder and $\widetilde{a}$ is a normalized action label. Robot demonstrations provide recorded controls; human demonstrations provide tracked or reconstructed hand motions and, where applicable, retargeted robot actions. A binary mask $M_a \in \{0,1\}^{K\times d_a}$ excludes padded action coordinates and timesteps.

Tracking and retargeting errors make these human action labels less reliable (Appendix C.5). We therefore regulate action supervision by source reliability and use observed futures as complementary outcome evidence.

We model future visual features and action chunks jointly as $p_{\psi,\theta}(y_v, y_a \mid c)$, with the action marginal defining the policy for the specified embodiment. At deployment, the policy conditions only on c; the observed future is available solely for training.

Mixture-of-Transformers (MoT). Our model comprises a pretrained VLM expert and two coupled Transformer experts for future vision and action. VLM features, current visual features, and available state and embodiment inputs form the conditioning inputs $C_\psi(c)$. The vision and action experts use modality-specific projections and feed-forward layers with joint bidirectional attention, and both read the context. The VLM retains causal attention and cannot read either generated stream, so future observations never enter $C_\psi(c)$. Student objectives nevertheless backpropagate through the conditioning features into the VLM. We denote the coupled experts and their conditioning projections by θ, and the VLM parameters by ψ.

Joint flow parameterization. Using the flow-matching framework (Lipman et al., 2023; Liu et al., 2023), we define our own separate interpolation paths for future vision and action:

$$x^v_{\tau_v} = (1-\tau_v)y_v + \tau_v\epsilon_v, \qquad x^a_{\tau_a} = (1-\tau_a)y_a + \tau_a\epsilon_a, \tag{2}$$

where ϵ_v and ϵ_a are independent standard Gaussian tensors, and $\tau_v, \tau_a \in [0,1]$ control corruption from the data endpoint at zero to Gaussian noise at one. Action expressions are restricted to valid coordinates. The coupled experts predict

$$(\widehat{u}_v, \widehat{u}_a) = F_\theta\big(x^v_{\tau_v}, x^a_{\tau_a}, C_\psi(c), \tau_v, \tau_a; M_a\big), \tag{3}$$

with velocity targets $u^\star_v = \epsilon_v - y_v$ and $u^\star_a = \epsilon_a - y_a$. Separate flow times control the information available in each stream, while joint attention lets the action prediction draw on the partially resolved future.

3.2 Reliability-Conditioned Corruption

We use flow time to allocate action supervision across noise levels. Ambient Diffusion Policy (ADP) (Wei et al., 2026) motivates this noise-dependent use of data: its high-noise region exploits distribution contraction, while its low-noise region exploits shared local motion. In our human–robot mixture, fine human motions need not transfer across hand morphologies, and robot demonstrations already provide substantial local-control supervision. We therefore favor stronger corruption for unreliable actions without introducing ADP's additional low-noise admission window. Each source d receives a reliability $q_d \in [0,1]$, where larger values indicate greater trust in its action labels. This training prior determines the corruption parameter through

$$\alpha_d = \alpha_{\min} + (1-q_d)(\alpha_{\max} - \alpha_{\min}), \qquad 0 < \alpha_{\min} < \alpha_{\max}. \tag{4}$$

We use $\alpha_{\min} = 0.5$ and $\alpha_{\max} = 3$, so lower reliability gives a larger α_d. The prior expresses relative trust in a source without requiring a calibrated error estimate or a known annotation-noise distribution.

We define a continuous joint corruption law

$$(\tau_v, \tau_a) \sim \mathcal{S}_{\mathrm{reg}}(q_d), \qquad 0 \le \tau_v \le \tau_a \le 1, \tag{5}$$

with marginals

$$\tau_v \sim \mathcal{U}(0,1), \qquad \tau_a \mid d \sim \mathrm{Beta}(\alpha_d + 1, 1). \tag{6}$$

The sampler preserves both marginals while keeping future vision less corrupted than action. All sources share the full action-time support $(0,1)$; reliability changes the density instead of imposing two hard admission thresholds. This fits mixed-domain VLA co-training by varying supervision

frequency smoothly across noise levels and avoiding entire intervals supervised by only a subset of domains. The effect on action inputs is explicit:

$$\mathbb{E}_{\mathrm{reg}}[1-\tau_a \mid d]=\frac{1}{\alpha_d+2}, \qquad \mathbb{E}_{\mathrm{reg}}[\tau_a-\tau_v \mid d]=\frac{\alpha_d}{2(\alpha_d+2)}. \tag{7}$$

Lower reliability reduces the expected contribution of the action label to the input and increases the vision–action corruption gap, leaving the vision marginal unchanged. This encourages the action expert to use the current context and the comparatively more informative future. Appendix A provides the sampler and derivation.

Flow matching learns a conditional average of velocity targets, so the contribution of annotation noise is governed by the resulting conditional field. Appendix A.1 analyzes how stronger corruption brings the annotation and reference fields closer under a simple noise model. We additionally multiply the direct action loss by q_d to regulate each source's contribution to learning. Timestep sampling and loss weighting thus determine where and how strongly each source supervises the joint model.

3.3 Privileged-Future Representation Distillation

The observed future provides supervision beyond action labels. At $\tau'_v = 0$, the action expert sees both the current observation and the demonstrated outcome, acting as a conditional inverse dynamics model (IDM). Trained on higher-quality demonstrations, its action estimates can be more reliable than labels affected by reconstruction or retargeting errors. An exponential-moving-average (EMA) teacher in this regime transfers outcome-informed action features to a student with a corrupted future.

The teacher maintains an EMA of the coupled experts,

$$\bar{\theta} \leftarrow \mu\bar{\theta} + (1-\mu)\theta, \qquad 0 \le \mu < 1, \tag{8}$$

while reusing the student's contextual features with gradients stopped. Given the student input, we construct a teacher input with the observed future and the corrupted action:

$$\tau'_v = 0, \qquad \tau'_a = \tau_a, \qquad x^a_{\tau'_a} = (1-\tau'_a)y_a + \tau'_a\epsilon_a. \tag{9}$$

The teacher receives y_v, $x^a_{\tau'_a}$, and the detached context $\mathrm{sg}(C_\psi(c))$, where sg denotes stop-gradient. It shares the student's action mask, target, and Gaussian noise, keeping both action inputs on the same interpolation path. The EMA covers the coupled experts and conditioning projections; VLM features are reused from the student. To calibrate this clean-future regime, we set $\tau_v = 0$ for a small fraction p_{edge} of student examples, retaining action flow and representation losses but applying *no vision flow loss*. Appendix A gives the complete sampler.

Adapting Self-Flow's projection-based cosine alignment between an earlier student layer and a deeper EMA-teacher layer (Chefer et al., 2026), we apply the mechanism to action tokens whose teacher representations are informed by the observed future. Let $h^{\ell_s}_{S,k}$ and $h^{\ell_t}_{T,k}$ denote the student and teacher action features at token k, with $\ell_s < \ell_t$. A trainable projection P_ϕ matches the features through a cosine loss:

$$\ell_{\mathrm{PFD},k} = 1 - \cos\Big(P_\phi(h^{\ell_s}_{S,k}), \mathrm{sg}(h^{\ell_t}_{T,k})\Big). \tag{10}$$

In our construction, the flow-time gap trains early student features to anticipate deeper representations informed by the observed future.

Two-stage data selection. To ground distillation in reliable supervision, we first learn visual–action correspondences from higher-quality robot demonstrations and curated human trajectories through joint flow matching and VLM co-training, without distillation. The resulting checkpoint initializes both the student and its EMA teacher. Stage two retains these sources and adds a much larger collection of action-annotated human videos, enabling both proposed components. Appendix C details the datasets and their allocation across stages.

3.4 Training Objective

Action labels are mapped to a canonical action space and normalized; the space and dataset-specific mappings are detailed in Appendix C.4. We train the joint flow model with mean

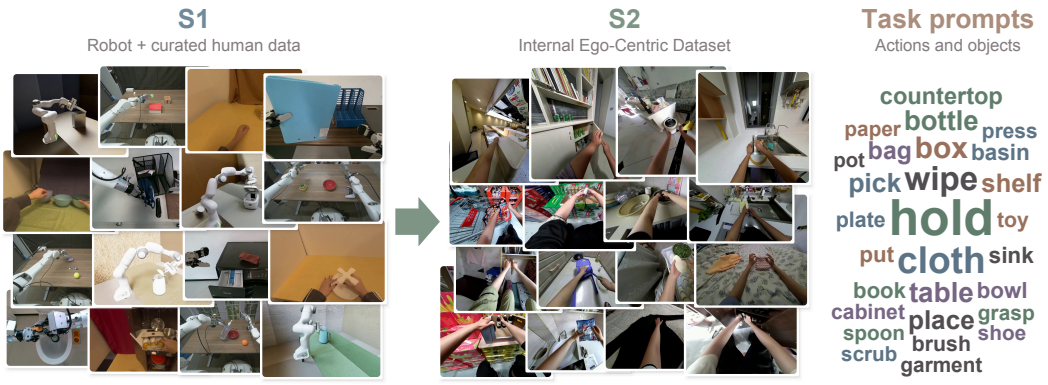


Figure 3: **Two-stage pretraining data.** Left: examples from AgiBot World, EgoDex, InternData-A1, and T-Rex used in S1. Center: examples from the Internal Ego-Centric Dataset, which adds 20,000 hours to S2 alongside all S1 sources. Right: task terms sized by frequency in 8,192 sampled instructions from this dataset, excluding stopwords and generic descriptors.

squared error (MSE):

$$\mathcal{L}_v^{\mathrm{FM}} = \mathbb{E}\big[\mathrm{MSE}(\widehat{u}_v, u_v^\star)\big], \qquad \mathcal{L}_a^{\mathrm{FM}} = \mathbb{E}\big[q_d\,\mathrm{MSE}(\widehat{u}_a, u_a^\star)\big]. \tag{11}$$

MSE averages over valid coordinates, and the vision loss excludes clean-future boundary examples. The distillation loss $\mathcal{L}_{\mathrm{PFD}}$ averages Equation (10) over valid action tokens without reliability weighting, preserving feature supervision for less reliable sources. Appendix B gives the exact batch reductions.

During continual adaptation to embodied tasks, we retain the autoregressive cross-entropy $\mathcal{L}_{\mathrm{AR}}$ on interleaved vision–language data to preserve general visual–language knowledge and reduce forgetting. Combining this objective with the flow and distillation losses gives

$$\min_{\psi,\theta,\phi}\ \lambda_{\mathrm{AR}}\mathcal{L}_{\mathrm{AR}} + \lambda_v\mathcal{L}_v^{\mathrm{FM}} + \lambda_a\mathcal{L}_a^{\mathrm{FM}} + \lambda_T\mathcal{L}_{\mathrm{PFD}}, \tag{12}$$

with nonnegative loss weights. Stage one uses unit action weights and $\lambda_T = 0$; stage two enables both proposed components. The teacher is updated only by EMA.

4 Experiments

Our experiments address three questions: (i) does reliability-conditioned corruption improve learning from noisy action labels (Section 3.2), (ii) does privileged-future representation distillation provide effective policy supervision (Section 3.3), and (iii) how does the complete EgoForesight compare with other policies on downstream tasks? We evaluate overall performance on RoboCasa and real-world tasks. Our ablations examine the two proposed mechanisms, pretraining-data scaling, the allocation of data across training stages, and the use of future information during generation.

4.1 Model Pretraining

We pretrain in two stages (Figure 3). S1 uses AgiBot World, EgoDex, InternData-A1, and T-Rex; its human videos undergo clip-level screening for visual quality, hand visibility, and global motion. S2 retains these sources and adds 20,000 hours of action-annotated egocentric human data, enabling reliability-conditioned corruption and privileged-future representation distillation after the S1 initialization. All manipulation sources provide action labels.

Our implementation selects a 4B backbone from the Qwen3.5 family (Qwen Team, 2026) and pairs it with our coupled vision and action experts; we use a DINOv3 encoder (Siméoni et al., 2025) as a frozen visual-target extractor. The action expert predicts 30-step chunks of normalized actions in a shared 130-dimensional canonical space, with unavailable coordinates masked. We use Muon (Jordan et al., 2024) for selected hidden-layer weight matrices and auxiliary AdamW for the remaining parameters. We train S1 for 100k updates and S2 for 200k updates, using a global batch size of 2,048

Table 1: **RoboCasa evaluation.** Success rates (%) over 18 pick-and-place (PnP), 6 articulated-object (Art.), and all 24 tasks (Avg.). Blue and green denote full-data and few-shot training; dark and light shading mark the best and second-best score per column within each setting. Baseline variants and sources are detailed in Appendix D.1.

Full data (1,000 demonstrations / task)				*Few-shot (100 demonstrations / task)*			
Method	PnP	Art.	Avg.	Method	PnP	Art.	Avg.
GR00T N1.6	55.4	24.2	47.6	GR00T-Qwen2.5	17.1	36.0	21.8
LDA-1B	55.1	56.3	55.4	GR00T-Qwen2.5-FT	26.9	41.7	30.6
RLDX-1	59.0	57.7	58.7	+ FLARE	47.7	64.7	51.9
JoyAI-RA	60.3	71.7	63.2	DIAL (robot-only)	56.0	65.3	58.3
DIAL (robot-only)	68.9	74.3	70.2	DIAL + EgoDex	60.8	62.0	61.1
EgoForesight (Ours)	**71.9**	**80.0**	**73.9**	**EgoForesight (Ours)**	**64.8**	**69.7**	**66.0**

and 5k learning-rate warmup steps in both stages. Appendix C provides the dataset table, filtering rules, architecture, action layout, and training settings.

4.2 Downstream Evaluation

RoboCasa. We use RoboCasa's simulator and household-task assets (Nasiriany et al., 2024) to evaluate dexterous manipulation with a humanoid robot. Following DIAL's GR1 Tabletop suite (Chen et al., 2026), we consider 24 tasks: 18 pick-and-place (PnP) tasks and 6 articulated-object tasks that combine placement with closing a cabinet, drawer, or microwave. We post-train on 100 demonstrations per task in the few-shot setting and 1,000 per task in the full-data setting, and evaluate each task over 50 episodes. Privileged-future distillation remains enabled in both settings. Appendix D.1 gives the training and evaluation details.

Real world. We evaluate four tasks on a Galaxea R1 Pro with Sharpa Wave hands, as shown in Figure 4. We control only the two arms and hands, totaling 58 DoF, using relative end-effector commands for the arms and joint commands for the hands. For EgoForesight, $\pi_{0.5}$ (Black et al., 2025a), and GR00T N1.7 (NVIDIA, 2026), we jointly post-train one model per method on 100 teleoperated demonstrations per task and evaluate full-task success over 20 trials per task. Appendix D.2 provides protocol details.

4.3 Overall Performance

Dexterous manipulation on RoboCasa. Table 1 shows that EgoForesight reaches 66.0% success with only 100 robot demonstrations per task, improving on robot-only DIAL by **7.7 percentage points** and on DIAL with EgoDex by **4.9 points**. With full data, our success rate rises to 73.9%, exceeding DIAL by 3.7 points. Reducing the robot demonstrations by 90% lowers our success rate by only 7.9 points, compared with 11.9 points for robot-only DIAL. These results highlight effective adaptation to dexterous tasks when robot supervision is limited. The comparison uses published baseline results with different pretraining data and optimization budgets, as detailed in Appendix D.1.

Real-world dexterous manipulation. EgoForesight demonstrates strong overall dexterous manipulation performance in Figure 4, achieving the highest mean success rate of 63.8% and progress score of 84.8. GR00T N1.7 was pretrained on R1 Pro–Sharpa, whereas ours saw only limited Sharpa-hand data on other robots, indicating strong cross-embodiment adaptation under the same downstream demonstration budget. Our advantage is most pronounced on bottle opening and syringe transfer, which demand precise manipulation, with the latter additionally requiring sustained coordination over a long sequence of tool-use stages. For $\pi_{0.5}$, the pretrained checkpoint supports only 32 action dimensions, insufficient for our 58-DoF embodiment, so we replace its input and output projection layers to accommodate the larger action space, as detailed in Appendix D.2. This change alters the pretrained action interface and may affect performance on difficult tasks, underscoring that transferring from grippers to dexterous hands is nontrivial.

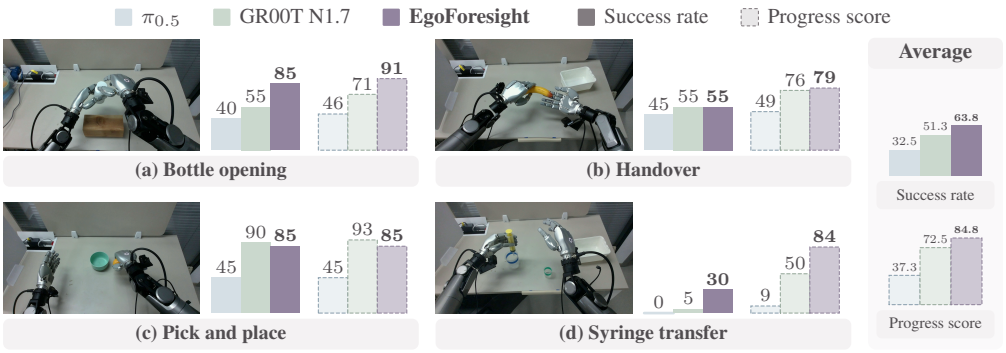


Figure 4: **Real-world dexterous manipulation.** Each task pairs a head-camera image with success rates (%, solid bars) and progress scores (0–100, dashed borders) on the R1 Pro with Sharpa Wave hands. The right column reports the equal-weight mean across the four tasks. The evaluation protocol uses 100 demonstrations and 20 trials per task.

4.4 Ablation

4.4.1 Pretraining

We evaluate pretraining components and data allocation through RoboCasa few-shot adaptation at one tenth of the pretraining data. We additionally vary the pretraining scale and measure real-world open-loop error and RoboCasa closed-loop success. Detailed settings are in Appendix D.3.

Table 2: **Pretraining components.** RoboCasa few-shot success (%) at $1/10$ pretraining scale. Ours combines both components; shading marks column maxima.

Variant	PnP	Art.	Avg.
Baseline	53.4	47.7	52.0
w/ reliability	53.1	58.0	54.3
w/ teacher	55.8	57.7	56.3
Ours	**57.0**	**63.3**	**58.6**

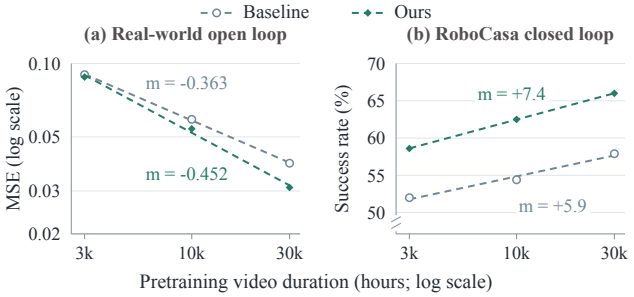


Figure 5: **Pretraining-data scaling.** Dashed lines show fits in the plotted coordinates; m denotes the slope.

Reliability and privileged teacher. Table 2 shows that reliability at $q_d = 0.2$ and the privileged teacher are complementary: individually, they raise reduced-scale average success from 52.0% to 54.3% and 56.3%; together, they reach 58.6%, exceeding either component alone. PnP primarily probes generalization across objects, where the complete method improves success by 3.6 percentage points at reduced scale. Art. additionally demands accurate action generation to manipulate articulated fixtures. Here, reliability alone yields a large gain from 47.7% to 58.0%, despite nearly unchanged PnP performance, and adding the teacher further raises success to 63.3%. This pattern supports assigning different trust to higher-quality robot action labels and noisier human labels, while using privileged future evidence for complementary supervision.

Average success is relatively insensitive to the precise reliability value within the tested range: human $q_d \in \{0.2, 0.5, 0.8\}$ gives 54.0–54.3%, consistently above 52.0% at $q_d = 1$, as shown in Table 9 of Appendix D.3. We therefore use simple source-level reliability rather than introducing task- or sample-specific estimates, which would add complexity.

Pretraining-data scaling. We subsample each pretraining source at $1/10$, $1/3$, and full scale, yielding nested 3k-, 10k-, and 30k-hour subsets with matched source proportions. Each scale follows the same two-stage recipe, retaining S1 replay in S2. Baseline and EgoForesight use identical subsets and optimization budgets at each scale; downstream data and adaptation budgets are fixed.

The baseline disables reliability conditioning and privileged-future distillation; video hours count unique data and exclude repeated sampling or S1 replay. In Figure 5, we fit $\log_{10}(\mathrm{MSE})$ and success percentage against $\log_{10}(h)$ by ordinary least squares over the three nonzero scales. The real-world open-loop MSE slope is -0.452 for ours versus -0.363 for the baseline; RoboCasa closed-loop success increases by 7.4 versus 5.9 percentage points per decade of data. ***Both slope magnitudes are about 25% larger***, indicating stronger returns from added data within the tested range. At 30k hours, MSE is 20.5% lower and success is 8.1 points higher. The MSE comparison uses the baseline's best checkpoint rather than its overfitted endpoint.

Data allocation. Table 3(a) compares S2→S1 ordering, joint S1+S2 training, and S1→S2. Our S1→S2 schedule replays S1 data throughout S2 alongside the additional egocentric data, rather than switching exclusively to the latter. We call the first ordering EgoScale-style (Zheng et al., 2026b) only in the sense that broad human data precede higher-quality aligned data; our stages, datasets, and objectives differ. S1→S2 achieves the highest average success of 58.6%, exceeding reverse ordering by 1.0 point and joint training by 4.3 points. It also yields the best PnP score, although reverse ordering is slightly better on Art. Ending with higher-quality S1 data emphasizes action precision, but may partially overwrite the broader object and scene knowledge acquired during S2, consistent with this trade-off. These results suggest a promising scaling strategy: establish a strong action model with higher-quality data, then expand its coverage using broad human data.

4.4.2 POST-TRAINING

Table 3(b) shows that privileged-future distillation improves average success during post-training from the same full-scale pretrained checkpoint. This supports its ability to learn better action representations during downstream adaptation. Using $\tau_v = \tau_a^{1.5}$ at inference then raises Art. success to 80.0% while retaining 71.9% on PnP, yielding the best average of 73.9%. This supports a vision-leading denoising path consistent with the training-time ordering $\tau_v < \tau_a$. The last two variants share policy weights and integration budget.

Table 3: **Ablation of data allocation and post-training.** RoboCasa success rates (%): **(a)** pretraining order in the few-shot setting; **(b)** teacher distillation and denoising path in the full setting. Blue shading marks the best score per column within each panel.

(a) Data allocation

Pretraining order	PnP	Art.	Avg.
S2 → S1	55.2	64.7	57.6
S1 + S2 (joint)	53.1	58.0	54.3
S1 → S2	57.0	63.3	58.6

(b) Post-training

Teacher distillation	Denoising path	PnP	Art.	Avg.
Off	$\tau_v = \tau_a$	67.0	81.7	70.7
On	$\tau_v = \tau_a$	72.0	75.0	72.8
On	$\tau_v = \tau_a^{1.5}$	71.9	80.0	73.9

5 CONCLUSION

We presented EgoForesight, which makes noisy egocentric experience more useful for policy learning through reliability-aware action supervision and privileged-future representation distillation. Within a joint visual–action flow model, these mechanisms regulate the influence of unreliable action labels while learning complementary structure from demonstrated outcomes. Scaling experiments and component ablations support this combination, with improved returns from added pretraining data and strong downstream performance in simulation and on real hardware. Together, these results suggest that effective scaling depends on supervision quality as well as data volume.

These findings motivate extending outcome-based supervision to videos without action annotations, using self-supervised objectives to learn interaction structure without reconstructed controls. At larger scales, a key question is whether broader experience enables compositional generalization to unseen combinations of objects, skills, and scenes. Studying how supervision quality and data diversity affect this capability could clarify when scaling produces reusable manipulation knowledge rather than better coverage of familiar combinations.

AI USE STATEMENT

Generative AI tools were used to polish the manuscript's writing, create the paper teaser and illustrative icons in the pipeline diagram, assist with coding, and support literature search and review. AI-generated visual content is limited to these illustrative elements and does not include dataset example images. The authors are responsible for the accuracy and integrity of the manuscript, code, references, and reported results.

ETHICS STATEMENT

This work studies policy learning from egocentric human video and robot demonstrations. Such data can raise privacy and data-governance concerns; the internal egocentric data used for pretraining cannot be publicly released due to applicable data-access and disclosure policies. This restriction should not be interpreted as granting permission to redistribute or repurpose these data. Learned robot policies can fail in ways that damage objects or cause injury, and the reported results do not establish safety beyond the evaluated tasks and operating conditions. Any deployment requires task-specific safety assessment, appropriate safeguards, and human oversight.

REPRODUCIBILITY STATEMENT

The main paper and appendix describe the method, training objectives, and two-stage pretraining procedure, together with the post-training settings and evaluation protocols. Full reproduction of pretraining is limited by the internal egocentric data, which cannot be publicly released due to policy restrictions. In contrast, the post-training procedure is fully specified to support reproduction from the pretrained checkpoints. We will open-source the method implementation and release model checkpoints, enabling researchers to reproduce post-training and build on the method without access to the restricted internal egocentric data.

REFERENCES

AgiBot-World Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, Shu Jiang, Yuxin Jiang, Cheng Jing, Hongyang Li, Jialu Li, Chiming Liu, Yi Liu, Yuxiang Lu, Jianlan Luo, Ping Luo, Yao Mu, Yuehan Niu, Yixuan Pan, Jiangmiao Pang, Yu Qiao, Guanghui Ren, Cheng Ruan, Jiaqi Shan, Yongjian Shen, Chengshi Shi, Mingkang Shi, Modi Shi, Chonghao Sima, Jianheng Song, Huijie Wang, Wenhao Wang, Dafeng Wei, Chengen Xie, Guo Xu, Junchi Yan, Cunbiao Yang, Lei Yang, Shukai Yang, Maoqing Yao, Jia Zeng, Chi Zhang, Qinglin Zhang, Bin Zhao, Chengyue Zhao, Jiaqi Zhao, and Jianchao Zhu. AgiBot World Colosseo: A Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3549–3556, 2025. doi: 10.1109/IROS60139.2025.11247088. URL https://ieeexplore.ieee.org/document/11247088.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1692–1717, 2023. URL https://proceedings.mlr.press/v202/bao23a.html.

Joshua Batson and Loic Royer. Noise2Self: Blind Denoising by Self-Supervision. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 524–533. PMLR, 2019. URL https://proceedings.mlr.press/v97/batson19a.html.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A Unified Latent Action World Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35101–35113, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Bi_Motus_A_Unified_Latent_Action_World_Model_CVPR_2026_paper.html.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. In Joseph Lim, Shuran Song, and Hae-Won Park (eds.), *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 2025a. URL https://proceedings.mlr.press/v305/black25a.html.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action Flow Model for General Robot Control. In *Robotics: Science and Systems XXI*, volume 21, 2025b. doi: 10.15607/RSS.2025.XXI.010. URL https://roboticsproceedings.org/rss21/p010.html.

Ashish Bora, Eric Price, and Alexandros G. Dimakis. AmbientGAN: Generative Models from Lossy Measurements. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hy7fDog0b.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S Ryoo, Grecia Salazar, Pannag R Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics Transformer for Real-World Control at Scale. In *Robotics: Science and Systems XIX*, 2023. doi: 10.15607/RSS.2023.XIX.025. URL https://www.roboticsproceedings.org/rss19/p025.html.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to Act Anywhere with Task-centric Latent Actions. In *Robotics: Science and Systems XXI*, 2025. doi: 10.15607/RSS.2025.XXI.014. URL https://www.roboticsproceedings.org/rss21/p014.html.

Remi Cadene, Simon Alibert, Alexander Soare, Quentin Gallouedec, Adil Zouitine, Steven Palma, Pepijn Kooijmans, Michel Aractingi, Mustafa Shukor, Dana Aubakirova, Martino Russi, Francesco Capuano, Caroline Pascal, Jade Choghari, Khalil Meftah, Maxime Ellerbach, Jess Moss, and Thomas Wolf. LeRobot: State-of-the-art machine learning for real-world robotics in PyTorch. https://github.com/huggingface/lerobot, 2024.

Junhao Cai, Zetao Cai, Jiafei Cao, Yilun Chen, Zeyu He, Lei Jiang, Hang Li, Hengjie Li, Yang Li, Yufei Liu, Yanan Lu, Qi Lv, Haoxiang Ma, Jiangmiao Pang, Yu Qiao, Zherui Qiu, Yanqing Shen, Xu Shi, Yang Tian, Bolun Wang, Hanqing Wang, Jiaheng Wang, Tai Wang, Xueyuan Wei, Chao Wu, Yiman Xie, Boyang Xing, Yuqiang Yang, Yuyin Yang, Qiaojun Yu, Feng Yuan, Jia Zeng, Jingjing Zhang, Shenghan Zhang, Shi Zhang, Zhuoma Zhaxi, Bowen Zhou, Yuanzhen Zhou, Yunsong Zhou, Hongrui Zhu, Yangkun Zhu, and Yuchen Zhu. InternVLA-A1: Unifying Understanding, Generation and Action for Robotic Manipulation. *arXiv preprint arXiv:2601.02456*, 2026. URL https://arxiv.org/abs/2601.02456.

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-Supervised Flow Matching for Scalable Multi-Modal Synthesis. In *Forty-Third International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65011. Accepted at ICML 2026; archival proceedings metadata unavailable as of September 23, 2026.

Xinyi Chen, Yilun Chen, Yanwei Fu, Ning Gao, Jiaya Jia, Weiyang Jin, Hao Li, Yao Mu, Jiangmiao Pang, Yu Qiao, Yang Tian, Bin Wang, Bolun Wang, Fangjing Wang, Hanqing Wang, Tai Wang, Ziqin Wang, Xueyuan Wei, Chao Wu, Shuai Yang, Jinhui Ye, Junqiu Yu, Jia Zeng, Jingjing Zhang, Jinyu Zhang, Shi Zhang, Feng Zheng, Bowen Zhou, and Yangkun Zhu. InternVLA-M1: A Spatially Guided Vision-Language-Action Framework for Generalist Robot Policy. *arXiv preprint arXiv:2510.13778*, 2025. URL https://arxiv.org/abs/2510.13778.

Yi Chen, Yuying Ge, Hui Zhou, Mingyu Ding, Yixiao Ge, and Xihui Liu. DIAL: Decoupling Intent and Action via Latent World Modeling for End-to-End VLA. *arXiv preprint arXiv:2603.29844*, 2026. URL https://arxiv.org/abs/2603.29844.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin C. M. Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Robotics: Science and Systems XIX*, 2023. doi: 10.15607/RSS.2023.XIX.026. URL https://www.roboticsproceedings.org/rss19/p026.html.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal Manipulation Interface: In-The-Wild Robot Teaching Without In-The-Wild Robots. In *Robotics: Science and Systems XX*, 2024. doi: 10.15607/RSS.2024.XX.045. URL https://www.roboticsproceedings.org/rss20/p045.html.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alex Dimakis, and Adam Klivans. Ambient Diffusion: Learning Clean Distributions from Corrupted Data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 288–313, 2023. doi: 10.52202/075280-0015. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/012af729c5d14d279581fc8a5db975a1-Abstract-Conference.html.

Giannis Daras, Adrian Rodriguez-Munoz, Adam Klivans, Antonio Torralba, and Constantinos Daskalakis. Ambient Diffusion Omni: Training Good Models with Bad Data. In *Advances in Neural Information Processing Systems*, volume 38, pp. 38064–38101, 2025. doi: 10.52202/085713-1274. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/369d566bd6419aeaf5a02c2be971391e-Abstract-Conference.html.

Danny Driess, Jost Tobias Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Z. Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, and Sergey Levine. Knowledge Insulating Vision-Language-Action Models: Train Fast, Run Fast, Generalize Better. In *Advances in Neural Information Processing Systems*, volume 38, pp. 113969–113990, 2025. doi: 10.52202/085713-3439. URL https://doi.org/10.52202/085713-3439.

Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual Foresight: Model-Based Deep Reinforcement Learning for Vision-Based Robotic Control. *arXiv preprint arXiv:1812.00568*, 2018. URL https://arxiv.org/abs/1812.00568.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An Open-Source Generalist Robot Policy. In *Robotics: Science and Systems XX*, volume 20, 2024. doi: 10.15607/RSS.2024.XX.090. URL https://www.roboticsproceedings.org/rss20/p090.html.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez,

David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022. doi: 10.1109/CVPR52688.2022.01842. URL https://openaccess.thecvf.com/content/CVPR2022/html/Grauman_Ego4D_Around_the_World_in_3000_Hours_of_Egocentric_Video_CVPR_2022_paper.html.

Jun Guo, Qiwei Li, Peiyan Li, Zilong Chen, Nan Sun, Yifei Su, Heyun Wang, Yuan Zhang, Xinghang Li, and Huaping Liu. Unified 4D world action modeling from video priors with asynchronous denoising. *arXiv preprint arXiv:2604.26694*, 2026. URL https://arxiv.org/abs/2604.26694.

Yasaman Haghighi, Bastien van Delft, Mariam Hassan, and Alexandre Alahi. LayerSync: Self-aligning intermediate layers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4itprlvbRQ.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a19744e268754fb0148b017647355b7b-Abstract.html.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning Dexterous Manipulation from Large-Scale Egocentric Video. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=FFxkFMU89E.

Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. Representation Alignment for Diffusion Transformers without External Components. In *International Conference on Learning Representations*, pp. 18884–18915, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/201e691e8a0940291dc591eb0439a789-Abstract-Conference.html.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An Optimizer for Hidden Layers in Neural Networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Siddharth Karamcheti, Suraj Nair, Annie S. Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-Driven Representation Learning for Robotics. In *Robotics: Science and Systems XIX*, volume 19, 2023. doi: 10.15607/RSS.2023.XIX.032. URL https://roboticsproceedings.org/rss19/p032.html.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. EgoMimic: Scaling Imitation Learning via Egocentric Video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233, 2025. doi: 10.1109/ICRA55743.2025.11127989. URL https://doi.org/10.1109/ICRA55743.2025.11127989.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul

Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems XX*, 2024. doi: 10.15607/RSS.2024.XX.120. URL https://www.roboticsproceedings.org/rss20/p120.html.

Dongyoung Kim, Huiwon Jang, Myungkyu Koo, Suhyeok Jang, Taeyoung Kim, Beomjun Kim, Byungjun Yoon, Changsung Jang, Daewon Choi, Dongsu Han, Donguk Lee, Heeseung Kwon, Hojin Jeon, Jaehyun Kang, Jaekyoung Bae, Jihyuk Lee, Jimin Lee, John Won, Joonwoo Ahn, Junhyeong Park, Junyoung Sung, Kyungmin Lee, Minseong Han, Minsung Yoon, Sejune Joo, Seonil Son, Seungcheol Park, Seunggeun Cho, Seungjun Moon, Seungku Kim, Yonghoon Dong, Yongjin Cho, Youngchan Kim, Chang Hwan Kim, Dohyeon Kim, Heecheol Kim, Heewon Lee, Hensen Ahn, Hyungkyu Ryu, Hyunsoo Choi, Hyunsoo Shin, Jaeheon Jung, Jaewoo Kim, Jinwook Kim, Joochul Chang, Joonsoo Kim, Junghun Park, Jungwoo Park, Junho Cho, Junhyeok Park, Junwon Lee, Kangwook Lee, Kwanghoon Kim, Kyoungwhan Choe, Manoj Bhadu, Nayoung Oh, Sangjun Kim, Sangwoo Kim, Seunghoon Shim, Seunghyun Kim, Seungjun Lee, Seungyup Ka, Sungryol Yang, Wook Jung, Yashu Shukla, Yeonjae Lee, Yeonwoo Bae, and Jinwoo Shin. RLDX-1 Technical Report. *arXiv preprint arXiv:2605.03269*, 2026. URL https://arxiv.org/abs/2605.03269.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An Open-Source Vision-Language-Action Model. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713, 2025. URL https://proceedings.mlr.press/v270/kim25c.html.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014. URL https://iclr.cc/archive/2014/conference-proceedings/.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2965–2974. PMLR, 2018. URL https://proceedings.mlr.press/v80/lehtinen18a.html.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for End-to-End Tuning of Latent Diffusion Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18262–18272, 2025. doi: 10.1109/ICCV51701.2025.01697. URL https://openaccess.thecvf.com/content/ICCV2025/html/Leng_REPA-E_Unlocking_VAE_for_End-to-End_Tuning_of_Latent_Diffusion_Transformers_ICCV_2025_paper.html.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-End Training of Deep Visuomotor Policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016. URL https://www.jmlr.org/papers/v17/15-522.html.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*, 2024. URL https://arxiv.org/abs/2408.03326.

Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgExaVtwr.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Luyao Zhang, Mingrui Yu, Zelin Gao, Nan Xue, Boyu Zhou, Xing Zhu, Mingyu Ding, Yujun Shen, and Yinghao Xu. Causal World Modeling for Robot Control. In *Robotics: Science and Systems XXII*, volume 22, 2026. doi: 10.15607/RSS.2026.XXII.016. URL https://www.roboticsproceedings.org/rss22/p016.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is Scalable for LLM Training. *arXiv preprint arXiv:2502.16982*, 2025. URL https://arxiv.org/abs/2502.16982.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Jiangran Lyu, Kai Liu, Xuheng Zhang, Haoran Liao, Yusen Feng, Wenxuan Zhu, Tingrui Shen, Jiayi Chen, Jiazhao Zhang, Yifei Dong, Wenbo Cui, Senmao Qi, Shuo Wang, Yixin Zheng, Mi Yan, Xuesong Shi, Haoran Li, Dongbin Zhao, Ming-Yu Liu, Zhizheng Zhang, Li Yi, Yizhou Wang, and He Wang. LDA-1B: Scaling Latent Dynamics Action Model via Universal Embodied Data Ingestion. In *Robotics: Science and Systems XXII*, 2026. doi: 10.15607/RSS.2026.XXII.210. URL https://www.roboticsproceedings.org/rss22/p210.html.

Haoxiang Ma, Junhao Cai, Xiaoxu Xu, Hao Li, Yuyin Yang, Yang Tian, Jiafei Cao, Hongrui Zhu, Zherui Qiu, Zhaxizhuoma, Yuqiang Yang, Jiaqi Peng, Xueyuan Wei, Yangkun Zhu, Jiahao Jiang, Xing Gao, Hanqing Wang, Feng Yuan, Kailin Li, Xueyue Zhu, Tai Wang, Yan Ding, Jiangmiao Pang, Jia Zeng, Jingjing Zhang, Bowen Zhou, Yao Mu, Chunhua Shen, and Weinan Zhang. InternVLA-A1.5: Unifying Understanding, Latent Foresight, and Action for Compositional Generalization. *arXiv preprint arXiv:2607.04988*, 2026. URL https://arxiv.org/abs/2607.04988.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1820–1864. PMLR, 2023. URL https://proceedings.mlr.press/v229/mandlekar23a.html.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A Universal Visual Representation for Robot Manipulation. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 892–909. PMLR, 2023. URL https://proceedings.mlr.press/v205/nair23a.html.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-Scale Simulation of Household Tasks for Generalist Robots. In *Robotics: Science and Systems XX*, volume 20, 2024. doi: 10.15607/RSS.2024.XX.050. URL https://roboticsproceedings.org/rss20/p050.html.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with Noisy Labels. In *Advances in Neural Information Processing Systems*, volume 26, pp. 1196–1204, 2013. URL https://papers.nips.cc/paper/5073-learning-with-noisy-labels.pdf.

Dantong Niu, Zhuoyang Liu, Zekai Wang, Boning Shao, Zhao-Heng Yin, Anirudh Pai, Yuvan Sharma, Stefano Saravalle, Ruijie Zheng, Jing Wang, Ryan Punamiya, Mengda Xu, Yuqi Xie, Yunfan Jiang, Letian Fu, Konstantinos Kallidromitis, Matteo Gioia, Junyi Zhang, Jiaxin Ge, Haiwen Feng, Fabio Galasso, Wei Zhan, David M. Chan, Yutong Bai, Roei Herzig, Jiahui Lei, Fei-Fei Li, Ken Goldberg, Jitendra Malik, Pieter Abbeel, Yuke Zhu, Danfei Xu, Linxi Fan, and Trevor

Darrell. T-Rex: Tactile-reactive dexterous manipulation. *arXiv preprint arXiv:2606.17055*, 2026. URL https://arxiv.org/abs/2606.17055.

NVIDIA. GR00T N1.6. NVIDIA Research, 2025. URL https://research.nvidia.com/labs/gear/gr00t-n1_6/.

NVIDIA. NVIDIA Isaac GR00T N1.7. Official model and reference implementation, 2026. URL https://github.com/NVIDIA/Isaac-GR00T. Accessed September 23, 2026.

Open X-Embodiment Collaboration. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477. URL https://doi.org/10.1109/ICRA57147.2024.10611477.

Maximus A. Pace, Prithwish Dan, Chuanruo Ning, Atiksh Bhardwaj, Audrey Du, Edward W. Duan, Wei-Chiu Ma, and Kushal Kedia. X-Diffusion: Training diffusion policies on cross-embodiment human demonstrations. *arXiv preprint arXiv:2511.04671*, 2025. URL https://arxiv.org/abs/2511.04671.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Patrini_Making_Deep_Neural_CVPR_2017_paper.html.

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. In *Robotics: Science and Systems XXI*, 2025. doi: 10.15607/RSS.2025.XXI.012. URL https://www.roboticsproceedings.org/rss21/p012.html.

Physical Intelligence, Bo Ai, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Greg Balke, Kevin Black, George Bokinsky, Shihao Cao, Thomas Charbonnier, Vedant Choudhary, Foster Collins, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Maitrayee Dhaka, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachlan Groom, Haroun Habeeb, Hunter Hancock, Karol Hausman, Gashon Hussein, Victor Hwang, Brian Ichter, Connor Jacobsen, Szymon Jakubczak, Rowan Jen, Tim Jones, Gregg Kammerer, Ben Katz, Liyiming Ke, Mairbek Khadikov, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Brendon LeCount, Sergey Levine, Xinyu Li, Adrian Li-Bell, Vladislav Lialin, Zhonglin Liang, Wallace Lim, Yao Lu, Enyu Luo, Vishnu Mano, Nandan Marwaha, Aikys Mongush, Liam Murphy, Suraj Nair, Tyler Patterson, Karl Pertsch, Allen Z. Ren, Gavin Schelske, Charvi Sharma, Baifeng Shi, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Jiaming Tang, Jimmy Tanner, Shalom Tekeste, Marcel Torne, Kyle Vedder, Quan Vuong, Anna Walling, Haohuan Wang, Jason Wang, XuDong Wang, Chris Whalen, Samuel Whitmore, Blake Williams, Charles Xu, Sukwon Yoo, Lili Yu, Wuming Zhang, Zhuoyang Zhang, and Ury Zhilinsky. $\pi_{0.7}$: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities. *arXiv preprint arXiv:2604.15483*, 2026. URL https://arxiv.org/abs/2604.15483.

Shiquan Qiu. AnyDexRetarget. Open-source software, 2025. URL https://github.com/qqsq12321/AnyDexRetarget. Revision a76a0e1.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Arijit Ray, Jiafei Duan, Ellis L Brown II, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Dynamic Spatial Aptitude Training for Multimodal Language Models. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=DW8U8ZWa1U.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to Reweight Examples for Robust Deep Learning. In *Proceedings of the 35th International Conference on Machine*

Learning, volume 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343, 2018. URL https://proceedings.mlr.press/v80/ren18a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026. URL https://arxiv.org/abs/2605.18324.

Yue Su, Sijin Chen, Haixin Shi, Mingyu Liu, Zhengshen Zhang, Ningyuan Huang, Weiheng Zhong, Zhengbang Zhu, Yuxiao Liu, and Xihui Liu. World Guidance: World Modeling in Condition Space for Action Generation. In *Forty-Third International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61757. Accepted at ICML 2026; archival proceedings metadata unavailable as of September 23, 2026.

Peijun Tang, Shangjin Xie, Baifu Huang, Binyan Sun, Haotian Yang, Kuncheng Luo, Weiqi Jin, Shilin Fang, and Jianan Wang. Towards Predictive, Aligned, and Scalable Robot Learning. *arXiv preprint arXiv:2607.11270*, 2026. URL https://arxiv.org/abs/2607.11270.

Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, Yaping Li, Ping Wang, Junhao Cai, Jia Zeng, Hao Dong, and Jiangmiao Pang. 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. URL https://openaccess.thecvf.com/content/CVPR2026/html/Tian_InternData-A1_Pioneering_High-Fidelity_Synthetic_Data_for_Pre-training_Generalist_Policy_CVPR_2026_paper.html.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, 11(110):3371–3408, 2010. URL https://www.jmlr.org/papers/v11/vincent10a.html.

Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A Dataset for Robot Learning at Scale. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1723–1736. PMLR, 2023. URL https://openreview.net/forum?id=f55MlAT1Lu.

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C. Karen Liu. DexCap: Scalable and portable mocap data collection system for dexterous manipulation. In *Robotics: Science and Systems XX*, volume 20, 2024. doi: 10.15607/RSS.2024.XX.043. URL https://www.roboticsproceedings.org/rss20/p043.html.

Mengmeng Wang, Dengyang Jiang, Liuzhuozheng Li, Yucheng Lin, Guojiang Shen, Xiangjie Kong, Yong Liu, Guang Dai, and Jingdong Wang. SRA 2: Variational Autoencoder Self-Representation Alignment for Efficient Diffusion Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32978–32987, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_SRA_2_Variational_Autoencoder_Self-Representation_Alignment__for_Efficient_Diffusion_CVPR_2026_paper.html.

Adam Wei, Nicholas Pfaff, Thomas Cohn, Arif Kerem Dayı, Constantinos Daskalakis, Giannis Daras, and Russ Tedrake. Ambient diffusion policy: Imitation learning from suboptimal data in robotics. *arXiv preprint arXiv:2606.12365*, 2026. URL https://arxiv.org/abs/2606.12365.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/2c37c5bcef24b9541550261dcd63261b-Paper-Conference.pdf.

Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, Yiyu Ren, Kejia Zhang, Hui Yu, Jingmei Zhao, Shuai Zhou, Zhenqi Qiu, Houlong Xiong, Ziyu Wang, Zechen Wang, Ran Cheng, Yong-Lu Li, Yongtao Huang, Xing Zhu, Yujun Shen, and Kecheng Zheng. A Pragmatic VLA Foundation Model. *arXiv preprint arXiv:2601.18692*, 2026. URL https://arxiv.org/abs/2601.18692.

Chendong Xin, Mingrui Yu, Yongpeng Jiang, Zhefeng Zhang, and Xiang Li. Analyzing Key Objectives in Human-to-Robot Retargeting for Dexterous Manipulation. *IEEE Robotics and Automation Practice*, 1:29–34, 2026. doi: 10.1109/RAP.2026.3656110. URL https://doi.org/10.1109/RAP.2026.3656110.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VYOe2eBQeh.

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

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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DJSZGGZYVi.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. RoboPoint: A Vision-Language Model for Spatial Affordance Prediction in Robotics. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4005–4020. PMLR, 2025. URL https://proceedings.mlr.press/v270/yuan25c.html.

Chubin Zhang, Jianan Wang, Zifeng Gao, Yue Su, Tianru Dai, Cai Zhou, Jiwen Lu, and Yansong Tang. CLAP: Contrastive latent action pretraining for learning vision-language-action models from human videos. *arXiv preprint arXiv:2601.04061*, 2026a. URL https://arxiv.org/abs/2601.04061.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. HaWoR: World-Space Hand Motion Reconstruction from Egocentric Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1805–1815, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_HaWoR_World-Space_Hand_Motion_Reconstruction_from_Egocentric_Videos_CVPR_2025_paper.html.

Tianle Zhang, Zhihao Yuan, Dafeng Chi, Peidong Liu, Dongwei Li, Kejun Hu, Likui Zhang, Junnan Nie, Ziming Wei, Zengjue Chen, Yili Tang, Jiayi Li, Zhiyuan Xiang, Mingyang Li, Tianci Luo, Hanwen Wan, Ao Li, Linbo Zhai, Zhihao Zhan, Xiaodong Bai, Jiakun Cai, Peng Cao, Kangliang

Chen, Siang Chen, Yixiang Dai, Shuai Di, Yicheng Gong, Chenguang Gui, Yucheng Guo, Peng Hao, Qingrong He, Haoyang Huang, Kunrui Huang, Zhixuan Huang, Shibo Jin, Yixiang Jin, Anson Li, Dongjiang Li, Jiawei Li, Ruodai Li, Yihang Li, Yuzhen Li, Jiaming Liang, Fangsheng Liu, Jing Long, Mingxi Luo, Xing Pan, Hui Shen, Xiaomeng Tian, Daming Wang, Song Wang, Junwu Xiong, Hang Xu, Wanting Xu, Zhengcheng Yu, He Zhang, Jiyao Zhang, Lin Zhao, Chen Zhou, Nan Duan, Yuzheng Zhuang, and Liang Lin. JoyAI-RA 0.1: A Foundation Model for Robotic Autonomy. *arXiv preprint arXiv:2604.20100*, 2026b. URL https://arxiv.org/abs/2604.20100.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=0u1LigJaab.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. FLARE: Robot Learning with Implicit World Modeling. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3952–3971. PMLR, 2025. URL https://proceedings.mlr.press/v305/zheng25a.html.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. EgoScale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026b. URL https://arxiv.org/abs/2602.16710.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified World Models: Coupling Video and Action Diffusion for Pretraining on Large Robotic Datasets. In *Robotics: Science and Systems XXI*, 2025. doi: 10.15607/RSS.2025.XXI.015. URL https://www.roboticsproceedings.org/rss21/p015.html.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

A RELIABILITY-CONDITIONED SAMPLING

This appendix specifies the corruption distribution $\mathcal{S}(q_d)$ used in Sections 3.2–3.3. It combines ordinary examples from $\mathcal{S}_{\mathrm{reg}}(q_d)$, with a strictly positive vision–action time gap, and clean-future boundary examples.

Ordinary examples. For source reliability $q_d \in [0,1]$, use the mapping in Equation (4), with $\alpha_{\min} = 0.5$ and $\alpha_{\max} = 3$. Independently sample

$$\tau_v \sim \mathcal{U}(0,1), \qquad g \sim \mathrm{Beta}(\alpha_d, 2), \qquad \tau_a = \tau_v + (1-\tau_v)g. \tag{13}$$

Since $0 < g < 1$ almost surely, this rule gives $0 < \tau_v < \tau_a < 1$ almost surely and assigns no point mass to $\tau_v = \tau_a$. At fixed τ_v, increasing α_d shifts the action time upward in first-order stochastic order. To see this, for $\alpha_2 > \alpha_1 > 0$, the ratio of the two gap densities is proportional to $g^{\alpha_2-\alpha_1}$, which is increasing on $(0,1)$. The affine map from g to τ_a preserves this ordering. Hence decreasing reliability increases action corruption while leaving the vision-time distribution unchanged.

Action-time marginal. Write $\alpha = \alpha_d$. The density of g is $\alpha(\alpha+1)g^{\alpha-1}(1-g)$. Changing variables from g to τ_a gives the joint density

$$f_{\tau_v,\tau_a}(v,a \mid d) = \frac{\alpha(\alpha+1)(a-v)^{\alpha-1}(1-a)}{(1-v)^{\alpha+1}}, \qquad 0 < v < a < 1. \tag{14}$$

For $0 < a < 1$, integrate over v and substitute $r = (a-v)/(1-v)$:

$$\begin{aligned} f_{\tau_a}(a \mid d) &= \int_0^a \frac{\alpha(\alpha+1)(a-v)^{\alpha-1}(1-a)}{(1-v)^{\alpha+1}}\, dv \\ &= \alpha(\alpha+1)\int_0^a r^{\alpha-1}\, dr = (\alpha+1)a^{\alpha}. \end{aligned} \tag{15}$$

Thus, for ordinary examples,

$$\begin{aligned} \tau_a \mid d &\sim \mathrm{Beta}(\alpha_d+1, 1), \\ \mathbb{E}[\tau_a \mid d] = \frac{\alpha_d+1}{\alpha_d+2}, &\qquad \mathbb{E}[\tau_a - \tau_v \mid d] = \frac{\alpha_d}{2(\alpha_d+2)}. \end{aligned} \tag{16}$$

The last identity also follows directly from independence of g and τ_v. Both the mean action time and the expected gap increase as reliability decreases.

Boundary examples and the complete distribution. Independently select a boundary indicator $e \sim \mathrm{Bernoulli}(p_{\mathrm{edge}})$. When $e = 0$, use Equation (13); when $e = 1$, sample

$$\tau_v = 0, \qquad u \sim \mathcal{U}(0,1), \qquad \tau_a = u^{1/\alpha_d} \sim \mathrm{Beta}(\alpha_d, 1). \tag{17}$$

Consequently, the full training distribution has marginals

$$\begin{aligned} \tau_v \mid d &\sim (1-p_{\mathrm{edge}})\mathcal{U}(0,1) + p_{\mathrm{edge}}\delta_0, \\ \tau_a \mid d &\sim (1-p_{\mathrm{edge}})\,\mathrm{Beta}(\alpha_d+1, 1) + p_{\mathrm{edge}}\,\mathrm{Beta}(\alpha_d, 1), \end{aligned} \tag{18}$$

where δ_0 is a point mass at zero. The Beta marginal in Equation (16) therefore applies only to ordinary examples. The shared p_{edge} preserves the source-independent vision marginal, and the action marginal remains stochastically increasing as reliability decreases. The boundary action distribution has positive density throughout $(0,1)$, covering the action times used by the privileged teacher without requiring an identical density. Boundary examples retain valid-action flow and representation supervision, but their vision flow loss is masked out. This calibrates action prediction conditioned on the observed future.

A.1 POPULATION INTERPRETATION OF ACTION CORRUPTION

We analyze the action flow loss at fixed flow times and conditioning $z = (c, x^v_{\tau_v}, \tau_v, M_a)$, including the partially corrupted future available to the action expert. For this calculation, write $y_a = y^0_a + \eta$, with $y^0_a \mid z \sim \mathcal{N}(m, \Sigma_p)$ and $\eta \mid z \sim \mathcal{N}(0, \Sigma_\eta)$ independent conditional on z. The reference and annotation laws, denoted by p and r, share the mean m, with $\Sigma_r = \Sigma_p + \Sigma_\eta$. All expressions restrict to valid action coordinates, and the interpolation noise ϵ_a is independent of both y^0_a and η.

Conditional flow field. Squared regression has the conditional expectation as its population minimizer. Using $\epsilon_a - y_a = (x^a_{\tau_a} - y_a)/\tau_a$, the optimal action field under r is

$$v^\star_r(x, \tau_a; z) = \mathbb{E}_r[\epsilon_a - y_a \mid x^a_{\tau_a} = x, z] = \frac{x - \mathbb{E}_r[y_a \mid x^a_{\tau_a} = x, z]}{\tau_a}, \tag{19}$$

for $\tau_a > 0$. The corresponding expression holds for p. Thus, the field comparison depends on the posterior action means under the two laws.

Noise-dependent agreement. Write $a = 1 - \tau_a$, $w = x - am$, and $C_s = a^2\Sigma_s + \tau_a^2 I$ for $s \in \{p, r\}$. Gaussian conditioning gives the posterior action mean

$$m_s(x, \tau_a; z) = m + a\Sigma_s C_s^{-1} w. \tag{20}$$

Substituting into Equation (19) and applying the inverse-matrix identity gives

$$\begin{aligned} v^\star_r - v^\star_p &= \frac{m_p - m_r}{\tau_a} = -\tau_a a\, C_r^{-1}\Sigma_\eta C_p^{-1} w, \\ \|v^\star_r - v^\star_p\|_2 &\le \frac{1 - \tau_a}{\tau_a^3}\|\Sigma_\eta\|_{\mathrm{op}}\|w\|_2. \end{aligned} \tag{21}$$

The bound follows from $C_s \succeq \tau_a^2 I$, with $\|\cdot\|_{\mathrm{op}}$ denoting the operator norm. For fixed x, z, the field discrepancy vanishes as $\tau_a \to 1$, even though individual velocity targets retain annotation variation. This manuscript-specific bound gives a conditional-flow interpretation for our setting. Its high-noise contraction is consistent with ADP's noise-dependent use of suboptimal data (Wei et al., 2026) and motivates our choice to assign less reliable sources stronger action corruption. The weight q_d complements this allocation by controlling each source's contribution to the shared training objective.

B Masked objectives and teacher implementation

Action validity. All manipulation examples have action labels. Let $M_{a,i,kj} \in \{0, 1\}$ identify a valid action coordinate j at step k of example i, excluding padding. Padded coordinates remain zero after corruption. The full action expressions underlying Equation (2) are

$$x^a_{\tau_a,i} = M_{a,i} \odot \big[(1 - \tau_a)y_{a,i} + \tau_a\epsilon_{a,i}\big], \qquad u^\star_{a,i} = M_{a,i} \odot (\epsilon_{a,i} - y_{a,i}). \tag{22}$$

An action token is valid when at least one of its coordinates is annotated: $b_{i,k} = \mathbf{1}\{\sum_j M_{a,i,kj} > 0\}$. Tokens with $b_{i,k} = 0$ are padding and are excluded from action attention and representation alignment.

Loss reductions. The MSE notation in Equation (11) suppresses padding and batch reductions: flow losses average over valid coordinates, and representation loss averages over valid action tokens. With source reliability applied inside the action average, the sampled losses for a manipulation batch $\mathcal{B}$ are

$$\widehat{\mathcal{L}}^{\mathrm{FM}}_v = \frac{\sum_{i\in\mathcal{B}}(1 - e_i)\|\widehat{u}_{v,i} - u^\star_{v,i}\|_F^2}{N_v d_v \max(1, \sum_{i\in\mathcal{B}}(1 - e_i))}, \tag{23}$$

$$\widehat{\mathcal{L}}^{\mathrm{FM}}_a = \frac{\sum_{i,k,j} q_{d_i} M_{a,i,kj}(\widehat{u}_{a,i,kj} - u^\star_{a,i,kj})^2}{\sum_{i,k,j} M_{a,i,kj}}, \tag{24}$$

$$\widehat{\mathcal{L}}_{\mathrm{PFD}} = \frac{\sum_{i,k} b_{i,k}\ell_{\mathrm{PFD},i,k}}{\sum_{i,k} b_{i,k}}. \tag{25}$$

Here e_i marks boundary examples. Vision loss averages only over ordinary examples and vanishes for an all-boundary batch. The $\mathcal{L}$'s in Equation (12) average over batches, Gaussian noise, and corruption times. Direct action loss normalizes by valid coordinates, retaining the effect of q_d in single-source batches. Distillation averages over valid tokens without reliability weighting, including boundary examples. The autoregressive cross-entropy $\mathcal{L}_{\mathrm{AR}}$ uses separately interleaved vision–language batches.

Teacher construction. The EMA covers the coupled vision–action module and its conditioning projections, excluding the VLM. The teacher reuses detached contextual, current visual, and state features, runs without gradients, and maintains FP32 parameters. For each example, we reconstruct the action input with the student's original ϵ_a, keeping the input consistent.

The action projection P_ϕ is a trainable MLP. Cosine similarity in Equation (10) is evaluated with norm stabilization,

$$\cos_\varepsilon(u, v) = \frac{u^\top v}{\max(\|u\|_2, \varepsilon)\max(\|v\|_2, \varepsilon)}, \qquad \varepsilon = 10^{-8}. \tag{26}$$

The current configuration uses $\mu = 0.999$, $p_{\text{edge}} = 0.02$, student layer $\ell_s = 5$, teacher layer $\ell_t = 11$, and a projection hidden width of 3072. Stage one uses synchronized times $\tau_v = \tau_a \sim \text{Beta}(1.5, 1)$, unit action-loss weights, and no boundary examples or teacher. Its coupled-expert weights initialize both the stage-two student and teacher.

C Model pretraining details

C.1 Data and quality filtering

Table 4 lists the manipulation and VLM sources. S1 combines real and simulated robot demonstrations with curated human trajectories. S2 retains this mixture and adds the larger internal egocentric collection. Human samples pair observed motion with reconstructed wrist/fingertip trajectories or retargeted hand commands. Task and subtask annotations form the language instruction, together with available control-mode information. The VLM data comprise LLaVA-OneVision and the RoboPoint and SAT subsets released through Molmo2-ER; separate batches from these sources provide $\mathcal{L}_{\text{AR}}$ in both stages.

Table 4: **Pretraining sources and data allocation.** Hours denote the retained manipulation data after filtering.

Dataset	Retained hours	Supervision used in this work
Stage 1: Higher-quality initialization		
AgiBot World (2025)[a]	2,392	EEF, arm joints, gripper, head/waist, selected base commands
EgoDex[b]	715	Wrist poses, fingertips, and retargeted gripper/hand targets
InternData-A1[c]	7,330	Simulated single-/dual-arm EEF, joint, and gripper targets
T-Rex[d]	50	Bimanual and dexterous-hand joints
Stage 2: Human-video expansion		
All Stage 1 sources	10,487	Same action annotations as S1
Internal Ego-Centric Dataset	20,000	Bimanual EEF and fingertip targets
VLM data: Shared across S1 and S2		
LLaVA-OneVision[e]	—	Visual question answering and instruction following
Molmo2-ER-RoboPoint[f]	—	Spatial affordance prediction and visual grounding
Molmo2-ER-SAT[g]	—	Spatial reasoning question answering

[a]AgiBot-World Contributors et al. (2025); [b]Hoque et al. (2026); [c]Tian et al. (2026); [d]Niu et al. (2026). [e]Li et al. (2024); [f]Yuan et al. (2025); [g]Ray et al. (2025).

Hours and selected action fields are outputs of our filtering and preprocessing rather than totals or schemas reported by the cited sources; unavailable coordinates are masked. VLM data supervise autoregressive text prediction through $\mathcal{L}_{\text{AR}}$; hours are not applicable. The Molmo2-ER row names refer to our selected re-releases of RoboPoint and SAT subsets, while the citations establish only the original datasets.

We screen the human videos used for S1 with a lightweight clip-level pipeline (Figure 6). Videos are divided into short clips and sparsely sampled to reject decoding failures, insufficient duration, excessive black or frozen frames, and severe blur. WiLoR hand detections measure the fraction of frames with visible hands, while frame-to-frame motion estimates identify excessive global movement. We retain clips with sufficient hand visibility and acceptable visual quality and motion, and

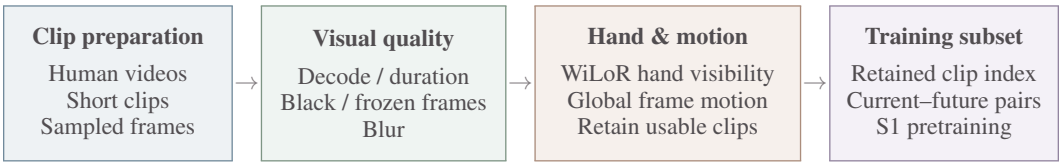


Figure 6: **Human-video filtering for S1.** Clips that pass visual-quality, hand-visibility, and global-motion checks define the retained time intervals from which training pairs are constructed.

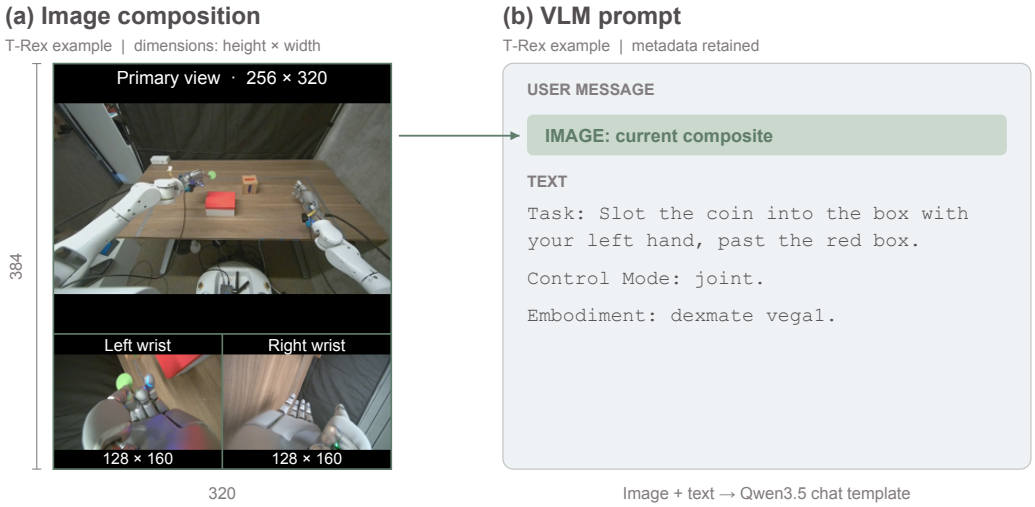


Figure 7: **Shared input format.** Input format across data sources, illustrated with a T-Rex example. Left: the primary view occupies the upper tile, with left and right wrist views below; each view is resized with its aspect ratio preserved and centered on black padding. Dimensions are height × width. Right: the VLM user message contains the composite image followed by task instructions and available metadata. Image labels and borders are explanatory overlays; prompt line breaks are added for readability.

export their time intervals as a dataset index. Training samples use current and future frames from these retained regions. This screening improves visual usability while leaving pose-estimation and retargeting noise to be handled during training.

Available action annotations are prepared before pretraining: for EgoDex, we construct wrist, fingertip, and retargeted hand targets, with the dexterous retargeting procedure detailed in Appendix C.6; robot sources provide EEF, joint, and gripper targets as applicable. Each source populates its corresponding canonical coordinates throughout training.

C.2 Model architecture

In our architecture, a 4B checkpoint from the Qwen3.5 family (Qwen Team, 2026) provides visual–language context to our 16-block coupled vision–action module with hidden width 1,536 and 24 attention heads. The two streams have separate projections and feed-forward networks, exchange information through joint attention, and use modality-specific flow-time conditioning.

Images are resized or composed from available camera views at 384×320 resolution. We use a frozen DINOv3 ViT-B/16 encoder (Siméoni et al., 2025); in our preprocessing, its 768-dimensional patch features are pooled to 120 tokens per image. The vision expert predicts future features and the action expert predicts 30-step action chunks. Conditioning uses the current observation, instruction, state, and source/control metadata, without observation history. The teacher follows Appendix B.

C.3 INPUT FORMAT

We use a shared input format across data sources, illustrated in Figure 7 with a T-Rex example. A 256×320 primary-view tile (e.g., a head or egocentric view) sits above two 128×160 wrist-view tiles, yielding a 384×320 composite (height $\times$ width). Views preserve their aspect ratios through centered black padding; unavailable wrist tiles remain black. The current composite precedes the instruction text in a user message processed by the Qwen3.5 chat template. The prompt combines task/subtask instructions with available control-mode, embodiment, and reference-frame metadata; the fields and their values depend on the source. Numeric state values enter the coupled module through a separate state encoder.

C.4 CANONICAL ACTION SPACE

All sources share a 130-dimensional interface for normalized continuous actions. Table 5 lists the used coordinates; unavailable coordinates and padded timesteps are masked. The current state uses the same layout with its own validity mask.

Table 5: **Canonical action space.** Used coordinates in the 130-dimensional canonical action space. Indices are zero-based and ranges are inclusive. Availability is source-specific; unused slots and unused capacity within a slot are omitted.

Field	Indices	Target convention
Arm joints	L: 0–6; R: 29–35	Up to 7 absolute joint positions per arm
EEF motion	L: 7–12; R: 36–41	3D translation and 3D rotation vector, relative to the current pose in camera coordinates
Gripper	L: 13; R: 42	One opening/control scalar per gripper
Fingertips	L: 14–28; R: 43–57	Five 3D positions per hand: thumb, index, middle, ring, little finger
Base velocity	58, 63	Forward velocity and yaw rate for selected AgiBot tasks
Head / waist	64–65 / 68–69	Two absolute coordinates each, in source order
Dexterous joints	L: 80–101; R: 105–126	22 absolute Sharpa joint angles per hand; shared order for human retargeting and T-Rex

EEF targets express motion relative to the current end-effector state in the current observation-camera frame. Let ${}^{A}T_{B}$ map coordinates from frame B to frame A, and let C_t be the camera frame at the current observation. We convert camera trajectories to a shared camera-to-world convention ${}^{W}T_{C_s}$. For an EEF target at time s stored in its contemporaneous camera frame, camera-motion compensation transports the target into C_t:

$$\bar{T}_s = \left({}^{W}T_{C_t}\right)^{-1} {}^{W}T_{C_s}\, {}^{C_s}T_{E_s} = \begin{bmatrix} \bar{R}_s & \bar{p}_s \\ 0 & 1 \end{bmatrix}. \tag{27}$$

For robot targets already expressed in a common world or robot-base frame W, we instead use $\bar{T}_s = ({}^{W}T_{C_t})^{-1} {}^{W}T_{E_s}$. Let $(R_t^{\mathrm{ref}}, p_t^{\mathrm{ref}})$ denote the current EEF state in C_t. The six-dimensional motion target is

$$\Delta p_{t,s} = \bar{p}_s - p_t^{\mathrm{ref}}, \qquad \Delta r_{t,s} = \mathrm{Log}\big(\bar{R}_s (R_t^{\mathrm{ref}})^{\top}\big)^{\vee}, \tag{28}$$

where $\mathrm{Log}(\cdot)^{\vee}$ yields an axis–angle rotation vector. All targets in a chunk share the current EEF reference and camera axes. With consistent pose estimates, this removes apparent EEF motion caused solely by camera movement; for a static camera, the camera-to-camera transform in Equation (27) is the identity. Future camera poses are used only to construct training labels.

Joint, gripper, and fingertip targets retain their absolute conventions before normalization. We align images and actions in time, construct the EEF increments in physical units, and then normalize actions per dataset using percentile or bounded-range statistics, mapping valid coordinates to $[-1, 1]$.

C.5 TRACKING AND RETARGETING ERRORS

Egocentric action labels depend on both camera tracking and hand reconstruction. For monocular video, moving hands and objects can interfere with camera-motion estimation, and recovering met-

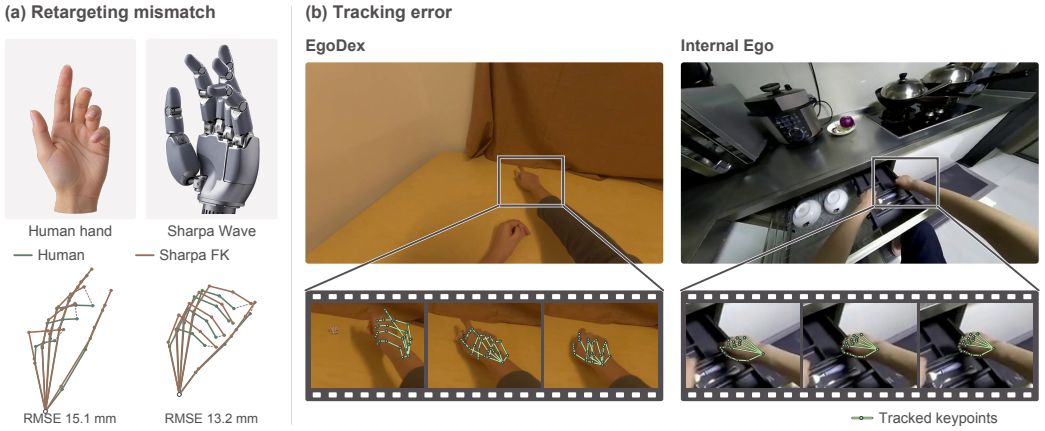


Figure 8: **Retargeting mismatch and tracking errors.** **(a)** Human–robot morphological differences lead to residual geometric mismatch when retargeting human motions to Sharpa Wave. **(b)** Tracking annotations can deviate from visible hand poses in both EgoDex and Internal Ego. Both effects introduce noise into egocentric action supervision.

ric scale can introduce additional errors; camera-pose errors then propagate into hand trajectories expressed in a common frame (Zhang et al., 2025). Hand tracking also degrades under occlusion and rapid motion. Calibrated multi-camera capture, as in EgoDex, provides multi-view geometric information, but EgoDex still reports imperfect hand annotations under heavy occlusion and rapid motion (Hoque et al., 2026). The resulting annotation noise therefore depends on the capture and annotation pipeline.

Retargeting introduces a further geometric mismatch even when the human pose is accurately known. Human and robot hands differ in palm size, finger-length proportions, and joint structure, so the same wrist pose, fingertip positions, and finger orientations need not be jointly realizable by the robot (Wang et al., 2024; Xin et al., 2026). Retargeting must compromise among these objectives; matching fingertips alone need not preserve the hand's contact geometry. This morphology-induced error is separate from upstream tracking errors and from the remaining visual and dynamical domain gap. Figure 8 illustrates retargeting mismatch and visible tracking errors.

C.6 Retargeting implementation

Starting from EgoDex's tracked wrist and fingertip data (Hoque et al., 2026), we retarget each hand to 22 absolute Sharpa Wave joint angles. This paper-specific conversion follows the joint-space representation used in EgoScale (Zheng et al., 2026b); our implementation uses the adaptive optimizer in AnyDexRetarget (Qiu, 2025), with its AVP-to-Sharpa configuration at revision `a76a0e1`. The two hands are processed independently, and optimizer and filter states are reset at every episode boundary.

Keypoints and coordinate alignment. From the 25 tracked transforms per hand, we select 21 positions: the wrist and four points along each finger, ordered thumb, index, middle, ring, and little finger. The four non-thumb metacarpal entries are omitted. AnyDexRetarget subtracts the wrist position, estimates a local frame from the wrist and the index and middle finger bases, and applies its handedness-specific axis convention. We then apply an additional rotation about the local x-axis of $+7^\circ$ for the left hand and -7° for the right hand. Finger fitting uses these positions and the distal segment directions derived from them. Tracked wrist poses are retained separately for the camera-frame EEF construction in Appendix C.4.

Adaptive geometric objective. Let $\theta \in \mathbb{R}^{22}$ be the robot joint angles, and let $v_{f,k}(\theta)$ denote the forward-kinematic vector from the wrist to point k on finger f. We fit the two intermediate points and fingertip, indexed by $k \in \{\mathrm{P}, \mathrm{D}, \mathrm{T}\}$. Human targets $\widehat{v}_{f,k}$ are constructed by accumulating scaled finger segments from the wrist; the wrist-to-finger-base vector is left unscaled. In thumb-to-little-

finger order, the three successive segment multipliers are

$$S = \begin{bmatrix} 1.251 & 1.259 & 1.214 \\ 1.108 & 1.148 & 1.160 \\ 1.101 & 1.115 & 1.128 \\ 1.117 & 1.141 & 1.155 \\ 1.212 & 1.290 & 1.305 \end{bmatrix}. \tag{29}$$

These are the fixed scales in the upstream AVP-to-Sharpa configuration. Let d_f be the human thumb-to-fingertip distance in centimeters. For each non-thumb finger, the blend weight is

$$\beta_f = \mathrm{clip}\left(\frac{4 - d_f}{2}, 0, 0.7\right), \qquad \beta_{\mathrm{th}} = \max_{f \neq \mathrm{th}} \beta_f. \tag{30}$$

Let $u_f(\theta)$ and $\widehat{u}_f$ be unit distal-segment directions. Using the Huber penalty $h_\delta(r) = r^2/2$ for $r \leq \delta$ and $h_\delta(r) = \delta(r - \delta/2)$ otherwise, define

$$L_f^{\mathrm{hand}}(\theta) = \frac{1}{3} \sum_{k \in \{\mathrm{P,D,T}\}} h_2(\|v_{f,k}(\theta) - \widehat{v}_{f,k}\|_2), \tag{31}$$

$$L_f^{\mathrm{tip}}(\theta) = 5h_2(\|v_{f,\mathrm{T}}(\theta) - \widehat{v}_{f,\mathrm{T}}\|_2) + h_{0.5}(\|u_f(\theta) - \widehat{u}_f\|_2). \tag{32}$$

All position residuals are evaluated in centimeters; direction residuals are dimensionless. With $f^\star = \arg\max_{f \neq \mathrm{th}} \beta_f$, each frame solves

$$\begin{aligned} \theta_t^\star = \arg \min_{\theta_{\min} \leq \theta \leq \theta_{\max}} \; & \sum_f \left[(1 - \beta_f) L_f^{\mathrm{hand}}(\theta) + \beta_f L_f^{\mathrm{tip}}(\theta)\right] \\ & + 6\beta_{f^\star} \|v_{\mathrm{th,T}}(\theta) - v_{f^\star,\mathrm{T}}(\theta)\|_2^2 + 0.04\|\theta - \theta_{\mathrm{prev}}^\star\|_2^2. \end{aligned} \tag{33}$$

The bounds come from the robot URDF. The contact term brings the thumb toward the most active pinch partner, while the cap $\beta_f \leq 0.7$ retains a posture penalty during pinching. The configuration uses `pinch_scaling`=1, so the tip target remains the segment-scaled target above. We use NLopt SLSQP with analytical gradients, at most 50 objective evaluations per frame, and absolute objective tolerance 10^{-4}. The previous optimizer solution supplies both the initialization and temporal reference; the first solve starts from the neutral pose and omits the temporal term.

Temporal filtering and invalid frames. Accepted solutions pass through an exponential filter with new-sample weight 0.2. Its output is then clipped to a maximum change of $12/30 = 0.4$ radians per joint per frame at 30 Hz and to the URDF joint limits. Before optimization, we check that all selected transforms are finite and geometrically valid. Wrist-to-finger-base distances must lie in $(0.005, 0.20)$ m, successive finger segments in $(0.002, 0.10)$ m, wrist-to-middle-tip distance in $(0.06, 0.35)$ m, and index-to-little-finger base distance in $(0.02, 0.18)$ m. Consecutive wrist displacements must be below 0.75 m. When confidence annotations are present, all scores must be finite, their median must be at least 0.05, and wrist confidence must be at least 0.10. Invalid tracking, a negative solver status, or non-finite joint output causes the exported joints to hold their previous values, using the neutral pose if no preceding output exists.

Action alignment and diagnostics. For a source sequence with T frames, the training export contains $T - 1$ rows: row t stores the filtered hand state at t and the absolute joint target at $t + 1$. It preserves the separately constructed wrist targets. The 22 joint values follow the URDF joint-name order and populate coordinates 80–101 and 105–126 in Table 5. We retain tracking-validity and solver-failure flags, fitting costs, pinch weights, and keypoint errors as diagnostics; these fields do not themselves define the training action-validity mask. The diagnostic RMSE compares 21 forward-kinematic points from the exported joints against the aligned, unscaled human points in meters, as illustrated in Figure 8(a). It therefore measures residual geometric mismatch rather than the weighted objective in Equation (33). This kinematic fit enforces joint limits but does not impose object-contact or collision constraints.

C.7 Training details

Table 6 summarizes the recipe. S1 trains for 100k updates with synchronized flow times and no privileged teacher. S2 trains for 200k updates on the expanded mixture with both proposed mechanisms enabled. Both stages use a global batch size of 2,048 and 5k learning-rate warmup steps, with

cosine learning-rate decay and BF16 mixed precision. The two-stage pretraining uses 22k NVIDIA H200 GPU-hours in total. Vision–language co-training provides the auxiliary language objective.

Following the Kimi/Moonlight learning-rate scaling (Liu et al., 2025), the orthogonalized Muon update for an $m \times n$ matrix is scaled by $0.2\sqrt{\max(m, n)}$ to approximately match AdamW's update RMS. The learning rates in Table 6 are the base rates before this matrix-dependent scaling.

Table 6: **Pretraining settings.** Settings for S1 initialization and S2 expansion. Global batch sizes count manipulation samples.

Setting	**S1**	**S2**
Optimizer updates	100k	200k
Global batch size	2,048	2,048
Learning-rate warmup	5k	5k
Peak base LR: VLM / coupled module	10^{-5} / 5×10^{-5}	10^{-5} / 5×10^{-5}
Flow-time sampling	Synchronized Beta(1.5, 1)	Reliability-conditioned
Action reliability q_d	1 for all sources	Robot: 1; human: 0.2
Vision / action loss coefficient	0.1 / 1	0.1 / 1
Distillation / AR loss coefficient	0 / 0.05	0.05 / 0.05

D Additional experimental protocol

D.1 RoboCasa GR1 Tabletop

Tasks and observations. We follow DIAL's 24-task protocol (Chen et al., 2026). The training trajectories come from the humanoid tabletop subset of PhysicalAI-Robotics-GR00T-X-Embodiment-Sim. PnP tasks require moving objects between supports or containers; articulated tasks additionally require closing the destination fixture after placement. DIAL uses language, a 224×224 egocentric image, and proprioception, with 16-step action chunks and a 47-dimensional learning representation that includes 18 end-effector-pose dimensions and 29 commanded joint dimensions. Our preprocessing instead forms the 384×320 composite used during pretraining, placing the head-camera view above two blank wrist-camera regions. The GR1 controller executes the same 29 joint commands: 14 for the arms, 12 for the hands, and 3 for the waist.

Post-training. We adapt the pretrained model jointly to all 24 tasks using the settings in Table 7. Few-shot training uses one tenth of the full robot demonstrations. Both settings retain the privileged-future distillation objective from Section 3.3; privileged future observations are used only during training.

Table 7: **RoboCasa post-training settings.** Settings for EgoForesight on RoboCasa GR1.

Setting	Few-shot	Full data
Demonstrations per task	100	1,000
Total robot trajectories	2,400	24,000
Global batch size	128	512
Training steps	70,000	200,000
Feature distillation	Enabled	Enabled
Evaluation episodes per task	50	50

Baseline variants and sources. The GR00T N1.6 (NVIDIA, 2025) category scores, the Qwen2.5 variants, and the robot-only DIAL results are from Figures 7–8 of Chen et al. (2026). The FLARE row is DIAL's FLARE-style extension of the fine-tuned GR00T-Qwen2.5 baseline (Chen et al., 2026; Zheng et al., 2025), rather than a different-backbone result. DIAL + EgoDex uses their additional human-data training and the in-distribution results in Figure 9. For the remaining category scores, we average the published per-task results: LDA-1B, Table VI of Lyu et al. (2026); RLDX-1, Table 14 of Kim et al. (2026); and JoyAI-RA 0.1, Table B of Zhang et al. (2026b).

Table 8: **Per-task results on RoboCasa.** Success rates (%) of EgoForesight on RoboCasa GR1. Each task is evaluated over 50 episodes in each setting.

Task	Full data	Few-shot
Pick-and-place (18 tasks)		
Cutting board → Basket	98.0	88.0
Cutting board → Cardboard box	62.0	52.0
Cutting board → Pan	90.0	84.0
Cutting board → Pot	90.0	90.0
Cutting board → Tiered basket	52.0	48.0
Placemat → Basket	88.0	78.0
Placemat → Bowl	66.0	52.0
Placemat → Plate	48.0	44.0
Placemat → Tiered shelf	44.0	28.0
Plate → Bowl	72.0	56.0
Plate → Cardboard box	54.0	42.0
Plate → Pan	50.0	48.0
Plate → Plate	90.0	80.0
Tray → Cardboard box	92.0	84.0
Tray → Plate	76.0	84.0
Tray → Pot	86.0	92.0
Tray → Tiered basket	86.0	70.0
Tray → Tiered shelf	50.0	46.0
Category average	**71.9**	**64.8**
Articulated objects (6 tasks)		
Bottle → Cabinet + close	92.0	80.0
Can → Drawer + close	96.0	84.0
Cup → Drawer + close	64.0	54.0
Milk → Microwave + close	70.0	70.0
Potato → Microwave + close	80.0	58.0
Wine → Cabinet + close	78.0	72.0
Category average	**80.0**	**69.7**
Overall average	**73.9**	**66.0**

Evaluation configuration. We evaluate the EMA teacher checkpoint using 20 flow integration steps, generating 16 action steps and executing all 16 steps per prediction.

D.2 Real-world dexterous manipulation

Hardware and recorded observations. Our real-world platform is a Galaxea R1 Pro equipped with two Sharpa Wave dexterous hands (Figure 9). The teleoperation recordings include RGB observations from a head-mounted RealSense L515 and an OAK-1-W camera at each wrist. The converted training data store all three views at 1280×720 resolution and 10 Hz. The recorded joint-state vector has 58 coordinates: seven joints for each arm and 22 joints for each hand. End-effector poses and hand-joint targets are also available in the recordings.

Teleoperation interface. The operator wears MANUS gloves, with a VIVE tracker mounted on the back of each hand (Figure 9b).

Tasks, completion criteria, and progress scores. We use four tabletop tasks with language instructions from the demonstration metadata. A trial succeeds only when the complete task goal is achieved. Progress scores lie in $[0, 1]$: bottle opening, handover, and pick and place receive the highest achieved milestone score (0 if none), whereas syringe transfer sums the credits for completed steps, each counted once:

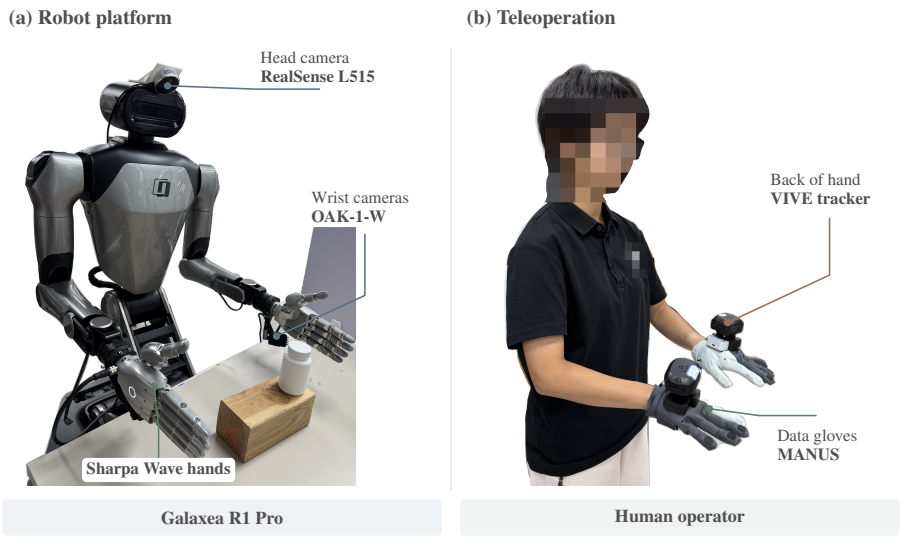


Figure 9: **Real-world hardware and teleoperation setup.** (a) Galaxea R1 Pro with two Sharpa Wave hands, a RealSense L515 head camera, and OAK-1-W wrist cameras. (b) MANUS gloves with a VIVE tracker on the back of each hand. The operator's face and clothing logo are pixelated.

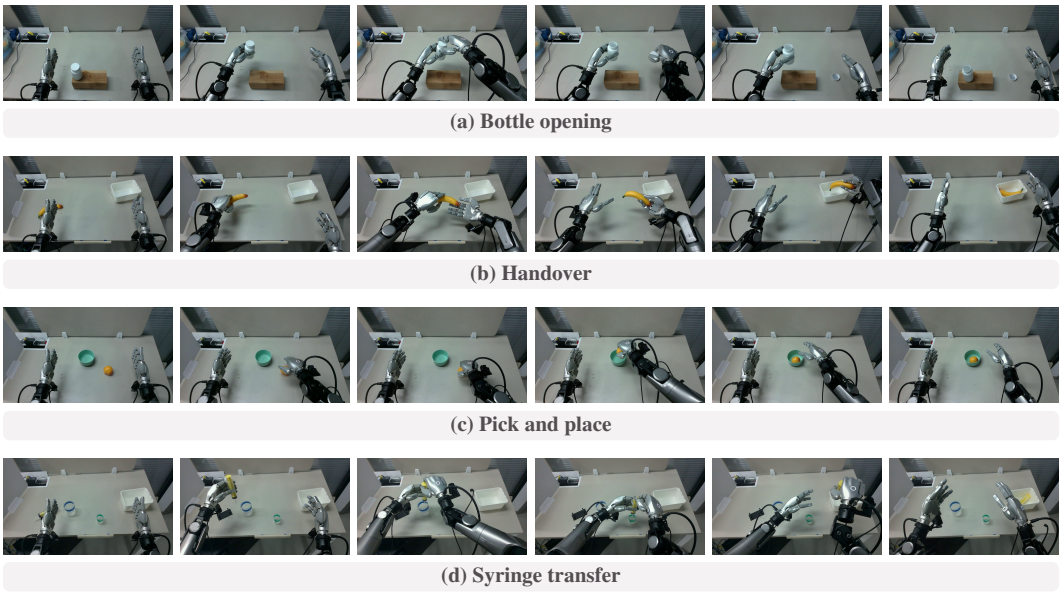
1. **Bottle opening:** unscrew the bottle lid, place the lid on the table, and return the bottle to the wooden block. Milestones: pick up the bottle (0.1), unscrew the cap with at least two rotation attempts (0.5), remove the cap (0.6), place the cap on the table (0.8), and return the bottle to the block (1.0).
2. **Handover:** pick up the banana with the left hand, transfer it to the right hand, and release it into the white box. Milestones: grasp the banana (0.2), lift it from the table (0.4), transfer it to the right hand (0.7), and place it into the box (1.0).
3. **Pick and place:** use the right hand to pick up the fruit and place it into the bowl. Milestones: lift the fruit from the table (0.5) and place it into the bowl (1.0).
4. **Syringe transfer:** draw material from the left beaker with the syringe, transfer it into the right beaker, and place the syringe into the white box. Additive credits: pick up the syringe (+0.1), aim at the left beaker (+0.1), pull the plunger to draw liquid (+0.2), aim at the right beaker (+0.1), push the plunger to inject liquid (+0.2), transfer the syringe to the right hand (+0.2), and place it into the box (+0.1).

Reported progress scores are averaged over trials and scaled to 0–100. Figure 4 illustrates the four tasks, and Figure 10 visualizes each task from its initial configuration to completion.

Demonstration budget and evaluation. We use 100 teleoperated demonstrations per task for post-training, totaling 400 demonstrations. We compare EgoForesight with $\pi_{0.5}$ (Black et al., 2025a) and GR00T N1.7 (NVIDIA, 2026) under this per-task demonstration budget. For each method, a single model is jointly post-trained on the combined demonstrations from all four tasks. Each method is evaluated in 20 trials per task. Task-level success is the percentage of trials that achieve the complete task goal (Figure 4).

Post-training settings. All methods are post-trained for 20k optimization steps with a batch size of 128. For EgoForesight, we use Muon (Jordan et al., 2024) with learning rates of 5×10^{-5} for the coupled vision–action MoT module and 1×10^{-5} for the VLM, and 100 learning-rate warmup steps. We use the LeRobot implementations (Cadene et al., 2024) of $\pi_{0.5}$ and GR00T N1.7, retaining the recommended default optimizer settings for each baseline.

Action-space adaptation for $\pi_{0.5}$. The released $\pi_{0.5}$ configuration uses a 32-dimensional action space. To accommodate our 58-DoF setting, we replace its action input and output layers with newly



(a) Bottle opening

(b) Handover

(c) Pick and place

(d) Syringe transfer

Figure 10: **Real-world task demostrations.** Each row presents a task from its initial configuration to completion, ordered from left to right.

initialized 58-dimensional layers before post-training. This adaptation changes the pretrained action interface and should be considered when interpreting its performance.

D.3 Additional ablations

Training setting. The reduced pretraining set contains approximately 1k hours of S1 data and 2k hours of additional S2 egocentric data, with S1 replay retained during S2. The default recipe uses 50k S1 updates and 100k S2 updates at batch size 256. Downstream adaptation follows the RoboCasa few-shot protocol in Appendix D.1. Pretraining teacher variants share the same adaptation recipe; post-training teacher variants start from the same full-scale pretrained checkpoint.

Pretraining-data scaling protocol. The scaling comparison uses nested subsets containing 0, $1/10$, $1/3$, or all of the pretraining data, with matched source proportions for the nonzero subsets. At each nonzero scale, we apply the same fraction to every S1 source and the additional S2 egocentric data, train both stages in order, and retain the selected S1 data for replay throughout S2. Baseline and EgoForesight share the same subset and optimization budget at each scale; the baseline disables reliability conditioning and privileged-future distillation. The zero-data control skips both pretraining stages. Each downstream benchmark keeps its adaptation data and budget fixed across scales. Real-world open-loop evaluation computes mean squared error between predicted action chunks and held-out teleoperated actions, averaged over timesteps and valid normalized action coordinates; these demonstrations are excluded from adaptation. RoboCasa evaluation follows the few-shot closed-loop protocol in Appendix D.1. Video hours count unique retained pretraining data, not repeated sampling or S1 replay in S2. We fit $\log_{10}(\mathrm{MSE})$ and success percentage by ordinary least squares against $\log_{10}(h)$ using only nonzero durations, and omit the zero-hour control from the figure. Thus, m is the dimensionless log–log slope for MSE or success-rate percentage points per decade, respectively. The nonzero durations are 3k, 10k, and 30k hours. For full-scale real-world MSE, the baseline value is the lowest observed along its checkpoint trajectory (0.039), rather than the later 0.042 endpoint shown in Figure 1(b); this avoids attributing a gain to the baseline's late-stage overfitting. Figure 5 uses a logarithmic 0.02–0.10 MSE axis and a linear 50–70% success axis; the success axis marks the omitted lower range with a break.

Reliability sensitivity and loss weighting. Table 9 varies human-data reliability while keeping robot-data reliability at one. The $q_d = 1$ control removes source-dependent corruption and down-weighting. All three tested values below one ($q_d = 0.8, 0.5, 0.2$) improve average success from

52.0% to 54.0–54.3%, a spread of only 0.3 percentage points. The gains come primarily from Art., which improves from 47.7% to 54.9–58.0%, while PnP remains near 53%. This limited sensitivity supports the use of coarse source-level reliability without more complex task- or sample-specific estimates. At $q_d = 0.2$, retaining reliability-conditioned corruption but using unit action-loss weights gives nearly identical results (54.2% versus 54.3% average success). Thus, the benefit persists without loss downweighting; its additional contribution is small in this setting.

Table 9: **Ablation of reliability and loss weighting.** RoboCasa few-shot success rates (%) after reduced-scale pretraining. The last row changes only the loss weight. Dark and light green shading mark the best and second-best scores per column, including ties.

Human q_d	Loss weight	PnP	Art.	Avg.
1.0	1.0	53.4	47.7	52.0
0.8	0.8	**53.7**	54.9	54.0
0.5	0.5	53.0	57.8	54.2
0.2	0.2	53.1	58.0	**54.3**
0.2	1.0	52.9	**58.1**	54.2

E Qualitative future prediction and IDM outputs

Figure 11 shows current-only future prediction and IDM outputs conditioned on current and GT future observations.

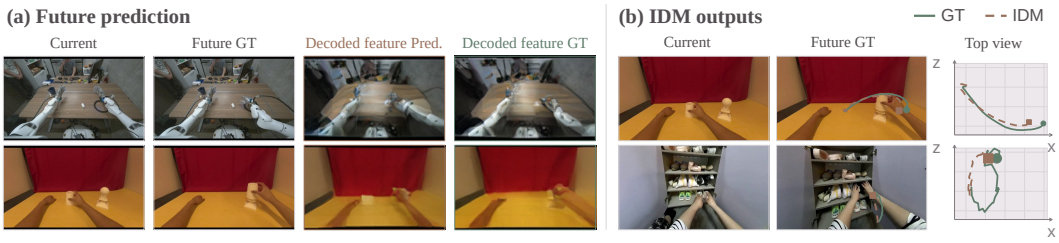


Figure 11: **Future prediction and IDM outputs. (a)** Current/future RGB and RAE reconstructions of pooled GT versus predicted DINO features, with identical decoding and cropping. **(b)** Relatively smooth IDM outputs (dashed) and GT (solid).

Pooled GT and predicted DINOv3-B/16 features (Siméoni et al., 2025) share a $12 \times 10 \times 768$ grid and the same frozen RAEv2 decoder (Singh et al., 2026). Only decoder positional embeddings are resized for the rectangular grid. These lossy reconstructions and selected pretraining clips are qualitative illustrations, not pixel-accurate futures or held-out evaluations.

F Limitations

The analysis in Appendix A.1 assumes conditionally independent, zero-mean Gaussian annotation noise. Systematic retargeting bias can shift conditional action means, leaving flow-field discrepancies even at maximal corruption; source weighting alone does not correct these shifts. Dataset-level reliability misses within-source variation. Future work could explore finer-grained estimates at the example, time-step, or action-coordinate level. Dependence on action annotations motivates self-supervised learning from unannotated videos. The tested scales leave compositional generalization to unseen object–skill–scene combinations unverified, motivating larger-scale studies of supervision quality and diversity. Predicted futures and EMA teachers may supply biased or action-irrelevant supervision, while adding computational cost.

G Additional Related Work

G.1 Embodied Data Pyramid

Robot learning draws on a data pyramid shaped by coverage, collection cost, and action availability. Human videos offer broad interaction coverage (Grauman et al., 2022; Hoque et al., 2026), whereas robot demonstrations tie observations to executable controls (Walke et al., 2023; Open X-Embodiment Collaboration, 2024; Khazatsky et al., 2024; AgiBot-World Contributors et al., 2025). Simulation expands trajectory coverage through data generation (Mandlekar et al., 2023), and handheld interfaces make transferable demonstrations easier to collect (Chi et al., 2024). Exploiting this diversity requires connecting visual experience to action supervision. Early use of human video focused on learning visual representations for control (Nair et al., 2023; Karamcheti et al., 2023). Later approaches make the connection more explicit by inferring latent actions from visual transitions (Ye et al., 2025; Bu et al., 2025; Zhang et al., 2026a). EgoMimic and EgoScale convert estimated or tracked human hand motion into robot supervision (Kareer et al., 2025; Zheng et al., 2026b), whereas X-Diffusion transfers action distributions across embodiments (Pace et al., 2025). These routes broaden the available training signal, but produce mixtures in which robot actions may be absent or noisy. We study joint future–action learning under this uneven supervision.

G.2 Representation Learning for Generation

Generative objectives have long provided a way to learn representations: denoising autoencoders extract useful features by reconstructing corrupted inputs (Vincent et al., 2010), while VAEs and latent diffusion learn the spaces in which generation takes place (Kingma & Welling, 2014; Rombach et al., 2022). REPA (Yu et al., 2025) exploits the reverse connection, using pretrained clean-image representations to guide intermediate denoiser features. This makes the source and form of representation guidance central design choices. Existing VAE features can supply alignment targets (Wang et al., 2026). REPA-E jointly tunes the VAE and latent diffusion transformer end to end (Leng et al., 2025), whereas RAE rebuilds the generative latent space around a pretrained representation encoder (Zheng et al., 2026a). Guidance can also come from within the generator itself. SRA and LayerSync use representations from the model's own layers as alignment targets, reducing dependence on external encoders (Jiang et al., 2026; Haghighi et al., 2026). Self-Flow (Chefer et al., 2026) develops this internal approach with an EMA teacher evaluated at a cleaner, noise-shifted input and aligns deeper teacher features to earlier student features; it also evaluates the mechanism on joint video–action prediction.

G.3 Learning from Noisy Data

Noisy supervision can bias optimization, motivating loss corrections based on assumptions about label corruption (Natarajan et al., 2013; Patrini et al., 2017). When trust is estimated from the data, meta-reweighting (Ren et al., 2018) and sample selection with peer learning or semi-supervision (Han et al., 2018; Li et al., 2020) adjust how examples contribute to training. For restoration and generation, the problem extends from unreliable labels to corruption of the observations themselves. Noise2Noise and Noise2Self establish that restoration remains possible without clean targets under suitable noise assumptions (Lehtinen et al., 2018; Batson & Royer, 2019), while AmbientGAN and Ambient Diffusion learn generative distributions from corrupted measurements (Bora et al., 2018; Daras et al., 2023). The usefulness of lower-quality data can also depend on the denoising stage. Ambient Omni (Daras et al., 2025) studies noise- and timestep-dependent use or weighting of data with different quality, and Ambient Diffusion Policy (Wei et al., 2026) carries this idea into policy learning by exploiting coarse structure or local motion at different stages. We jointly adapt action corruption and loss weighting to source reliability.