LATENT RL PRIORS FOR DATA-EFFICIENT DEXTEROUS MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning dexterous manipulation from limited visual demonstrations is challenging because policies must acquire both perceptual grounding and precise control, while reinforcement learning (RL) in simulation can acquire rich control behaviors from large-scale interaction with privileged information. We introduce **Latent RL Prior**, a two-stage framework that reuses this control knowledge for data-efficient visual policy learning. During privileged pretraining, we distill large-scale RL rollouts into a flow-matching action decoder conditioned on proprioception and latent task context. During visual adaptation, we freeze the decoder and learn a visual interface that maps images to its latent context using action supervision alone, without privileged labels or latent targets. The same decoder can further be shared across arms by learning coordinated visual contexts for bimanual manipulation. Across three simulated tasks, our method outperforms full-parameter fine-tuning of $\pi_{0.5}$ and GR00T N1.7 in the low-data regime; on door lifting and bimanual assembly, it surpasses full-data baselines using $10\times$ and $5\times$ fewer demonstrations, respectively. Further analysis shows that direct visual-to-latent conditioning improves adaptation performance and training stability over routing visual predictions through the pretrained state encoder. Robustness to unseen mass variation and greater linear accessibility of contact-force information in decoder representations further provide evidence of retained RL-acquired control knowledge. Training and evaluation code and datasets will be released.

1 INTRODUCTION

Dexterous manipulation requires coordinated arm and finger motion under changing contacts and object dynamics. Learning these behaviors from limited visual demonstrations is challenging because the policy must acquire both perceptual grounding and precise action generation (Zhao et al., 2023; Chi et al., 2023). Pretrained vision-language-action (VLA) models offer transferable priors from broad visual, language, and robot datasets (Zitkovich et al., 2023; Kim et al., 2024; Black et al., 2024). We investigate a complementary source of prior knowledge for dexterous control: behaviors acquired through large-scale reinforcement learning (RL) in simulation.

Simulated interaction enables RL policies to acquire manipulation strategies across diverse object geometries and physical properties (Andrychowicz et al., 2020; Singla et al., 2024; Kedia et al., 2026). However, these policies often rely on privileged information, such as object and goal states, that is not directly available at deployment. Reusing their control capabilities requires an interface that connects visual observations to the pretrained controller. Our premise is that preserving pretrained proprioceptive processing and action generation can reduce the visual demonstrations needed to adapt these capabilities to downstream tasks.

We introduce Latent RL Prior, a simple yet effective framework for this transfer. We distill rollouts from an RL teacher trained on procedural pose alignment into a flow-matching action decoder (Lipman et al., 2023). The decoder receives proprioceptive history through a dedicated pathway and latent context encoding privileged task information. During visual adaptation, we freeze the decoder and learn a visual interface that maps images directly to this context. This *latent transfer* uses only the action objective on image–proprioception–action demonstrations, without requiring corresponding privileged information, pretrained latent targets, or paired trajectories between the pretraining and adaptation datasets.

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.

The same design extends a single-arm RL prior to multiple arms. We share the frozen decoder across arms and learn coordinated visual contexts, each paired with the corresponding arm's proprioceptive history. Bimanual adaptation thus reuses the pretrained control prior while learning inter-arm coordination through the visual interface.

We evaluate the shared prior on target pose alignment, door lifting, and bimanual assembly in simulation. With limited visual demonstrations, our method outperforms full-parameter fine-tuning of $\pi_{0.5}$ (Physical Intelligence et al., 2025) and GR00T N1.7 (NVIDIA, 2026). On door lifting and bimanual assembly, it surpasses both baselines trained on the full task datasets with $10\times$ and $5\times$ fewer demonstrations, respectively. Comparisons with the same visual policy architecture trained from scratch further support the value of the pretrained action prior.

Further investigations examine how to access and adapt this prior. Latent transfer improves performance and training stability over explicit transfer through the pretrained state encoder. Separate adaptation studies favor conservative updates to the image encoder and a frozen action decoder, particularly for precise alignment. We also investigate the control behaviors and physical information accessible after transfer. The adapted policies exhibit robustness to mass variation absent from the visual demonstrations, and their decoder representations make contact-force information more linearly accessible than those learned from scratch. These observations are consistent with retained control knowledge from large-scale RL pretraining.

Our contributions are:

- a framework for adapting a frozen RL action prior through latent visual conditioning with action supervision alone, including reuse of a single-arm decoder for multi-arm manipulation;
- empirical evidence of improved visual-demonstration efficiency across three dexterous tasks, together with evaluations of the transfer interface and adaptation choices; and
- behavioral and representational analyses of mass robustness and contact-force accessibility after visual adaptation.

2 Related Work

Dexterous control from reinforcement learning. RL with dynamics randomization has enabled dexterous in-hand manipulation (Andrychowicz et al., 2020). Parallel simulation and scalable policy optimization support learning from large amounts of interaction (Makoviychuk et al., 2021; Singla et al., 2024). SimToolReal trains a single policy to manipulate procedurally generated objects toward goal poses and deploys it using object-pose estimates and goal-pose sequences (Kedia et al., 2026). We build on its released teacher, distill rollouts into an action decoder, and investigate reuse of the resulting control prior through visual conditioning learned from limited downstream demonstrations.

Transfer with privileged information. Asymmetric actor–critic training uses privileged information in the critic to learn policies with restricted observations (Pinto et al., 2018). Policy distillation transfers behavior between policies (Rusu et al., 2016), while Learning by Cheating uses a privileged teacher to train a visual student (Chen et al., 2020). Closer to our interface, Rapid Motor Adaptation (RMA) and its extension to in-hand rotation retain an RL controller while estimating latent representations of physical properties from proprioceptive and action histories (Kumar et al., 2021; Qi et al., 2023). Their adaptation modules are trained to match embeddings computed from corresponding privileged information. Our visual adaptation uses action supervision alone: the decoder distilled from RL rollouts remains frozen, while its visual conditioning is learned without privileged information or latent targets in the downstream demonstrations. We reuse this prior across tasks and extend it to bimanual manipulation through a shared single-arm decoder.

Visual imitation and generalist robot policies. Action chunking and generative action modeling support learning visual manipulation from demonstrations (Zhao et al., 2023; Chi et al., 2023). Generalist policies and VLA models transfer knowledge from diverse robot and vision–language datasets (Zitkovich et al., 2023; Ghosh et al., 2024; Kim et al., 2024). The π_0 and GR00T families combine pretrained vision–language representations with flow-matching action models (Black et al., 2024; NVIDIA et al., 2025). Our approach combines DINOv2 visual features (Oquab et al.,

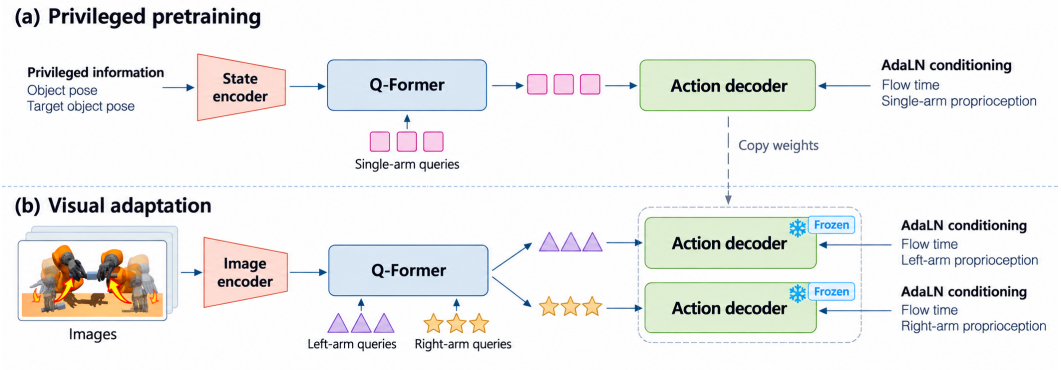

Figure 1: Latent RL Prior. (a) Privileged pretraining: a state tokenizer and Q-Former supply task context to an action decoder conditioned on flow time and single-arm proprioception. (b) Visual adaptation: an image encoder and Q-Former learn context tokens while the decoder remains frozen. For bimanual assembly, left- and right-arm queries interact within the same Q-Former; each output group conditions the shared decoder with its corresponding arm's proprioception.

2024) with a control prior distilled from privileged RL rollouts. We investigate how this source of control experience supports adaptation to dexterous tasks with limited visual demonstrations, comparing against full-parameter fine-tuning of $\pi_{0.5}$ (Physical Intelligence et al., 2025) and GR00T N1.7 (NVIDIA, 2026).

3 Latent RL Prior

Large-scale reinforcement learning (RL) in simulation can acquire complex dexterous manipulation behavior through abundant interaction across diverse object geometries, masses, and physical properties. However, these policies often rely on privileged state information, such as object and goal states, that is unavailable at deployment. Therefore, in this section, we investigate how to transfer such dexterous behaviors acquired by privileged RL into a reusable *RL prior* for downstream visual tasks. Our goal is to preserve dexterous capabilities learned from large-scale simulated experience while adapting a visual interface from limited demonstrations using only deployable observations.

3.1 Problem Formulation

We consider dexterous manipulation tasks in which, at time t, the robot observes its proprioceptive state $\boldsymbol{r}_t$ and synchronized images $\boldsymbol{I}_t^{1:C}$ from C cameras, and executes an action $\boldsymbol{a}_t$. Using a history of H observations, the policy predicts a K-step action chunk. We denote

$$\mathcal{P}_t = \boldsymbol{p}_{t-H+1:t}, \quad \mathcal{R}_t = \boldsymbol{r}_{t-H+1:t}, \quad \mathcal{I}_t = \boldsymbol{I}_{t-H+1:t}^{1:C}, \quad \mathcal{A}_t = [\boldsymbol{a}_t, \ldots, \boldsymbol{a}_{t+K-1}].$$

We represent task information with M latent context tokens $\boldsymbol{Z}_t \in \mathbb{R}^{M\times d}$, each of dimension d. Together with the proprioceptive history $\mathcal{R}_t$, these tokens condition the action decoder to predict the action chunk $\mathcal{A}_t$. The source of $\boldsymbol{Z}_t$ depends on the training stage: it is derived from the privileged history $\mathcal{P}_t$ during pretraining and from the visual history $\mathcal{I}_t$ during visual adaptation.

We use a large privileged RL dataset $\mathcal{D}_{\mathrm{RL}} = \{(\mathcal{P}_t, \mathcal{R}_t, \mathcal{A}_t)\}$ and a smaller downstream visual dataset $\mathcal{D}_{\mathrm{vis}} = \{(\mathcal{I}_t, \mathcal{R}_t, \mathcal{A}_t)\}$. The visual demonstrations contain no privileged-state labels and need not be paired with the RL trajectories. Our goal is to reuse the action prior learned from $\mathcal{D}_{\mathrm{RL}}$ while replacing privileged conditioning with visual conditioning for downstream tasks.

3.2 Privileged Pretraining

Large-scale teacher rollouts. Following Kedia et al. (2026), we first train an RL teacher on a broad object pose-alignment task, where the robot aligns procedural objects with target poses. We apply domain randomization over object shape, size, and density during training to expose the policy

to diverse manipulation conditions. We then roll out the trained policy at scale and construct $\mathcal{D}_{\mathrm{RL}}$ from the collected trajectories by extracting privileged histories $\mathcal{P}_t$, proprioceptive histories $\mathcal{R}_t$, and future action chunks $\mathcal{A}_t$.

Flow-matching action prior. A state tokenizer T_η followed by a Q-Former Q_ψ (Li et al., 2023) maps the privileged history to the latent context

$$\boldsymbol{Z}_t^{\mathrm{priv}} = Q_\psi(T_\eta(\mathcal{P}_t)) \in \mathbb{R}^{M\times d}.$$

The action decoder v_θ is a transformer (Vaswani et al., 2017) whose action tokens cross-attend to this context. The proprioceptive history $\mathcal{R}_t$ and flow time τ condition each block through adaptive layer normalization (AdaLN) (Peebles & Xie, 2023). This separation allows the source of the latent context to change while preserving the pretrained processing of proprioception.

We train the decoder with conditional flow matching (Lipman et al., 2023). For a demonstrated action chunk $\mathcal{A}_t^0 := \mathcal{A}_t$, we sample $\mathcal{A}_t^1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\tau \sim \mathcal{U}[0,1]$, and form $\mathcal{A}_t^\tau = (1-\tau)\mathcal{A}_t^0 + \tau\mathcal{A}_t^1$. The decoder learns the corresponding velocity field by minimizing

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{(\mathcal{R}_t, \mathcal{A}_t^0)\sim\mathcal{D}_{\mathrm{RL}},\, \mathcal{A}_t^1\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\, \tau\sim\mathcal{U}[0,1]} \left[\left\| v_\theta(\mathcal{A}_t^\tau, \tau, \boldsymbol{Z}_t, \mathcal{R}_t) - \left(\mathcal{A}_t^1 - \mathcal{A}_t^0\right) \right\|_2^2 \right]. \tag{1}$$

3.3 Visual Adaptation

Having pretrained the action prior from privileged RL rollouts, we next adapt it to downstream visual observations, where privileged task information is unavailable. We learn a visual conditioning module from limited demonstrations while keeping the pretrained action decoder fixed. Specifically, an image encoder E_{vis}, instantiated with DINOv2 (Oquab et al., 2024), encodes the visual history as $\boldsymbol{F}_t = E_{\mathrm{vis}}(\mathcal{I}_t)$. We keep the pretrained action decoder v_θ fixed and train the visual pathway on $\mathcal{D}_{\mathrm{vis}}$ using the same flow-matching objective in Equation 1. The action loss provides the only supervision; no privileged-state labels or pretrained latent targets are required.

Latent and explicit transfer. We consider two interfaces for conditioning the pretrained action prior from visual observations. In *latent transfer*, a visual Q-Former Q_ϕ maps image features directly to the decoder context,

$$\boldsymbol{Z}_t^{\mathrm{latent}} = Q_\phi(\boldsymbol{F}_t) \in \mathbb{R}^{M\times d}.$$

This replaces the pretrained state tokenizer and state Q-Former, while preserving the action decoder and its proprioceptive conditioning pathway. We optimize E_{vis} and Q_ϕ with $\boldsymbol{Z}_t = \boldsymbol{Z}_t^{\mathrm{latent}}$.

In *explicit transfer*, a bridge B_ρ first maps visual features to a privileged-shaped history, $\widehat{\mathcal{P}}_t = B_\rho(\boldsymbol{F}_t)$, which is then processed by the pretrained state encoder. The two interfaces are therefore

$$\boldsymbol{Z}_t^{\mathrm{latent}} = Q_\phi(\boldsymbol{F}_t), \qquad \boldsymbol{Z}_t^{\mathrm{explicit}} = Q_\psi\left(T_\eta(\widehat{\mathcal{P}}_t)\right). \tag{2}$$

For explicit transfer, we freeze T_η, Q_ψ, and v_θ, and optimize only E_{vis} and B_ρ.

Extension to multi-arm manipulation. Many bimanual manipulation tasks build on dexterous skills that are already required at the single-arm level, which motivates reusing the pretrained single-arm *RL prior* for both arms, while learning the additional inter-arm coordination from bimanual demonstrations. Rather than introducing a separate bimanual action decoder, we retain the shared single-arm decoder and extend its visual conditioning to produce coordinated, arm-specific contexts.

Specifically, as shown in Figure 1, we assign M queries to each arm and process the two query groups jointly with the visual Q-Former. Self-attention between the groups allows each arm's context to incorporate information from the other arm before action generation. The resulting contexts $\boldsymbol{Z}_t^L$ and $\boldsymbol{Z}_t^R$ are paired with the corresponding proprioceptive histories $\mathcal{R}_t^L$ and $\mathcal{R}_t^R$. The same frozen decoder v_θ processes the two context–proprioception pairs separately to generate arm–hand actions for the left and right arms, respectively. Thus, both arms reuse the action-generation and proprioceptive processing learned during privileged pretraining, while cross-arm coordination is introduced through the jointly inferred visual contexts.

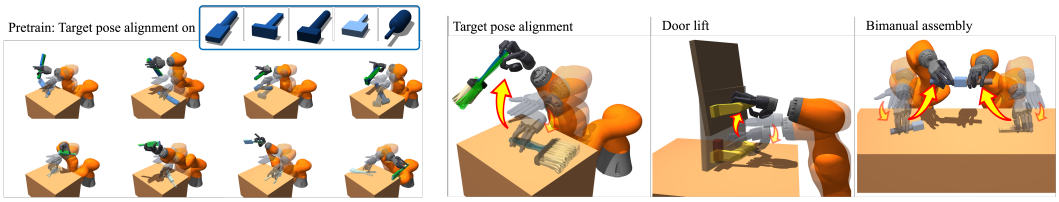

Figure 2: Pretraining and downstream tasks. Left: the pretraining task of target pose alignment with varied procedural objects; the inset shows example object templates. Right: the downstream tasks of target pose alignment with a held-out brush, door lifting, and bimanual assembly (left to right). The same pretrained action decoder is reused for visual adaptation across all three downstream tasks.

4 Experiments

We first evaluate whether the pretrained action prior improves visual-demonstration efficiency across downstream dexterous manipulation tasks. We then compare latent transfer, which maps visual features directly to decoder context tokens, with explicit transfer, which conditions the decoder through the pretrained state encoder, in terms of adaptation performance and training stability. Next, we examine the effects of fine-tuning the image encoder E_{vis} and the action decoder v_θ. Finally, we investigate what control behaviors and physical information are retained from large-scale RL training after visual adaptation, using mass-variation tests and contact-force probes.

4.1 Experimental Setup

Task design and evaluation. We conduct the experiments in MuJoCo Warp (MJWarp) (Google DeepMind & NVIDIA, n.d.) using KUKA iiwa arms equipped with SHARPA dexterous hands. Figure 2 shows the pretraining task (Section 3.2) and three downstream tasks.

Pretraining. Following the SimToolReal experimental setup (Kedia et al., 2026), a single arm–hand system reorients procedural object templates with randomized shapes, sizes, and densities to target poses. We measure translation and orientation error using four corresponding keypoints from the clean object and target poses: $d_{\max} = \max_{j\in\{1,\dots,4\}} \|\mathbf{k}_j^{\text{obj}} - \mathbf{k}_j^{\text{goal}}\|_2$. Success requires $d_{\max} \le \epsilon = 0.015$ m on 10 policy steps, not necessarily consecutive (Figure 3).

Downstream tasks. (1) In *target pose alignment*, the robot grasps a blue brush excluded from pretraining and aligns it with a target pose. We use the same keypoint criterion and report strict ($\epsilon = 0.015$ m) and relaxed ($\epsilon = 0.05$ m) success rates. (2) In *door lifting*, the robot lifts a door to reveal a cube behind it. We randomize the cube's height, color, and size, as well as the door's position and orientation. Success requires lifting the door above the cube's height. (3) In *bimanual assembly*, two arm–hand systems grasp separate objects from a tabletop, reorient them in hand, and insert a cubic protrusion on one object into a matching cavity in the other. We randomize the objects' masses and sizes, as well as the cavity size. Success requires full insertion, measured by insertion depth relative to cavity depth. Blue-brush alignment evaluates transfer to an unseen object within the pretraining task, whereas door lifting and bimanual assembly evaluate transfer to different tasks.

Data and implementation. We fine-tune the released SimToolReal teacher checkpoint (Kedia et al., 2026) on procedural target pose alignment and collect 100,000 successful rollouts for $\mathcal{D}_{\text{RL}}$. We train a separate SAPG policy (Singla et al., 2024) for each downstream task and collect its a small amount rollouts for $\mathcal{D}_{\text{vis}}$. Table 1 shows the dataset details. Unless otherwise noted, we use latent transfer (Section 3.3) with DINOv2-Small (ViT-S/14) as E_{vis} and reuse the same pretrained action decoder across downstream tasks.

Baselines. We compare our method with (1) full-parameter fine-tuning of the vision-language-action models $\pi_{0.5}$ (Physical Intelligence et al., 2025) and GR00T N1.7 (NVIDIA, 2026); (2) a scratch baseline that uses the same visual policy architecture without the pretrained action prior; and (3) explicit transfer, as defined in Section 3.3.

Training and evaluation protocol. For the main comparisons, we fine-tune GR00T N1.7 and $\pi_{0.5}$ for 30,000 optimization steps using their officially recommended fine-tuning hyperparameters. We

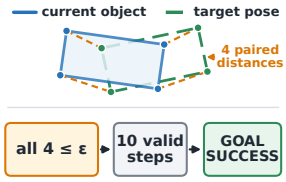

Figure 3: Target pose alignment success criterion.

Table 1: Datasets for privileged pretraining and visual adaptation. F: front; S: side; H: in-hand (L/R: left/right).

Dataset	Episodes	Arm–hand systems	Images	Privileged information
Pretraining	100,000	single	–	Yes
Target pose alignment	4,000	single	F, S, H	No
Door lift	500	single	F, S, H	No
Bimanual assembly	5,000	bimanual	F, H_L, H_R	No

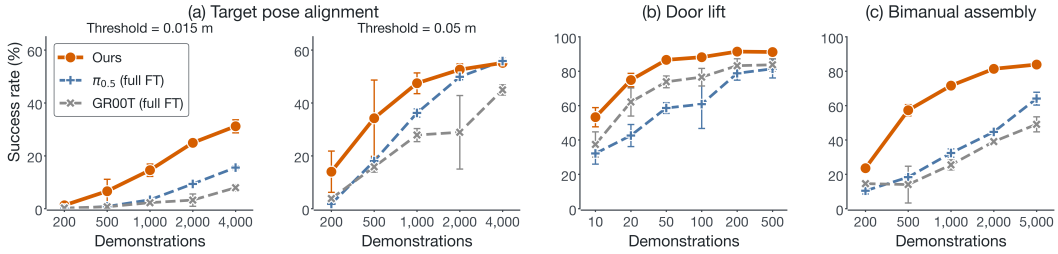

Figure 4: Comparison with full-parameter fine-tuning of GR00T N1.7 and $\pi_{0.5}$ across demonstration budgets. (a) Target pose alignment at strict ($\epsilon = 0.015$ m) and relaxed ($\epsilon = 0.05$ m) thresholds; (b) door lifting; (c) bimanual assembly. We report average success rate and standard deviation across three random seeds.

train our policy for 100,000 steps. For all methods, we save checkpoints every 5,000 steps and retain the final five for evaluation. We evaluate these checkpoints on held-out environment configurations, which means environment seeds distinct from those used for dataset generation. Each checkpoint is evaluated over 256 policy rollouts. Following prior work (Wang et al., 2024; Chi et al., 2023), we report the highest success rate among the five evaluated checkpoints for each run. Unless otherwise noted, we use three training seeds and report the mean and standard deviation of these per-run maximum success rates. The training seeds control demonstration subsampling and the PyTorch random-number generator such as the initialization of newly introduced parameters.

4.2 Latent RL Prior Enables Data-Efficient Visual Adaptation

Comparison with VLA baselines. As shown in Figure 4, our method achieves higher success rates than GR00T N1.7 and $\pi_{0.5}$ with limited visual demonstrations across all three tasks. The advantage is particularly pronounced in the low-data regime. On door lifting, our method trained with only 50 demonstrations outperforms both VLA baselines trained on the full 500-demonstration dataset, using $10\times$ fewer demonstrations. On bimanual assembly, it likewise surpasses both full-data baselines with 1,000, a $5\times$ reduction. For target pose alignment, under the strict 0.015-m criterion, our method maintains an advantage across demonstration budgets, consistent with the pretrained RL prior retaining control capabilities useful for precise alignment. Under the relaxed 0.05-m criterion, the advantage remains clear in the low-data regime but narrows as more demonstrations become available, with performance becoming numerically comparable to $\pi_{0.5}$ at 4,000 demonstrations. Overall, these results indicate that the RL prior is most beneficial when visual demonstrations are scarce or when the task demands precise control.

Contribution of the pretrained RL prior. To isolate the contribution of the pretrained RL prior, we compare latent transfer with training from scratch using the same visual policy architecture and training setup. As shown in Figure 5, latent transfer with the image encoder frozen consistently achieves higher success rates than the from-scratch baseline across the evaluated demonstration budgets and success thresholds. Since the visual encoder remains fixed, these gains cannot be explained by improved visual representation learning during adaptation. The results therefore support that reusing the pretrained RL action prior substantially improves learning from limited visual demonstrations, providing a key source of the data efficiency observed in our method.

4.3 Latent Transfer Is More Effective under Action-Only Supervision

We compare latent and explicit transfer under action-only supervision, considering both adaptation performance and training stability.

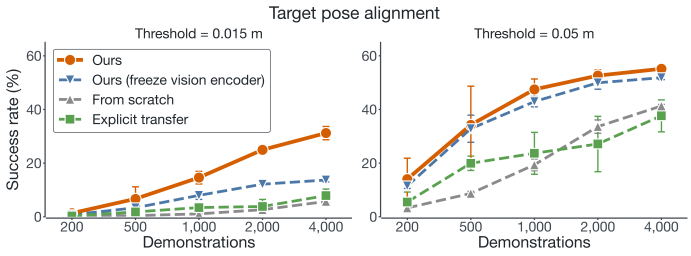

Figure 5: Target pose-alignment ablations with the action decoder frozen for both latent-transfer variants. Results show mean success rates and standard deviations over three random seeds, except for explicit transfer, which uses five seeds.

Adaptation performance. As shown in Figure 5, latent transfer achieves substantially higher target pose-alignment success than explicit transfer across the evaluated demonstration budgets under both the strict and relaxed success criteria (note that the image encoder is frozen). These results favor latent transfer for visual adaptation under action-only supervision. Interestingly, under the strict 0.015-m criterion, explicit transfer still outperforms training from scratch across most demonstration budgets, suggesting that the pretrained RL prior retains useful capabilities for precise manipulation even when accessed through the explicit interface. Under the relaxed 0.05-m criterion, however, explicit transfer performs similarly to the from-scratch baseline for much of the data range, with a clearer advantage only in the low-data regime. This pattern is consistent with the results in Section 4.2, where the benefit of the RL prior is most pronounced with limited demonstrations and under stricter precision requirements. More importantly, the substantially stronger and more consistent performance of latent transfer indicates that effectively reusing the RL prior depends not only on its presence, but also on how visual observations interface with the pretrained controller.

Training stability. As shown in Figure 6, with the image encoder frozen, latent transfer exhibits rapid performance improvement together with relatively consistent learning trajectories across random seeds. In contrast, explicit transfer shows greater run-to-run variability, while training from scratch is comparatively stable but converges to lower performance. Thus, latent transfer combines consistent learning with stronger performance.

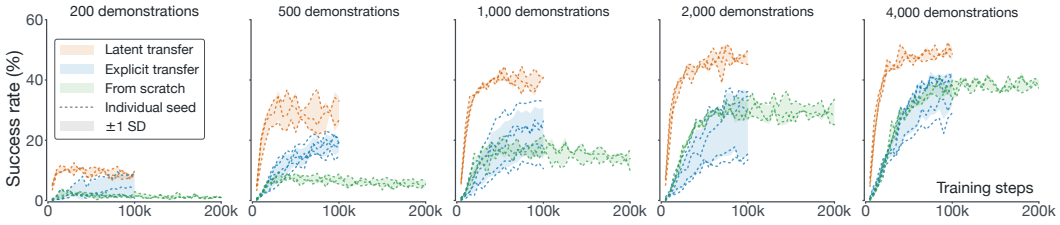

Figure 6: Training stability on target pose alignment at a relaxed ($\epsilon = 0.05$ m) threshold. Dotted curves show individual runs for latent transfer, explicit transfer, and scratch training; shaded regions show the mean $\pm$ one standard deviation. Explicit transfer uses five random seeds, while latent transfer and scratch training use three seeds each.

4.4 Ablations on Visual and Action Adaptation

To examine how visual adaptation should modify the pretrained policy, we independently sweep the learning rates of the image encoder E_{vis} and action decoder v_θ, while keeping the visual Q-Former Q_ϕ trainable throughout.

Image-encoder adaptation. As shown in Figure 7(a), with the action decoder frozen, conservative updates to the image encoder improve strict pose-alignment success across demonstration budgets, whereas larger updates degrade performance. This effect is particularly pronounced in the low-data regime: with 500 demonstrations, increasing the learning-rate scale leads to both lower mean

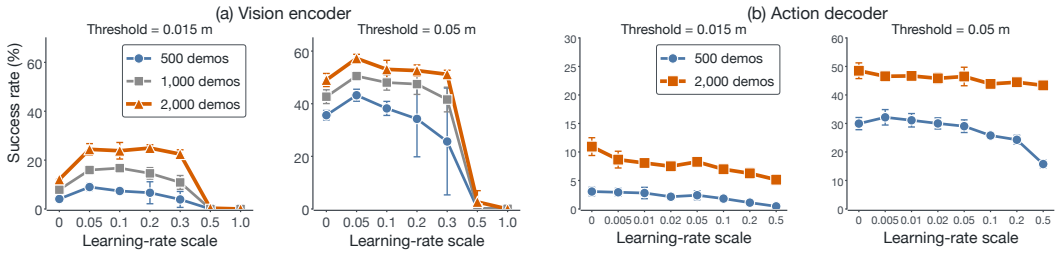

Figure 7: Learning-rate ablations for (a) the image encoder with the action decoder frozen and (b) the action decoder with the image encoder frozen. Panels report strict ($\epsilon = 0.015$ m) and relaxed ($\epsilon = 0.05$ m) pose-alignment success. We report average success rate and standard deviation across three random seeds. Each scale multiplies the tested module's original learning rate; zero freezes it.

success and greater variation across seeds, while the 1,000- and 2,000-demonstration settings are comparatively less sensitive. Under the relaxed criterion, the benefit of image-encoder adaptation is less uniform. These results suggest that limited-data adaptation benefits from small, conservative updates to the visual representation, especially when precise alignment is required.

Action-decoder adaptation. As shown in Figure 7(b), fine-tuning the action decoder provides no benefit across the tested learning-rate scales and demonstration budgets. Keeping the pretrained decoder frozen performs best under the strict criterion, while larger updates reduce performance. We hypothesize that fine-tuning the decoder on a much smaller downstream dataset may lead to overfitting and reduce the benefit of the control prior learned from large-scale privileged pretraining with $\mathcal{D}_{\mathrm{RL}}$ in Table 1. These results favor keeping the action decoder frozen during visual adaptation.

4.5 Control Behaviors and Physical Information

Having shown that the pretrained RL prior enables data-efficient visual adaptation, we next investigate what control behaviors and physical information from large-scale RL pretraining remain after visual adaptation. These retained properties may help explain the observed data-efficiency gains.

Generalization to increased object mass. RL pretraining includes randomization over object density, whereas the visual demonstrations contain only a single object mass for each task. We therefore test whether generalization to mass variation learned during RL pretraining remains after visual adaptation by evaluating the policies on increasingly heavy objects without further training (see results in Figure 8). For door lifting, our policy uses only 50 demonstrations compared with 500 for $\pi_{0.5}$; for pose alignment, the corresponding budgets are 1,000 and 2,000. At the reference mass, the two methods achieve similar performance. As mass increases, however, our policy degrades more gradually: for example, on door lifting, it retains 32.6% success at 5.0 kg compared with 12.9% for $\pi_{0.5}$. These results provide evidence that robustness acquired during RL pretraining can remain useful after visual adaptation.

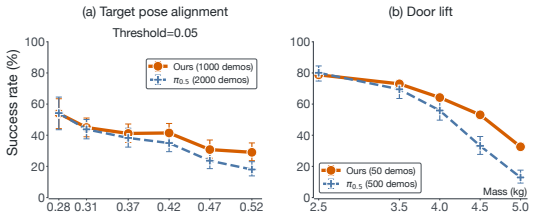

Figure 8: Generalization to increased object mass. (a) Target pose alignment at a relaxed ($\epsilon = 0.05$ m) thresholds. (b) Door lift. The leftmost point in each plot corresponds to the reference mass used for downstream visual training, e.g., for $\mathcal{D}_{\mathrm{vis}}$.

Linear probing of contact-force information. Contact forces are central to dexterous manipulation, and large-scale interaction during privileged RL pretraining may induce representations that capture contact-related physical information. To examine whether such information is retained after visual adaptation, we probe the hidden representations of the action decoder for contact forces. Specifically, for target pose alignment, we collect a separate force-annotated dataset $\mathcal{D}_{\mathrm{probe}}$ and fit linear ridge probes to the hidden state at each decoder layer, while keeping the adapted policy

frozen. We compare the RL-prior policy with the same architecture trained from scratch. As shown in Figure 9, contact-force information is substantially more linearly decodable from the RL-prior representations: the RL-prior model achieves lower probe RMSE than the scratch model across all evaluated layers and demonstration budgets. More notably, with only 200 visual demonstrations, the RL-prior model already matches or outperforms the scratch model trained with 4,000 demonstrations across the force targets, with substantial margins on most metrics. This result suggests that large-scale privileged RL pretraining implicitly learns structured contact-related physical representations that are retained after visual adaptation, but are difficult to acquire from limited visual demonstrations alone. Such retained physical knowledge may therefore contribute to the strong data efficiency provided by the RL prior.

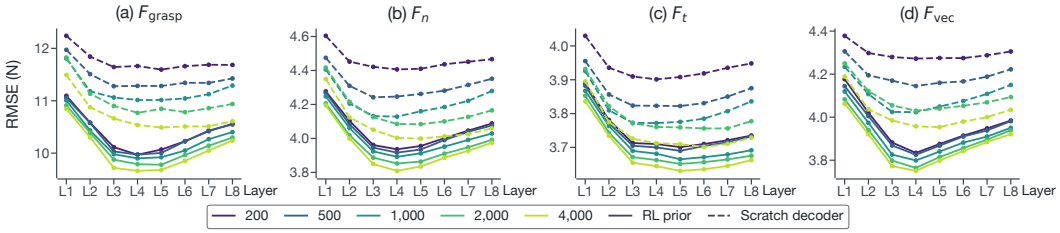

Figure 9: Linear contact-force probes of the visual policies. Panels predict (a) total grasp normal force, (b) per-finger normal forces, (c) per-finger tangential forces, and (d) per-finger Cartesian force vectors from each layer of the action decoder. Colors denote visual-demonstration budgets; solid and dashed curves denote the frozen pretrained and trainable scratch action decoders. Both policies fine-tune the image encoder at $0.2\times$ its base learning rate for 100k steps. Curves average three policy seeds on the same IID probe test episodes.

5 Limitations

Our results concern downstream visual-demonstration efficiency, not total data or compute efficiency: the method relies on a capable privileged teacher and substantial simulated experience. The frozen prior may inherit the teacher's failure modes and restrict adaptation when new tasks require different control behaviors. All tasks use the same arm–hand system; bimanual transfer reuses the single-arm decoder and does not establish transfer to different robot morphologies. The evaluation is entirely in simulation, including visual demonstration collection, so physical-robot performance remains untested. Finally, mass robustness and force readout support the proposed interpretation but do not isolate a causal mechanism for the demonstration-efficiency gains.

6 Conclusion

We presented Latent RL Prior, a simple yet effective framework that transfers RL control behavior to visual dexterous manipulation by adapting latent conditioning around a frozen action decoder. This adaptation uses action supervision alone, without privileged information in the visual demonstrations, and sharing the decoder extends a single-arm prior to bimanual manipulation. Across three simulated tasks, the method improves visual-demonstration efficiency, surpassing the evaluated VLA baselines trained on the full task datasets with $10\times$ and $5\times$ fewer demonstrations on door lifting and bimanual assembly, respectively. Further investigations favor latent transfer over the tested explicit interface for performance and training stability, while separate adaptation studies favor conservative visual updates and a frozen decoder. Robustness to mass variation absent from visual demonstrations and greater linear accessibility of contact-force information than in policies trained from scratch provide evidence consistent with retained control knowledge. These findings support preserving RL control priors and adapting their visual interfaces for data-efficient dexterous manipulation.

AI USE STATEMENT

We used generative AI tools to help draft and refine manuscript text based on author-provided drafts and an author-defined narrative, and to assist with coding and figure preparation. We did not use AI to propose research ideas, design experiments, or analyze or interpret results. The manuscript was reviewed by the authors, who take responsibility for the final content.

ETHICS STATEMENT

This work studies dexterous manipulation for robotics. No human data are used.

REPRODUCIBILITY STATEMENT

The experiment results are evaluated using at least three random seeds. The main text describes the method and experimental setup for reproducibility. We plan to publicly release the training and evaluation code and datasets.

REFERENCES

Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020. doi: 10.1177/0278364919887447.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, 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, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Proceedings of the Conference on Robot Learning*, volume 100, pp. 66–75. PMLR, 2020.

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 *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.026.

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 *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.090.

Google DeepMind and NVIDIA. MuJoCo Warp (MJWarp). `https://github.com/google-deepmind/mujoco_warp`, n.d. Software repository, accessed September 17, 2026.

Kushal Kedia, Tyler Ga Wei Lum, Jeannette Bohg, and C. Karen Liu. SimToolReal: An object-centric policy for zero-shot dexterous tool manipulation, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag 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, 2024.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. In *Proceedings of Robotics: Science and Systems*, 2021. doi: 10.15607/RSS.2021.XVII.011.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19730–19742. PMLR, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac Gym: High performance GPU-based physics simulation for robot learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

NVIDIA. NVIDIA Isaac GR00T N1.7: A foundation model for generalist robots. Hugging Face model card, 2026. URL `https://huggingface.co/nvidia/GR00T-N1.7-3B`. Early Access release.

NVIDIA, Johan Bjorck, Fernando Castaneda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots, 2025.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael 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, 2025.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. In *Proceedings of Robotics: Science and Systems*, 2018. doi: 10.15607/RSS.2018.XIV.008.

Haozhi Qi, Ashish Kumar, Roberto Calandra, Yi Ma, and Jitendra Malik. In-hand object rotation via rapid motor adaptation. In *Proceedings of the 6th Conference on Robot Learning*, volume 205, pp. 1722–1732. PMLR, 2023. URL `https://proceedings.mlr.press/v205/qi23a.html`.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. In *International Conference on Learning Representations*, 2016.

Jayesh Singla, Ananye Agarwal, and Deepak Pathak. SAPG: Split and aggregate policy gradients. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 45759–45772. PMLR, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Yixiao Wang, Yifei Zhang, Mingxiao Huo, Ran Tian, Xiang Zhang, Yichen Xie, Chenfeng Xu, Pengliang Ji, Wei Zhan, Mingyu Ding, et al. Sparse diffusion policy: A sparse, reusable, and flexible policy for robot learning. *arXiv preprint arXiv:2407.01531*, 2024.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.016.

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 *Proceedings of the 7th Conference on Robot Learning*, volume 229, pp. 2165–2183. PMLR, 2023.