DO YOU NEED PROPRIOCEPTIVE STATES FOR SPATIAL GENERALIZATION IN VISUOMOTOR POLICIES?

Anonymous authors
Paper under double-blind review

ABSTRACT

Visuomotor policies commonly combine images with proprioceptive states, yet such states can limit transfer to unseen spatial layouts. We empirically investigate State-free Policies, which remove proprioceptive state inputs, in the setting of spatial generalization from demonstrations collected under constrained spatial layouts. We study how state-input choices, action representations, and task-relevant visual coverage affect generalization. We identify that State-free Policies consistently outperforms the evaluated state-based alternatives, suggesting that reliance on state cues associated with training layouts can limit spatial generalization. This benefit depends on action representation, with relative end-effector (EEF) actions emerging as a key enabling condition. With this action representation, sufficient visual coverage for capturing task-relevant regions further supports State-free Policies' spatial generalization. Under these conditions, State-free Policies shows an increase in average success from 0% to 86% for height generalization and from 6% to 64% for horizontal generalization, compared with matched state-based policies. These findings characterize when state removal supports spatial generalization from demonstrations with constrained spatial coverage, and how action representation and visual coverage shape this benefit. *Real-world robot videos are in supplementary materials.*

1 INTRODUCTION

Visuomotor policies (Chi et al., 2023; Zhao et al., 2023; Black et al., 2024; Barreiros et al., 2025; Zhao et al., 2025; O'Neill et al., 2024; Khazatsky et al., 2024; Walke et al., 2023; Jones et al., 2025) commonly combine visual observations with proprioceptive states to predict robot actions (Torabi et al., 2019; Black et al., 2024; Kuang et al., 2025). Visual observations describe the scene and robot–object interactions, while proprioception provides information about the robot's current pose. When demonstrations only cover a constrained set of spatial layouts (Lin et al., 2024), proprioceptive states can correlate strongly with task progress and demonstrated actions (Chen et al., 2025; Zeng et al., 2021; Yin et al., 2023). A policy can exploit these training-layout-specific correlations to fit the demonstrations, yet generalize poorly to unseen spatial layouts.

We focus on spatial generalization, where the policy's intrinsic ability to perform the same task in layouts beyond those represented in task-specific training demonstrations. Prior work has explored relative proprioceptive states (Chi et al., 2024), perception alignment (Wang et al., 2026), and optimization strategies (Lu et al., 2026) for vision–proprioception learning. While, we ask whether removing proprioceptive state inputs can further improve spatial generalization when training demonstrations have constrained spatial coverage, and what the key conditions for this improvement are.

We empirically investigate State-free Policies, which remove explicit proprioceptive inputs from the policy model. We observe that, with relative end-effector (EEF) actions, removing state inputs already improves the spatial generalization under conventional wrist views. But performance can remain suboptimal when limited visual coverage misses task-relevant interaction regions, as shown in Figure 1(a). This observation motivates examining how action representation supports spatial transfer and how visual coverage supports successful execution.

Our action-representation evaluation clarifies the role of action representations in State-free Policies. All evaluated action representations succeed in the in-domain training layout, but only relative EEF

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.

(a) Spatial Generalization Overview

	Limited Visual Coverage	Sufficient Visual Coverage
State-based	Limited	Limited
State-free	Weak	**Strong**

(b) Expanding Visual Coverage Implement

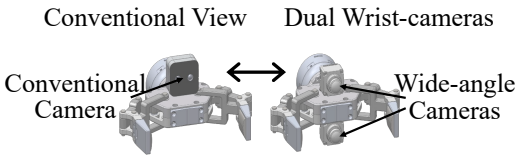


Figure 1: **State-free Policies spatial generalization and visual coverage.** (a) Summary of comparisons with state-based policies and limited visual coverage, using the relative action representation. Removing absolute state inputs improves spatial generalization, while broader wrist views further improve task success when conventional views miss relevant objects. (b) A conventional wrist camera and dual wide-angle wrist-cameras used to expand task-relevant visual coverage.

actions generalize to unseen layouts, which is a key condition of the spatial generalization for State-free Policies. This representation allows similar visual robot–object relationships to be associated with similar local motions across spatial layouts.

Because State-free Policies rely solely on visual observations (and task instructions) to predict actions, we find that capturing task-relevant objects and interaction regions needed to predict the current action chunk is a key condition for spatial generalization. We refer to this condition as *sufficient task-relevant visual coverage*. In our camera comparison, wider wrist views reveal the task-relevant regions missed by conventional wrist views and improve success (Figure 1(b)).

Our evaluation combines real-world and simulation experiments. We test height generalization through table height changes, and horizontal generalization through shifts in object or robot placement within the horizontal plane. In overall real-world evaluation, average success increases from 0% to 86% for height generalization and from 6% to 64% for horizontal generalization, over matched state-based policies. A controlled state-sensitivity analysis further links this failure to state shifts. Additional evaluations in representative settings reproduce the spatial generalization benefit across policy architectures, and show an advantage over a range of state-based alternatives. Supplementary studies further show better performance in low-data regime and better cross-embodiment adaptation, with the adapted State-free Policy also retaining spatial generalization on the new embodiment.

Our study yields three main empirical findings:

- **Proprioceptive inputs can hinder spatial generalization.** With constrained spatial coverage in demonstrations, State-free Policies achieve substantial gains over matched state-based policies while retaining comparable in-domain performance. Comparisons against a range of state-based alternatives further support this finding, even when they reduce absolute-coordinate information in proprioceptive inputs.
- **The benefit depends on action representation.** On a representative task, every evaluated representation learns to succeed in the training layout, but only relative EEF actions show the spatial generalization to unseen layouts. The action representation is therefore a key condition for spatial generalization.
- **Visual coverage conditions the state-free generalization.** State-free Policies predict actions solely relying on visual observations and task instructions. When conventional views miss task-relevant regions, expanding their visual coverage with dual wide-angle cameras improves success rates, complementing state removal and relative EEF actions.

2 Related Works

Input Design in Visuomotor Policies Visuomotor policies (Bjorck et al., 2025; Liu et al., 2024b; Zitkovich et al., 2023; Brohan et al., 2022; Team et al., 2024; Barreiros et al., 2025; Liu et al., 2024a; Lin et al., 2025; Zhang et al., 2025; Wang et al., 2025) have been widely adopted for robot learning, including Diffusion Policy (Chi et al., 2023), ACT (Zhao et al., 2023), and π_0 (Black et al., 2024). These policies commonly combine visual observations with proprioceptive states. HPT (Wang et al., 2024) studies joint visual–proprioceptive pretraining across heterogeneous embodiments and reports

Table 1: **State inputs and visual coverage in height generalization.** Results on *Pick Pen* after fine-tuning π_0 with relative EEF actions. "Hack state" applies a table-height-dependent offset to the policy's state input at evaluation. "Noised state" adds random noise in $[-5, 5]$ cm to the state height coordinate during fine-tuning. ID and OOD denote in-domain and out-of-domain table heights settings. ✓ indicates a success rate above 90%, ✗ indicates below 10%, and weak indicates in between.

Wrist-camera		Conventional camera			Dual wide-angle cameras
State input	w/ state	Hack state	Noised state	w/o state	w/o state
h=80cm (ID)	✓	✓	✓	✓	✓
h=72cm (OOD)	✗	✓	weak	✓	✓
h=90cm (OOD)	✗	✓	weak	weak	✓

benefits from both modalities. RT-2 (Brohan et al., 2023) and OpenVLA (Kim et al., 2024) formulations instead predict actions from images and language without explicit proprioceptive inputs, but not isolating how state-input choices and task-relevant visual coverage affect spatial transfer.

Proprioception and Spatial Generalization UMI (Chi et al., 2024) represents historical EEF poses relative to the current pose, providing motion history instead of absolute states. PALM (Wang et al., 2026) combines local visual alignment with proprioception that omits planar position. GAP (Lu et al., 2026) modulates proprioceptive gradients according to motion-transition phases. These approaches retain proprioceptive inputs while modifying their representation or use during learning. Our findings show that, when demonstrations have constrained spatial coverage, removing state inputs can yield stronger spatial generalization than the evaluated alternatives that retain them.

3 State-free Policies

We study State-free Policies, which omit explicit proprioceptive states from the policy model inputs. We first examine the sensitivity of spatial generalization to state inputs, then investigate how action representation and visual coverage support State-free Policies' spatial generalization.

3.1 State Inputs under Spatial Distribution Shifts

When demonstrations cover constrained spatial layouts, states can correlate with these trajectories. We examine it through a controlled table-height study, involving placing a pen into a pen holder. We fine-tune π_0 using demonstrations collected at a table height of $80\,\mathrm{cm}$, with relative EEF actions.

As shown in Table 1, the policy with state inputs fails at unseen table heights. For example, at 72 cm, we observe that the policy repeatedly attempts to grasp noticeably above the pen, consistent with the state-action relationships in the training layout. Simply applying a table-height-dependent offset to the state at evaluation recovers high success without any other change. Together, these observations indicate that the failure arises from reliance on state associated with the training layout, rather than an inability to execute the task at unseen layouts. This motivates removing state inputs and comparing state removal with alternatives that retain the state input.

3.2 Relative EEF Actions

Removing state inputs leaves an important question: how should the policy represent actions so that similar visual task geometry can support similar motions across spatial layouts? Consider two spatial layouts where the target is at different positions, while the EEF reaches the same pose relative to it at corresponding stages of the interaction. The visual observations $o_t^{(1)}$ and $o_t^{(2)}$ can depict similar local EEF-object relationships even when the underlying joint configurations differ.

We denote a local EEF pose increment by $\Delta\mathbf{P}_t = [\Delta\mathbf{x}_t^\top, \Delta\boldsymbol{\theta}_t^\top]^\top \in \mathbb{R}^6$, where $\Delta\mathbf{x}_t$ and $\Delta\boldsymbol{\theta}_t$ are the translation increment and rotation vector, respectively, expressed in a common Cartesian frame. For the corresponding action, the desired increment is shared across the two layouts:

$$\Delta\mathbf{P}_t^{\star,(1)} = \Delta\mathbf{P}_t^{\star,(2)} = \Delta\mathbf{P}_t^{\star}. \tag{1}$$

Let $f_\theta(o_t)$ denotes the action predicted by the State-free Policy, with sampling randomness held fixed when applicable. We analyze the case where visual observations with similar task-relevant content lead to similar actions predicted by the policy,

$$\widehat{\mathbf{a}}_t^{(1)} = f_\theta(o_t^{(1)}) \approx f_\theta(o_t^{(2)}) = \widehat{\mathbf{a}}_t^{(2)}. \tag{2}$$

Whether these predictions produce the required motion depends on the action representation.

Relative joint actions. In this representation, the predicted action specifies a joint increment. Let $\mathbf{q}_t^{(i)} \in \mathbb{R}^n$ denote the joint configuration in layout i, and let $J(\mathbf{q}_t^{(i)}) \in \mathbb{R}^{6\times n}$ denote its geometric Jacobian. To isolate the effect of configuration, consider the same small predicted joint increment $\Delta\mathbf{q}_t$ in both layouts. The resulting EEF pose increments satisfy

$$\Delta\mathbf{P}_t^{(i)} \approx J(\mathbf{q}_t^{(i)})\Delta\mathbf{q}_t, \qquad i \in \{1, 2\}. \tag{3}$$

Suppose this joint action produces the desired motion in the first layout, $J(\mathbf{q}_t^{(1)})\Delta\mathbf{q}_t \approx \Delta\mathbf{P}_t^\star$. Applying the same command in the second layout yields

$$\Delta\mathbf{P}_t^{(2)} - \Delta\mathbf{P}_t^\star \approx \left[J(\mathbf{q}_t^{(2)}) - J(\mathbf{q}_t^{(1)})\right]\Delta\mathbf{q}_t. \tag{4}$$

Because the Jacobian depends on the current joint configuration, this error can be nonzero even though both layouts require the same EEF motion. A joint action that correctly moves the EEF relative to the target in one layout can therefore produce an incorrect motion in another. Making joint actions relative does not eliminate this configuration dependence.

Relative EEF actions. In this representation, the predicted action $\widehat{\mathbf{a}}_t \in \mathbb{R}^6$ directly specifies the desired local EEF pose increment. Under the similar predictions in Eq. 2, an action that specifies the required motion in the first layout also specifies that motion in the second,

$$\widehat{\mathbf{a}}_t^{(1)} \approx \Delta\mathbf{P}_t^\star \quad \Longrightarrow \quad \widehat{\mathbf{a}}_t^{(2)} \approx \Delta\mathbf{P}_t^\star. \tag{5}$$

The controller uses the current robot state to execute these EEF actions (such a state is excluded from the policy inputs), which allows similar visual EEF-target relationships to be associated with the same task-appropriate motion action across translated layouts.

Our following action-representation evaluations in Section 4.4 support relative EEF actions as a key enabling condition for state-free spatial generalization in the evaluated setting. Note that this derivation assumes that spatial shifts remain within a range where the desired EEF action is shared across layouts; sufficiently large shifts may violate the assumption in Eq.2, and we discuss the corresponding visual-input setting in SectionF.

3.3 Sufficient Visual Coverage

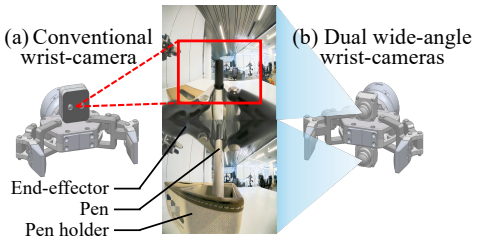


Figure 2: (a) The conventional wrist-camera view misses the pen holder beneath the EEF. (b) Dual wide-angle wrist cameras expand the visual coverage to include the pen holder in the illustrated placement configuration.

With relative EEF actions, removing state inputs already improves the spatial generalization under conventional wrist views, but success remains suboptimal at unseen settings, as shown in Table 1. We observe that failures are concentrated on placing the pen into the pen holder, and find that the conventional wrist view does not capture the pen holder position, as illustrated in Figure 2. Although our default setting includes an overhead view for global context, it provides limited detail of the object interaction region. We therefore examine sufficient task-relevant visual coverage as a complementary factor affecting the success, i.e. where the policy can observe the task-relevant objects and interaction regions needed to predict the current action chunk. This does not require every scene object to remain visible throughout the trajectory.

In State-free Policies, information about the environment and task targets is available only through visual observations. When targets are outside the available views, visual observations may leave the required action ambiguous. For example, during pen placement, a conventional wrist view can miss the pen holder beneath the EEF (Figure 2(a)). We address this by expanding task-relevant visual coverage using dual wide-angle wrist cameras mounted above and below the EEF (Figure 2(b)). Each wide-angle camera has a field of view of $120^\circ \times 120^\circ$, compared with $87^\circ \times 58^\circ$ for the conventional RealSense wrist camera. The expanded wrist views expose the region beneath the EEF and improve the success rates at $90\,\mathrm{cm}$ while maintaining high success at $72\,\mathrm{cm}$ (Table 1). Further camera comparisons in Section 4.5 examine single and dual wrist cameras with conventional and wide-angle fields of view. These results support expanding task-relevant visual coverage to improve state-free execution when conventional views miss relevant objects.

Note that dual wide-angle wrist cameras provide one implementation of this condition; conventional wrist views may already provide sufficient visual coverage, depending on the task. We also emphasize that, even when proprioceptive inputs are retained, they do not necessarily compensate for missing task-relevant visual information in spatial generalization when visual observations become unreliable, e.g., due to occlusion. This is because state-action associations learned from the training layouts can become unreliable when target locations change.

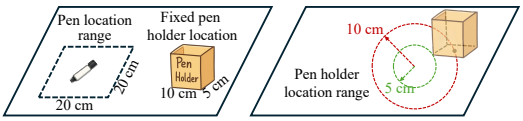
Summary The relative EEF actions is a key enabling condition for spatial generalization in State-free Policies. Under this condition, missing task-relevant visual information can still limit generalization success. When conventional views omit task-relevant regions, we expand visual coverage with dual wide-angle wrist cameras. We next evaluate spatial generalization across systematic evaluations, and examine how state inputs, action representations, and visual coverage affect performance.

4 Experiments

4.1 Setup

Task and Data We investigate the performance of State-free Policies on real-world robots, complemented by simulations. Our real-world tasks include 3 *"Pick & Place"* tasks and 2 challenging long-horizon tasks, *"Fold Shirt"* and *"Fetch Bottle"* (on a whole-body robot with torso controlled). These tasks cover different manipulation skills, including precise grasping, accurate placement, deformable-object manipulation, and long-horizon manipulation with multiple sub-actions. The detailed task descriptions are as follows (more details can be found in our demo videos):

- **Pick Pen.** Pick up a pen and place it into a pen holder on the table.
- **Pick Bottle.** Grasp the bottle cap and remove the bottle from the step.
- **Put Lid.** Pick up the lid and accurately place it on the teacup on the table.
- **Fold Shirt.** Fold the shirt that is laid flat on the table.
- **Fetch Bottle (whole-body).** Open the refrigerator, take out the bottle, and close the door.



(a) Location in training data (in-domain)

(b) Locations in horizontal generalization evaluation (out-of-domain)

Figure 3: Object locations in *"Pick Pen"* task. During training, the pen is placed within a small region and the pen holder is fixed. For horizontal generalization, we keep the pen range unchanged and shift the pen holder by 5 cm and 10 cm.

Figure 4 presents representative tasks across three robot embodiments, including a human-like dual-arm robot, an Arx5 dual-arm system, and a whole-body robot. We collect real-world demonstrations via teleoperation with professional operators. Data amounts are detailed in Appendix A. During collection, we fix the table height and constrain object locations to a limited 2D range. For example, in "*Pick Pen*" (Figure 3), the pen holder location is fixed during training and shifted by 5 cm and 10 cm for horizontal evaluation. This ensures that spatial generalization reflects the learned policy rather than diverse training data.

In simulation, we use LIBERO (Liu et al., 2023) for in-domain and RoboCasa365 (Nasiriany et al., 2026) for spatial generalization evaluation. Results are reported in Appendix D.



Figure 4: Overview of our representative tasks and embodiments, including "*Pick & Place*" tasks, challenging "*Fold Shirt*" and "*Fetch Bottle*" tasks. These tasks span three kinds of embodiments: a 2×8 DoF human-like robot, a 2×7 DoF Arx5 dual-arm system, and a 26 DoF whole-body robot.

Model Structures In main evaluations, we use the classic π_0 (Black et al., 2024) policy and follow its released fine-tuning recipe. This setting allows us to focus on native spatial generalization under task-specific fine-tuning, with less influence from the broader spatial and state coverage potentially provided by more diverse pretraining data (e.g., $\pi_{0.5}$). Implementation and evaluation details, including independent training runs, proprioceptive-state inputs, and controlled camera comparisons, are provided in Appendix B. In Appendix E, we also evaluate other architectures, including ACT (Zhao et al., 2023), Diffusion Policy (Chi et al., 2023), and $\pi_{0.5}$ (Intelligence et al., 2025), showing that State-free Policies generalize across model architectures.

Evaluation Metric We evaluate spatial generalization along height and horizontal dimensions.

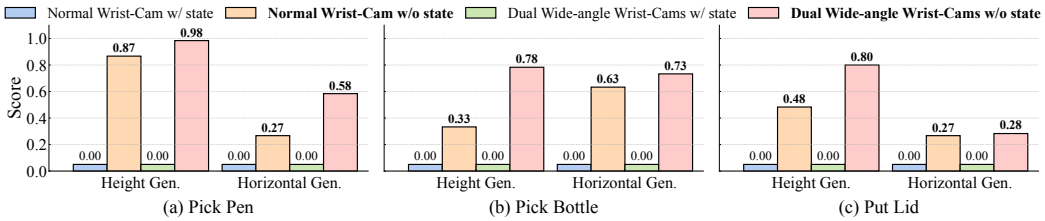
For height generalization, *"Pick & Place"* data are collected at an 80 cm table height, and the score is the average success rate over 60 trials at two OOD heights, 72 cm and 90 cm. As shown in Figure 4, *"Fold Shirt"* and *"Fetch Bottle (whole-body)"* are not applicable because the Arx5 arms are fixed to the table and the refrigerator height cannot be adjusted.

For horizontal generalization, we shift target objects in *"Pick & Place"* and *"Fetch Bottle (whole-body)"* within 2D ranges of 5 cm and 10 cm, including the pen holder, step, teacup, and refrigerator, as illustrated by *"Pick Pen"* in Figure 3. For *"Fold Shirt"*, we shift either one Arx5 arm by 15 cm or both arms by 15 cm in opposite directions, with illustrations in video demos. Each setting (shifting 5 cm or 10 cm) has 30 trials; thus each task has 60 trials to compute the average success rate.

4.2 Overall Performance of State-free Policies

We first present the overall performance of State-free Policies on real-world tasks. As shown in Figure 5 and Table 2, removing proprioceptive state consistently improves spatial generalization under constrained training layouts. In three pick-and-place tasks, state-based policies fail under both generalization settings, whereas State-free Policies achieve substantially higher success rates. The same trend extends to the more challenging "*Fold Shirt*" and "*Fetch Bottle (whole-body)*" tasks, where horizontal-generalization success increases from 18.3% to 83.3% and from 11.7% to 78.3%, respectively (as discussed in Section 4.1, height generalization is not applicable to these two tasks).

The results also separate the effect of state removal from that of visual coverage. With conventional wrist views alone, removing state already yields considerable spatial generalization, while expanding task-relevant visual coverage with dual wide-angle wrist cameras further improves performance. Furthermore, *Fold Shirt* and *Fetch Bottle* achieve strong State-free generalization using only conventional wrist cameras, suggesting that the required visual coverage is task-dependent rather than tied to a particular camera configuration.



(a) Pick Pen (b) Pick Bottle (c) Put Lid

Figure 5: Spatial generalization performances on real-world *"Pick & Place"* tasks. With full task observation, State-free Policies show much stronger spatial generalization than state-based policies.

Table 2: Horizontal generalization performances across 2 challenging tasks, *"Fold Shirt"* and *"Fetch Bottle (whole-body)"*.

	Fold Shirt	*Fetch Bottle*
w/ state	0.183	0.117
w/o state	**0.833**	**0.783**

In corresponding in-domain evaluations, both state-based and State-free Policies achieve consistently at least 90.0% success (detailed results in Appendix C). Complementary simulation results, including LIBERO for in-domain evaluation and RoboCasa365 for spatial generalization, are reported in Appendix D. Results across additional model structures are provided in Appendix E. Beyond spatial generalization, State-free Policies also show better performance in low-data regime and better cross-embodiment adaptation, as further examined in Appendix G. In the following sections, we move beyond aggregate success rates and conduct controlled empirical studies to identify the factors underlying these spatial-generalization behaviors.

4.3 Does Removing Absolute Coordinates Works in Spatial Generalization?

The preceding results show a spatial generalization gap between state-based and State-free Policies under constrained training spatial layouts. We analyze whether this gap arises from the absolute coordinate information in proprioceptive states. We therefore examine whether alternative state representations or training strategies can narrow this gap while retaining proprioceptive inputs.

Table 3: Height generalization on *Pick Pen* with alternative uses of proprioceptive state. All methods are provided with sufficient visual coverage.

Backbone	State treatment	Success
π_0	State-based Policy	0.0
	State-hacked Policy	0.967
	PALM	0.117
	Relative State (UMI)	0.217
	Noised State Aug	0.633
	State-free Policy	**0.983**
$\pi_{0.5}$	State-based Policy	0.517
	GAP	0.617
	State-free Policy	**0.967**

We compare absolute-state inputs with state-hacked policy, UMI-style relative state history (Chi et al., 2024), state-noise augmentation, and PALM (Wang et al., 2026), which weaken absolute coordinate information, as well as GAP (Lu et al., 2026), which regulates the use of states, and State-free Policies. For the UMI-style baseline, we follow the original two-frame observation setting, expressing EEF pose history relative to the current pose. State-hacked policy applies a table-height-dependent offset to the state height coordinate only at evaluation. State-noise augmentation perturbs proprioceptive inputs during training. GAP modulates proprioceptive gradient updates according to estimated motion-transition phases (implement details in Appendix B). We include PALM to test whether aligned proprioception that retains height transfers to unseen table heights. We also include $\pi_{0.5}$ (Intelligence et al., 2025), a model pretrained on large-scale and diverse data.

We evaluate them on the representative "*Pick Pen*" height-generalization setting, where height shifts provide a controlled real-world perturbation with clear success criteria. All methods achieve at least 95.0% in-domain success. As shown in Table 3, state-retaining alternatives improve generalization to varying degrees but remain below State-free Policies. Among them (exclude the manually corrected state-hacked policy), state-noise augmentation performs best, while UMI, PALM, and GAP provide smaller gains. The state-hacked policy approaches the performance of State-free Policies.

We further conduct a sensitivity analysis of the State-hacked Policy by varying the manual correction applied to the input state toward the evaluated table height. As the residual height offset decreases, the policy behavior progressively recovers, reflected in closer approaches to the pen during grasping and more appropriate lifting heights during placement. Success increases markedly when the residual height offset falls within approximately 2 cm. These results indicate that, in this evaluation setting, state-based policy 's actions are strongly conditioned on the supplied state coordinates, making its behavior sensitive to coordinate shifts encountered at test time.

In summary, removing proprioceptive inputs yields the highest height-generalization performance among the evaluated strategies on each backbone under limited-layout training. Its advantage persists when compared with relative state history, state-noise augmentation, and phase-guided gradient adjustment. These results suggest that the limitation can not be fully addressed by reducing absolute coordinate information in state; retaining proprioceptive states can still encourage the policy to rely on state cues associated with training layouts and is consistent with failure in spatial generalization.

We next examine two conditions for translating State-free Policies into effective spatial generalization, including action representation and task-relevant visual coverage.

4.4 How Does Action Representation Affect Spatial Generalization?

Having examined how proprioceptive inputs affect spatial generalization, we next ask whether removing state alone is sufficient. As analyzed in Section 3.2, different action representations impose different dependencies on the robot configuration and spatial reference frame. We therefore compare relative EEF, absolute EEF, relative joint, and absolute joint actions on the same "*Pick Pen*" height-generalization setting, using dual wide-angle wrist-cameras to provide sufficient visual coverage.

Table 4: In-domain and height-generalization performance of State-free Policies across different action spaces. "rel." and "abs." denote relative and absolute, respectively.

Action space	In-Domain	Height Gen.
rel. EEF	1.0	0.983
abs. EEF	1.0	0.0
rel. joint	1.0	0.0
abs. joint	1.0	0.0

As shown in Table 4, all four action representations achieve perfect in-domain success, confirming that the policies can learn the task under the training layout. Their behavior differs sharply under the unseen height shift. Relative EEF actions retain a success rate of 98.3%, whereas absolute EEF and both joint-space representations drop to zero. The difference therefore reflects spatial transfer rather than failure to acquire the in-domain task.

These results are consistent with Section 3.2. Relative EEF actions express the desired local end-effector motion without requiring the policy to reproduce an absolute target pose or a configuration-dependent joint-space action. Thus, state removal alone is not sufficient; the action representations also play an important role in the spatial generalization of State-free Policies.

4.5 How Does Visual Coverage Affect State-free Spatial Generalization?

Table 5: Spatial generalization of State-free Policies under different visual-coverage configurations, varying the number and field of view of wrist cameras.

Wrist-camera number	Wrist-camera type	Overhead camera	Height Gen.	Horizontal Gen.
N/A	N/A	✓	0.217	0.133
Single	Conventional	✓	0.867	0.267
Dual	Conventional	✓	0.917	0.400
Single	Wide-angle	✓	0.917	0.500
Dual	Wide-angle	✓	0.983	0.583

State removal and relative EEF actions alone do not guarantee successful spatial generalization, as the policy must still observe the task-relevant objects and interaction regions needed to predict the

current action. We therefore examine how visual coverage affects State-free Policies by systematically varying the number and field of view of wrist cameras.

As shown in Table 5, spatial generalization improves as the wrist-camera observations cover more of the manipulation region. With the overhead camera alone, height- and horizontal-generalization success rates are 21.7% and 13.3%, respectively. Adding a conventional wrist camera raises them to 86.7% and 26.7%, while dual wide-angle wrist cameras further increase them to 98.3% and 58.3%. These comparisons show that wrist observations, and broader coverage of the interaction region, improve performance under the tested shifts.

Scope of State-free Policies Spatial Generalization Our analysis in Section 3.2 assumes that spatial shifts remain within a range where similar local EEF-target configurations produce similar visual observations, and therefore similar relative EEF actions. In practice, State-free Policies begin to fail when the spatial shift becomes sufficiently large. We find substantial changes in the fixed overhead-camera observations, and provide visualizations of this effect in Appendix Figure 6.

To examine this failure mode, we evaluate "*Pick Pen*" under three more challenging spatial shifts: 100 cm table height, double-height pen holder, and 20 cm pen holder shift from the training position.

Table 6: Success rates of State-free Policies on task"*Pick Pen*" under increasingly challenging spatial shifts, with dual wide-angle wrist-cameras.

Overhead Camera	Table 72 & 90 cm	Table 100 cm	Double Height Pen Holder	Shifted Pen Holder 20 cm
✓	0.983	0.0	0.467	0.0
✗	1.0	1.0	0.867	0.8

As shown in Table 6, State-free Policies with the overhead view degrade substantially, whereas removing it yields higher success rates. This pattern is consistent with visual-input consistency contributing to performance across layouts. Wrist cameras move with the end-effector and can preserve relatively similar observations of the local interaction region, while a fixed overhead camera can undergo much larger appearance changes as objects move substantially across the workspace.

These results suggest that spatial generalization of State-free Policies only on depends not whether task-relevant information is adequately visible, but also can be affected by whether visual observations remain sufficiently consistent across spatial layouts. We do not claim that overhead observations are generally unnecessary, as they can provide useful global context or additional task-relevant information in other tasks. We therefore retain the overhead camera in our default setting, treat this sensitivity to large visual shifts as the scope of State-free Policies, and leave the design of more robust visual-input configurations for State-free Policies to future work.

5 Conclusion

We investigate State-free Policies under task-specific training with constrained spatial coverage. Our experiments show improved spatial generalization over matched state-based policies while maintaining comparable in-domain performance. Comparisons with alternative state representations and training strategies suggest that these gains extend beyond avoiding absolute-coordinate shifts, reflecting reduced reliance on state cues tied to training layouts. Relative EEF actions support spatial transfer, while expanded task-relevant visual coverage improves success when conventional views omit relevant information. Further experiments highlight visual consistency across layouts also as an important factor in the spatial generalization of these policies.

Limitation Our study are limited to the native spatial generalization arising from task-specific training with constrained spatial coverage. This ability is effective within a certain range of spatial generalization, whose boundary is not yet quantitatively characterized. And our experiments do not address tasks inherently requiring the joint-space action representation, which we expect to represent a relatively limited subset of manipulation tasks. Our results suggest that maintaining task-relevant visual consistency across layouts may be important, and we leave the study of more suitable visual-input configurations for State-free Policies to future work.

AI USE STATEMENT

Large Language Model tools were used to assist with language editing and revision of the manuscript. All AI-assisted contents were critically reviewed by the authors. The experimental results were obtained from author-conducted experiments, and the final methodological choices, analyses, interpretations, and claims were determined and verified by the authors. The authors take full responsibility for the content of this work.

ETHICS STATEMENT

This work studies visuomotor policies and does not involve human-subject experiments. Real-world robot demonstrations are collected through teleoperation by professional operators; the operators themselves are not the subject of the study, and no conclusions are drawn about human behavior or personal attributes. We are not aware of additional ethical concerns specific to the proposed method.

REPRODUCIBILITY STATEMENT

Our main experiments use the released π_0 implementation and follow its officially released fine-tuning recipe. Appendix B provides the implementation details specific to this study, where the states are directly removed rather than replaced with zeros.

REFERENCES

Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, et al. A careful examination of large behavior models for multitask dexterous manipulation. *arXiv preprint arXiv:2507.05331*, 2025.

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

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π0: A vision-language-action flow model for general robot control. corr, abs/2410.24164, 2024. doi: 10.48550. *arXiv preprint ARXIV.2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Yifei Chen, Yuzhe Zhang, Giovanni D'urso, Nicholas Lawrance, and Brendan Tidd. Improving generalization ability of robotic imitation learning by resolving causal confusion in observations. *arXiv preprint arXiv:2507.22380*, 2025.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, pp. 02783649241273668, 2023.

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

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. URL https://arxiv.org/abs/2504.16054.

Joshua Jones, Oier Mees, Carmelo Sferrazza, Kyle Stachowicz, Pieter Abbeel, and Sergey Levine. Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding. *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5961–5968, 2025.

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

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

Fuhang Kuang, Jiacheng You, Yingdong Hu, Tong Zhang, Chuan Wen, and Yang Gao. Adapt your body: Mitigating proprioception shifts in imitation learning. *arXiv preprint arXiv:2506.23944*, 2025.

Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. *ArXiv*, abs/2410.18647, 2024.

Fanqi Lin, Ruiqian Nai, Yingdong Hu, Jiacheng You, Junming Zhao, and Yang Gao. Onetwovla: A unified vision-language-action model with adaptive reasoning. *ArXiv*, abs/2505.11917, 2025.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qian Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *ArXiv*, abs/2306.03310, 2023.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *ArXiv*, abs/2410.07864, 2024a.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024b.

Jingxian Lu, Wenke Xia, Yuxuan Wu, Zhiwu Lu, and Di Hu. When would vision-proprioception policies fail in robotic manipulation? *arXiv preprint arXiv:2602.12032*, 2026.

Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. Robocasa365: A large-scale simulation framework for training and benchmarking generalist robots. *arXiv preprint arXiv:2603.04356*, 2026.

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

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Faraz Torabi, Garrett Warnell, and Peter Stone. Imitation learning from video by leveraging proprioception. *arXiv preprint arXiv:1905.09335*, 2019.

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

Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *Advances in neural information processing systems*, 37:124420–124450, 2024.

Ruiyu Wang, Zheyu Zhuang, Danica Kragic, and Florian T Pokorny. Palm: Enhanced generalizability for local visuomotor policies via perception alignment. *IEEE Robotics and Automation Letters*, 2026.

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

Zhao-Heng Yin, Yang Gao, and Qifeng Chen. Spatial generalization of visual imitation learning with position-invariant regularization. In *RSS 2023 Workshop on Symmetries in Robot Learning*, 2023.

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

Di Zhang, Chengbo Yuan, Chuan Wen, Hai Zhang, Junqiao Zhao, and Yang Gao. Kinedex: Learning tactile-informed visuomotor policies via kinesthetic teaching for dexterous manipulation. *ArXiv*, abs/2505.01974, 2025.

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

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

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

APPENDIX

A AMOUNT OF REAL-WORLD DATA

We employ professional data collectors to collect the real-world demonstration data using the teleoperation. For each *"Pick & Place"* task, 300 trajectory episodes, around 5 hours of data are collected.

For the other 2 tasks, we do not report the number of episodes, since they are long-horizon tasks and each episode does not cover a complete task execution. For the challenging *"Fold Shirt"* task, around 20 hours of data are collected (counting only the folding process from the flattened state to the folded state). For the challenging *"Fetch Bottle (whole-body)"* task, around 80 hours of data are collected due to its long-horizon nature with multiple challenging sub-skills, while the refrigerator position remains fixed during data collection.

B ADDITIONAL IMPLEMENTATION AND EVALUATION DETAILS

Proprioceptive State Input. The proprioceptive state consists of the EEF pose expressed in the world coordinate frame, with rotation represented in axis-angle form. The way the state is incorporated follows the corresponding policy architecture. For π_0, the state is embedded and placed before the noised-action embeddings in the action-expert input sequence. For $\pi_{0.5}$, the state is provided to the VLM branch. For State-free Policies, we directly remove the state input and its corresponding embedding rather than replacing it with zeros.

Controlled Camera Comparisons. The compared settings use the same demonstration trajectories and data processing, differing only in the visual observations by cropping the recorded images. This controlled setup isolates the effect of camera configuration and task-relevant visual coverage.

Independent Training Runs and Checkpoint Selection. To examine the variability across independent training runs, we fine-tune the State-free Policy on "*Pick Pen*" with three different random seeds. For each run, we use the checkpoint after training has converged for evaluation. The variance of the height-generalization success rate across the three runs is 0.00191, whereas all three state-based runs achieve 0 success. This indicates that the observed performance is stable across independent training runs.

GAP Adaptation to $\pi_{0.5}$. We implement GAP (Lu et al., 2026) adaptation on $\pi_{0.5}$ following its original procedure. we use change-point detection on EEF pose and gripper trajectories to supervise a temporal estimator of motion-transition probabilities, and use the batch-averaged probability to modulate gradients of the proprioceptive input branch in $\pi_{0.5}$'s VLM. We use $\alpha = 1$, $\beta = 2 \times 10^{-3}$, and $\lambda = 0.3$, applying gradient modulation during the first half of fine-tuning.

C REAL-WORLD IN-DOMAIN PERFORMANCE

Table 7: In-domain success rates in different real-world tasks with and without state input.

Task name	w/ state input	w/o state input
Pick Pen	1.0	1.0
Pick Bottle	1.0	1.0
Put Lid	1.0	1.0
Fold Shirt	1.0	0.967
Fetch Bottle (whole-body)	0.900	0.933

In this section, we report the in-domain performance of both state-based policies and State-free Policies in our real-world tasks. In Table 7, we report the in-domain success rates for different real-world tasks using policies with and without state input, with dual wide-angle wrist-cameras. Even after removing the state input, the policies still maintain comparable performance on in-domain tasks, as the distribution of visual observations remains fully consistent with training.

D SIMULATION EVALUATIONS

We also conduct experiments in simulation to evaluate both the in-domain performance and spatial generalization ability of State-free Policies. Since defining new wide-angle camera configurations and modifying the table height in simulation require great engineering effort, our spatial generalization evaluation follows the original simulation setup. Under this setting, we mainly evaluate the horizontal generalization ability of simple State-free Policies.

D.1 SIMULATION ON THE LIBERO BENCHMARK

Table 8: Simulation success rates on the Libero benchmark, using policies with and without state input.

Evaluation suite	w/ state input	w/o state input
Libero Goal	0.942	0.956
Libero Object	0.964	0.962
Libero Spatial	0.968	0.976
Libero 10	0.876	0.886
Average	0.938	0.945

First, we compare the in-domain performance of the π_0 policy with and without state input on the LIBERO (Liu et al., 2023) benchmark. As reported in Table 8, State-free Policies achieve performance as perfect as state-based policies, and in some cases even surpass them.

D.2 SIMULATION ON THE ROBOCASA365 BENCHMARK

On the RoboCasa365 (Nasiriany et al., 2026) benchmark, we compare the $\pi 0.5$ policy with and without state input. During evaluation, object placements are out-of-domain relative to the training data, primarily evaluating the horizontal generalization.

Table 9: Simulation success rates on the RoboCasa365 benchmark, using policies with and without state input.

Task	w/ state input	w/o state input
CloseFridge	0.620	0.840
CloseToasterOvenDoor	0.200	0.220
NavigateKitchen	0.420	0.460
OpenDrawer	0.220	0.240
OpenStandMixerHead	0.560	0.640
TurnOnSinkFaucet	0.100	0.140
Average	0.353	0.423

We evaluate the policy on six tasks: *CloseFridge*, *CloseToasterOvenDoor*, *NavigateKitchen*, *OpenDrawer*, *OpenStandMixerHead*, and *TurnOnSinkFaucet*. Table 9 reports the task success rates with and without robot proprioceptive state inputs. The State-free Policy consistently outperforms the state-based policy across all evaluated tasks. In particular, significant improvements are observed on *CloseFridge* and *OpenStandMixerHead*.

E Evaluations Across Policy Architectures

Table 10: Spatial generalization of policies with and without state input, using different model structures. "DP" refers to Diffusion Policy.

Model	State	Height Gen.	Horizontal Gen.
π_0	✓	0	0
	✗	0.983	0.583
ACT	✓	0	0.083
	✗	0.933	0.517
DP	✓	0	0
	✗	0.867	0.533
$\pi_{0.5}$	✓	0.517	0.533
	✗	0.967	0.883

We also evaluate different policies, including π_0 (Black et al., 2024), ACT (Zhao et al., 2023), Diffusion Policy (DP) (Chi et al., 2023), and $\pi_{0.5}$ (Intelligence et al., 2025). ACT uses action chunking to model temporally extended behaviors; Diffusion Policy models action distributions through diffusion dynamics; $\pi_{0.5}$ is pre-trained on large-scale diverse data and naturally exhibits a certain degree of spatial generalization. All evaluations use dual wide-angle wrist-cameras for full task observation. In Table 10, State-free Policies consistently outperform state-based policies across architectures, indicating their effectiveness is not tied to a specific policy implementation.

F Visualization of Overhead-View Distribution Shift



Table height = 100 cm

In-domain setting
Table height = 80 cm

Shifting pen holder 20 cm

Figure 6: Overhead-camera observations of Pick Pen in the training layout, at a 100 cm table height, and with the pen holder shifted by 20 cm. These examples illustrate changes in the fixed overhead view under larger spatial shifts.

Figure 6 visualizes the overhead-camera observations under increasingly large spatial shifts in the *Pick Pen* task. As the object layout moves further beyond the training distribution, the overhead view exhibits substantial appearance changes, while the wrist-camera observations remain more consistent around the local manipulation region.

G Further Benefits of State-free Policies

In this section, we demonstrate additional advantages of State-free Policies, including the higher data efficiency and better cross-embodiment adaptation.

In-domain Performance with Fewer Demonstrations. We evaluate "*Pick Pen*" using 50, 100, 200, and 300 fine-tuning demonstrations, with success measured after two and four epochs. As shown in Figure 7, State-free Policies achieve higher **in-domain** success at several smaller data budgets and evaluated checkpoints, because state-based policies are more prone to overfitting the training trajectories with limited data. At 300 demonstrations, both variants reach the same training loss of 0.003, while State-free Policies achieve a lower evaluation loss (0.005 vs. 0.009). This result shows an advantage in in-domain performance under the tested data budgets and training schedule.

Cross-embodiment Adaptation State-free Policies also benefit cross-embodiment fine-tuning. For state-based policies, adaptation to a new embodiment requires aligning a new state space; even EEF states can differ in reference-frame definitions. In contrast, State-free Policies avoid this issue.

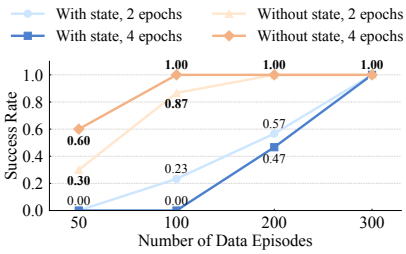


Figure 7: **In-domain** success rates on *"Pick Pen"* task with varying fine-tuning data amounts.

Table 11: Success rates on *"Fold Shirt"* task using the human-like robot. Each policy is fine-tuned from its corresponding checkpoint pre-trained on Arx5 data.

Setting	State input	Fine-tune 5k steps	Fine-tune 10k steps
In domain	✓	0.333	0.767
Height Gen.	✓	0.0	0.0
In domain	✗	**0.700**	**0.967**
Height Gen.	✗	0.617	0.900

We validate this on the *"Fold Shirt"* task. Policies are first trained on dual-arm Arx5, where the EEF space is defined in the table frame, and then adapted to a human-like dual-arm robot, where the EEF space is robot-centric. We collect 100 demonstrations on the human-like robot and fine-tune the π_0 policy with and without state input, each initialized from its corresponding Arx5 checkpoint. As shown in Table 11, State-free Policies adapt faster and achieve higher in-domain success rates than state-based policies. They also retain strong performance under height generalization on the new embodiment, suggesting that the state-free formulation also benefits cross-embodiment adaptation.