Do You Need Proprioceptive States in Visuomotor Policies?

Anonymous CVPR submission

Paper ID 17

Abstract

Large-scale imitation-learning-based visuomotor policies have been widely used in robot manipulation, where both visual observations and proprioceptive states are typically adopted together for precise control. However, whether proprioceptive state is necessary for learning robust policies remains unclear, and it can also make the policy overly reliant on the proprioceptive state. This leads to overfitting to training trajectories and poor spatial generalization. In this study, we investigate the State-free Policy, removing the proprioceptive state input completely. The State-free Policy is built in the relative end-effector action space, and more importantly, we find that making a State-free Policy work well requires sufficient task-relevant visual observations (ensured by dual wide-angle wrist cameras). Empirical results demonstrate that the State-free Policy achieves significantly stronger spatial generalization than the state-based policy. Across multiple real-world tasks and robot embodiments, the average success rate improves from 0% to 85% in height generalization and from 6% to 64% in horizontal generalization. Furthermore, it also shows advantages in data efficiency and cross-embodiment adaptation, suggesting a promising direction for building more scalable robot learning systems in the real world.

1. Introduction

Imitation-learning-based visuomotor policies [2, 4, 7, 34, 35] have been widely used in robotic manipulation. More recently, the rise of large-scale robot learning has demonstrated the potential of scaling data, models, and training pipelines to improve policy capability across diverse real-world tasks [10, 11, 23, 27]. As robot learning systems continue to scale, an increasingly important question is not only how to train larger models, but also what policy designs are fundamentally easier to scale in real-world settings.

For complete information input, existing visuomotor policies typically incorporate not only visual observations but also the robot proprioceptive state (hereafter referred to as *state*) inputs [4, 13, 26], such as end-effector poses and joint angles. While such designs bring complete information, whether the state is necessary for learning robust policies remains unclear. More importantly, the state can make the policy overfitting by simply memorizing training trajectories. As a result, spatial generalization [6, 15, 31, 32] becomes severely limited, which is a crucial capability for large-scale robotic policies.

In this study, we investigate completely removing the state input in visuomotor policies to enhance their spatial generalization ability, hereafter referred to as "**State-free Policies.**" Such design is built upon two conditions:

- **Relative end-effector (EEF) action space [8]:** The visuomotor policies predict relative displacements of the end-effector based on the current observation. Among different action spaces, the relative EEF action space most naturally supports the spatial generalization of policies.
- **Full task observation:** Prior findings suggest that State-free Policies can be suboptimal in practice [20], and we observe a similar trend (Figure 1(a)). In this study, we identify that the key to making State-free Policies work well is the sufficient task-relevant visual observations, i.e., full task observation. This enables visuomotor policies to fully "see" the objects in the task.

As shown in Figure 1(b), we ensure the full task observation with dual wide-angle wrist-cameras (field of view $120^\circ \times 120^\circ$) mounted on the top and bottom of the end-effector, which provides full task observation for State-free Policies even in complex scenarios.

This mechanism of State-free Policies forces the policy to develop a deeper understanding of the task environment rather than simply memorizing the trajectories in training data. Therefore, State-free Policies can achieve advantages that state-based policies cannot provide:

- **Spatial Generalization:** Since State-free Policies do not rely on the state input, they avoid overfitting to training trajectories. Therefore, they exhibit strong height and horizontal generalization abilities, where height refers to variations of objects' location in the vertical direction, and horizontal refers to variations of location in 2D plane.
- **Data efficiency:** Even in in-domain settings, state-based policies require diverse demonstrations to avoid overfit-

(a) Spatial Generalization Performance

	Conventional View	Full Task Observation
with state input	✗	✗
without state input	Weak	**Strong**

(b) Full Task Observation Implement

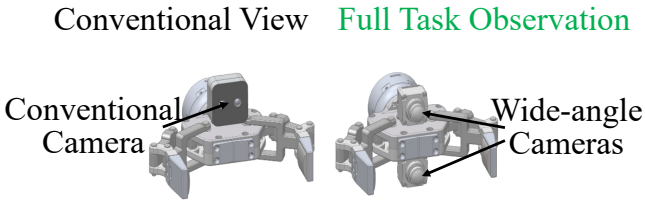


Figure 1. (a) With relative EEF action space and full task observation, State-free Policies demonstrate improved spatial generalization ability compared to state-based policies. (b) A key to making State-free Policies work well is to provide full task observation. In some complex scenarios, a conventional camera setting is insufficient to provide it; therefore, here it's ensured by dual wide-angle wrist-cameras.

ting to specific trajectories. In contrast, removing the state input eliminates this dependence on trajectory diversity, allowing State-free Policies to be fine-tuned with less demonstration data. This reduces the cost of data collection, which is often a major bottleneck in deploying real-world visuomotor policies.

- **Cross-embodiment adaptation:** Since State-free Policies rely only on visual inputs and predict actions in the relative EEF space, they exhibit stronger cross-embodiment adaptation ability than state-based policies. They do not require additional adaptation to different state spaces, so the same task can be easily adapted to new embodiments with fewer fine-tuning steps.

We have conducted extensive experiments across a diverse range of tasks, robot embodiments, and policy architectures. In both real-world and simulation environments, State-free Policies can achieve comparably great in-domain performance to state-based policies. Most importantly, when trained on strictly collected real-world data (i.e., the object location has a constrained initial distribution range), State-free Policies exhibit significantly stronger spatial generalization ability than state-based policies. For further benefits, e.g., data efficiency and cross-embodiment adaptation ability, they also demonstrate advantages over state-based policies, suggesting that removing the state input can be a simple yet effective design choice for building more scalable, robust, and practical robot learning systems.

2. Related Works

2.1. Visuomotor Policies

Imitation-learning-based visuomotor policies [2–5, 7, 12, 16, 18, 19, 24, 29, 33, 35, 38] have been widely adopted for robotic manipulation, achieving remarkable performance across diverse tasks by directly mapping the observation information to action outputs in an end-to-end manner. Recent advances such as ACT (Action Chunking with Transformers) [35], Diffusion Policy [7], and π_0 [4] highlight the effectiveness of combining large-scale trajectory datasets with powerful model architectures. These policies often adopt an action chunk mechanism [35], predicting a short horizon of future actions at each step rather than a single action. Such chunked action prediction better captures temporally extended behaviors in manipulation and often improves training stability and efficiency.

2.2. Proprioceptive State in Visuomotor Polices

A common practice in the above policies is to incorporate proprioceptive state (hereafter referred to as *state*) inputs alongside visual observations to improve the policy performance [2, 4, 7, 12, 13, 26, 35]. However, whether such state inputs are necessary for learning robust policies remains unclear, and they can create a shortcut for the policy: instead of reasoning from visual cues, the policy can simply memorize training trajectories tied to specific states [6, 13, 30–32]. As a result, the policy overfits the training trajectories and cannot adapt to spatial layout changes, limiting its spatial generalization. To address this issue, prior efforts often improve the spatial generalization in either real-world [23, 36] or simulated environments [21, 22, 25, 37] through data-driven strategies that increase the state coverage and diversity. However, the high cost of real-world data collection and the persistent sim-to-real gap in simulation mean that neither direction reliably works in practice for real-world deployment.

State-free Policies have been explored for use in pre-training methods [12, 14, 38], due to their simplicity and minimal state requirements. Despite their potential for improved spatial generalization, state-free designs have not been widely adopted for real-world robot manipulation deployment [9, 20, 28]. In this study, we identify a key reason: the lack of full task observation. Once full task observation is ensured, State-free Policies can match state-based counterparts in in-domain performance, while additionally delivering strong spatial generalization and other benefits.

Table 1. Exploration of table height generalization, using the "*Pick a pen into pen holder*" task as the representative example. Applying simple hacks on z state and noise on the state both improve the spatial generalization, indicating the state input as a bottleneck. This motivates removing the state input. ✓ indicates a success rate above 90%, ✗ indicates below 10%, and weak indicates in between; the quantitative results can be found in Appendix Section A.2 and 4.

Wrist-camera setting	Conventional camera				Dual wide-angle cameras
Method on state	simply with state	z state hack	noised state	without state	without state
h=80cm (In-domain)	✓	✓	✓	✓	✓
h=72cm (Out-of-domain)	✗	✓	weak	✓	✓
h=90cm (Out-of-domain)	✗	✓	weak	weak	✓

3. State-free Policies

3.1. Preliminary

3.1.1. Imitation-Learning-Based Visuomotor Policies

We consider visuomotor policies mapping raw observations to low-level control actions. At time t, the observation is $o_t \in \mathcal{O}$ (camera images and, typically, states), and the policy with trainable parameters θ is defined as: $\pi_\theta(a_t \mid o_t)$, where a_t denotes the low-level control action. In imitation learning [1], the policy is trained on demonstration data $\mathcal{D}$ by minimizing the negative log-likelihood of actions:

$$\mathcal{L}_{\text{IL}}(\theta) = -\sum_{(o_t,a_t)\in\mathcal{D}} \log \pi_\theta(a_t \mid o_t). \tag{1}$$

During deployment, π_θ takes online observations o_t and outputs the actions a_t, which will be executed on the robot to control its motion.

3.1.2. Action Representation Space

We consider two common action representation spaces: relative EEF action and relative joint-angle action.

In the relative EEF action space, the end-effector pose at time t is $p_t = [x_t, q_t]$, where $x_t \in \mathbb{R}^3$ is the Cartesian position and $q_t \in SO(3)$ is the orientation. The policy outputs a relative displacement:

$$a_t = \Delta p_t = [\Delta x_t, \Delta q_t], \tag{2}$$

where Δx_t and Δq_t denote the Cartesian translation and rotation. The next end-effector pose is updated by: $p_{t+1} = p_t \oplus \Delta p_t$, where $\oplus$ denotes composition of the Cartesian translation and rotation.

In the relative joint-angle action space, the policy drives the end-effector motion by predicting relative joint changes $\Delta\theta_t$. In this case, the end-effector displacement depends on both $\Delta\theta_t$ and the current joint pose θ_t, i.e.,

$$\Delta p_t = f(\Delta\theta_t, \theta_t), \tag{3}$$

where f denotes the forward kinematics mapping.

3.2. Spatial Generalization Challenge of State Input

Proprioceptive states provide direct and accurate robot configuration, but may act as shortcuts, where the policy directly associates the state input with the training trajectories. Consequently, the policy tends to overfit to the training trajectories and fails to adapt to spatial layout changes.

We validate this with a real-world height generalization evaluation as an example. Specifically, we collect "*Pick a pen into pen holder*" demonstrations at a fixed 80 cm table height (the overview is illustrated in Figure 3) and fine-tune the π_0 [4] policy using the relative EEF action space.

As shown in Table 1, state-based policies completely fail to generalize across different table heights. However, adding human-designed hacks (manually shifting the state input according to table height) effectively improves the height generalization ability, indicating that state input is a critical factor limiting spatial generalization. At the same time, adding random noise augmentation ($[-5\text{ cm},\ 5\text{ cm}]$) on the state height dimension z improves height generalization without affecting in-domain performance, which indicates that state input may not be necessary. These motivate us to consider removing the state input.

3.3. What Makes State-free Policies Work Well?

Prior findings suggest that purely state-free designs can be difficult to deploy in real-world manipulation [20]. In this study, we revisit this setting and identify two key conditions that make State-free Policies work well: the **relative EEF action space** and **full task observation**. These help State-free Policies improve their performance and spatial generalization (Table 1) without requiring additional architectural changes or costly diverse data collection.

3.3.1. Relative EEF Action Space

To study the spatial generalization, we begin by clarifying the action space. First, we exclude all absolute actions, where the policy predicts absolute poses. It learns fixed mappings tied to training trajectories and thus fails to adapt to new spatial layouts. Then, we consider two common relative action spaces: relative joint-angle and relative EEF:

CVPR #17

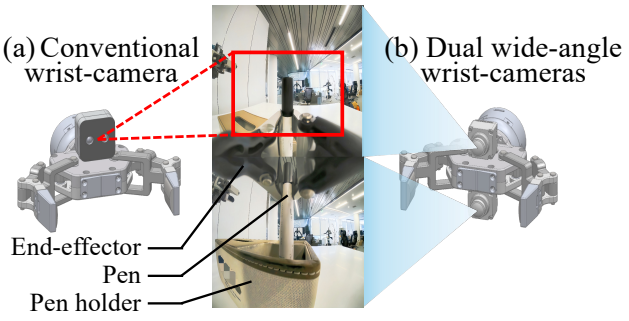


Figure 2. (a) Conventional wrist-camera setting. Some target objects (e.g., pen holder) may not be visible. (b) Dual wide-angle wrist-cameras setting. It provides the full task observation even in complex scenarios.

- **Relative Joint-angle Action Space:** In spatial generalization tasks such as height generalization, as the table height changes, when the end-effector is at the same relative position with respect to the table, the robot receives the same visual observations. In this case, the policy predicts the same $\Delta\theta_t$, but the joint configurations θ_t are different. As shown in Eq. 3, this results in different end-effector displacements Δp_t, leading to incorrect actions.
- **Relative EEF Action Space:** As shown in Eq. 2, the policy predicts relative end-effector motions directly from observations. The action Δp_t depends only on the observations, not on the absolute pose, so identical observations yield the same displacement regardless of absolute robot poses. This invariance allows relative EEF actions to naturally support spatial generalization across heights and horizontal positions.

Since the relative EEF action space naturally supports the spatial generalization, our State-free Policies will be built upon this action representation space.

3.3.2. Full Task Observation

The key condition making State-free Policies work well is full task observation. With state input, the policy can directly learn shortcut associations, such as what action to take once the robot reaches a certain configuration, rather than relying on visual information. In contrast, without the state input, the policy has to make decisions entirely from vision, which requires sufficient task-relevant visual information, i.e., the full task observation. While in simple scenarios a conventional view field may be enough for the full task observation, many real-world manipulation scenarios are more complex; for example, in the "*Pick a pen into pen holder*" task, the policy needs to see the pen holder beneath the end-effector. This motivates us to equip the end-effector with a broader view field for a wide range of scenarios.

Our camera system consists of an overhead camera and wrist-cameras. In the conventional wrist-camera setting, a single conventional-view wrist-camera is mounted on top of the end-effector (in this study with view field $87° \times 58°$), as illustrated in Figure 2(a). To achieve full task observation even in complex scenarios, we adopt dual wide-angle wrist-cameras (field of view $120° \times 120°$) mounted on the top and bottom of the end-effector, as illustrated in Figure 2(b). This setting expands the view and exposes the workspace beneath the end-effector (note that in tasks with simple scenarios, e.g., involving a single task-relevant object, the conventional wrist-camera setting can already be enough).

3.4. Summary

As shown in Table 1, in the exploration on *"Pick a pen into pen holder"* task, with relative EEF action and full task observation, State-free Policies achieve comparable in-domain performance with state-based policies while delivering improved height generalization.

In the following sections, we will conduct extensive evaluations to validate and further analyze State-free Policies. Meanwhile, we demonstrate their further benefits, including higher data efficiency and better cross-embodiment adaptation. In addition, we also demonstrate an interesting finding that removing the overhead camera can further enhance the policy's spatial generalization ability.

4. Performance Across Various Tasks

To evaluate the performance of State-free Policies, we conduct extensive evaluations across various tasks. Performances of the state-based policy and several optimizing strategies on it can be found in Appendix Section A.2.

4.1. Setup

4.1.1. Task

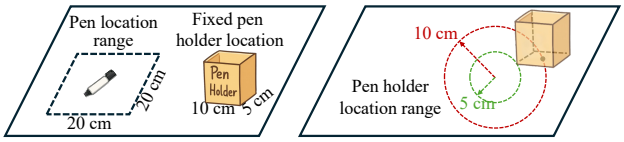
Our real-world tasks include 3 *"Pick & Place"* tasks, a more challenging *"Fold Shirt"* task, and a difficult task *"Fetch Bottle"* (on a whole-body robot with torso, waist, and leg motions controlled by a 6-dimensional torso pose vector, representing position and orientation in the EEF form). The detailed task descriptions are as follows:

- **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 door, take out the bottle, and close the refrigerator door.

As shown in Figure 3, we present the overviews of the robot embodiments and representative tasks in our evaluations.

We also conduct evaluations in the simulation environment on the LIBERO benchmark [17], where we fine-tune

Figure 3. Overview of our robot embodiments and representative tasks, including "*Pick & Place*" tasks, more challenging "*Fold Shirt*" and "*Fetch Bottle*" task. These tasks span a wide range of robot embodiments: a 2×8 DoF human-like dual-arm robot, a 2×7 DoF dual-arm Arx5 robotic arm system, and a 26 DoF whole-body robot.



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

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

Figure 4. Object locations illustration in task *"Pick Pen"*. In training data, the pen is randomly placed within a small region and the pen holder is fixed on the table. For horizontal generalization evaluation, we keep the pen location range unchanged and shift the pen holder by 5 cm and 10 cm to compute the average success rate.

the policy separately in each suite and subsequently evaluate it in the corresponding test suite.

4.1.2. Real-world Data

We employ professional data collectors to collect the real-world demonstration data using teleoperation. Details of the data amount can be found in Appendix Section A.1. Importantly, during data collection, we fix the table height and limit the object locations within a constrained 2D range. Taking the task *"Pick Pen"* as an example, as shown in Figure 4, in training data, the pen holder location is fixed, and in horizontal evaluation, we shift its location. This design ensures that the spatial generalization ability originates from the policy itself rather than from diverse data.

4.1.3. Evaluation Metric

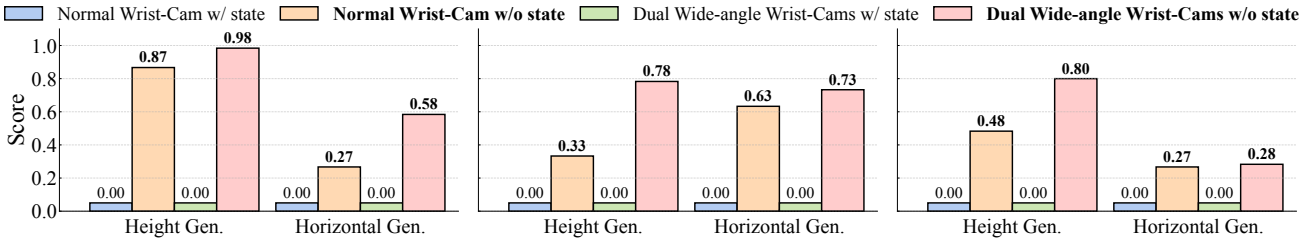
We evaluate the spatial generalization of visuomotor policies along two dimensions: height and horizontal generalization. Each real-world evaluation consists of 30 trials, with success counted only if the entire trajectory is completed. A trial is marked as a failure if the policy takes no reasonable action within 30 seconds or if any action fails.

Height Generalization Evaluation The *"Pick & Place"* data are collected at 80 cm table height. The height generalization score is computed as the average success rate of the total 60 trials at 72 cm and 90 cm table heights. And as shown in Figure 3, since the Arx5 arms are fixed to the table and the refrigerator height cannot be adjusted, the *"Fold Shirt"* and *"Fetch Bottle (whole-body)"* tasks are not applicable for height generalization evaluation.

Horizontal Generalization Evaluation In *"Pick & Place"* and *"Fetch Bottle (whole-body)"* tasks, the target objects (pen holder, step, teacup, and refrigerator) are shifted within 2D ranges of 5 cm and 10 cm, as illustrated by the *"Pick Pen"* example in Figure 4. In the *"Fold Shirt"* task, we evaluate by laterally shifting a single arm by 15 cm, as well as shifting both arms by 15 cm in opposite directions. We run 30 trials for each setting (shifting objects by 5 cm and 10 cm, shifting a single Arx5 arm and shifting both arms). Thus, each task is evaluated with 60 trials, and the horizontal generalization score is the average success rate across them.

4.1.4. Model

In our main evaluations, we use π_0 [4] policy, following its released fine-tuning recipe. π_0 is widely regarded as one of the most powerful policies in the community. In addition, our detailed analysis in Section 5.3 further evaluates



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

Figure 5. The height and horizontal generalization (written as Gen.) performances across 3 real-world *"Pick & Place"* tasks. With full task observation, State-free Policies show significantly improved 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**

Table 3. Spatial generalization evaluation of State-free Policies using different action representations.

Action space	Height Gen.	Horizontal Gen.
relative EEF	0.983	0.583
absolute EEF	0	0
relative joint	0	0
absolute joint	0	0

different policy architectures, including ACT [35], which employs action chunking to model temporally extended behaviors, and Diffusion Policy [7], which models actions as a distribution via diffusion dynamics. These demonstrate that the effectiveness of State-free Policies generalizes policy model structures.

4.2. Real-world Evaluations

Here we report the spatial generalization evaluations, including height and horizontal generalization, on 5 real-world tasks. We report their in-domain performance and the simulation evaluations in Appendix Section A.3.

In Figure 5, we report the height and horizontal generalization performance in 3 real-world *"Pick & Place"* tasks. Compared to state-based policies, State-free Policies exhibit significant improvement in both height and horizontal generalization: taking the *"Pick Pen"* task as an example, the success rate in height generalization rises from 0 to 0.98, and in horizontal generalization from 0 to 0.58. And compared with the conventional wrist-camera setting, the full task observation helps height generalization success rate improve from 0.87 to 0.98, and the horizontal generalization from 0.27 to 0.58.

In task *"Fold Shirt"*, folding a shirt is difficult due to the deformable nature of fabric which makes the folding manipulation challenging. And task *"Fetch Bottle (whole-body)"* is more challenging because the robot's torso motions are not directly observable. As discussed in Section 4.1.3, height generalization evaluation is not applicable to these two tasks. In addition, due to hardware limitations, the dual wide-angle wrist-cameras cannot be mounted in these embodiments. In Table 2, we report the horizontal generalization performance on these two tasks. Even in these 2 challenging tasks, State-free Policies still achieve significantly stronger spatial generalization ability. Moreover, this also reflects that for simpler scenarios (i.e., with simple task-relevant objects, even when the task motion is challenging), the conventional wrist-camera setting can still provide full task observation.

5. Detailed Analysis on State-free Policies

For deeper insights, we conduct more detailed analysis of State-free Policies, examining how their behavior varies under different conditions. In this section, we mainly focus on the *"Pick Pen"* task as the example, which is both intuitive and typical, making it well-suited for detailed analysis. Unless otherwise specified, all evaluations in this section use the π_0 policy and the relative EEF action space.

5.1. Action Representation

As discussed in Section 3.3.1, the relative EEF action space most naturally supports the generalization ability of State-free Policies. In this section, we evaluate alternative action representations, including the absolute EEF, both absolute and relative joint-angle action spaces. Evaluations are performed using the dual wide-angle wrist-cameras setting, ensuring full task observation.

As reported in Table 3, the relative EEF action space achieves the best performance in height and horizontal generalization settings, whereas others (e.g., absolute EEF, both absolute and relative joint-angle space) show disastrous per-

formance in spatial generalization. These results highlight that the relative EEF action space most naturally supports the spatial generalization ability of State-free Policies.

5.2. Full Task Observation

Another key condition making State-free Policies work well is full task observation. Our camera system includes an overhead camera and wrist-cameras. As illustrated in Figure 2, for full task observation, each end-effector is equipped with two wide-angle wrist-cameras on the top and bottom. By cropping image regions or masking one of the inputs, we create different levels of task observation to demonstrate the critical role of full task observation in State-free Policies.

Table 4. Spatial generalization performances of State-free Policies on different task observation levels, implemented by varying the camera settings.

Wrist-cam number	Wrist-cam 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
Dual	**Wide-angle**	**N/A**	**1.0**	**1.0**

As reported in Table 4, the spatial generalization ability of State-free Policies gradually improves as the field of view expands, indicating that full task observation is important for achieving better performance in State-free Policies.

Moreover, an interesting finding shows that even without the overhead camera, the dual wide-angle wrist-cameras alone enable the best spatial generalization. This indicates that, in the current task, they provide completely full task observation for the entire trajectory, while the overhead camera is not only unnecessary but can even be harmful (we will discuss further in Section 7).

5.3. Policy Architecture

We also evaluate different model architectures without state input, including ACT and Diffusion Policy. All use the dual wide-angle wrist-cameras setting for full task observation. As reported in Table 5, the results are consistent across architectures: State-free Policies exhibit stronger spatial generalization than state-based policies, indicating their effectiveness is independent of specific policy implementations, representing a general and universal conclusion.

6. Further Benefits of State-free Policies

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

Table 5. 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

6.1. Higher Data Efficiency

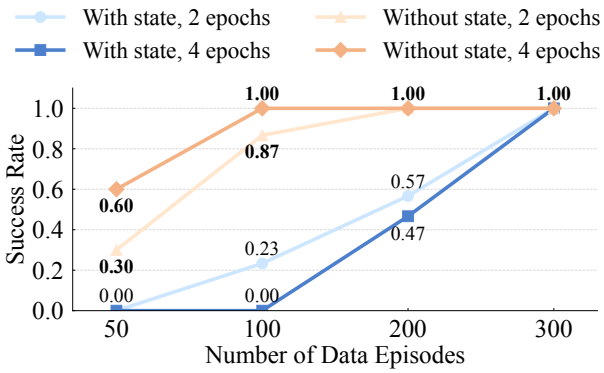


Figure 6. Evaluation success rates (in-domain) on the *"Pick Pen"* task with varying amounts of fine-tuning data.

Even in in-domain settings, state-based policies require diverse demonstrations to avoid overfitting to specific trajectories, greatly raising data collection costs. While, State-free Policies are less prone to memorizing specific trajectories and can achieve comparable performance with fewer fine-tuning data, thereby enhancing data efficiency and practicality for their real-world deployment.

We validate this on the in-domain *"Pick Pen"* task with dual wide-angle wrist-cameras for full task observation, varying fine-tuning data to 300, 200, 100, and 50 episodes, and measuring after 2 and 4 fine-tuning epochs. As shown in Figure 6, reducing data leads state-based policies to overfit and lose success, while State-free Policies maintain much higher performance.

6.2. Better Cross-embodiment Adaptation

We find that State-free Policies also benefit the cross-embodiment fine-tuning. For state-based policies, cross-embodiment adaptation requires aligning with a new state

space, and even with EEF-based states, differences in reference frame definitions across embodiments still create gaps. In contrast, State-free Policies avoid this issue: with similar camera setups, they only adapt to minor image shifts, enabling more efficient cross-embodiment fine-tuning.

Table 6. Success rates in in-domain *"Fold Shirt"* task using the human-like robot. Each policy is fine-tuned from its corresponding checkpoint pre-trained on data collected using Arx5 arms.

State input	Fine-tune 5k steps	Fine-tune 10k steps
✓	0.333	0.767
✗	0.700	0.967

We validate this on the *"Fold Shirt"* task (in-domain setting). Policies are first trained on dual-arm Arx5 (the EEF space is in table frame) and then adapted to a human-like dual-arm robot (the EEF space is in robot-centric frame). 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 6, State-free Policies adapt much faster across embodiments, achieving substantially higher success rates than state-based policies under the same fine-tuning epochs. This indicates that State-free Policies have a better cross-embodiment ability than state-based policies.

7. Rethinking the Overhead Camera

After removing the state input, we consider the overhead camera might be another potential bottleneck to spatial generalization. Changes in object locations can induce distribution shifts in overhead camera images. This will degrade performance in extreme cases, e.g., 100 cm table height. In contrast, since the end-effector can move along with the object, the wrist-camera can still capture observations consistent with those in training, avoiding the out-of-domain issues. During manipulation, the dual wide-angle wrist-cameras can already provide full task observation, the overhead camera may not only be unnecessary but even harmful to the spatial generalization ability.

We evaluate this through experiments on the *"Pick Pen"* task under more challenging settings:

- Raising the table height to 100 cm.
- Raising the pen holder to double its height, changing its relative hieght with respect to the table.
- Shifting the pen holder 20 cm away from its position in training data.

As reported in Table 7, State-free Policies with the overhead camera show terrible performance across all 3 more challenging setting. While without the overhead camera, success rates remain consistently high, confirming that dual wide-angle wrist-cameras alone are sufficient, while the overhead view introduces harmful shifts. This finding motivates us to rethink sensor design, perhaps removing the overhead camera, for future visuomotor policies.

Table 7. Success rates with and without the overhead camera in more challenging *"Pick Pen"* generalization settings, with dual wide-angle wrist-cameras. These settings include: raising the table height further to 100 cm, raising the pen holder to double its height, and shiftiing the pen holder 20 cm away from its initial position.

Overhead camera	Table height 100 cm	Raising pen holder height	Shifting pen holder 20 cm
✓	0	0.467	0
✗	1.0	0.867	0.800

8. Conclusion

In this study, we investigate the State-free Policies, removing the proprioceptive state input completely. State-free Policies are built in the relative EEF action space. And we identify that the key to making State-free Policies work well is the full task observation. Without state input, these policies maintain perfect in-domain performance while achieving significant improvements in spatial generalization. Importantly, the benefits of State-free Policies go beyond generalization alone. By reducing reliance on trajectory-specific state inputs, they alleviate the need for costly real-world data with broad state coverage and enable more efficient cross-embodiment adaptation. These properties make State-free Policies a promising design choice for scalable robot learning, where robustness, data efficiency, and transferability are all critical.

Limitation

State-free Policies also remain some limitations. First, vision-only policies might exhibit sensitivity to the background: changing the background (e.g., relocating the robot and table) may require additional fine-tuning to restore performance. And in dual-arm settings, if only one arm is used for working, distribution shifts in the unused arm's visual input may occasionally lead to unexpected movements of the unused arm.

References

[1] Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine intelligence 15*, pages 103–129, 1995. 3

[2] 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. 1, 2

[3] 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.

[4] 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. 1, 2, 3, 5

[5] 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. 2

[6] 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. 1, 2

[7] 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*, page 02783649241273668, 2023. 1, 2, 6

[8] 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. 1

[9] Lin Cong, Hongzhuo Liang, Philipp Ruppel, Yunlei Shi, Michael Görner, Norman Hendrich, and Jianwei Zhang. Reinforcement learning with vision-proprioception model for robot planar pushing. *Frontiers in Neurorobotics*, 16: 829437, 2022. 2

[10] 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)*, pages 5961–5968, 2025. 1

[11] 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. 1

[12] 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. 2

[13] 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. 1, 2

[14] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024. 2

[15] 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. 1

[16] 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. 2

[17] 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. 4

[18] 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, 2024. 2

[19] 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*, 2024. 2

[20] Jingxian Lu, Wenke Xia, Yuxuan Wu, Zhiwu Lu, and Di Hu. When would vision-proprioception policies fail in robotic manipulation?, 2026. 1, 2, 3

[21] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, N. Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning. *ArXiv*, abs/2108.10470, 2021. 2

[22] OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Ma teusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas A. Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei M. Zhang. Solving rubik's cube with a robot hand. *ArXiv*, abs/1910.07113, 2019. 2

[23] 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)*, pages 6892–6903. IEEE, 2024. 1, 2

[24] 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. 2

[25] Joshua Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and P. Abbeel. Domain randomization for

transferring deep neural networks from simulation to the real world. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017. 2

[26] Faraz Torabi, Garrett Warnell, and Peter Stone. Imitation learning from video by leveraging proprioception. *arXiv preprint arXiv:1905.09335*, 2019. 1, 2

[27] 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*, pages 1723–1736. PMLR, 2023. 1

[28] 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. 2

[29] 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. 2

[30] Quantao Yang, Michael C Welle, Danica Kragic, and Olov Andersson. S^2-diffusion: Generalizing from instance-level to category-level skills in robot manipulation. *arXiv preprint arXiv:2502.09389*, 2025. 2

[31] 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. 1

[32] 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*, pages 726–747. PMLR, 2021. 1, 2

[33] 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. 2

[34] 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)*, pages 1702–1713, 2025. 1

[35] 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. 1, 2, 6

[36] Tony Z. Zhao, Jonathan Tompson, Danny Driess, Pete Florence, Kamyar Ghasemipour, Chelsea Finn, and Ayzaan Wahid. Aloha unleashed: A simple recipe for robot dexterity. In *Conference on Robot Learning*, 2024. 2

[37] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 737–744, 2020. 2

[38] 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*, pages 2165–2183. PMLR, 2023. 2

A. Appendix

A.1. Amount of Real-world Data

In this study, 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 challenging tasks, we do not report the episode number, since they are long-horizon tasks and each episode does not cover a complete task execution. For the *"Fold Task"* task, around 20 hours of data are collected (counting only the process from the flattened state to the folded state). For the *"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.

A.2. Challenges With State Input

Table 8. Height generalization performance of state-based policies under different optimization strategies.

Optimization	Height generalization
w/o state	**0.983**
Noise augmentation	0.633
Diverse data	0.117
Task-mixed	0
LoRA fine-tune	0

In the *"Pick Pen"* task, we evaluate several strategies to improve height generalization of state-based policies with dual wide-angle wrist-cameras, including: (1) adding random noise ($[-5$ cm, 5 cm$]$) to the height component of the state, (2) collecting diverse data at table heights 75–84 cm (30 episodes for each 1 cm interval), (3) task-mixed training on *"Pick Pen"* and *"Pick Bottle"*, and (4) LoRA fine-tuning. As reported in Table 8, none of these methods yields fundamental improvement, confirming that state inputs limit spatial generalization.

A.3. In-domain Performance

Table 9. In-domain success rates in different real-world tasks using policies with and without state input, using dual wide-angle wrist-cameras for full task observation.

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 will report the in-domain performance of both state-based policies and State-free Policies in our real-world tasks. At the same time, we also report their performance on the LIBERO benchmark.

In Table 9, 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.

Table 10. Simulation evaluations 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

In the simulation environment, we compare the in-domain performance of the π_0 policy with and without state input on the LIBERO benchmark. As reported in Table 10, State-free Policies achieve performance as perfect as state-based policies, and in some cases even surpass them.