BeTTER: Diagnose the Illusion of Embodied Reasoning in Vision-Language-Action Models

Haiweng Xu[1,2], Sipeng Zheng[2], Hao Luo[1,2], Wanpeng Zhang[1,2], and Zongqing Lu[1,2⋆]

[1] Peking University, Beijing, China
[2] BeingBeyond, Beijing, China
haiwengxu@stu.pku.edu.cn, zhengsipeng27@gmail.com,
lh2000@pku.edu.cn, wpzhang.edu@gmail.com, zongqing.lu@pku.edu.cn

Abstract. Recent Vision-Language-Action (VLA) models report impressive success rates on robotic benchmarks, but whether these scores reflect genuine embodied reasoning remains questionable. To address this gap, we introduce **BeTTER**, a diagnostic **Be**nchmark for **T**esting **T**rue **E**mbodied **R**easoning. By applying targeted causal interventions (e.g., spatial layout shifts and temporal extrapolation) under strict kinematic isolation, BeTTER decouples high-level cognitive failures from low-level execution errors. Systematic evaluations reveal that state-of-the-art VLAs break down under these interventions, exhibiting lexical-kinematic shortcuts, semantic feature collapse, behavioral inertia, and causal state-tracking failures. Our analysis traces these failures to two coupled sources: real-time deployment constraints, such as capacity compression and myopic perception, which weaken semantic representations, and behavioral cloning, which can amplify predictive but non-causal correlations into shortcut policies under static training distributions. We further show that highly static evaluations mask these defects by permitting overfitting to sensorimotor priors. Real-world robotic validation confirms that these failures are not simulation artifacts, highlighting the need for future VLA paradigms to preserve semantic and causal reasoning while maintaining efficient continuous control. Code and benchmark: https://github.com/BeingBeyond/BeTTER.

Keywords: Vision-Language-Action Benchmark · Embodied Reasoning

1 Introduction

Over the past few years, the field has shifted its focus from visual understanding of human activities [5, 39, 41] to enabling robots to mimic and generalize human behaviors. Recent advances in Vision-Language-Action (VLA) models [2, 13, 22, 26] have demonstrated impressive performance on robotic manipulation benchmarks, hinting at the emergence of general-purpose embodied reasoning. Concurrently, embodied QA-oriented VLMs achieve strong results on

⋆ Corresponding author.

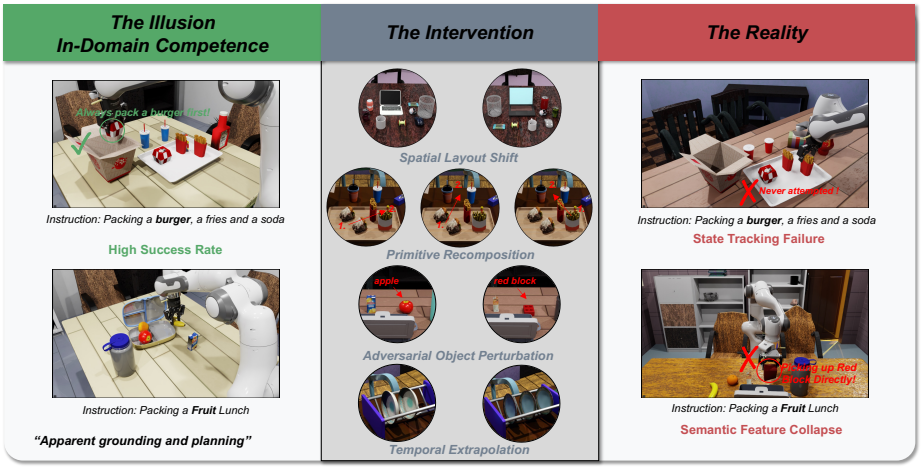


Fig. 1: Unmasking the illusion of embodied reasoning. (Left) The Illusion: State-of-the-art VLA models achieve impressive success rates on standard, visually static manipulation benchmarks, projecting an illusion of robust semantic grounding and planning. **(Middle) The Intervention:** We introduce systematic cognitive stress tests to break this illusion by applying controlled interventions such as spatial layout shift. **(Right) The Reality:** Model performance collapses catastrophically, exposing state tracking failure and feature collapse.

reasoning tasks [12, 15]. Together, these successes suggest physical intelligence may be closer than anticipated.

However, such progress can be misleading. While supervision signals in embodied QA benchmarks [28, 31] improve reasoning metrics, they often fail to yield corresponding downstream action gains [36]. This disconnect suggests that existing action evaluations may reward behavioral competence without validating the underlying reasoning process. As illustrated in Fig. 1, this discrepancy creates an *illusion of embodied reasoning*: policies appear competent in stable settings, yet may rely on shortcut correlations rather than grounded, causal, and compositional task understanding. Recent robustness-oriented benchmarks, such as LIBERO-Plus [9] and LIBERO-PRO [43], take important steps toward evaluating policies under distribution shifts. However, their reliance on fixed simulation assets limits the ability to systematically vary task semantics, interaction structures, and underlying task logic. As VLAs become increasingly proficient at dense action generation, behavioral competence becomes easier to imitate, while the reasoning behind the behavior remains opaque.

To address this gap, we introduce BeTTER, a diagnostic **Be**nchmark for **T**esting **T**rue **E**mbodied **R**easoning in robotic policies. BeTTER constructs controlled task variations along multiple reasoning axes, including spatial layout shift, primitive recomposition, adversarial object perturbation, and temporal extrapolation, logging privileged states to enable interpretable failure analysis.

Across our diagnostic stress tests, state-of-the-art VLAs exhibit clear brittleness despite near-saturated baseline performance [20]. They consistently fail

under reasoning interventions, manifesting as recurring modes including lexical-kinematic shortcuts, semantic feature collapse, behavioral inertia, and causal state-tracking failures. We trace these limitations to two coupled sources. First, real-time deployment constraints, such as capacity compression and myopic perceptual inputs, weaken the semantic representations needed for grounding and planning. Second, the behavioral cloning objective provides no direct pressure to recover causal task structure; when non-causal correlations are predictive under static training distributions, BC can reinforce them as shortcut policies. Consequently, policies may overfit to sensorimotor regularities rather than develop robust grounded, causal, and compositional reasoning. This resolves the paradox between high benchmark scores and poor generalization. Crucially, real-world stress tests confirm that these failures persist in physical control, showing that they are not simulation artifacts.

Summary of Contributions: (1) **BeTTER Benchmark:** We introduce a template-driven, extensible diagnostic benchmark that generates controlled task variations to evaluate embodied reasoning while minimizing low-level execution confounders. (2) **Mechanistic Diagnosis:** We reveal that current VLAs fail through shortcut-driven modes, such as lexical-kinematic shortcuts, semantic feature collapse, behavioral inertia, and causal state-tracking failures, and trace them to the coupled effects of deployment constraints and behavioral cloning. (3) **Real-World Validation:** We conduct physical robot stress tests showing that these reasoning failures persist beyond simulation, confirming that the illusion of embodied reasoning remains present in real-world continuous control.

2 Related Work

Vision-Language-Action Models. Early VLAs [3, 16, 44] formulate robotic control as autoregressive next-token prediction, enabling high-level reasoning but limiting inference speed. To support high-frequency reactive control, recent architectures [2,40] replace autoregressive action tokenization with efficient MLP [21] or diffusion/flow-matching heads [19], often with substantially reduced model capacity. However, such compression risks eroding semantic priors inherited from vision-language models [1, 18, 38]. Although recent approaches [10, 13] incorporate dense spatial grounding and VQA data [15,34] to mitigate this degradation, preserving genuine reasoning under action imitation remains an open challenge.

Robot Manipulation Benchmarks. Early benchmarks such as RLBench [14], CALVIN [24], and LIBERO [20] established foundational protocols for motor proficiency, language-conditioned control, and semantic generalization. However, as VLAs increasingly benefit from large-scale pre-training and robot data [4,27], these benchmarks can become saturated: stronger initialization allows policies to fit benchmark demonstrations more effectively, especially when evaluation environments exhibit limited variation. As a result, high success rates may reflect adaptation to stable visual layouts and action patterns rather than genuine embodied reasoning. Recent benchmarks further probe generalization through sim-to-real transfer, environment diversity, task difficulty, or controlled pertur-

bations [11, 17, 25, 30, 35, 37]. Notably, LIBERO-Plus [9] and LIBERO-Pro [43] expose VLA fragility under structural and visual shifts. Yet many robustness-oriented protocols still rely on fixed task assets or surface-level variations, making it difficult to disentangle perceptual confounders from high-level cognitive failures.

This work introduces BeTTER, shifting the focus from perceptual robustness to genuine embodied reasoning. Rather than applying surface-level perturbations, our framework systematically manipulates underlying task semantics and causal logic, enabling the precise isolation of cognitive deficits from low-level execution errors.

3 BeTTER: A Diagnostic Benchmark for True Embodied Reasoning

To address the limitations of static evaluation protocols, we introduce **BeTTER**, a diagnostic benchmark for testing true embodied reasoning in robotic policies. Rather than measuring only whether a policy completes a task under familiar conditions, BeTTER constructs controlled task variations that manipulate task semantics, spatial relations, object identity, and temporal structure. These interventions are designed to expose whether a policy grounds its actions in the current scene and task logic, or instead relies on shortcut correlations learned from demonstrations.

A key design goal of BeTTER is to support such interventions without being restricted to a small set of fixed simulation assets. To this end, BeTTER combines modular task templates, open-vocabulary 3D asset integration, and controlled variation synthesis to generate semantically meaningful diagnostic scenarios. In addition, the benchmark records privileged physical states during execution, enabling fine-grained behavioral analysis beyond binary success rates. In our instantiation, BeTTER contains 10 base manipulation tasks spanning diverse physical constraints, from loose containment to precision insertion, and expands them into 60 diagnostic variations along our intervention axes. The following sections describe the benchmark construction pipeline and the associated data collection and logging procedure.

3.1 Asset Curation and Modular Task Design

To overcome the limitations of fixed simulation assets, BeTTER supports dynamic construction of interactive environments. Tasks are specified through a modular, template-based design, which enables scalable generation of semantically meaningful scenarios while preserving physical feasibility. The pipeline consists of three components.

1) Template-driven Task Generation. We first define high-level task templates that capture common interaction structures, such as manipulation, placement, containment, and sequential assembly. A VLM is then used to instantiate these templates into concrete task specifications with realistic objects,

goals, and physical constraints. This design allows BeTTER to generate diverse tasks while maintaining coherent task logic.

2) Open-vocabulary Environment Construction. Given the object descriptions from task instantiation, BeTTER retrieves compatible 3D assets from large-scale repositories such as Objaverse [7]. The retrieved assets are normalized for simulation compatibility, including object centering and collision-bound adjustment. For contact-rich tasks where precise geometry is critical, we apply lightweight human verification to refine object placement and physical compatibility.

3) Controlled Diagnostic Variations. After a base scenario is instantiated, BeTTER synthesizes diagnostic variations through controlled perturbations of spatial layout, object identity, task composition, and temporal structure. These variations are designed to preserve physical validity while selectively changing the reasoning requirements of the task, enabling systematic evaluation of whether policies rely on grounded task understanding or shortcut correlations.

3.2 Data Collection and Privileged State Logging

To support scalable evaluation across diagnostic variations, BeTTER combines efficient trajectory generation with dense privileged state logging. This design provides executable demonstrations for policy training and detailed state information for analyzing failure modes beyond binary success rates.

1) Trajectory Amplification. For each base task, we collect a small set of canonical human demonstrations via teleoperation. Instead of manually recording demonstrations for every variation, we expand the data through procedural amplification methods such as MimicGen [23], which retarget demonstrations to new object poses, layouts, and instances. This enables efficient generation of diverse and physically valid manipulation trajectories with limited human effort.

2) Privileged State Logging and VQA Generation. During execution, the simulator records synchronized privileged states at each timestep, including depth maps, 2D/3D bounding boxes, and object segmentation masks. These signals allow us to inspect whether failures arise from incorrect object grounding, spatial mislocalization, or task-progress confusion. We further use the logged states to construct task-oriented VQA annotations and scene descriptions, complementing action trajectories with structured multimodal annotations. Representative examples and implementation details are provided in the Supplementary Material.

4 Unmasking the Illusion of Embodied Reasoning

Embodied reasoning, as claimed by recent VLAs [6,29], can obtain improvement through pre-training evidenced by high success rates on standard downstream manipulation tasks. However, it remains unclear for existing evaluation protocols to distinguish whether this apparent proficiency stems from true reasoning or the mere exploitation of shortcut behaviors memorized from training data. To

address this, we conduct a diagnostic study using our BeTTER benchmark, with the goal to identify the reasoning weakness overlooked in prior VLAs.

We evaluate three representative VLAs: $\pi_{0.5}$ [13], GR00T-N1.6 [26], and Being-H0.5 [22]. To systematically probe their reasoning limitations, we introduce four categories of targeted semantic and physical interventions at test time: **1) Spatial Layout Shift**. We rearrange object positions or alter their spatial relationships with nearby distractors, testing the model's ability to adapt to out-of-distribution configurations. **2) Primitive Recomposition**. We assess compositional generalization by recombining learned action primitives, e.g., evaluating $B \rightarrow C$ after training on $A \rightarrow B$ and $A \rightarrow C$. **3) Adversarial Object Perturbation**. We replace targets or distractors with visually similar but semantically distinct objects, examining whether models rely on true semantic grounding rather than superficial visual cues. **4) Temporal Extrapolation**. We extend task horizons by chaining subgoals or introducing repeated actions, evaluating the model's ability to maintain consistent state tracking and generalize across longer time horizons. To preserve physical and logical validity, all perturbations are applied based on the specific semantics of each task. Below, we present key findings that expose fundamental limitations of current VLAs in embodied reasoning.

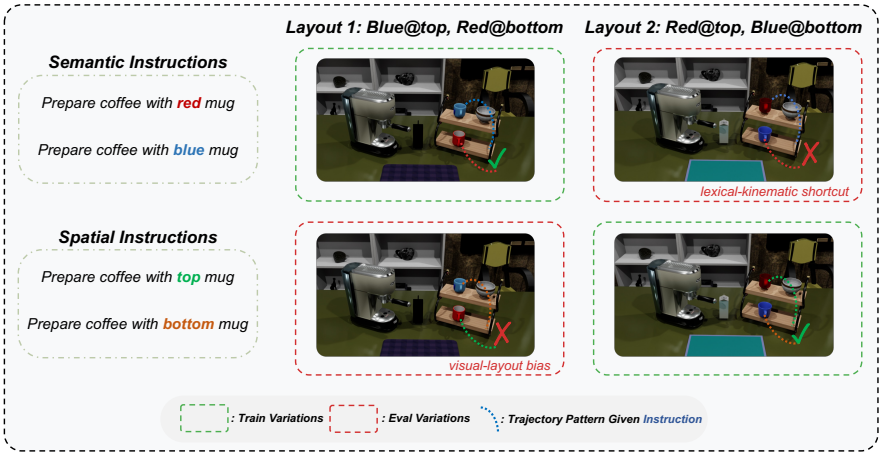


Fig. 2: Shortcut learning in instruction following. We recombine spatial (*top/bottom*) and semantic (*red/blue*) instructions under counterfactual layouts. Dashed lines denote observed trajectory patterns. The diagonal cases expose two shortcuts: lexical-kinematic mapping from language tokens to motion primitives, and visual-layout bias toward salient spatial positions.

Finding 1: Apparent Instruction Understanding without True Grounding. Recent VLAs demonstrate strong instruction-following on benchmarks [24]. However, such benchmarks often use simple instructions and do not fully con-

Table 1: Apparent instruction understanding. We report target-selection success under recombined layout–instruction pairings. Polarized performance relative to the 50% random baseline indicates shortcut-based instruction following rather than visual–semantic grounding.

Model	**Spatial (L1)**		**Semantic (L2)**	
	top	*bottom*	*red*	*blue*
Random Guess	50.0%	50.0%	50.0%	50.0%
$\pi_{0.5}$	65.0%	50.0%	70.0%	35.0%
GR00T-N1.6	100.0%	100.0%	5.0%	5.0%
Being-H0.5	100.0%	30.0%	95.0%	45.0%

trol cross-modal confounders, making it unclear whether policies ground language in visual semantics or exploit spurious language–action correlations. To test this, we design *Preparing Morning Coffee* (Figure 2), where spatial relations (*top/bottom*) and semantic attributes (*red/blue*) are systematically recombined into novel instruction–scene pairings. All objects, layouts, and motor primitives are observed during training, ensuring that failures arise from instruction grounding rather than motion unfamiliarity.

As shown in Table 1, model accuracies deviate from random guessing, indicating that policies do respond to instructions. Yet the strong polarization across variations reveals shortcut learning rather than reliable visual–semantic grounding. GR00T-N1.6 succeeds perfectly on spatial instructions but fails almost entirely on semantic ones, exhibiting a lexical-kinematic shortcut that maps tokens such as "red" directly to a learned motion primitive without visual verification. Being-H0.5 instead shows a pronounced visual-layout bias, often defaulting to the upper object even when the instruction specifies otherwise. $\pi_{0.5}$ is less polarized but frequently stalls near counterfactual cases, suggesting conflicts between learned language-conditioned motions and visual evidence. Together, these results show that apparent instruction following can emerge without true grounding between language, vision, and action.

Finding 2: Failure of Subgoal Composition Generalization. To evaluate whether models can flexibly decompose and recombine learned behaviors, we design the *Packing Fast Food Order* task (Figure 3, Top). During training, models observe sequences such as $A \to B$ (pack burger, then fries) and $A \to C$ (pack burger, then drink). At test time, they must execute the unseen composition $B \to C$ (pack fries, then drink), which cannot be solved by replaying a previously seen trajectory.

As shown in Table 2, performance drops sharply on the unseen composition, revealing a systematic failure we term *behavioral inertia*. Because all training sequences begin with packing the burger (A), models acquire a strong bias toward this initial primitive. When instructed to execute $B \to C$, the policy often maps the initial visual state to the most frequent first action, ignoring the compositional structure of the instruction. This shows that the illusion of embodied

Fig. 3: Reasoning breakdowns in sequential planning. In *Packing Fast Food Order*, we expose two failure modes. **Top:** models trained on $A \to B$ and $A \to C$ fail to execute the novel composition $B \to C$, instead defaulting to the frequent starting primitive A. **Bottom:** models use the visual cue "one burger remaining" as a proxy for task progress, causing them to skip the first step when the episode starts from this state.

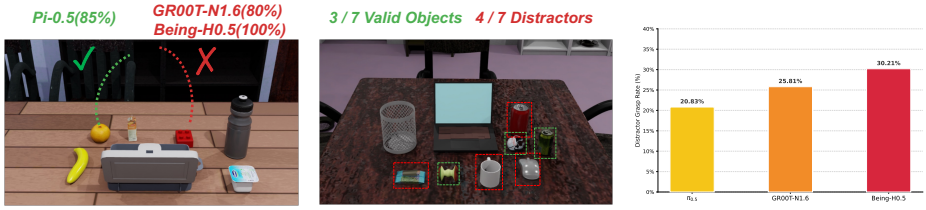
reasoning extends beyond spatial grounding to temporal planning and subgoal composition.

Table 2: Subgoal composition generalization. Models perform substantially worse on the unseen composition $B \to C$ than on seen sequences, indicating reliance on memorized trajectory patterns rather than compositional reasoning.

Model	Seen		Novel	Δ SR
	$A \to B$	$A \to C$	$B \to C$	
$\pi_{0.5}$	60.0	45.0	5.0	-47.5
GR00T-N1.6	75.0	40.0	15.0	-42.5
Being-H0.5	65.0	40.0	0.0	-52.5

Finding 3: Failure of Causal State Tracking. A robust reasoning agent must maintain an accurate internal representation of task progress. However, imitation-based policies often rely on superficial visual cues as proxies for causal state tracking. To expose this limitation, we introduce a targeted variation in the same *Packing Fast Food Order* task (Figure 3, Bottom).

During training, all episodes begin with two burgers, and the operator consistently packs one first. As a result, the visual state 'one burger remaining" becomes spuriously associated with the next subgoal, 'pack the fries." At evaluation time, we initialize the scene with only one burger while keeping the instruction unchanged. Instead of packing the burger first, models skip this step and directly attempt to pack the fries. This reveals a form of causal confusion: policies infer task progress from non-causal visual patterns rather than tracking



(a) Instance-Level Distractor **(b)** Scene-Level Clutter **(c)** Distractor Grasp Rate

Fig. 4: Fine-grained semantic grounding under adversarial distractors. (a) Instance-level substitutions, such as replacing a red apple with a red block, expose semantic feature collapse toward superficial attributes. **(b)** In cluttered scenes, policies must distinguish valid targets from functional distractors without relying on spatial priors. **(c)** Distractor Grasp Rate (DGR) quantifies invalid grasps under novel layouts, revealing a "blind grasping" heuristic toward salient objects.

which subgoals have actually been completed. Consequently, they may hallucinate task advancement despite having access to the correct instruction and valid motor primitives.

Finding 4: Failure of Fine-Grained Semantic Grounding. A truly generalizable policy must distinguish objects that share superficial visual attributes, such as color or shape, but differ in identity or affordance. However, current embodied policies often exhibit *semantic feature collapse*, relying on low-level visual salience rather than robust object grounding.

We first examine this failure at the instance level. In the *Packing a Fruit Lunch* task, the target 'red apple" is replaced with a visually similar 'red block" at the same location (Figure 4a). Policies frequently execute invalid grasps on the distractor, indicating that spatial memorization and superficial color matching can override object identity. Notably, $\pi_{0.5}$ rejects the distractor in 85% of trials, suggesting partial semantic discrimination in simple, uncluttered settings. However, this robustness does not extend to more complex scenes. In the *Tidying Office Desk* task, policies must distinguish valid targets, such as crumpled napkins or cans, from functional distractors, such as computer mice or intact mugs (Figure 4b).

To isolate semantic grounding from spatial priors, we evaluate only novel layouts. We quantify failures using the **Distractor Grasp Rate (DGR)**, defined as the proportion of grasp attempts directed at invalid objects, excluding null or aborted actions. As shown in Figure 4c, all models exhibit substantial DGRs under unseen layouts. While such errors may appear moderate in classification, they are catastrophic in physical deployment, e.g., discarding a functional computer mouse instead of a crumpled napkin. These results show that, under clutter and without reliable spatial priors, current policies fail to maintain fine-grained semantic grounding and instead revert to a "blind grasping" heuristic, interacting with salient objects regardless of task intent.

5 Root Causes of Embodied Reasoning Degradation

The diagnostic findings above reveal a pervasive "illusion of embodied reasoning": under targeted causal interventions, state-of-the-art VLAs do not merely suffer from occasional execution errors, but systematically regress to lexical-kinematic shortcuts, spatial heuristics, behavioral inertia, and causal state-tracking failures. These failures expose a central paradox. Modern VLAs are often initialized from powerful vision-language models with strong semantic and reasoning priors, yet such high-level capabilities become fragile once the model is adapted for continuous physical control.

We argue that this degradation arises from two coupled sources. First, the architectural constraints required for real-time deployment—including capacity compression and low-resolution perception—weaken the visual-semantic representations needed for grounding and planning. Second, the behavioral cloning objective provides no direct pressure to recover causal structure: when strong non-causal correlations exist in the training distribution, imitation learning can reinforce them as shortcut policies. Thus, architectural bottlenecks make reasoning brittle, while the BC paradigm turns this brittleness into systematic failures.

In the following sections, we first examine how VLM-to-VLA deployment constraints degrade the representational substrate for embodied reasoning. We then show why such degraded policies can still achieve saturated performance on standard static benchmarks, and how this benchmark paradox is explained by shortcut optimization under behavioral cloning.

5.1 Architectural Bottlenecks in VLA Deployment

We first examine how the representational substrate for embodied reasoning is weakened during the VLM-to-VLA transition. To separate such architectural effects from continuous-control failures, we evaluate intermediate model configurations on static embodied reasoning benchmarks, including spatial understanding [8], fine-grained grounding [42], and sequential planning [28]. As summarized in Table 3, VLA deployment introduces a cascade of compromises: scaling down model capacity weakens both grounding and planning, mixed VLM–VLA co-training partially restores spatial grounding but does not recover planning, and enforcing a single low-resolution observation severely erases the recovered grounding ability.

The ablation reveals three effects. First, capacity compression imposes a strong upper bound on reasoning: reducing the InternVL-3.5 backbone from 8B to 2B substantially degrades both semantic grounding (RefSp. 27.50 $\rightarrow$ 6.00) and sequential planning (EgoPlan 40.80 $\rightarrow$ 33.38). Second, VLM–VLA co-training partially restores spatial grounding (RefSp. 6.00 $\rightarrow$ 29.50), suggesting that multimodal data can help preserve perceptual semantics under compression. However, this improvement does not transfer to sequential planning, whose score slightly decreases from 33.38 to 31.95. This asymmetry suggests that the compressed model can recover local perceptual alignment more easily than causal-relational reasoning.

Table 3: Reasoning degradation under VLA deployment constraints. We report aggregate scores on embodied spatial understanding (EmbSp. [8]), fine-grained reference grounding (RefSp. [42]), and sequential planning (EgoPlan [28]). Capacity compression and low-resolution perception weaken the semantic representations required for embodied reasoning, even before continuous action generation is introduced.

Model / Configuration	EmbSp.	RefSp.	EgoPlan
Reference VLMs			
Gemini-2.5-Pro	78.74	38.16	42.85
Mimo-Embodied [12] (7B)	76.24	48.00	43.00
RoboBrain-2.0 [32] (32B)	78.57	54.00	57.23
Ablating VLA Constraints on InternVL-3.5 [33]			
8B Base	74.96	27.50	40.80
4B Base	72.08	23.00	36.83
2B Base	60.19	6.00	33.38
2B + VLM + VLA Data	57.47	29.50	31.95
2B + VLM + VLA Data (Single 224px)	52.42	7.50	31.95

Finally, real-time control introduces a myopic perceptual constraint. While modern VLMs typically process high-resolution observations through multiple image patches, VLAs often rely on a single low-resolution input to satisfy control-frequency requirements. When we enforce this single-224×224 setting, the previously recovered grounding ability collapses again (RefSp. $29.50 \rightarrow 7.50$). This result links the semantic collapse observed in Section 4 to a concrete architectural bottleneck: fine-grained object semantics become fragile when high-level perception is compressed into a low-latency control interface.

5.2 Shortcut Amplification in Static Benchmarks

The architectural bottlenecks above explain why the semantic substrate for embodied reasoning becomes fragile. However, they do not fully explain why degraded policies can still achieve near-perfect success on standard robot benchmarks. We argue that this benchmark paradox arises from the interaction between behavioral cloning and limited distributional variation. Under BC, any cue that reliably predicts demonstrated actions can be reinforced, even if it is not causally related to the task—for example, fixed object locations, repeated subgoal orders, or proprioceptive progress signals. When standard evaluation environments preserve these same correlations, the resulting shortcut policies remain effective under the benchmark distribution while leaving causal reasoning untested.

Table 4 illustrates this effect. On LIBERO [20], all pre-training recipes saturate around 97% success, including the no-pre-training baseline. This does not imply that all representations are equally strong; rather, the benchmark distribution is sufficiently static that policies can succeed through memorized tra-

Table 4: Static benchmarks mask shortcut learning. All recipe variants achieve saturated performance on in-distribution LIBERO, where static layouts allow policies to exploit memorized sensorimotor patterns. In contrast, CALVIN ABC→D introduces out-of-distribution environmental shifts that weaken such shortcuts. The improvement from VLA-only pre-training to VLA+VLM pre-training is therefore more visible in CALVIN than in LIBERO, suggesting that preserved multimodal semantics matter most when memorized trajectories no longer suffice.

Configuration	LIBERO	CALVIN ABC→D					Avg. Len.
	Avg.	1	2	3	4	5	↑
Reference Models							
GR00T-N1.6 [26]	97.0	96.3	91.4	85.3	79.7	71.7	4.24
Being-H0.5 [22]	98.9	93.4	87.0	82.0	77.7	73.7	4.14
Pre-training Recipe Ablations							
No Pre-training	97.4	89.9	80.1	71.8	66.4	58.7	3.67
VLA (OXE)	96.3	93.1	84.6	76.2	69.4	62.7	3.86
VLA + VLM	97.2	93.5	85.9	81.0	77.1	71.1	4.09

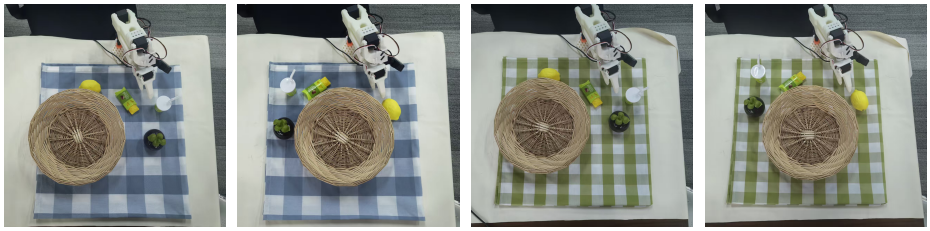
jectories and local sensorimotor regularities. Under such conditions, standard success rates cannot distinguish semantic grounding from shortcut imitation.

The contrast becomes clearer on CALVIN ABC→D [24], where environmental shifts invalidate many layout-specific heuristics. Here, VLA-only pre-training improves over no pre-training, but VLA+VLM pre-training further increases the average completed task length from 3.86 to 4.09. This gap is modest on saturated in-distribution evaluation, but becomes meaningful under out-of-distribution shifts, where policies must rely more on visual semantics to adapt their actions.

This comparison clarifies the role of behavioral cloning in the illusion of reasoning. BC does not explicitly optimize for causal state tracking, compositional planning, or counterfactual grounding; it rewards any representation that predicts demonstrated actions. When semantic representations are weakened by architectural bottlenecks, the optimization process can instead favor easier predictive cues, such as spatial priors, frequent first actions, or spurious phase signals. These cues are highly effective in static benchmarks, but they are precisely the failure points exposed by BeTTER's causal interventions. Thus, high benchmark success should not be interpreted as evidence of embodied reasoning unless the evaluation actively breaks the shortcut correlations available during imitation.

6 Real-World Experiments

The simulation diagnostics (Section 4) and mechanistic analysis (Section 5) reveal substantial reasoning failures in state-of-the-art VLAs. We next ask whether these failures persist in physical deployment, rather than arising from simulation artifacts. To this end, we conduct targeted stress tests on the SO101 robotic platform. The results show that the illusion of embodied reasoning remains present in real-world continuous control.



(a) Train: Canonical Short-horizon Layout 1 Blue tablecloth | **(b) Train: Canonical** Short-horizon Layout 2 Blue tablecloth | **(c) Train: Canonical** Long-horizon Layout 1 Green tablecloth | **(d) Train: Canonical** Long-horizon Layout 2 Green tablecloth

Fig. 5: Real-world evaluation protocol. We fine-tune the policy on four canonical configurations spanning two task horizons (short vs. long) and two spatial layouts. At test time, we apply targeted causal interventions, such as partial target absence or pre-placed objects, while keeping the required motions within known reachable configurations. This setup isolates reasoning failures from low-level execution errors in physical deployment.

6.1 Experimental Setup

As shown in Figure 5, the workspace contains a fruit basket, two target fruits (a lemon and a mangosteen), and two distractors. We construct four canonical training configurations by varying task horizon and spatial layout: a short-horizon instruction, '*Put the lemon into the fruit basket*," and a long-horizon instruction, '*Put all the fruits into the fruit basket*," each under two layouts. Using $\pi_{0.5}$ [13], we collect 100 teleoperated trajectories (25 demonstrations per configuration) and fine-tune the policy for 10,000 steps.

Preliminary trials show that the policy can execute the canonical demonstrations but is highly sensitive to object placement, often reverting to memorized grasp locations when targets are moved. To avoid conflating such failures with unreachable motions, all subsequent diagnostics enforce *kinematic isolation*: targets are placed within configurations covered by the learned motor primitives. Thus, failures under our causal interventions can be attributed to higher-level breakdowns such as poor semantic grounding, state tracking, or sequential reasoning, rather than to low-level execution constraints.

6.2 The Illusion of Competence in Short-Horizon Diagnostics

We first evaluate the policy with short-horizon instructions, such as "*Put the lemon into the fruit basket*." Under kinematic isolation, the policy appears competent in standard configurations: it grasps the target when the lemon is present and stops when the lemon is already in the basket (Table 5). This confirms that the policy has acquired the necessary physical primitives for the canonical task.

However, such success does not imply genuine embodied reasoning. Short-horizon tasks can be solved by reactive heuristics, such as mapping a salient

Table 5: Short-horizon diagnostics. For "*Put the lemon into the fruit basket*," the policy succeeds when the target is present or already completed, but behaves ambiguously when the target is absent. This shows that short-horizon success can be achieved through reactive heuristics and is therefore an unreliable proxy for embodied reasoning.

Target	Distractor		
	On Table	**In Basket**	**Absent**
On Table	**Success** Grasp target	**Success** Grasp target	**Success** Grasp target
In Basket	**Correct Freeze** Recognizes done	**Correct Freeze** Recognizes done	**Correct Freeze** Recognizes done
Absent	**Ambiguous** Grasps distractor	**Ambiguous** Freeze / air grasp	**Ambiguous** Air grasp

Table 6: Long-horizon diagnostics. For "*Put all the fruits into the fruit basket*," partial-absence and pre-placement interventions expose sequential reasoning failures. Red cells denote shortcut-driven failures; the gray cell violates the task premise.

Lemon (T1)	Mangosteen (T2)		
	On Table	**In Basket**	**Absent**
On Table	**Success** Packs L → M	**Success** Packs L → End	**Inertia** Packs L → air grasp
In Basket	**Phase Conflict I** Freeze / skip	**Success** Recognizes done	**Phase Conflict II** Oscillate / air grasp
Absent	**Success** Skips L, packs M	**Correct Freeze** Recognizes done	**Ambiguous** Air grasp

lemon-like visual cue to a grasp action or using basket occupancy as a termination signal. When the target is absent, the policy may grasp distractors, freeze, or perform an air grasp, revealing that its behavior is not grounded in a robust semantic understanding of task validity. Yet these target-absent cases also partially violate the task premise, making them insufficient as standalone evidence of reasoning failure. Thus, short-horizon diagnostics mainly expose why physical evaluations can look successful: they verify execution competence, but do not require causal state tracking or sequential planning. We therefore turn to long-horizon interventions, where reactive shortcuts are no longer sufficient.

6.3 Deconstructing Cognitive Breakdowns in Long-Horizon Tasks

We next evaluate the long-horizon instruction "*Put all the fruits into the fruit basket*," where the policy must track task progress across multiple subgoals. As shown in Table 6, targeted causal interventions disrupt the expected execution chain and expose failures that are hidden in short-horizon tasks.

Behavioral inertia. When the lemon is present but the mangosteen is absent, the policy first packs the lemon correctly, but then performs an air grasp at

the expected location of the missing mangosteen. This indicates that completing the first subgoal triggers a memorized next-step prior, rather than a visual verification of whether the next target actually exists.

Phase conflict. When the lemon is pre-placed in the basket, the visual observation corresponds to an intermediate task state, while the robot starts from the initial physical pose. This mismatch leads to freezing, skipped actions, or oscillatory motions depending on the state of the second target. In both cases, the policy fails to causally track task progress from the current scene and instead relies on rigid temporal assumptions learned from demonstrations.

Together, these physical stress tests show that the failures revealed by BeTTER are not simulation artifacts. Even when low-level motions are reachable, the policy breaks down under causal interventions that require semantic verification and sequential state tracking. This provides real-world evidence that current VLA policies can exhibit strong execution competence while lacking robust embodied reasoning.

7 Conclusion

In this work, we unmask the "illusion of embodied reasoning" in state-of-the-art VLAs through BeTTER, a diagnostic benchmark that isolates high-level cognitive breakdowns from low-level execution failures. Across controlled causal interventions, we show that strong benchmark performance often hides shortcut-driven behavior: current VLAs exhibit semantic feature collapse, behavioral inertia, and causal state-tracking failures once familiar correlations are broken. Our analysis traces these failures to two coupled sources. Architectural constraints required for real-time control, such as capacity compression and myopic perceptual downsampling, weaken the semantic representations needed for grounding and planning. Meanwhile, behavioral cloning can amplify non-causal correlations into reliable action shortcuts under static training distributions, without requiring the policy to learn causal task structure. This explains why standard benchmarks can appear saturated while out-of-distribution evaluations and BeTTER interventions reveal severe brittleness. Finally, real-world stress tests confirm that these reasoning failures persist in physical deployment rather than arising from simulation artifacts. Together, our findings provide a diagnostic lens for distinguishing genuine embodied reasoning from superficial behavioral competence, and highlight the need for future VLA paradigms to preserve semantic and causal state representations while maintaining efficient continuous control.

Acknowledgments

This work was supported by NSFC in part under Grant 62450001 and 62476008. The authors would like to thank the anonymous reviewers for their valuable comments and advice.

References

1. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al.: Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726 (2024)
2. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
3. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022)
4. Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., Gao, S., He, X., Hu, X., Huang, X., et al.: Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. arXiv preprint arXiv:2503.06669 (2025)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
6. Chen, X., Chen, Y., Fu, Y., Gao, N., Jia, J., Jin, W., Li, H., Mu, Y., Pang, J., Qiao, Y., et al.: Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy. arXiv preprint arXiv:2510.13778 (2025)
7. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023)
8. Du, M., Wu, B., Li, Z., Huang, X.J., Wei, Z.: Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 346–355 (2024)
9. Fei, S., Wang, S., Shi, J., Dai, Z., Cai, J., Qian, P., Ji, L., He, X., Zhang, S., Fei, Z., et al.: Libero-plus: In-depth robustness analysis of vision-language-action models. arXiv preprint arXiv:2510.13626 (2025)
10. Feng, Y., Zhang, W., Wang, Y., Luo, H., Yuan, H., Zheng, S., Lu, Z.: Spatial-aware vla pretraining through visual-physical alignment from human videos. arXiv preprint arXiv:2512.13080 (2025)
11. Garcia, R., Chen, S., Schmid, C.: Towards generalizable vision-language robotic manipulation: A benchmark and llm-guided 3d policy. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 8996–9002. IEEE (2025)
12. Hao, X., Zhou, L., Huang, Z., Hou, Z., Tang, Y., Zhang, L., Li, G., Lu, Z., Ren, S., Meng, X., et al.: Mimo-embodied: X-embodied foundation model technical report. arXiv preprint arXiv:2511.16518 (2025)
13. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: $\pi_{0.5}$: A vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054 (2025)
14. James, S., Ma, Z., Arrojo, D.R., Davison, A.J.: Rlbench: The robot learning benchmark & learning environment. IEEE Robotics and Automation Letters **5**(2), 3019–3026 (2020)
15. Ji, Y., Tan, H., Shi, J., Hao, X., Zhang, Y., Zhang, H., Wang, P., Zhao, M., Mu, Y., An, P., et al.: Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)

16. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
17. Li, X., Hsu, K., Gu, J., Pertsch, K., Mees, O., Walke, H.R., Fu, C., Lunawat, I., Sieh, I., Kirmani, S., Levine, S., Wu, J., Finn, C., Su, H., Vuong, Q., Xiao, T.: Evaluating real-world robot manipulation policies in simulation. arXiv preprint arXiv:2405.05941 (2024)
18. Li, Z., Chen, G., Liu, S., Wang, S., VS, V., Ji, Y., Lan, S., Zhang, H., Zhao, Y., Radhakrishnan, S., et al.: Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. arXiv preprint arXiv:2501.14818 (2025)
19. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
20. Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: Libero: Benchmarking knowledge transfer for lifelong robot learning. Advances in Neural Information Processing Systems **36**, 44776–44791 (2023)
21. Luo, H., Feng, Y., Zhang, W., Zheng, S., Wang, Y., Yuan, H., Liu, J., Xu, C., Jin, Q., Lu, Z.: Being-h0: vision-language-action pretraining from large-scale human videos. arXiv preprint arXiv:2507.15597 (2025)
22. Luo, H., Wang, Y., Zhang, W., Zheng, S., Xi, Z., Xu, C., Xu, H., Yuan, H., Zhang, C., Wang, Y., et al.: Being-h0. 5: Scaling human-centric robot learning for cross-embodiment generalization. arXiv preprint arXiv:2601.12993 (2026)
23. Mandlekar, A., Nasiriany, S., Wen, B., Akinola, I., Narang, Y., Fan, L., Zhu, Y., Fox, D.: Mimicgen: A data generation system for scalable robot learning using human demonstrations. arXiv preprint arXiv:2310.17596 (2023)
24. Mees, O., Hermann, L., Rosete-Beas, E., Burgard, W.: Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics and Automation Letters **7**(3), 7327–7334 (2022)
25. Nasiriany, S., Maddukuri, A., Zhang, L., Parikh, A., Lo, A., Joshi, A., Mandlekar, A., Zhu, Y.: Robocasa: Large-scale simulation of everyday tasks for generalist robots. arXiv preprint arXiv:2406.02523 (2024)
26. Nvidia, J.B., Castaneda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al.: Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734 (2025)
27. O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., 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)
28. Qiu, L., Chen, Y., Ge, Y., Ge, Y., Shan, Y., Liu, X.: Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. arXiv preprint arXiv:2412.04447 (2024)
29. Qu, D., Song, H., Chen, Q., Chen, Z., Gao, X., Ye, X., Lv, Q., Shi, M., Ren, G., Ruan, C., et al.: Eo-1: Interleaved vision-text-action pretraining for general robot control. arXiv preprint arXiv:2508.21112 (2025)
30. Sedlacek, M., Yefanov, P., Ponimatkin, G., Bardhan, J., Pilc, S., Fourmy, M., Kazakos, E., Snoek, C.G., Sivic, J., Petrik, V.: Realm: A real-to-sim validated benchmark for generalization in robotic manipulation. arXiv preprint arXiv:2512.19562 (2025)
31. Sermanet, P., Ding, T., Zhao, J., Xia, F., Dwibedi, D., Gopalakrishnan, K., Chan, C., Dulac-Arnold, G., Maddineni, S., Joshi, N.J., et al.: Robovqa: Multimodal long-

horizon reasoning for robotics. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 645–652. IEEE (2024)
32. Team, B.R., Cao, M., Tan, H., Ji, Y., Chen, X., Lin, M., Li, Z., Cao, Z., Wang, P., Zhou, E., et al.: Robobrain 2.0 technical report. arXiv preprint arXiv:2507.02029 (2025)
33. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
34. Yuan, W., Duan, J., Blukis, V., Pumacay, W., Krishna, R., Murali, A., Mousavian, A., Fox, D.: Robopoint: A vision-language model for spatial affordance prediction for robotics. arXiv preprint arXiv:2406.10721 (2024)
35. Zhang, B., Li, J., Shen, J., Cai, Y., Zhang, Y., Chen, Y., Dai, J., Ji, J., Yang, Y.: Vla-arena: An open-source framework for benchmarking vision-language-action models. arXiv preprint arXiv:2512.22539 (2025)
36. Zhang, J., Chen, X., Wang, Q., Li, M., Guo, Y., Hu, Y., Zhang, J., Bai, S., Lin, J., Chen, J.: Vlm4vla: Revisiting vision-language-models in vision-language-action models. In: International Conference on Learning Representations (2026)
37. Zhang, S., Xu, Z., Liu, P., Yu, X., Li, Y., Gao, Q., Fei, Z., Yin, Z., Wu, Z., Jiang, Y.G., et al.: Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11142–11152 (2025)
38. Zhang, W., Xie, Z., Feng, Y., Li, Y., Xing, X., Zheng, S., Lu, Z.: From pixels to tokens: Byte-pair encoding on quantized visual modalities. In: The Thirteenth International Conference on Learning Representations (2025)
39. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2914–2923 (2017)
40. Zheng, J., Li, J., Wang, Z., Liu, D., Kang, X., Feng, Y., Zheng, Y., Zou, J., Chen, Y., Zeng, J., et al.: X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. arXiv preprint arXiv:2510.10274 (2025)
41. Zheng, S., Chen, S., Jin, Q.: Few-shot action recognition with hierarchical matching and contrastive learning. In: European conference on computer vision. pp. 297–313. Springer (2022)
42. Zhou, E., An, J., Chi, C., Han, Y., Rong, S., Zhang, C., Wang, P., Wang, Z., Huang, T., Sheng, L., et al.: Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. arXiv preprint arXiv:2506.04308 (2025)
43. Zhou, X., Xu, Y., Tie, G., Chen, Y., Zhang, G., Chu, D., Zhou, P., Sun, L.: Libero-pro: Towards robust and fair evaluation of vision-language-action models beyond memorization. arXiv preprint arXiv:2510.03827 (2025)
44. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023)