TRANSFERRING THE INTELLIGENCE OF VLMS TO ROBOTIC CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans can seamlessly adapt to both physical and digital worlds, suggesting that while a digital-to-real gap exists in embodiment, environment and task, human intelligence itself may transfer across this gap. This naturally raises a fundamental question: can the intelligence of vision-language models (VLMs) similarly generalize from the digital world to the physical world for robotic control? We investigate this question through **RoboDawn**, a human-intuitive interface that exposes robotic control to an agentic VLM through a compact set of discrete translation, rotation, and gripper commands. Using this interface, the VLM controls a robot in a closed loop: it observes the current visual state, reasons about the next action, executes it, and adapts subsequent decisions to the resulting state. Furthermore, we introduce an in-context learning (ICL) scheme that uses a few demonstrations to ground the VLM in both interface usage and task-solving strategies. Experiments on RoboTwin 2.0 C2R and RoboDojo demonstrate that **RoboDawn** achieves strong performance without task-specific robot training. In the zero-shot setting, **RoboDawn** outperforms several strong policies trained on benchmark-specific robot data, while a single in-context demonstration further yields substantial performance gains and establishes state-of-the-art (SOTA) results. On RoboTwin 2.0 C2R, the success rate increases from 53.2% zero-shot to 73.6% one-shot, exceeding the solid baseline $\pi_{0.5}$ (46.0%). Similar gains are observed on RoboDojo, where success rate improves from 35.67% zero-shot to 47.17% one-shot. The same framework also transfers to real-world robots, performing block-in-basket and block stacking on Franka.

1 INTRODUCTION

Humans can routinely transfer knowledge and skills between digital and physical worlds. Despite substantial differences in embodiment, environment, and task, the underlying capabilities for perception, learning, reasoning, and decision-making often remain reusable. For example, a person can readily adapt to controlling unfamiliar virtual players in games or robotic embodiments in the physical world across diverse environments and tasks. We refer to this ability to reuse such general capabilities across variations in embodiment, environment, and task as **intelligence transfer**. This raises a natural question: can VLMs, whose intelligence has emerged predominantly from the digital world, exhibit a similar form of transfer and directly control robots in the physical world?

Some existing approaches pursue this goal by adapting pretrained models such as VLMs to robotics through explicit training on robot data. Vision-language-action (VLA) models (Kim et al., 2024; Zitkovich et al., 2023) and world-action models (WAMs) (Ye et al., 2026; Yuan et al., 2026; Bi et al., 2026; Li et al., 2026) typically learn mappings from visual and linguistic observations to robot actions using large-scale collections of paired data, which are costly to collect, embodiment-specific, and difficult to scale. More importantly, recent evidence suggests a deeper limitation: training VLMs for action prediction may lead to degradation in general-purpose capabilities. including instruction following and reasoning (Hancock et al., 2026; Yang et al., 2026), This is particularly concerning because robot data remain orders of magnitude scarcer than the language-vision data used to build modern VLMs, raising the risk that large-scale parameter updates overfit to narrow robot data distributions and compromise the model's broadly generalizable intelligence. As shown in Fig. 1, these observations motivate a question: rather than adapting a VLM with large-scale robot training

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

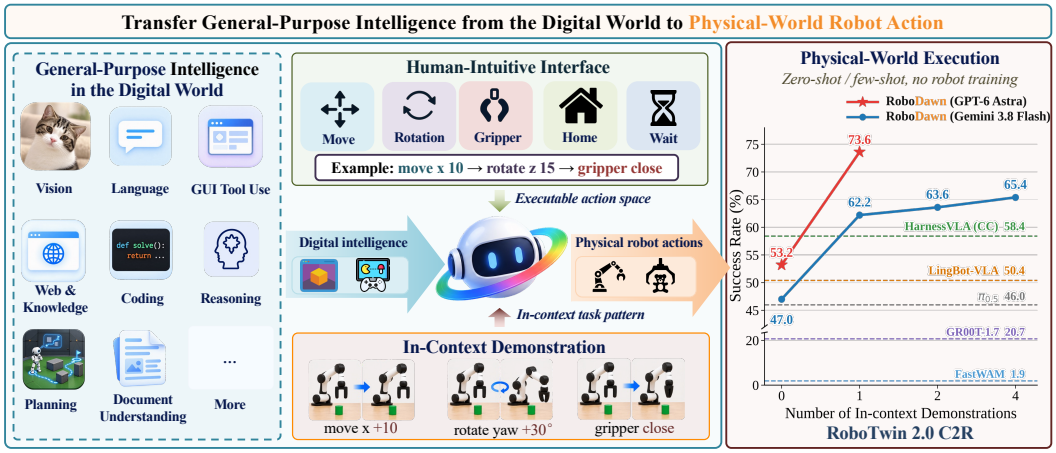


Figure 1: **RoboDawn** transfers general-purpose VLM intelligence from the digital world to physical robot control through a human-intuitive action interface and in-context demonstrations.

data, can we achieve **intelligence transfer** with a lightweight interface and only a few in-context demonstrations?

In this work, we investigate this hypothesis with **RoboDawn**, a human-intuitive interface that exposes robotic control to an agentic VLM through a compact set of discrete motion primitives. The interface provides translation, rotation, and gripper actions, each corresponding to a discrete incremental change in robot state, such as moving the end effector by multiples of a unit distance along a specified direction or rotating it by multiples of a unit angle. Instead of predicting low-level continuous controls directly, the VLM interacts with the robot through these semantically interpretable actions. It repeatedly observes the current visual state, reasons about the next action, executes that action, and observes its consequence before deciding what to do next. Robotic manipulation is thus formulated as a closed-loop visual decision-making process that is structurally similar to interactive environments in which modern multimodal models already exhibit strong reasoning capabilities.

A simple interface along with language rules, however, does not eliminate all ambiguity Yu & Mooney (2022). The model must understand the semantics and granularity of individual actions, infer how actions alter the observed scene, and learn interactions that may not be obvious from the action names themselves. Much like humans, who may struggle to fully understand certain rules from written instructions alone but can grasp them much more easily after observing a few demonstrations, VLMs may also benefit substantially from example-based guidance. We complement **RoboDawn** with an ICL scheme, using a few demonstrations to illustrate interface interaction. Without parameter updates, these examples provide task and interface context that helps the VLM infer effective action sequences. Success would suggest that intelligence need not be learned from scratch, but may already exist and simply require the right mechanism to unlock it.

We evaluate **RoboDawn** across diverse simulated and real-world manipulation settings spanning tasks, environments, and robot embodiments. In simulation, we study both RoboTwin 2.0 C2R and RoboDojo under zero-shot and one-shot settings. On RoboTwin 2.0 C2R, **RoboDawn** achieves 53.2% success zero-shot, already outperforming strong robot-trained policies such as $\pi_{0.5}$ (46.0%) and LingBot-VLA (50.4%), and improves substantially to 73.6% with only one in-context demonstration. Consistent improvements are observed on RoboDojo, with the success rate rising from 35.67% zero-shot to 47.17% one-shot. We further validate the framework on a Franka robot for block-in-basket and stacking tasks.Together, these results suggest that pretrained VLMs' intelligence can be translated into physical action through an appropriate interface and in-context samples.

Our results suggest a different perspective on the role of foundation models in robotics. Much of the current effort in generalist robotics focuses on expanding robot datasets and training specific action models. These directions are important, especially for low-level dexterity and high-frequency control. Our findings are complementary: for a broad class of manipulation problems, a strong pretrained VLM can already serve as the high-level decision-making engine, while a simple action

interface and a few demonstrations can transfer its intelligence to robotic control. From this perspective, the key bottleneck may not always be acquiring new embodied intelligence from scratch, but rather designing effective interfaces and online lessons that enable existing intelligence to transfer, act, and improve in the physical world.

We make the following contributions:

- We introduce **RoboDawn**, a human-intuitive interface that enables agentic VLMs to perform closed-loop robotic manipulation through discrete motion primitives without task-specific robot training. It offers a new blueprint for general embodied intelligence through **intelligence transfer** from pretrained multimodal models to embodied control.
- We develop an interface-aligned ICL scheme that enables frozen VLMs to learn how to act from only a few demonstrations, jointly grounding primitive action semantics and task-level strategies. Remarkably, even a single demonstration can substantially improve robotic manipulation performance without any parameter updates.
- Experiments show that **RoboDawn** achieves SOTA performance on both RoboTwin 2.0 C2R and RoboDojo with only a single in-context demonstration, surpassing prior robot-trained and agentic methods without task-specific parameter updates.

2 Related Work

2.1 From Task-Specific Intelligence to Transferable Intelligence

The development of artificial intelligence has shifted from task-specific competence toward increasingly transferable (or generalized) intelligence. Early deep learning systems (LeCun et al., 1998; Krizhevsky et al., 2012) achieved remarkable performance on individual proxy tasks, such as image classification (He et al., 2016; Simonyan & Zisserman, 2014), object detection (Redmon et al., 2016; Ren et al., 2016), and machine translation (Sutskever et al., 2014; Bahdanau et al., 2014), but the resulting capabilities were largely tied to the particular task and objective on which they were trained. The emergence of large language models (LLMs) (Brown et al., 2020; Ouyang et al., 2022) fundamentally changed this paradigm: a single pretrained model can perform and rapidly adapt to a broad range of tasks (Liu et al., 2023; Chen et al., 2021; Yao et al., 2022) through language instruction, without task-specific training. Such generalization suggests a form of intelligence that is increasingly reusable across tasks, resembling human intelligence in its ability to transfer previously acquired perception, knowledge, learning, and reasoning capabilities to new problems. In this work, we focus on such transferable intelligence and ask whether intelligence acquired predominantly in the digital world can further transfer to the physical world without specific training.

2.2 End-to-end Robotic Policy

End-to-end robotic policies directly predict executable actions from robot observations and task conditions. This paradigm dates back to early deep visuomotor control and large-scale robot learning (Levine et al., 2016; Pinto & Gupta, 2016). With the development of transformer- and diffusion-based architectures, end-to-end policies have gradually evolved from task-specific policy learning toward increasingly generalist robot control (Brohan et al., 2023; Ghosh et al., 2024; Liu et al., 2025), with larger and more diverse robot datasets enabling improved generalization across tasks, environments, and embodiments. VLA models leverage pretrained VLMs and further adapt their semantic and reasoning capabilities to robot action generation (Zitkovich et al., 2023; Kim et al., 2024; Zheng et al., 2025; Black et al., 2024; Bjorck et al., 2025; Team, 2025; Zhou et al., 2025). Meanwhile, WAMs, building on recent world models (Yang et al., 2024; Bruce et al., 2024), further unify future prediction and robot action generation (Ye et al., 2026; Team et al., 2026a). Despite these advances, these approaches still require training on action data. In contrast, we investigate whether the intelligence already present in pretrained VLMs can be directly transferred to closed-loop robotic control through an appropriate action interface, without task-specific parameter updates.

2.3 Agentic Methods for Robotic Control

An alternative line of work uses pretrained VLMs as online agents for robotic control, rather than relying solely on monolithic end-to-end policies. Earlier approaches primarily employ foundation

models for high-level planning and decision making, grounding their predictions into executable robot behaviors through pretrained skill libraries, environmental feedback, or programmatic control interfaces (Ichter et al., 2022; Liang et al., 2023; Huang et al., 2022; 2024). Recently, agentic robotic systems have increasingly incorporated memory, verification, replanning, and closed-loop interaction, while most still rely on pretrained VLA policies for low-level action execution (Yang et al., 2025; Liu et al., 2026b; Fu et al., 2026; Zhang et al., 2026c). Concurrent to our work, Show-Harness (Chen et al., 2026b) similarly demonstrates that frontier VLMs can directly perform closed-loop robot control through a compact semantic action interface. Compared with Show-Harness, **RoboDawn** goes beyond designing a harness for VLMs and also focuses on developing an effective ICL strategy co-design, which significantly improves their performance.

3 ROBODAWN

3.1 OVERALL FRAMEWORK

Figure 2 provides an overview of **RoboDawn**. A manipulation task can be specified by a natural-language instruction L. For each task, at decision round t, the VLM receives annotated visual observations I_t, measured robot state x_t, execution feedback F_{t-1} from the previous rounds, and interaction memory M_t. In addition, the model is conditioned on two forms of context that remain fixed throughout an episode: a robot–environment profile E and an in-context demonstration set D. The profile describes interface conventions such as workspace constraints, camera, grid-based localization, and gripper properties, whereas D provides in-context examples of how the interface can be used to interact with the environment. The closed-loop interaction can be written as:

$$\begin{aligned} (y_t, \mathbf{a}_t) &= \pi_\theta\left(L, E, D; I_t, x_t, F_{t-1}, M_t\right), \\ (s_{t+1}, F_t) &= \mathcal{E}_P\left(s_t, \mathbf{a}_t\right), \\ (I_{t+1}, x_{t+1}) &= \mathcal{O}_P(s_{t+1}), \\ M_{t+1} &= \mathcal{U}\left(M_t, \mathbf{a}_t, y_t, F_t, x_{t+1}\right). \end{aligned} \tag{1}$$

Here, π_θ denotes a pretrained VLM whose parameters remain frozen. The model outputs a sequence of semantic action commands $\mathbf{a}_t$ together with a structured response y_t. The structured response y_t contains the estimate of task progress, the current plan, and a compact scratchpad. The execution operator $\mathcal{E}_P$ parses these commands, grounds them into robot motions, executes them, and converts the resulting physical outcome into feedback F_t. The observation operator $\mathcal{O}_P$ then constructs the next visual and proprioceptive observation, while $\mathcal{U}$ updates the interaction memory from both model-generated and execution-derived information. The environment state s_t is introduced only to formalize physical state transitions and is not directly exposed to the VLM. In particular, online control does not rely on privileged object poses. Throughout an episode, θ, E, and D are fixed; adaptation arises from the evolving observation, execution feedback, and memory.

3.2 HUMAN-INTUITIVE INTERFACE

The first component of **RoboDawn** is a human-intuitive interface that bridges multimodal reasoning and physical robot control. Our goal is to expose robotic interaction in a form that is both easy for a pretrained VLM to interpret and straightforward to ground into physical execution. We firstly define the *gripper interaction point* (GIP) as the midpoint between the two fingertips of the gripper, which directly represents the point at which the robot interacts with objects. We use this GIP consistently for visual annotations, robot-state reporting, and motion commands, avoiding ambiguity between the wrist pose used by the low-level controller and the actual interaction point of the gripper.

Semantic action interface. We adopt a game-like action interface built from simple, semantically meaningful spatial primitives. Such commands are intuitive to humans and resemble interaction abstractions commonly used in games, teleoperation, and instructional content, making them more likely to align with concepts already encountered during web-scale VLM pretraining. As shown in Fig. 2(a), instead of directly predicting joint-level actions or high-frequency continuous end-effector controls, the VLM interacts with the robot through a compact vocabulary of parameterized semantic

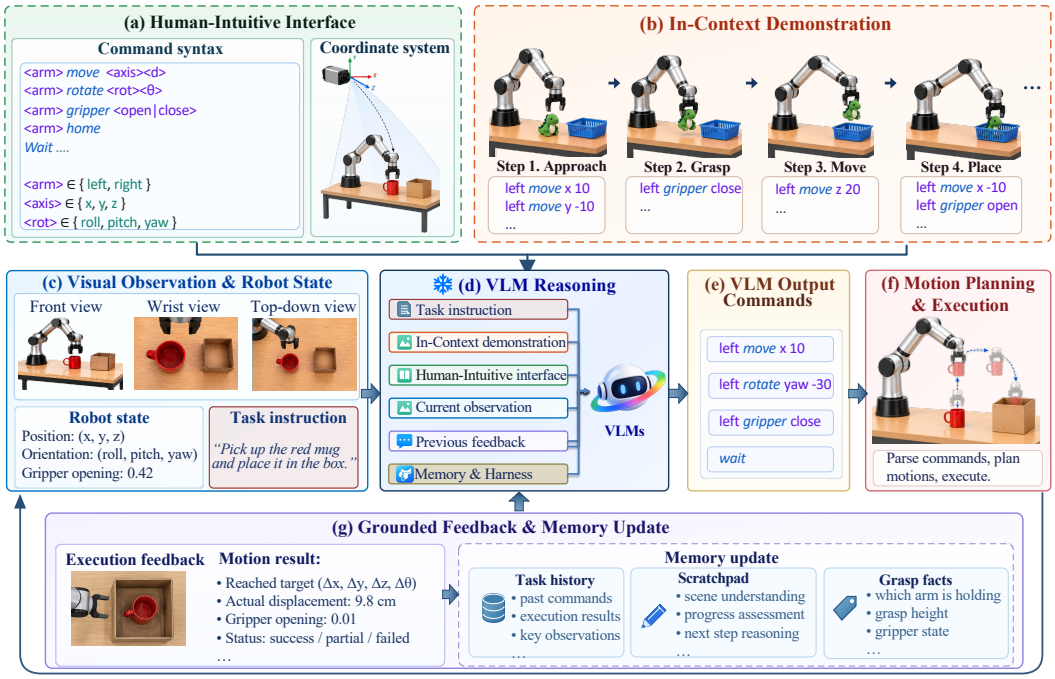


Figure 2: Overview of **RoboDawn**. **RoboDawn** conditions a pretrained VLM on (a) a human-intuitive semantic action interface and (b) in-context demonstrations. At each interaction round, the VLM receives (c) multi-view visual observations and robot state together with task context, execution feedback, and memory, performs (d) reasoning, and generates (e) semantic action commands. The commands are grounded into (f) planned robot motions and executed in the environment; the resulting physical outcomes are then converted into (g) grounded feedback and memory updates for the next round. This loop enables continual observation, reasoning, execution, and adaptation without task-specific parameter updates.

commands. The complete command grammar is

$$\begin{aligned}\mathcal{A} = \{&\texttt{<arm> move <axis> <d>},\\ &\texttt{<arm> rotate <rot> <}\theta\texttt{>},\\ &\texttt{<arm> point <pose>},\\ &\texttt{<arm> gripper <g>},\\ &\texttt{<arm> home, wait, done}\}.\end{aligned} \tag{2}$$

Here, `arm` $\in$ $\{$`left,right`$\}$; translation axes are $\{$`x,y,z`$\}$ and rotation axes are $\{$`roll,pitch,yaw`$\}$, both referring to the world frame; orientation presets include `down`, `forward`, and `down45`; and the gripper argument is `open`, `close`, or a normalized opening in $[0, 1]$. All spatial commands are defined with respect to the GIP and specify incremental changes of its pose: `move` translates the GIP by d along the given axis while preserving its orientation, whereas `rotate` turns it by θ about the given axis while preserving its position. The `point` command provides several common gripper-orientation presets, while `home`, `wait`, and `done` provide utilities for arm reset, environment settling, and task-completion checking.

From semantic commands to physical motion. Each semantic motion command is translated into a complete planned motion to a target GIP pose and executed until the robot reaches a stationary state, abstracting away low-level trajectory generation and control from the VLM.

3.3 In-Context Learning Design

A simple interface together with language-based action rules, however, does not eliminate all ambiguity Yu & Mooney (2022). The model must still understand the semantics and granularity of

individual actions, infer how each action changes the observed scene, and learn interaction conventions that may not be obvious from the command names alone. Much like humans, who may struggle to fully understand unfamiliar rules from written instructions but can grasp them more easily after observing a few examples, VLMs can benefit substantially from in-context demonstrations. This is particularly useful for action concepts such as gripper rotations and orientation changes, which may be less explicitly represented in web-scale pretraining data than simple translational motions; demonstrations provide direct visual examples that help the VLM learn their effects and incorporate them into task-level reasoning. We decompose the demonstration context as

$$D = D_{\text{prim}} \oplus D_{\text{task}}, \tag{3}$$

where D_{prim} is a shared command primer illustrating all basic command effects, and

$$D_{\text{task}} = \{\mathcal{D}^{(m)}\}_{m=1}^{N_D} \tag{4}$$

contains task-level demonstrations. This formulation does not assume a fixed number of examples: $N_D = 0$, $N_D = 1$, and $N_D > 1$ correspond to zero-shot, one-shot, and few-shot settings. Together, D_{prim} and D_{task} provide a two-level demonstration context: the former establishes how each primitive command affects the robot, while the latter shows when and how these primitives are composed into complete task-solving behaviors.

Interface-compatible demonstrations. Raw expert trajectories are not directly suitable as in-context examples, since they consist of continuous low-level actions that differ from those available to **RoboDawn**. We therefore express each trajectory in the semantic command space of Sec. 3.2. A trajectory is first reduced to a sequence of end-effector waypoints and gripper states. Each waypoint is then reached with a short sequence of translation, rotation, and gripper commands, which turns the entire trajectory into a command sequence that the online model could have issued itself.

Complete in-context demonstrations. Demonstrations are collected in scenes disjoint from those used for evaluation, from the benchmark's scripted expert in simulation and from human teleoperation on real robots; in simulation, these are the same trajectories used to train the robot policies we compare against. Each demonstration consists of N_m interaction rounds,

$$\mathcal{D}^{(m)} = \left\{ \left(I_j^{(m)}, x_j^{(m)}, r_j^{(m)}, \mathbf{a}_j^{(m)}, f_j^{(m)} \right) \right\}_{j=1}^{N_m}, \tag{5}$$

where $I_j^{(m)}$ is the visual observation at round j, $x_j^{(m)}$ the robot state, $\mathbf{a}_j^{(m)}$ the commands issued, $f_j^{(m)}$ their physical effect, and $r_j^{(m)}$ a short rationale. The physical effect $f_j^{(m)}$ is derived from consecutive states as the change in GIP pose and gripper opening. Note that $I_j^{(m)}$ may be empty for some rounds, while retaining the full textual trajectory for planning guidance. This sparsification is applied to long-horizon RoboDojo trajectories, where the number of rounds can substantially exceed the in-context image budget. We set the upper bound on the in-context image budget to 16 observations per round. When this budget is exceeded, we retain images only for semantically informative rounds, such as grasping, rotation, and task completion, while omitting visually redundant transition rounds. The rationale $r_j^{(m)}$ is written afterwards by a VLM, which reviews the recorded episode with a task-agnostic prompt and states, for each round, the plan behind its commands in the same format as the responses of online model.

4 Experiments

4.1 Simulation Benchmark Test

4.1.1 RoboTwin 2.0

We evaluate on RoboTwin 2.0 C2R Chen et al. (2025), covering 50 bimanual tasks. Following the official protocol, robot-trained baselines are post-trained on 50 clean demonstrations per task and evaluated under domain randomization. In contrast, **RoboDawn** is evaluated zero- or few-shot, with clean demonstrations provided only as in-context examples. We report average success rate over 10 runs per task, measuring generalization from clean to randomized scenes.

Table 1: Evaluation on RoboTwin 2.0 C2R. Full Set follows the official protocol, jointly post-training one policy on all 50 tasks with 50 clean demonstrations per task. 1-Shot denotes using 1 demonstration from the clean set. Success rates (%) are reported under the domain-randomized evaluation setting. HarnessVLA uses the pretrained LingBot-VLA as the default VLA.

Method	Number of Shots	Success rate (%)
FastWAM Yuan et al. (2026)	Full Set	1.9
StarVLA Community (2026)	Full Set	10.6
GalaxeaVLA Jiang et al. (2025)	Full Set	12.7
Xiaomi-Robotics-0 Cai et al. (2026)	Full Set	18.2
GR00T-1.7Bjorck et al. (2025)	Full Set	20.7
X-VLA Zheng et al. (2025)	Full Set	20.9
X-WAM Guo et al. (2026)	Full Set	25.8
$\pi_{0.5}$ Intelligence et al. (2025)	Full Set	46.0
LingBot-VLA Wu et al. (2026)	Full Set	50.4
HarnessVLA (Codex) Zhang et al. (2026c)	Full Set	58.0
HarnessVLA (CC) Zhang et al. (2026c)	Full Set	58.4
RoboDawn(Ours, Gemini-3.8-Flash)	0	47.0
RoboDawn(Ours, Gemini-3.8-Flash)	1	62.2
RoboDawn(Ours, GPT-6 Astra)	0	53.2
RoboDawn(Ours, GPT-6 Astra)	1	**73.6**

Tab. 1 compares **RoboDawn** with representative robot-trained and agentic baselines on RoboTwin 2.0 C2R. While prior methods are trained or post-trained on the full benchmark training set, **RoboDawn** performs robot control without task-specific parameter updates and uses at most a single demonstration as in-context guidance. Even in the zero-shot setting, **RoboDawn** with GPT-6 Astra achieves a 53.2% success rate, outperforming several strong robot policies including $\pi_{0.5}$ (46.0%) and LingBot-VLA (50.4%). With only one in-context demonstration, its success rate further increases to 73.6%, exceeding the SOTA agentic VLA method, HarnessVLA (Claude Code), by 15.2 percentage points. These results highlight that a pretrained VLM can effectively translate its general-purpose reasoning capabilities into robotic manipulation through a lightweight action interface, while benefiting substantially from few-shot in-context adaptation.

Ablation Study. As shown in Tab. 3, first, increasing the number of demonstrations generally improves performance: with Gemini-3.8-Flash, the success rate rises from 47.0% in the zero-shot setting to 62.2% with one demonstration, and further reaches 65.4% with four demonstrations. The marginal gain becomes smaller beyond one shot, while performance slightly drops to 62.7% with eight demonstrations, suggesting that excessively long context may interfere with effective VLM understanding, which is a commonly observed challenge in long-context VLM reasoning Ge et al. (2025). Second, **RoboDawn** benefits substantially from stronger VLMs. Under the same one-shot setting, the success rate increases from 14.4% with GPT-5.6-Luna to 43.2% with GPT-5.6-Sol, 62.2% with Gemini-3.8-Flash, and 73.6% with GPT-6 Astra, indicating that the effectiveness of **RoboDawn** scales with the underlying model capability. Finally, we ablate several components of the harness under the zero-shot setting. Removing reasoning reduces the success rate from 47.0% to 34.8%, while removing grid-based localization causes a larger drop to 32.4%. Removing the command primer leads to a smaller but still noticeable decrease to 44.0%. These results show that explicit reasoning and spatial grounding are particularly important for reliable robot control, while the command primer provides additional guidance for interpreting the action interface.

Execution Efficiency. To evaluate efficiency, we run each of the 50 tasks in RoboTwin 2.0 and report the average cost. Considering practical factors such as network latency, we use Seed-2.1-Pro as the VLM for latency evaluation. Tab. 2 compares the execution efficiency of **RoboDawn** with representative VLA and WAM policies. Beyond reporting inference latency alone, we additionally measure the inference-to-motion ratio to characterize the potential for pipelined execution between model reasoning and physical robot motion. In particular, when this ratio is below 0.5, inference can potentially be sufficiently overlapped with motion execution to enable streaming inference.

Table 2: Inference and execution efficiency of representative VLA and WAM policies. **Infer.**: inference latency; **Motion**: motion execution time; **I/M**: inference-to-motion time ratio. Seed-2.1-Pro on Volcano Engine is used as the VLM of **RoboDawn**, considering network latency. cmds denotes commands.

Model	Infer.	Actions	Motion	I/M
$\pi_{0.5}$	101 ms	45.0 steps	2.70 s	0.037
StarVLA	70 ms	16.0 steps	0.96 s	0.073
X-VLA	143 ms	28.6 steps	1.71 s	0.084
FastWAM	523 ms	27.3 steps	1.64 s	0.32
Motus	1.93 s	16.0 steps	0.96 s	2.01
LingBot-VA	8.89 s	22.2 steps	1.33 s	6.67
RoboDawn	9.74 s	3.4 cmds	2.09 s	4.65

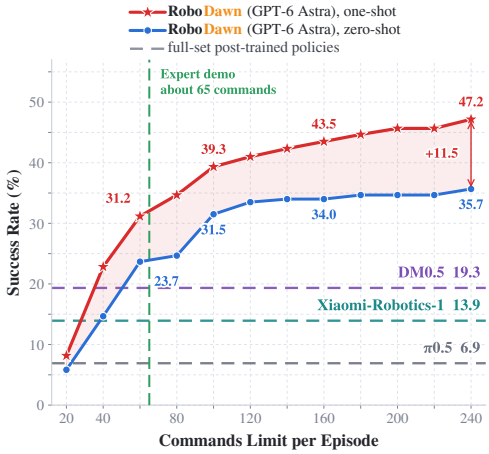


Figure 3: RoboDojo success rate under varying per-episode command budgets. **RoboDawn** exhibits clear test-time scaling.

Table 3: Ablation studies of **RoboDawn** on RoboTwin 2.0.

(a) **Number of shots (Gemini-3.8-Flash).**

# Shots	SR (%)
0	47.0
1	62.2
2	63.6
4	**65.4**
8	62.7

(b) **Models. Results use 1-shot by default**

Model	SR (%)
GPT-5.6-Luna	14.4
GPT-5.6-Sol	43.2
Seed-2.1-Pro	45.0
Gemini-3.8-Flash	62.2
GPT-6 Astra	**73.6**

(c) **Harness (Gemini 3.8 Flash, zero-shot).)**

Harness	SR (%)
Full	**47.0**
w/o reasoning	34.8
w/o grids	32.4
w/o primer	44.0

4.1.2 RoboDojo

RoboDojo Chen et al. (2026a) is a comprehensive benchmark designed to test generalist manipulation across generalization, memory, long-horizon, precision, and open capabilities. Its simulation suite contains 42 tasks on a bimanual robot platform. Following the benchmark protocol, we report both success rate and progress score. Here, we report the average results over 5 runs.

As shown in Tab. 4, **RoboDawn** achieves strong performance on RoboDojo without task-specific post-training. In the zero-shot setting, it attains a success rate of 35.67%. Providing only a single demonstration further improves the success rate to 47.17%. This result substantially surpasses full-set post-trained policies such as DM0.5, which achieves a success rate of 19.34%.

Beyond final task performance, Fig. 3 reveals a clear test-time scaling behavior: as the command budget increases, the success rate consistently improves. In particular, the one-shot success rate increases from 31.2% with 60 commands to 47.2% with 240 commands, while the zero-shot variant improves from 23.7% to 35.7%. This consistent scaling suggests that **RoboDawn** can effectively translate additional test-time computation into improved task success. We attribute this behavior to **RoboDawn**'s strong failure-recovery capability and memory mechanism, which enable it to adapt its subsequent actions based on previous observations and execution outcomes.

4.2 Real-World Robotics Deployment

Real-world evaluation. We further deploy **RoboDawn** with Gemini 3.8 Flash on real robots to evaluate whether the proposed harness transfers beyond simulation under zero-shot setting. As shown in Tab. 5, **RoboDawn** achieves promising performance on real-world block-in-basket and block-stacking tasks, demonstrating that the same harness can be directly applied to physical robots

Table 4: Evaluation on RoboDojo Chen et al. (2026a). Full Set follows the official RoboDojo training protocol, in which a single policy is post-trained on the full RoboDojo-specific training set. Score denotes the task progress score. Compared with the full set of 42 tasks, the 34 tasks setting excludes 8 open tasks. For these 8 tasks, other methods may not have corresponding training data.

Method	# Shots	Score	Success rate (%)
LingBot-VLA Wu et al. (2026)	Full Set	5.50	2.96
X-VLA Zheng et al. (2025)	Full Set	10.13	6.52
$\pi_{0.5}$ Intelligence et al. (2025)	Full Set	11.41	6.91
Hy-Embodied-0.5-VLA Zhang et al. (2026a)	Full Set	13.07	8.80
Xiaomi-Robotics-1 Team et al. (2026b)	Full Set	20.07	13.93
GalaxeaVLA (G0.5) Liu et al. (2026a)	Full Set	20.23	14.88
DM0.5 Dexmal (2026)	Full Set	24.90	19.34
GPT-6 Astra Zhang et al. (2026b)	0	28.97	22.58
RoboDawn(Ours, GPT-6 Astra, all tasks)	0	39.92	35.67
RoboDawn(Ours, GPT-6 Astra, all tasks)	1	**54.63**	**47.17**
RoboDawn(Ours, GPT-6 Astra, 34 tasks)	0	39.02	33.96
RoboDawn(Ours, GPT-6 Astra, 34 tasks)	1	**51.29**	**43.33**

Table 5: **Real-world evaluation.** Success rates on three real-world manipulation tasks.

	Block in Basket (Franka)	Block Stacking (Franka)	Cloth Folding (Piper)
Success Rate	9/10	5/10	0/10

in zero-shot setting. Block stacking achieves a lower success rate than block-in-basket, mainly because it requires more precise perception and spatial alignment during placement. Cloth folding remains considerably more challenging. We hypothesize that this is mainly because the task involves substantial end-effector rotation and orientation adjustment, which may be less frequently represented in the web-scale data used to train VLMs and are therefore harder for the model to reason about reliably. In fact, we observed several near-successful cases in the cloth-folding task. However, the final folds were not sufficiently neat, so we did not count these trials as successful.

5 Limitations

Rome was not built in a day. This work emphasizes generalization rather than a complete solution to robotic manipulation, positioning **RoboDawn** as an early step toward more generalizable embodied intelligence. Several limitations remain. **(a) Inference efficiency.** Iterative VLM reasoning and closed-loop interaction make **RoboDawn** slower than end-to-end VLA policies. **(b) Uneven DoF difficulty.** Translational motions are more reliable than precise arm and end-effector rotations; ICL helps but does not eliminate this gap. **(c) Fine-grained control.** The discrete semantic action interface can be too coarse for contact-rich interactions requiring precise grasping, placement, or local adjustment. **(d) Deployment safety.** Incorrect actions may cause collisions, object damage, or risks to humans, making safety critical for broader deployment.

6 Conclusion

We introduce **RoboDawn**, a framework that transfers pretrained VLM intelligence to robotic control through a human-intuitive action interface and in-context demonstrations, without task-specific robot training. Experiments on RoboTwin 2.0, RoboDojo, and real robots show strong zero-shot performance and substantial gains from only a single demonstration. Our results suggest that general-purpose embodied intelligence may emerge not only from scaling robot data, but also from better interfaces and lessons that unlock and transfer capabilities already present in foundation models.

7 AI USE STATEMENT

Generative AI tools such as GPT were used to assist with writing and language polishing. All AI-generated or AI-assisted content was carefully reviewed, verified, and revised as necessary by the authors.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pp. 35101–35113. Computer Vision Foundation, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Bi_Motus_A_Unified_Latent_Action_World_Model_CVPR_2026_paper.html.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith LLontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: an open foundation model for generalist humanoid robots. *CoRR*, abs/2503.14734, 2025. doi: 10.48550/ARXIV.2503.14734. URL https://doi.org/10.48550/arXiv.2503.14734.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. *CoRR*, abs/2410.24164, 2024. doi: 10.48550/ARXIV.2410.24164. URL https://doi.org/10.48550/arXiv.2410.24164.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong T. Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: robotics transformer for real-world control at scale. In Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu (eds.), *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. doi: 10.15607/RSS.2023.XIX.025. URL https://doi.org/10.15607/RSS.2023.XIX.025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal M. P. Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott E. Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp.

4603–4623. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Rui Cai, Jun Guo, Xinze He, Piaopiao Jin, Jie Li, Bingxuan Lin, Futeng Liu, Wei Liu, Fei Ma, Kun Ma, et al. Xiaomi-robotics-0: An open-sourced vision-language-action model with real-time execution. *arXiv preprint arXiv:2602.12684*, 2026.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Tianxing Chen, Yue Chen, Zixuan Li, Junyuan Tang, Kailun Su, Haoran Lu, Weijie Wan, Baijun Chen, Songling Liu, Haowen Yan, et al. Robodojo: A unified sim-and-real benchmark for comprehensive evaluation of generalist robot manipulation policies. *arXiv preprint arXiv:2607.04434*, 2026a.

Yanzhe Chen, Zechen Bai, Zhijun Cao, Wenzheng Zeng, Kevin Qinghong Lin, Yiqi Lin, Guoqiang Liang, Kevin Yuchen Ma, Qiming Huang, and Mike Zheng Shou. Show-harness: Just a vlm agent can play robots, 2026b. URL https://arxiv.org/abs/2609.10522.

StarVLA Community. Starvla: A lego-like codebase for vision-language-action model developing. *arXiv preprint arXiv:2604.05014*, 2026.

Dexmal. Dm0.5: Designed for the open world, where generalization emerges. *https://www.dexmal.com/blog/dm0.5*, 2026.

Max Fu, Justin Yu, Karim El-Refai, Ethan Kou, Haoru Xue, Huang Huang, Wenli Xiao, Guanzhi Wang, Feifei Li, Guanya Shi, Jiajun Wu, Shankar S. Sastry, Yuke Zhu, Ken Goldberg, and Linxi "Jim" Fan. Cap-x: A framework for benchmarking and improving coding agents for robot manipulation. *CoRR*, abs/2603.22435, 2026. doi: 10.48550/ARXIV.2603.22435. URL https://doi.org/10.48550/arXiv.2603.22435.

Junqi Ge, Ziyi Chen, Jintao Lin, Jinguo Zhu, Xihui Liu, Jifeng Dai, and Xizhou Zhu. V2pe: Improving multimodal long-context capability of vision-language models with variable visual position encoding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21070–21084. IEEE, 2025.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In Dana Kulic, Gentiane Venture, Kostas E. Bekris, and Enrique Coronado (eds.), *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, 2024. doi: 10.15607/RSS.2024.XX.090. URL https://doi.org/10.15607/RSS.2024.XX.090.

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

Asher Hancock, Xindi Wu, Lihan Zha, Olga Russakovsky, and Anirudha Majumdar. Actions as language: Fine-tuning vlms into vlas without catastrophic forgetting. In *International Conference on Learning Representations*, volume 2026, pp. 75333–75360, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In Karen Liu, Dana Kulic, and Jeffrey Ichnowski (eds.), *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pp. 1769–1782. PMLR, 2022. URL https://proceedings.mlr.press/v205/huang23c.html.

Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4573–4602. PMLR, 2024. URL https://proceedings.mlr.press/v270/huang25g.html.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as I can, not as I say: Grounding language in robotic affordances. In Karen Liu, Dana Kulic, and Jeffrey Ichnowski (eds.), *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318. PMLR, 2022. URL https://proceedings.mlr.press/v205/ichter23a.html.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *CoRR*, abs/2504.16054, 2025. doi: 10.48550/ARXIV.2504.16054. URL https://doi.org/10.48550/arXiv.2504.16054.

Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea open-world dataset and g0 dual-system vla model. *arXiv preprint arXiv:2509.00576*, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Paul Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2024. URL https://proceedings.mlr.press/v270/kim25c.html.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.*, 17:39:1–39:40, 2016. URL https://jmlr.org/papers/v17/15-522.html.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pp. 9493–9500. IEEE, 2023. doi: 10.1109/ICRA48891.2023.10160591. URL https://doi.org/10.1109/ICRA48891.2023.10160591.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1B: a diffusion foundation model for bimanual manipulation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=yAzN4tz7oI.

Yicheng Liu, Zibin Dong, Baijun Ye, Tianyuan Yuan, Tao Jiang, Anqi Yang, Shicheng Cao, Haonan Liu, Yue Sun, Zihan Guo, et al. G0. 5: One autoregressive stream for robot reasoning and action. *arXiv preprint arXiv:2608.11739*, 2026a.

Zhen Liu, Xinyu Ning, Zhe Hu, XinXin Xie, Weize Li, Zhipeng Tang, Chongyu Wang, Zejun Yang, Hanlin Wang, Yitong Liu, and Zhongzhu Pu. Goal2skill: Long-horizon manipulation with adaptive planning and reflection. *CoRR*, abs/2604.13942, 2026b. doi: 10.48550/ARXIV.2604.13942. URL https://doi.org/10.48550/arXiv.2604.13942.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In Danica Kragic, Antonio Bicchi, and Alessandro De Luca (eds.), *2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-21, 2016*, pp. 3406–3413. IEEE, 2016. doi: 10.1109/ICRA.2016.7487517. URL https://doi.org/10.1109/ICRA.2016.7487517.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Gemini Robotics Team. Gemini robotics: Bringing AI into the physical world. *CoRR*, abs/2503.20020, 2025. doi: 10.48550/ARXIV.2503.20020. URL https://doi.org/10.48550/arXiv.2503.20020.

MotuBrain Team, Chendong Xiang, Fan Bao, Haitian Liu, Hengkai Tan, Hongzhe Bi, James Li, Jiabao Liu, Jingrui Pang, Kiro Jing, Louis Liu, Mengchen Cai, Rongxu Cui, Ruowen Zhao, Runqing Wang, Shuhe Huang, Yao Feng, Yinze Rong, Zeyuan Wang, and Jun Zhu. Motubrain: An advanced world action model for robot control. *CoRR*, abs/2604.27792, 2026a. doi: 10.48550/ARXIV.2604.27792. URL https://doi.org/10.48550/arXiv.2604.27792.

Xiaomi Robotics Team, Jun Guo, Piaopiao Jin, Jason Li, Peiyan Li, Yingyan Li, Futeng Liu, Wanli Peng, Optimus Qin, Yifei Su, Nan Sun, Qiao Sun, Runze Suo, Heyun Wang, Yunhong Wang, Rujie Wu, Caoyu Xia, Lina Zhang, Jack Zhao, Guoliang Chen, Wenlong Chen, Xinze He, Bin Li, Qing Li, Zhuorong Li, Heng Qu, Wenxuan Song, Diyun Xiang, Yifan Xie, Peiran Xu, Hangjun

Ye, Wen Ye, Han Zhao, and Quanyun Zhou. Xiaomi-robotics-1: Scaling vision-language-action models with over 100k hours of real-world trajectories, 2026b. URL https://arxiv.org/abs/2607.15330.

Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, et al. A pragmatic vla foundation model. *arXiv preprint arXiv:2601.18692*, 2026.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=sFyTZEqmUY.

Shuai Yang, Hao Li, Bin Wang, Yilun Chen, Yang Tian, Tai Wang, Hanqing Wang, Feng Zhao, Yiyi Liao, and Jiangmiao Pang. Vision-language-action instruction tuning: From understanding to manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 152896–152943, 2026.

Zhejian Yang, Yongchao Chen, Xueyang Zhou, Jiangyue Yan, Dingjie Song, Yinuo Liu, Yuting Li, Yu Zhang, Pan Zhou, Hechang Chen, and Lichao Sun. Agentic robot: A brain-inspired framework for vision-language-action models in embodied agents. *CoRR*, abs/2505.23450, 2025. doi: 10.48550/ARXIV.2505.23450. URL https://doi.org/10.48550/arXiv.2505.23450.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies. *CoRR*, abs/2602.15922, 2026. doi: 10.48550/ARXIV.2602.15922. URL https://doi.org/10.48550/arXiv.2602.15922.

Albert Yu and Raymond J Mooney. Using both demonstrations and language instructions to efficiently learn robotic tasks. *arXiv preprint arXiv:2210.04476*, 2022.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

He Zhang, Lingzhu Xiang, Haitao Lin, Zeyu Huang, Minghui Wang, Dingyan Zhong, Yubo Dong, Yihao Wu, Yongming Rao, Dongsheng Zhang, et al. Hy-embodied-0.5-vla: From vision-language-action models to a real-world robot learning stack. *arXiv preprint arXiv:2606.14409*, 2026a.

Wenbo Zhang, Kaixuan Wang, Yutao Ouyang, Xiaoyu Huang, Liyang Li, Kailun Su, Weiyang Jin, Wenhao Chai, Haotian Liang, Zhiyang Dou, Yue Chen, and Tianxing Chen. An unexpected robot policy: Early evaluations of gpt-6 astra on robodojo and beyond, 2026b. URL https://robodojo-benchmark.com/gpt6-astra-robodojo. Equal contribution: Wenbo Zhang, Kaixuan Wang, Yutao Ouyang. Project lead: Wenbo Zhang. Corresponding: Yue Chen, Tianxing Chen.

Yixian Zhang, Huanming Zhang, Feng Gao, Xiao Li, Zhihao Liu, Chunyang Zhu, Jiaxing Qiu, Yuchen Yan, Jiyuan Liu, Wenhao Tang, Zhengru Fang, Yi Nie, Changxu Wei, Yu Wang, Wenbo Ding, and Chao Yu. Harness VLA: steering frozen vlas into reliable manipulation primitives via memory-guided agents. *CoRR*, abs/2607.08448, 2026c. doi: 10.48550/ARXIV.2607.08448. URL https://doi.org/10.48550/arXiv.2607.08448.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, Ya-Qin Zhang, Jiangmiao Pang, Jingjing Liu, Tai Wang, and Xianyuan Zhan. X-VLA: soft-prompted transformer as scalable cross-embodiment vision-language-action model. *CoRR*, abs/2510.10274, 2025. doi: 10.48550/ARXIV.2510.10274. URL https://doi.org/10.48550/arXiv.2510.10274.

Enshen Zhou, Cheng Chi, Yibo Li, Jingkun An, Jiayuan Zhang, Shanyu Rong, Yi Han, Yuheng Ji, Mengzhen Liu, Pengwei Wang, et al. Robotracer: Mastering spatial trace with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2512.13660*, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

A Appendix

A.1 Failure Analysis.

We identify three representative failure modes in RoboDojo, as illustrated in Fig. 4. (a) Insufficient manipulation precision. Although the model can infer the correct high-level intention, it may fail at the final stage when the task requires very accurate positioning, such as inserting an object into a small target region. This reflects the limited fine-grained control of the current semantic action interface. (b) mode comes from IK-related execution errors. In these cases, the predicted action sequence is reasonable at the semantic level, but the realized robot motion collides with nearby objects during execution, leading to scene disturbance and eventual failure. This reveals a mismatch between high-level semantic planning and low-level physically safe execution. (c) mode is incorrect success judgment. The model pours liquid into the cup but stops before a sufficient amount is transferred. This reveals a misalignment between the model's notion of task completion and the benchmark's success criterion.

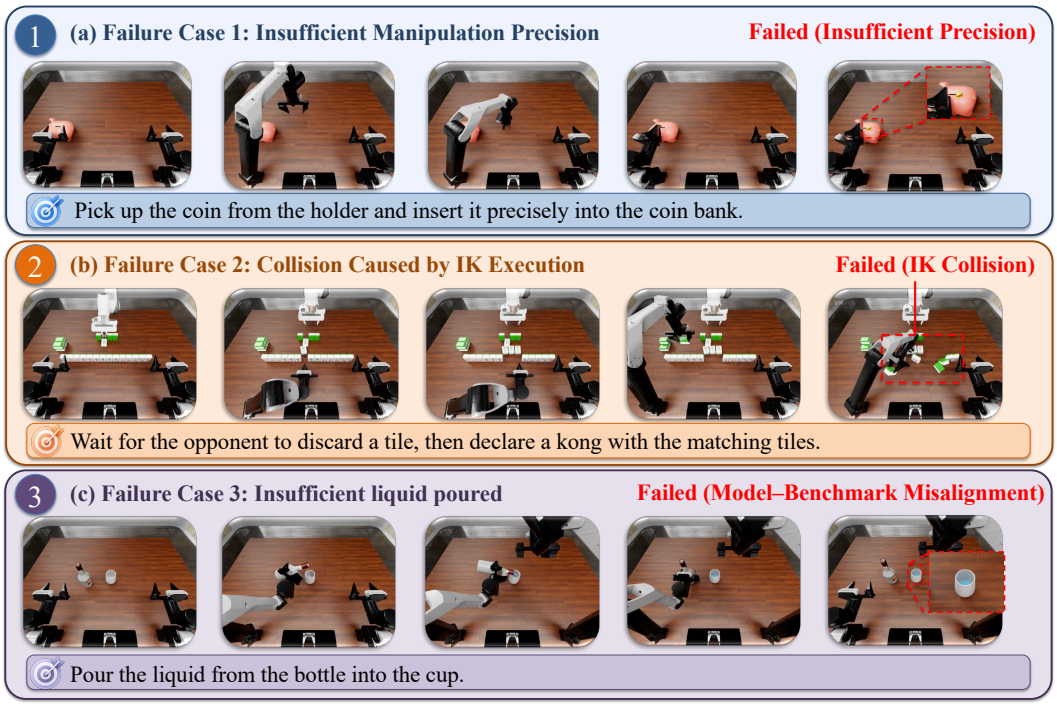


Figure 4: Representative failure cases of **RoboDawn** on RoboDojo.