Institute of Trustworthy Embodied AI
Fudan University
OpenMOSS

Learning to Move Before Learning to Do: Task-Agnostic pretraining for VLAs

Junhao Shi[1,2] **Siyin Wang**[1,2] **Xiaopeng Yu**[1] **Li Ji**[1] **Jingjing Gong**[2,†] **Xipeng Qiu**[1,2,†]

24110240071@m.fudan.edu.cn

[1]Fudan University [2]Shanghai Innovation Institute

Abstract

Vision-Language-Action (VLA) models are fundamentally bottlenecked by the scarcity of expert demonstrations—triplets of observations, instructions, and actions that are costly to collect at scale. We argue that this bottleneck stems from conflating two distinct learning objectives: acquiring *physical competence* (how to move) and acquiring *semantic alignment* (what to do). Crucially, only the latter requires language supervision. Building on this *Decomposition Hypothesis*, we propose **Task-Agnostic Pretraining (TAP)**, a two-stage framework that first learns transferable motor priors from cheap, unlabeled interaction data—including discarded off-task trajectories and autonomous robot play—via a self-supervised Inverse Dynamics objective. A lightweight second stage then grounds these priors in language using minimal expert data. On the SIMPLER benchmark, TAP matches models trained on over 1M expert trajectories while using orders of magnitude less labeled data, yielding a 10% absolute gain over standard behavior cloning. On a real-world WidowX platform, TAP retains 25% success under camera perturbations where internet-scale baselines collapse to 0%, demonstrating that task-agnostic pretraining produces robust, transferable physical representations and offers a scalable path forward for Embodied AI.

Homepage: https://sjh0354.github.io/task_agnostic_pretrain
Github Repo: https://github.com/sjh0354/Task-Agnostic-Pretrain
HF Models: https://huggingface.co/collections/Michael0354/task-agnostic-pretrain

1 Introduction

The development of Vision-Language-Action (VLA) models [1–5] has opened promising avenues for building general-purpose robots. However, training these models requires an enormous amount of high-quality robotic data. The predominant approach to acquiring such data relies on human teleoperation, where expert operators guide robot movements while providing language based task annotations [6–8]. This paradigm is inherently *passive* from the robot's perspective: the robot serves merely as a vessel for capturing human demonstrations, contributing no exploration of its own. Beyond being prohibitively expensive and labor-intensive, this data collection scheme is fundamentally *unnatural*—it conflates the embodied experience of a robot with the disembodied intentions of a human operator. Crucially, it is also *non-scalable*: the rate of data acquisition is bottlenecked by the availability and endurance of human operators, making it impractical to

†Corresponding Authors.

collect the diverse, large-scale datasets required for truly general-purpose manipulation.

Consider, by contrast, how biological agents acquire motor competence. A human infant does not learn to grasp, manipulate, and interact with objects by passively receiving demonstrations from an expert. Instead, infants engage in spontaneous, *task-agnostic* exploration—reaching, touching, dropping, and observing the consequences of their actions [9]. This curiosity-driven self-exploration is philosophically *task-unaware*: the infant is not optimizing for any specific goal but is rather building an internal model of how the world responds to its actions [10]. Through this process, the infant develops a rich understanding of physics, affordances, and sensorimotor contingencies [11] long before any explicit task instruction is given. The grounding in "how to move" emerges naturally from active interaction, decoupled from "what to do."

Inspired by this developmental perspective, we argue that robots should similarly benefit from *active*, task-agnostic data collection. Such data is abundant and inexpensive to acquire: robots can autonomously generate vast amounts of interaction trajectories through random play, without human supervision or task-specific annotations. Yet, this valuable resource remains largely underutilized in current VLA training pipelines, which discard any trajectory that lacks explicit task instructions. We observe that while task instructions are necessary for learning "what to do," they are not required for learning "how to move"—the fundamental dynamics and physical affordances of manipulation.

In this work, we demonstrate how to unlock the value of task-agnostic data for VLA learning. We identify two abundant sources of such data: (1) *task-irrelevant trajectories*—existing demonstrations collected for unrelated tasks that are typically discarded, and (2) *autonomous random play*—interaction data generated by robots exploring their environment without human supervision. To extract physical knowledge from this unlabeled data, we employ an Inverse Dynamics objective [12], where the model learns to predict the action a_t required to transition from observation o_t to a future state o_{t+1}. This self-supervised formulation forces the model to attend to dynamic elements—end-effector motion, object displacement—while ignoring static background noise, thereby acquiring "physical common sense" without any language annotations. With this grounding in place, only a minimal set of expert demonstrations is needed to align the learned affordances with task-specific linguistic instructions. We call this approach **Task-Agnostic Pretraining (TAP)**.

We evaluate our approach on both the Simpler benchmark [13] and real-world WidowX 250s robot experiments. In Simpler, by repurposing task-irrelevant trajectories from the Bridge dataset [7]for inverse dynamics training, we significantly improve performance on downstream tasks compared to standard training baselines. In the real world, we demonstrate that pretraining on autonomously generated random trajectories reduces the dependency on expensive expert teleoperation. Our results show that our method achieves superior sample efficiency and generalization, effectively transforming "useless" task-agnostic data—collected in the spirit of infant-like self-exploration—into a valuable resource for scaling Embodied AI.

2 Related Works

2.1 Large-scale pretraining for Vision-Language-Action Models

For years, visuomotor control was dominated by task-specific policies trained within constrained environments. While effective for isolated skills, these methods struggled to generalize to novel objects or unstructured language instructions[14–16].

Inspired by the scaling laws of natural language processing, RT-1 [17] and RT-2[18]pioneered the shift toward "generalist" agents, demonstrating that unifying perception, language, and action into a single Transformer and scaling data diversity could unlock emergent generalization capabilities. To harness the scaling potential of these evolving architectures, the community has increasingly focused on aggregating massive, multi-embodiment datasets. This effort culminated in the Open X-Embodiment (OXE) [4] dataset, which unifies diverse data sources such as BridgeData [7] and DROID [19]. Fueled by these millions of expert-teleoperated trajectories, the latest generation of VLA models has achieved remarkable success. Systems such as OpenVLA [2], π_0 [5], $\pi_{0.5}$ [1], and Gen-0 [20] integrate internet-scale VLM backbones with diffusion

or flow-matching action heads, achieving unprecedented levels of dexterity and real-time control through large-scale pretraining. Furthermore, RoboOmni [21] extends this paradigm by exploring and expanding the native end-to-end omni-modal capabilities of VLAs.

Despite their impressive performance, current state-of-the-art VLAs still suffer from a fundamental bottleneck: the "data wall." Their generalization is strictly bounded by the scale of expert human demonstrations, where generalization is constrained by the prohibitive cost of scaling expert demonstrations. Our work challenges this brute-force scaling regime by pretraining manipulation priors using cheap, task-agnostic data, thereby significantly reducing the dependency on expensive expert demonstrations.

2.2 Dynamics Learning in Robotics

Dynamics learning equips robot policies with physical reasoning by modeling state transitions, typically categorized into *forward dynamics* ($\hat{s}_{t+1} \leftarrow f_{\text{fwd}}(s_t, a_t)$) and *inverse dynamics* ($\hat{a}_t \leftarrow f_{\text{inv}}(s_t, s_{t+1})$). Prior works have extensively leveraged these objectives to pretrain visual representations. Explicit modeling approaches, such as MIDAS [22], SMART [23], and PACT [24], directly predict future states or actions to capture local physical laws and environmental transitions. Conversely, implicit methods like Vi-PRoM [25] and MaskDP [26] internalize dynamics through temporal reordering or masked reconstruction tasks without explicit state prediction. Furthermore, video-based frameworks such as VPT [27] and GR-1 [28] scale these objectives to large unlabeled datasets, utilizing inverse dynamics primarily for pseudo-labeling internet videos or anticipating future frames to refine action prediction.

Most prior methods treat dynamics learning either as an auxiliary objective or a tool for pseudo-labeling data. In contrast, we employ inverse dynamics as a *standalone pretraining phase* specifically to unlock the value of massive, task-agnostic action data. By learning physical priors—such as object affordances and kinematics—*before* encountering any task semantics, our method provides a robust structural foundation that significantly enhances downstream learning efficiency and performance.

3 Task-Agnostic Data for Physical Grounding

The fundamental bottleneck in VLA learning is the scarcity of aligned triplets (o, l, a)—observations paired with both language instructions and expert actions. We propose to bypass this bottleneck by exploiting a vastly more abundant resource: **task-agnostic interaction data**. Our key insight is a *Decomposition Hypothesis*: action generation can be factorized into (1) perceiving physical affordances ("how to move") and (2) grounding semantic intent ("what to do"). Crucially, the former can be learned entirely from task-agnostic data, without any language annotations.

As shown in Figure 1, our framework operationalizes this hypothesis by first harvesting massive task-agnostic data (Sec. 3.1) to learn physical priors via Inverse Dynamics (Sec. 3.2), and subsequently aligning these priors with semantic instructions via finetuning (Sec. 3.3).

3.1 Harnessing Task-Agnostic Data

We define **task-agnostic data** as any robot interaction trajectory $\tau = (o_0, a_0, o_1, a_1, \ldots, o_T)$ that lacks explicit task semantics—i.e., no language instruction l is associated with the trajectory. Such data captures valid physical interactions (the robot moved, objects responded) but carries no human-defined "purpose."

Sources of Task-Agnostic Data. This data is abundant, cheap, and exists in two primary forms:

- **Repurposed Existing Datasets.** Large-scale robotic datasets (e.g., BridgeData [7], Open X-Embodiment [4]) contain thousands of trajectories collected for tasks *irrelevant* to a target deployment. For instance, if the target downstream task is "put carrot on plate," trajectories of "open drawer" or "wipe table" are traditionally discarded. We argue these contain rich physical priors—grasping dynamics, collision responses, end-effector control—that transfer across tasks.

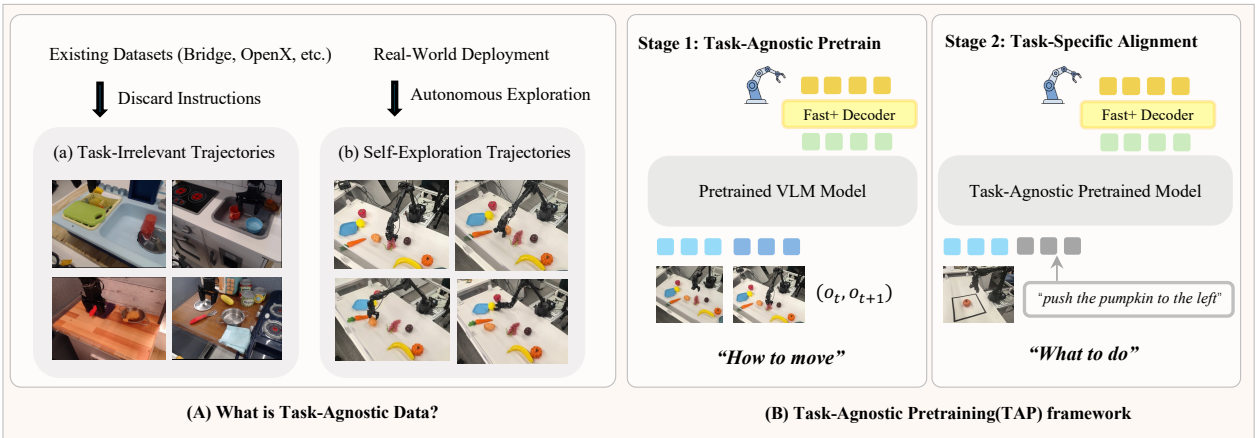

Figure 1 Overview of the Proposed Task-Agnostic Pretraining (TAP) Framework. (a) Data Sources: We leverage massive amounts of cheap, unlabeled interaction data—sourced from existing heterogeneous datasets (e.g., Bridge) or autonomous robot self-exploration —discarding any original task labels. **(b) Stage 1 (Task-Agnostic Pretraining):** The model is pretrained using a self-supervised Inverse Dynamics (ID) objective. By predicting the action a_t required to transition between frames o_t and o_{t+1}, the model learns robust physical affordances and motor control ("how to move") without human supervision. **(c) Stage 2 (Task-Specific Alignment):** The pretrained model is then finetuned on a small set of language-annotated expert demonstrations. This stage aligns the prelearned physical priors with high-level semantic instructions ("what to do"), achieving high performance with significantly improved data efficiency.

- **Autonomous Random Play.** Robots can generate unlimited interaction data through self-supervised exploration. By executing randomized end-effector commands, the robot pushes, sweeps, topples, and grasps objects without any human involvement. This "play" data is virtually free to collect and captures the robot's specific embodiment and workspace.

Autonomous Collection Pipeline. To ensure the robot collects meaningful, contact-rich data safely, we procedurally generate trajectories grounded in a verified spatial prior. First, an operator teleoperates the robot without specific tasks to densely cover the reachable workspace. We filter this data and apply Voxel Grid Downsampling to construct a uniform, discrete safe pose library $\mathcal{P}_{safe}$. Next, we stochastically sample waypoints from this library to form continuous trajectories. To prevent the end-effector from hovering and force meaningful interactions (e.g., pushing, sliding), we apply a contact heuristic that forces a descent if the trajectory remains above an elevation threshold (z_{thresh}) for too long. Finally, boundary-aware Gaussian noise is injected to maximize diversity before the robot executes the trajectory.ng). Detailed pipeline setups are provided in Appendix A.

3.2 Stage 1: Task-Agnostic Pretraining

To extract physical knowledge from unlabeled trajectories, we formulate a self-supervised **Inverse Dynamics (ID)** objective. Given two observations (o_t, o_{t+1}), the model predicts the action a_t that caused the transition:

$$p(a_t \mid o_t, o_{t+1}) \tag{1}$$

Why Inverse Dynamics? Predicting the action a_t that caused a state transition requires the model to focus on *what changed* between frames—the motion of the end-effector and the displacement of manipulated objects—while ignoring static background elements (lighting, textures, clutter). This forces the visual encoder to learn *dynamics-aware* representations that encode "how the world changes" rather than "how the world looks."

Training Objective. We instantiate f_θ using a Vision-Language Model (VLM) backbone. During inverse dynamics training, we construct a visual-only input sequence by treating the future observation o_{t+1} as an

Algorithm 1 Constrained Procedural Trajectory Generation

Require: Raw teleoperation poses $\mathcal{P}_{raw}$, safety bounds $\mathcal{B}$, voxel size v_{size}, min distance d_{min}, elevation threshold z_{thresh}, max high-elevation steps c_{max}, noise scale σ.
Ensure: Procedural trajectory dataset $\mathcal{D}_{play}$

Phase 1: Safe Pose Library Initialization
$\mathcal{P}_{valid} \leftarrow \{p \in \mathcal{P}_{raw} \mid p \in \mathcal{B}\}$
$\mathcal{P}_{safe} \leftarrow \text{VoxelGridDownsample}(\mathcal{P}_{valid}, v_{size})$
Phase 2: Autonomous Data Collection
$\mathcal{D}_{play} \leftarrow \emptyset$
while True do
  $\mathcal{W} \leftarrow \text{SampleWaypoints}(\mathcal{P}_{safe}, d_{min})$ {{Stochastic sampling}}
  $\mathcal{W}_{contact} \leftarrow \text{ContactHeuristic}(\mathcal{W}, z_{thresh}, c_{max})$ {{Bound consecutive high-Z points}}
  $\tau \leftarrow \text{CosineInterpolate}(\mathcal{W}_{contact})$
  $\tau \leftarrow \text{Clip}(\tau + \mathcal{N}(0, \sigma), \mathcal{B})$ {{Inject boundary-aware exploration noise}}
  Execute τ and append recorded transitions to $\mathcal{D}_{play}$
  if human intervention triggered (e.g., $\Delta t \geq 30$ mins) **then**
    Shuffle or swap objects in the workspace
  end if
end while
return $\mathcal{D}_{play}$

implicit *visual goal*. Let $\phi : \mathcal{O} \to \mathbb{R}^{L \times d}$ denote a visual encoder that maps an observation to a sequence of L tokens of dimension d.

We optimize the model parameters θ by minimizing the mean squared error between predicted and ground-truth actions:

$$\hat{a}_t \leftarrow f_\theta(\phi(o_t), \phi(o_{t+1})) \tag{2}$$

$$\mathcal{L}_{\text{ID}}(\theta) = \mathbb{E}_{(o_t, a_t, o_{t+1}) \sim \mathcal{D}_{\text{TAP}}} \left[\|\hat{a}_t - a_t\|_2^2 \right] \tag{3}$$

Upon convergence, the model has acquired robust physical priors—spatial reasoning, affordance detection, and motor coordination—purely from task-agnostic interactions, without any language supervision.

3.3 Stage 2: Task-Specific Alignment

Once the model understands "how to move," we align it with "what to do" using a minimal set of expert demonstrations $\mathcal{D}_{\text{expert}} = \{(o_t, l, a_t)\}_{t=1}^{N_{\text{expert}}}$, where each sample includes a language instruction $l \in \mathcal{L}$.

Input Representation. The input structure shifts from visual-goal conditioning to language-instruction conditioning. Let $\psi : \mathcal{L} \to \mathbb{R}^{M \times d}$ denote a text encoder that maps a language instruction to a sequence of M tokens. The model now receives:

$$\hat{a}_t \leftarrow f_\theta(\phi(o_t), \psi(l)) \tag{4}$$

Why Does This Work? Although the conditioning signal changes from a future observation o_{t+1} to a language instruction l, the backbone f_θ and action head are reused. The model has already learned to map visual contexts to motor outputs during the task-agnostic phase. This stage essentially learns a lightweight projection from semantic space to the pre-established dynamics space—requiring significantly fewer labeled samples than training from scratch.

Training Objective. The model is finetuned via standard behavior cloning:

$$\mathcal{L}_{\text{BC}}(\theta) = \mathbb{E}_{(o_t, l, a_t) \sim \mathcal{D}_{\text{expert}}} \left[\left\| f_\theta(\phi(o_t), \psi(l)) - a_t \right\|_2^2 \right] \tag{5}$$

Table 1 Comparison of Training Paradigms and Data Scale. Traditional foundational models rely on massive, expensive, task-labeled expert demonstrations. In contrast, our approach leverages cheap, task-agnostic data via a self-supervised Inverse Dynamics (ID) objective.

Model	Pretraining Data	Data Scale	Objective (Labels Required)
RT-1-X [4]	Open X-Emb.	~1.0M	BC (Yes)
OpenVLA [2]	Open X-Emb.	~970k	BC (Yes)
Nora [29]	Open X-Emb.	~1.0M	BC (Yes)
Octo [3]	Open X-Emb.	~800k	Masked BC (Yes)
π_0 [5]	Multi-Emb.	Massive	BC (Yes)
TAP (Stage 1)	Task-Agnostic (Irrelevant Bridge / Self-Exploration)	**20k** (Sim) / **30h** (Real)	**Inverse Dyn. (No)**
TAP (Stage 2)	Task-Specific Expert Data	**5k** (Sim) / **0.2k** (Real)	BC (Yes)

4 Experiments

We design our experiments to verify whether self-supervised physical priors can effectively bypass the expert data bottleneck. We structure our analysis around three hypothesis-driven questions:

RQ1 (Effectiveness & Efficiency): Can task-agnostic interaction data, combined with inverse dynamics pretraining, match or exceed the performance of models trained on massive expert datasets while using significantly less labeled data?

RQ2 (Mechanism): Does task-agnostic pretraining improve low-level physical affordances (e.g., grasping, contact), as evidenced by sub-goal success rates and learned visual representations?

RQ3 (Robustness): Does pretraining on diverse, autonomous exploration data improve resilience to real-world distribution shifts, including visual perturbations and environmental clutter?

4.1 Experimental Setup

To rigorously evaluate our Decomposition Hypothesis, we benchmark TAP across both simulated (SIMPLER [13]) and real-world (WidowX 250) environments.

Model & Baselines. We instantiate our framework using a Qwen2.5-VL (3B) backbone coupled with a SigLIP visual encoder. As detailed in Table 1, we compare TAP against two distinct categories of baselines. It is crucial to clarify the intended role of each:

(1) **Standard BC**: An identical architecture trained from scratch purely on limited expert data. *This serves as our primary experimental comparison* to rigorously isolate and prove the value of task-agnostic pretraining.

(2) **Large-scale Pretrained VLAs**: SOTA foundational models (OpenVLA, NORA, Octo, and π_0) pretrained on the massive Open X-Embodiment (OXE)[4] dataset. OXE comprises over 1 million expert-teleoperated, language-annotated trajectories—representing an enormous investment of human labor and annotation cost. By contrast, TAP relies merely on 30 hours of autonomous random play requiring minimal human effort. Therefore, we include these VLAs not as direct competitors, but as *approximate upper-bound references* to demonstrate what internet-scale, high-quality expert pretraining can achieve.

Action Representation. We adopt a **delta-pose end-effector action space**, where $a_t \in \mathbb{R}^7$ encodes the relative position change $(\Delta x, \Delta y, \Delta z)$, orientation change (represented as a 3D axis-angle vector), and a scalar gripper command. Predicting relative motion rather than absolute poses enables the model to learn local interaction dynamics that are invariant to global workspace coordinates—a property critical for transferring physical priors across different robot configurations.

Evaluation Protocols. In simulation, we evaluate four distinct manipulation tasks, reporting success rates averaged over 50 episodes per checkpoint. In the real world, we conduct over 600 physical trials across five testing conditions (spanning from in-domain setups to severe out-of-distribution scenarios) to probe the boundaries of model robustness. Detailed implementation specifics are provided in Appendix B.4.

Table 2 Success Rates on the SIMPLER Benchmark (WidowX Environment). We report task-specific success rates for intermediate sub-goals (*Part.*) and full task completion (*Ent.*). Summary metrics include **Avg-Partial** (mean success rate of object grasping), **Avg-Entire** (mean full completion), and **Avg-All** (aggregate mean). The top two performances are in bold. **Fine-tuning Fairness & Context:** To ensure a rigorous evaluation of task-specific adaptation, OpenVLA, NORA, and π_0 were fine-tuned using the *exact same subset* of Stage 2 expert data as our TAP model (i.e., 5k trajectories for simulation). Conversely, the results for RT-1-X and Octo were directly cited from the original SIMPLER benchmark paper [13] to provide a broader context. Overall, our task-agnostic pretraining significantly boosts physical grounding, frequently matching or exceeding foundational models trained on 1M+ trajectories.

Type	Model Name	Spoon on cloth		Carrot on plate		Stack Blocks		Eggplant in Basket		Avg-Partial	Avg-Entire	Avg-All
		Part.	Ent.	Part.	Ent.	Part.	Ent.	Part.	Ent.			
Reference	RT-1-X [4]	4.2%	0.0%	16.7%	0.0%	0.0%	0.0%	3.3%	0.0%	6.05%	0.00%	3.03%
	OpenVLA [2]	4.1%	0.0%	33.0%	0.0%	12.5%	0.0%	8.3%	4.1%	14.48%	1.03%	7.75%
	Nora [29]	37.5%	16.7%	48.0%	0.0%	41.7%	12.5%	4.17%	0.0%	32.84%	7.29%	20.06%
	Octo [3]	**50.0%**	33.0%	**50.0%**	**25.0%**	29.2%	0.0%	**40.0%**	**23.3%**	42.30%	20.33%	31.31%
	π_0 [5]	45.8%	29.1%	25.0%	0.0%	50.0%	**16.7%**	**91.6%**	**62.5%**	**53.10%**	**27.05%**	**40.08%**
Baseline	Standard BC	41.7%	33.3%	48.0%	8.0%	37.5%	**16.7%**	0.0%	0.0%	31.79%	14.50%	23.15%
Ours	**TAP-8k episodes**	**50.0%**	**37.5%**	37.5%	8.3%	**58.3%**	4.2%	0.0%	0.0%	36.45%	12.50%	24.47%
	TAP-14k episodes	41.7%	33.3%	**50.0%**	**16.7%**	**83.3%**	12.5%	4.2%	0.0%	44.80%	15.62%	30.21%
	TAP-20k episodes	**66.7%**	**58.3%**	**50.0%**	0.0%	**58.3%**	**16.7%**	8.3%	8.3%	**45.82%**	**20.82%**	**33.32%**

4.2 Simulation Results: Effectiveness and Physical Grounding

To address **RQ1** and **RQ2**, Table 2 decomposes performance into *Partial* success (successful grasping) and *Entire* success (full task completion, including precise placement) across four manipulation tasks. This two-level decomposition allows us to disentangle low-level physical competence from high-level semantic execution, and thereby attribute observed gains to specific stages of our framework.

RQ1: Effectiveness and Efficiency. Our TAP-20k model achieves an Avg-All success rate of 33.32%, significantly outperforming massive foundational models like OpenVLA (7.75%) and RT-1-X (3.03%). Notably, these large-scale models frequently suffer from a 0% Entire success rate on complex tasks like *Spoon on cloth* or *Eggplant in Basket*, indicating severe cross-embodiment degradation when fine-tuned on limited data. When compared against the Standard BC baseline (23.15%), which shares an identical architecture and Stage 2 dataset, our pretraining yields a +10% absolute gain in Avg-All performance. Furthermore, Table 2 reveals a monotonic improvement across TAP's pretraining trajectory (from 8k $\rightarrow$ 14k $\rightarrow$ 20k episodes; 24.47% $\rightarrow$ 30.2% $\rightarrow$ 33.32%), rigorously confirming that deeper task-agnostic physical exposure directly translates into higher downstream task proficiency, and suggesting that the scaling law of TAP has not yet saturated.

RQ2: Mechanism via Partial Success Analysis. Analyzing the Partial versus Entire metrics illuminates *how* TAP drives these improvements and overcomes the fundamental behavioral bottleneck. Our model registers an Avg-Partial success of 45.82%, mirroring Octo (42.30%) and approaching π_0 (53.10%). Partial success heavily depends on low-level physical competencies—end-effector alignment, precision reaching, and stable grasping—which are entirely independent of high-level task semantics. The fact that TAP matches these foundational models specifically on the *physical* sub-metric, while using neither language supervision nor task-specific data in Stage 1, provides direct empirical evidence that physical competence can be acquired in isolation from semantic grounding.

This decomposition strongly validates our core hypothesis: Stage 1 pretraining successfully teaches the model **"how to move"** by coercing the visual encoder to infer actions from unlabeled observation pairs ($o_t \rightarrow o_{t+1}$). The model internalizes fine-grained motor control before ever encountering a language instruction. In standard pipelines, if the policy fails at the initial physical affordance bottleneck (e.g., dropping the object), semantic execution becomes impossible, as no amount of language conditioning can rescue a failed grasp. TAP explicitly resolves this bottleneck: by establishing a deep physical grounding during Stage 1, the model's representational capacity is entirely freed to focus on semantic goals (e.g., navigating to the specific target plate) during Stage 2 finetuning, seamlessly converting high Partial success into superior Entire success.

Table 3 Real-World Evaluation Success Rates (%). Models are trained on only 200 expert demonstrations. Our TAP model is pretrained on 30 hours of autonomous self-exploration. **Bold** indicates the best performance. TAP demonstrates remarkable resilience, surpassing the internet-scale NORA baseline in dynamic tasks with clutter ("Visual Distractors") and consistently outperforming all baselines under severe visual perturbations ("Background Texture Shift" and "Viewpoint Variation").

Evaluation Scenario	Task: Put the carrot on the plate			Task: Push the pumpkin to the left		
	From scratch	**TAP (Ours)**	NORA (SOTA)	From scratch	**TAP (Ours)**	NORA (SOTA)
Standard Setup	20%	40%	**65%**	55%	75%	**85%**
Initial State Perturbation	20%	30%	**65%**	45%	75%	**80%**
Visual Distractors	5%	30%	**40%**	5%	**65%**	60%
Background Texture Shift	0%	**25%**	10%	0%	**65%**	55%
Viewpoint Variation	0%	**15%**	0%	0%	**25%**	0%
Average	9%	28%	**36%**	21%	**61%**	56%

4.3 Real-World Experiments: Robustness via Self-Exploration

To answer **RQ3**, we evaluate our method on a physical WidowX 250s robot. As shown in Figure 2, we select two complementary tasks that probe distinct aspects of physical understanding:

- **Put the carrot on the plate** tests precision grasping in a classic pick-and-place scenario. Success requires the robot to grasp the carrot and release it stably onto the plate, evaluating *geometric alignment* and *grasping affordance*.
- **Push the pumpkin to the left** tests dynamic, non-prehensile manipulation. We define a 30×30 cm^2 arena with the pumpkin initialized at center (±1cm). A trial succeeds only if the robot pushes the object such that more than 50% of it exits the boundary in the specified direction. Unlike grasping, pushing demands *sustained contact* and an implicit understanding of *object dynamics*—the spherical pumpkin will spin and deviate if force is not applied precisely through its center.

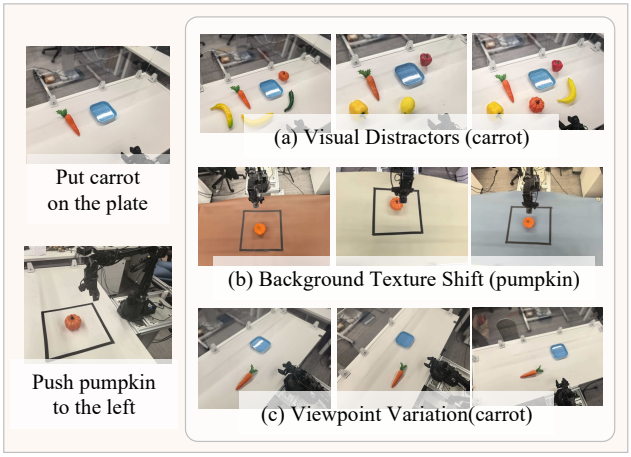

Figure 2 Real-World Evaluation Setup and Robustness Protocols. We evaluate our method on a physical WidowX 250 robot across two manipulation tasks: *Put carrot on plate* and *Push pumpkin*. To rigorously quantify resilience to distribution shifts (RQ3), we introduce systematic environmental perturbations: **(a) Visual Distractors** introduce unseen clutter (e.g., diverse fruits) to test semantic attention mechanisms; **(b) Background Texture Shifts** alter surface materials to evaluate visual invariance; and **(c) Viewpoint Variations** perturb camera extrinsics to assess geometric generalization. Combined with **Initial State Perturbations**, these scenarios probe the boundaries of model robustness beyond the training distribution.

To rigorously assess out-of-distribution robustness, we establish a tiered evaluation protocol comprising four distinct categories of environmental variation: *Initial State Perturbations*, *Visual Distractors*, *Background Texture Shifts*, and *Viewpoint Variations*. These perturbations are designed to simulate real-world unpredictability and verify whether the policy relies on robust physical priors rather than spurious visual correlations. Detailed configurations for each scenario are provided in Appendix B.4.

Data Collection & Baselines. To simulate severe data scarcity, we strictly limit human supervision to only 200 expert trajectories per task. For Stage 1, the robot autonomously collects 30 hours of task-agnostic random play data (Algorithm 1). We compare TAP against the **Standard BC** (trained from scratch) and **NORA** [29] (finetuned from massive OXE pretraining).

RQ3: Robustness from Self-Exploration. The results in Table 3 highlight two critical advantages of our framework, dissecting how physical priors combat real-world unpredictability:

1) Overcoming Spurious Correlations in Clutter. While NORA dominates in clean, standard setups (e.g., 85% on pushing), TAP exhibits vastly superior adaptability in chaotic environments. In the cluttered pushing task equipped with unseen fruits, standard BC drops to a near-random 5%, and NORA decays to 60%. In contrast, TAP maintains a 65% success rate. This indicates that without localized physical pretraining, policies easily overfit to spurious visual correlations in the background. TAP's self-exploration phase forces the model to attend to *causal interactive dynamics* (the relationship between the gripper and the manipulated object), rendering static visual distractors semantically invisible.

2) Resilience to Severe Spatial and Textural Shifts. The divergence is most profound under structural perturbations. When camera extrinsics are shifted significantly, NORA and the Standard BC suffer catastrophic spatial misalignment, frequently grasping at empty space (0% success on both tasks). Conversely, TAP retains robust functionality (15% and 25% success). Background texture shifts follow a similar pattern: replacing the wooden table with a colored cloth causes NORA's pushing performance to plummet to 55%, while TAP remains highly robust at 65%.

Overall Performance & The Value of TAP. Averaged across all demanding scenarios, TAP rivals or exceeds the overall success rate of the NORA baseline (e.g., 61% vs. 56% in the pushing task). Against this approximate upper-bound reference—which required a massive investment of over a million human-annotated trajectories—the fact that TAP achieves comparable overall proficiency and strictly superior robustness in out-of-distribution environments, utilizing merely 30 hours of autonomous random play, definitively proves the immense potential and scalability of task-agnostic physical pretraining.

4.4 Ablation Study: How Task-Agnostic Data Shapes Learning

Having established the effectiveness of our approach, we now investigate the *mechanisms* through which task-agnostic data improves downstream performance. We analyze three complementary aspects: convergence dynamics, data scaling behavior, and learned visual representations.

Overcoming Early Saturation. A fundamental limitation of standard Behavior Cloning is early saturation—models trained on limited expert data plateau quickly, unable to generalize beyond memorized trajectories. Figure 3 tracks validation success rates throughout Stage 2 finetuning.

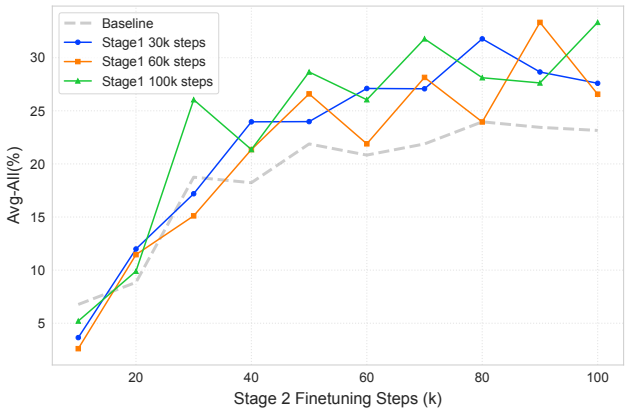

Figure 3 Convergence Dynamics. Avg-All success rates during Stage 2 finetuning. Initial learning rates are comparable across methods, but the Baseline (dashed) plateaus early while pretrained models (solid) achieve substantially higher final performance—demonstrating that task-agnostic pretraining expands learning capacity rather than accelerating convergence.

Notably, pretrained models (solid lines) and the Baseline (dashed) exhibit similar initial learning rates, indicating that task semantics are acquired at comparable speeds. The critical divergence occurs later: while the Baseline saturates around 23% and oscillates, pretrained models continue climbing to exceed 30%. This pattern suggests that task-agnostic pretraining does not speed up the acquisition of task semantics—instead, it raises the upper bound on achievable performance. The physical priors acquired in Stage 1 prevent convergence to poor local optima, enabling the model to extract more value from identical expert data. In other words, Stage 1 reshapes the loss landscape rather than the optimization trajectory, providing a structural advantage that compounds rather than competes with task-specific finetuning.

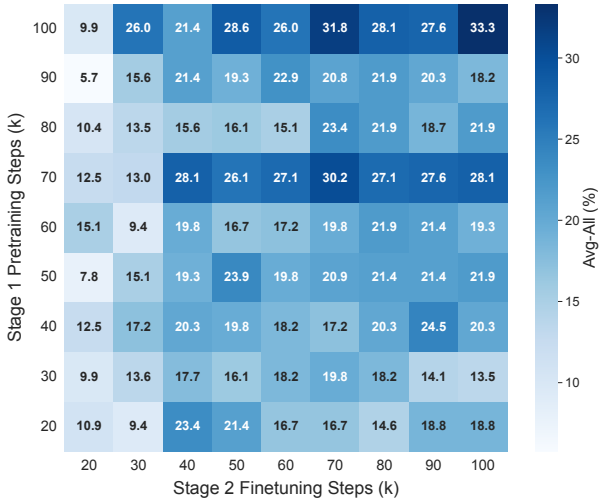

Figure 4 Data Scaling Analysis for Overall Success (Avg-All). Heatmap of Avg-All success rates on the SIMPLER benchmark across a joint sweep of Stage 1 (task-agnostic pretraining) and Stage 2 (task-specific finetuning) durations. The upward gradient along the pretraining axis dominates the horizontal gradient along the finetuning axis, revealing that the scale of task-agnostic exposure sets the achievable performance ceiling.

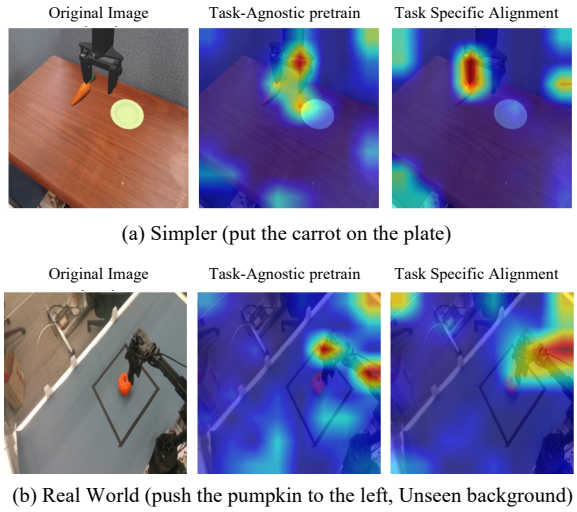

Figure 5 Attention Map Analysis. Comparison across simulation (top) and real-world deployment (bottom). *Middle:* Task-agnostic pretraining yields attention focused on manipulable entities (gripper, objects) without any language input. *Right:* Instruction tuning refines attention toward robotic agents (grippers). The consistency across domains confirms that learned physical priors transfer robustly to novel environments.

Data Scaling: A Necessary Foundation. We next examine how volumes of task-agnostic pretraining (Stage 1) and task-specific finetuning (Stage 2) jointly determine performance. Figure 4 presents a systematic sweep across both axes.

The heatmap reveals a strict dependency: *downstream performance is bounded by pretraining scale*. With minimal Stage 1 exposure (20k steps), extending Stage 2 yields diminishing returns—performance stagnates near 18% regardless of finetuning duration. This suggests that without sufficient diversity in task-agnostic interactions, the model lacks the generalizable physical representations required for robust manipulation. Conversely, scaling Stage 1 to 100k steps unlocks success rates exceeding 30%. The optimal regions (dark blue) confirm that abundant task-agnostic data acts as a regularizer, preventing overfitting to the limited expert trajectories available during finetuning.

Visualizing Learned Physical Priors. To verify that task-agnostic data instills meaningful physical understanding, we visualize Grad-CAM [30] attention maps from the model's final layer (Figure 5). We compare attention patterns after Stage 1 (no language input) and after Stage 2 (with task instructions) across both simulation and real-world settings.

Stage 1: Emergent Physical Saliency. Without any text prompt, the pretrained model's attention (middle column) automatically concentrates on the robot gripper and nearby objects—the carrot in simulation, the pumpkin in the real world. Background elements (wood texture, floor) are suppressed. This behavior emerges directly from the inverse dynamics objective: predicting actions from observation pairs (o_t, o_{t+1}) forces the encoder to track end-effector kinematics and object interactions. The result is an implicit *affordance map* that identifies manipulable entities without task specification.

Stage 2: Semantic Grounding and Execution Focus. Upon receiving a language instruction (right column), we observe a distinct shift in attention dynamics: the heatmap becomes **intensely concentrated on the robotic gripper**. Unlike Stage 1, where attention is distributed across multiple potential affordances, the language prompt acts as a strictly constraining filter. It effectively "prunes" away irrelevant physical possibilities (distractors), forcing the model to lock its visual processing resources onto the *agent of action* (the gripper) to

ensure precise motor execution. This confirms that while Stage 1 builds a broad space of physical possibilities, Stage 2 collapses this space into a singular, focused point of execution.

Cross-Domain Transfer. The bottom row demonstrates robustness under domain shift. Despite deployment in a real-world environment with novel backgrounds and lighting, the pretrained attention maps maintain consistent focus on the gripper and interactive objects. This suggests that the learned physical representations capture domain-invariant structure rather than overfitting to simulation textures.

4.5 Error Analysis

To provide a comprehensive understanding of our TAP method, we systematically analyzed the failure cases encountered during our real-world WidowX experiments. Grounded in our Decomposition Hypothesis, we categorize the failures into two primary modes: Execution Failures (deficits in "how to move") and Semantic Failures (deficits in "what to do").

Execution and Dynamics Failures (Approx. 25% of failures): These errors occur when the policy correctly identifies the target object and attempts the right sub-task, but fails during fine-grained physical contact. Common manifestations include the end-effector slipping off the object, millimetric pre-grasp misalignment, or depth ambiguity caused by singular camera viewpoints. In the task 'push the pumpkin to the left', failures may also occur due to While our task-agnostic pretraining significantly mitigates these issues compared to standard BC, purely reactive VLA models still struggle with complex, out-of-distribution 3D spatial reasoning under extreme visual shifts.

Semantic and Reasoning Failures (Approx. 75% of failures): These errors are characterized by flawless physical execution directed at the wrong semantic goal. For instance, in the presence of visual distractors, the robot might execute a perfectly smooth grasp on a distractor object rather than the target instruction. Alternatively, in longer horizon sequences, the model occasionally experiences "freezing" or repetitive looping, losing track of the overarching linguistic instruction. This suggests that while the lower-level execution capabilities are robust, the implicit reasoning capacity of a singular, reactive VLA model remains a bottleneck.

5 Conclusion

We introduced **Task-Agnostic Pretraining (TAP)**, a two-stage framework that decouples the learning of physical affordances from semantic task understanding in Vision-Language-Action models. Our key insight—the *Decomposition Hypothesis*—posits that "how to move" can be learned entirely from cheap, unlabeled interaction data, reserving expensive expert demonstrations for teaching "what to do."

Our experiments yield three principal findings. First, **task-agnostic data is surprisingly effective**: by pretraining on off-task trajectories or autonomous random play via an Inverse Dynamics objective, TAP achieves a 10% absolute gain over standard Behavior Cloning and matches models trained on 1M+ expert trajectories —using orders of magnitude less labeled data. Second, **physical grounding transfers across tasks**: the boost in partial success from 31.8% to 45.8% confirms that self-supervised pretraining instills generalizable motor competencies rather than task-specific behaviors. Third, **self-exploration breeds robustness**: in real-world experiments, TAP retains 15–25% success under camera perturbations that cause catastrophic failure (0%) in internet-scale baselines, demonstrating that diverse physical experience yields domain-invariant representations.

These results challenge the prevailing assumption that scaling expert data is the only path to capable embodied agents. Instead, we show that active, task-agnostic interaction, similar to infant-like motor babbling, provides a complementary and cost-effective foundation for robot learning.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62521004).

Additionally, we would like to express our sincere gratitude to Chunbiao Feng and Hongbo Tang for their invaluable assistance with the hardware setup and control implementations.

References

[1] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054, 2025.

[2] 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, editors, Conference on Robot Learning, 6-9 November 2024, Munich, Germany, volume 270 of Proceedings of Machine Learning Research, pages 2679–2713. PMLR, 2024. URL https://proceedings.mlr.press/v270/kim25c.html.

[3] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. CoRR, abs/2405.12213, 2024. doi: 10.48550/ARXIV.2405.12213. URL https://doi.org/10.48550/arXiv.2405.12213.

[4] Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar,

Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

[5] 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.

[6] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In IEEE International Conference on Robotics and Automation (ICRA), pages 6892–6903, 2024.

[7] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In Conference on Robot Learning (CoRL), 2023.

[8] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, P Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Ye Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sung Yul Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean-Pierre Mercat, Abdul Rehman, Pannag R. Sanketi, Archit Sharma, C. Blake Simpson, Quang Uyen Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Da Ling Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosa Maria Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J. Lim, Jitendra Malik, Roberto Mart'in-Mart'in, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. ArXiv, abs/2403.12945, 2024. URL https://api.semanticscholar.org/CorpusID:268531351.

[9] Justine E Hoch, Sinclaire M O'Grady, and Karen E Adolph. It's the journey, not the destination: Locomotor exploration in infants. Developmental science, 22(2):e12740, 2019.

[10] Celeste Kidd and Benjamin Y Hayden. The psychology and neuroscience of curiosity. Neuron, 88(3):449–460, 2015.

[11] Karen E Adolph and Justine E Hoch. Motor development: Embodied, embedded, enculturated, and enabling. Annual review of psychology, 70(1):141–164, 2019.

[12] David Brandfonbrener, Ofir Nachum, and Joan Bruna. Inverse dynamics pretraining learns good representations for multitask imitation. ArXiv, abs/2305.16985, 2023. URL https://api.semanticscholar.org/CorpusID:258947266.

[13] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, Conference on Robot Learning, 6-9 November 2024, Munich, Germany, volume 270 of Proceedings of Machine Learning Research, pages 3705–3728. PMLR, 2024. URL https://proceedings.mlr.press/v270/li25c.html.

[14] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai, 2026. URL https://arxiv.org/abs/2405.14093.

[15] Dapeng Zhang, Jing Sun, Chenghui Hu, Xiaoyan Wu, Zhenlong Yuan, Rui Zhou, Fei Shen, and Qingguo Zhou. Pure vision language action (vla) models: A comprehensive survey, 2025. URL https://arxiv.org/abs/2509.19012.

[16] Yifan Zhong, Fengshuo Bai, Shaofei Cai, Xuchuan Huang, Zhang Chen, Xiaowei Zhang, Yuanfei Wang, Shaoyang Guo, Tianrui Guan, Ka Nam Lui, Zhiquan Qi, Yitao Liang, Yuanpei Chen, and Yaodong Yang. A survey on vision-language-action models: An action tokenization perspective, 2025. URL https://arxiv.org/abs/2507.01925.

[17] 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, editors, 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.

[18] 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 T. 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, editors, Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA, volume 229 of Proceedings of Machine Learning Research, pages 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

[19] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R. Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J. Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A large-scale in-the-wild robot manipulation dataset. In Dana Kulic, Gentiane Venture, Kostas E. Bekris, and Enrique Coronado, editors, Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024, 2024. doi: 10.15607/RSS.2024.XX.120. URL https://doi.org/10.15607/RSS.2024.XX.120.

[20] Generalist AI Team. Gen-0: Embodied foundation models that scale with physical interaction. Generalist AI Blog, 2025. https://generalistai.com/blog/preview-uqlxvb-bb.html.

[21] Siyin Wang, Jinlan Fu, Feihong Liu, Xinzhe He, Huangxuan Wu, Junhao Shi, Kexin Huang, Zhaoye Fei, Jingjing Gong, Zuxuan Wu, Yu-Gang Jiang, See-Kiong Ng, Tat-Seng Chua, and Xipeng Qiu. Roboomni: Proactive robot manipulation in omni-modal context, 2025. URL https://arxiv.org/abs/2510.23763.

[22] Kieran Rendall, Alexios Mylonas, Stilianos Vidalis, and Dimitris Gritzalis. MIDAS: multi-layered attack detection architecture with decision optimisation. Comput. Secur., 148:104154, 2025. doi: 10.1016/J.COSE.2024.104154. URL https://doi.org/10.1016/j.cose.2024.104154.

[23] Yanchao Sun, Shuang Ma, Ratnesh Madaan, Rogerio Bonatti, Furong Huang, and Ashish Kapoor. SMART: self-supervised multi-task pretraining with control transformers. In The Eleventh International Conference on Learning

Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. URL https://openreview.net/forum?id=9piH3Hg8QEf.

[24] Rogerio Bonatti, Sai Vemprala, Shuang Ma, Felipe Frujeri, Shuhang Chen, and Ashish Kapoor. PACT: perception-action causal transformer for autoregressive robotics pre-training. In IROS, pages 3621–3627, 2023. doi: 10.1109/IROS55552.2023.10342381. URL https://doi.org/10.1109/IROS55552.2023.10342381.

[25] Ya Jing, Xuelin Zhu, Xingbin Liu, Qie Sima, Taozheng Yang, Yunhai Feng, and Tao Kong. Exploring visual pre-training for robot manipulation: Datasets, models and methods. In IROS, pages 11390–11395, 2023. doi: 10.1109/IROS55552.2023.10342201. URL https://doi.org/10.1109/IROS55552.2023.10342201.

[26] Fangchen Liu, Hao Liu, Aditya Grover, and Pieter Abbeel. Masked autoencoding for scalable and generalizable decision making. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/51fda94414996902ddaaa35561b97294-Abstract-Conference.html.

[27] Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (VPT): learning to act by watching unlabeled online videos. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9c7008aff45b5d8f0973b23e1a22ada0-Abstract-Conference.html.

[28] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. 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=NxoFmGgWC9.

[29] Chia-Yu Hung, Qi Sun, Pengfei Hong, Amir Zadeh, Chuan Li, U-Xuan Tan, Navonil Majumder, and Soujanya Poria. NORA: A small open-sourced generalist vision language action model for embodied tasks. CoRR, abs/2504.19854, 2025. doi: 10.48550/ARXIV.2504.19854. URL https://doi.org/10.48550/arXiv.2504.19854.

[30] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pages 618–626, 2017.

[31] Junhao Shi, Zhaoye Fei, Siyin Wang, Qipeng Guo, Jingjing Gong, and Xipeng Qiu. World-aware planning narratives enhance large vision-language model planner, 2025. URL https://arxiv.org/abs/2506.21230.

[32] Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. Libero-plus: In-depth robustness analysis of vision-language-action models, 2025. URL https://arxiv.org/abs/2510.13626.

Appendix

A Details of Autonomous Random Play Data Collection

To ensure that autonomous exploration yields safe, contact-rich physical interactions rather than redundant free-space motions, we implement a two-phase constrained procedural generation framework.

Phase 1: Safe Workspace Initialization. The first step for autonomous play is to establish a sufficient safe spatial prior. An operator first teleoperates the robot without specific task instructions to perform regular movements and densely cover the reachable workspace. After filtering out kinematically unsafe or out-of-bound poses, we apply Voxel Grid Downsampling (e.g., with a leaf size of 5cm^3) to the retained poses. This mitigates spatial density bias and yields a uniform, safe, and discrete pose library $\mathcal{P}_{safe}$.

Phase 2: Constrained Trajectory Generation and Execution. To procedurally generate trajectories, we stochastically sample sequences of waypoints from $\mathcal{P}_{safe}$ under a minimum distance constraint. To guarantee contact-rich interactions (e.g., pushing, sliding) and prevent the end-effector from hovering, we apply a contact-forcing heuristic: any generated sequence that remains above a specified elevation threshold (z_{thresh}) for more than c_{max} consecutive steps is geometrically adjusted to force a descent, ensuring frequent engagement with the tabletop.

The modified waypoints are then connected via cosine interpolation, and boundary-aware Gaussian noise is injected to enhance trajectory diversity. During execution, the robot continuously performs these trajectories, with human intervention strictly limited to periodic resets (e.g., every 30 minutes) to add, remove or shuffle interactive objects.

During data collection, raw trajectories are typically collected at high control frequencies (e.g., 25Hz), yielding minimal visual displacement between adjacent frames. At such rates, the inverse dynamics task becomes ill-posed: the action signal is dominated by sensor noise rather than meaningful motion. To address this, we downsample all training data to **5Hz**, ensuring that (1) the visual change between o_t and o_{t+1} is perceptible and causally attributable to the executed action, and (2) the model learns semantically meaningful action primitives (e.g., "approach," "grasp," "lift") rather than micro-adjustments.

B Training and Evaluation Details

B.1 Model Architecture

We instantiate our framework using **Qwen2.5-VL** (3B parameters) as the VLM backbone. The visual encoder is a ViT-based **SigLIP** (400M parameters), which processes input images at 224×224 resolution and produces a sequence of visual tokens. The action head is a lightweight 2-layer MLP that projects the VLM's last hidden state to the 7-dimensional action space. During Stage 1, we freeze the visual encoder and train only the VLM backbone and action head; in Stage 2, all parameters are jointly finetuned.

B.2 Action Representation

We adopt a **delta-pose end-effector action space**, where $a_t \in \mathbb{R}^7$ encodes the relative position change $(\Delta x, \Delta y, \Delta z)$, orientation change (represented as a 3D axis-angle vector), and a scalar gripper command. Predicting relative motion rather than absolute poses enables the model to learn local interaction dynamics that are invariant to global workspace coordinates—a property critical for transferring physical priors across different robot configurations.

B.3 Training Details.

We train our model for $100,000$ steps on a single node equipped with 8 NVIDIA H100 GPUs. The training is implemented using the Hugging Face Accelerate library to ensure efficient distributed execution. We use

a global batch size of 128 (16 per GPU). The model is optimized using the AdamW optimizer with a weight decay of 0.05 and standard β parameters ($\beta_1 = 0.9, \beta_2 = 0.999$). The learning rate is initialized at 5×10^{-5} and follows a cosine decay schedule with a warmup ratio of 0.05 (warming up for the first $5,000$ steps). To stabilize training, we apply global gradient clipping with a max norm of 1.0. For computational efficiency and numerical stability, all training runs are conducted in $\mathrm{bfloat16}$ precision.

B.4 Evaluation Details

Simpler Evaluation For Simpler Evaluation, we follow the standard SIMPLER protocol, reporting success rates averaged over 50 episodes per checkpoint. All comparing models are trained with the same amount of Bridge data with same training settings.

Real World Evaluation Details. To further validate the effectiveness and robustness of the **TAP** framework, we conduct real-world experiments on two manipulation tasks: *"Put Carrot on Plate"* and *"Push Pumpkin to Left"*. Inspired by the generalization aspects introduced by previous works[31, 32], we evaluate the policy across five distinct scenarios designed to test generalization: In-Domain setups, Initial State Perturbations, Visual Distractors, Background Texture Shifts, and Viewpoint Variations.

For each task, we conduct 20 evaluation trials per scenario. The detailed protocols for each scenario are listed as follows:

- **Standard Setup (In-Domain):** The experimental environment strictly replicates the training setting, ensuring no variations in object positions, lighting, or background. This serves as the baseline for assessing the upper bound of model performance.
- **Initial State Perturbation:** To evaluate the model's robustness to initial conditions, we apply random spatial perturbations to the robot's home pose (starting position). This tests the policy's ability to recover and complete the task from unseen starting configurations.
- **Visual Distractors:** We introduce unseen objects to test robustness against visual clutter. For this setting, we create five distinct combinations, each containing three to five random distractor objects placed on the table. Crucially, these distractors are positioned without obstructing the manipulation trajectory to ensure the task remains physically feasible. Each combination is tested over 4 trials (totaling 20 trials), and the average success rate is reported.
- **Background Texture Shift:** We evaluate the model's invariance to background changes by placing four different tablecloths (layers) with varying textures on the table, while maintaining the original object arrangements. Each background texture is evaluated over 5 trials (totaling 20 trials).
- **Viewpoint Variation:** To assess robustness against camera calibration noise, we apply minor shifts to the extrinsic parameters (e.g., angle, pitch) of the third-person view camera. These perturbations generate four distinct camera viewpoints that differ slightly from the training view while preserving the main visual semantics. Each viewpoint configuration is tested over 5 trials (totaling 20 trials).

Finally, the overall average success rate is calculated as the arithmetic mean across all five evaluation scenarios.

C Data Efficiency and Computational Cost Analysis

A core motivation of our work is to democratize generalist robot learning by reducing the dependency on massive, curated expert datasets and prohibiting computational budgets. In this section, we provide a detailed comparison of data requirements and training costs between our proposed TAP framework and current state-of-the-art Large-Scale VLA baselines.

Table 4 Comparison of Data Scale and Computational Resources. We compare our TAP against leading VLA baselines. *Expert Data* refers to human-teleoperated or curated demonstrations. *Cheap Data* refers to autonomous, unlabeled interactions (e.g., self-play). Note that our method achieves competitive results using significantly less expert data and manageable compute resources compared to models trained on the massive Open X-Embodiment (OXE) dataset.

Model	Pretraining Dataset	Expert Data Scale	Task-Agnostic (Cheap) Data	Compute Infrastructure	Training Objective
Baselines					
RT-1-X [4]	Open X-Embodiment	~1M Trajectories	None	TPU v4 Pods	BC (Categorical)
Octo [3]	Open X-Embodiment	~800k Trajectories	None	TPU v4-128	Diffusion
OpenVLA [2]	Open X-Embodiment	~970k Trajectories	None	64 × A100	Llama-2 Finetuning
NORA [29]	Open X-Embodiment	~970k Trajectories	None	-	VLA Finetuning
Ours					
TAP	**Self-Generated Play**	**< 1k Trajectories**	**~100k Steps**	**8 × H100**	**Inverse Dynamics + BC**

C.1 Data Comparison: Expert vs. Task-Agnostic

As illustrated in Table 4, standard VLA models (e.g., OpenVLA, Octo, NORA) rely heavily on the Open X-Embodiment (OXE) dataset, which aggregates over 2 million expert trajectories across varying embodiments. While effective, curating and standardizing such datasets requires immense human effort.

In contrast, our TAP framework minimizes the reliance on expert data.

- **Stage 1 (Pretraining):** We utilize purely autonomous, task-agnostic interaction data (e.g., random exploration or play). This data is ”free” in terms of human labeling cost.
- **Stage 2 (Finetuning):** We achieve competitive performance using only a fraction of the expert demonstrations (e.g., 200 trajectories in real-world experiments) compared to the millions seen by baselines.

Quantitatively, our approach reduces the demand for expert data by several orders of magnitude while maintaining comparable manipulation proficiency.

C.2 Computational Budget

Training foundation models like OpenVLA typically necessitates industrial-scale infrastructure (e.g., TPU v4 Pods or clusters of A100s) running for weeks. Our method is designed for academic-scale resources. As detailed in Table 4, our pretraining and finetuning can be completed on a single node with 8×H100 GPUs within a reasonable timeframe (approx. 24 GPU hours), making the reproduction and iteration of VLA policies significantly more accessible.

D Qualitative Analysis and Case Studies

To better understand the mechanisms driving the quantitative improvements, we conduct a qualitative analysis comparing our TAP framework against Standard BC and Nora baselines across both simulation and real-world environments, as visualized in Figure 6.

D.1 Simulation: Unlocking Task-Irrelevant Data in SIMPLER

A core claim of our Decomposition Hypothesis is that the physical affordances required for manipulation can be extracted from task-agnostic data. Figure 6(a) illustrates a comparison on the SIMPLER “Put the carrot on the plate” task.

When expert data is scarce, the Standard BC model struggles to ground the linguistic instruction in precise 3D geometry. As shown in the top row, the BC policy navigates to the general vicinity of the carrot but halts, failing to execute the fine-grained contact dynamics required for a successful grasp. Both the Nora baseline and our TAP demonstrate robust physical execution, successfully grasping the carrot and placing it on the plate. For TAP, this confirms that “how to move” (precise pre-grasp alignment and contact) transfers effectively even when the model is pretrained on discarded, task-irrelevant trajectories via Inverse Dynamics.

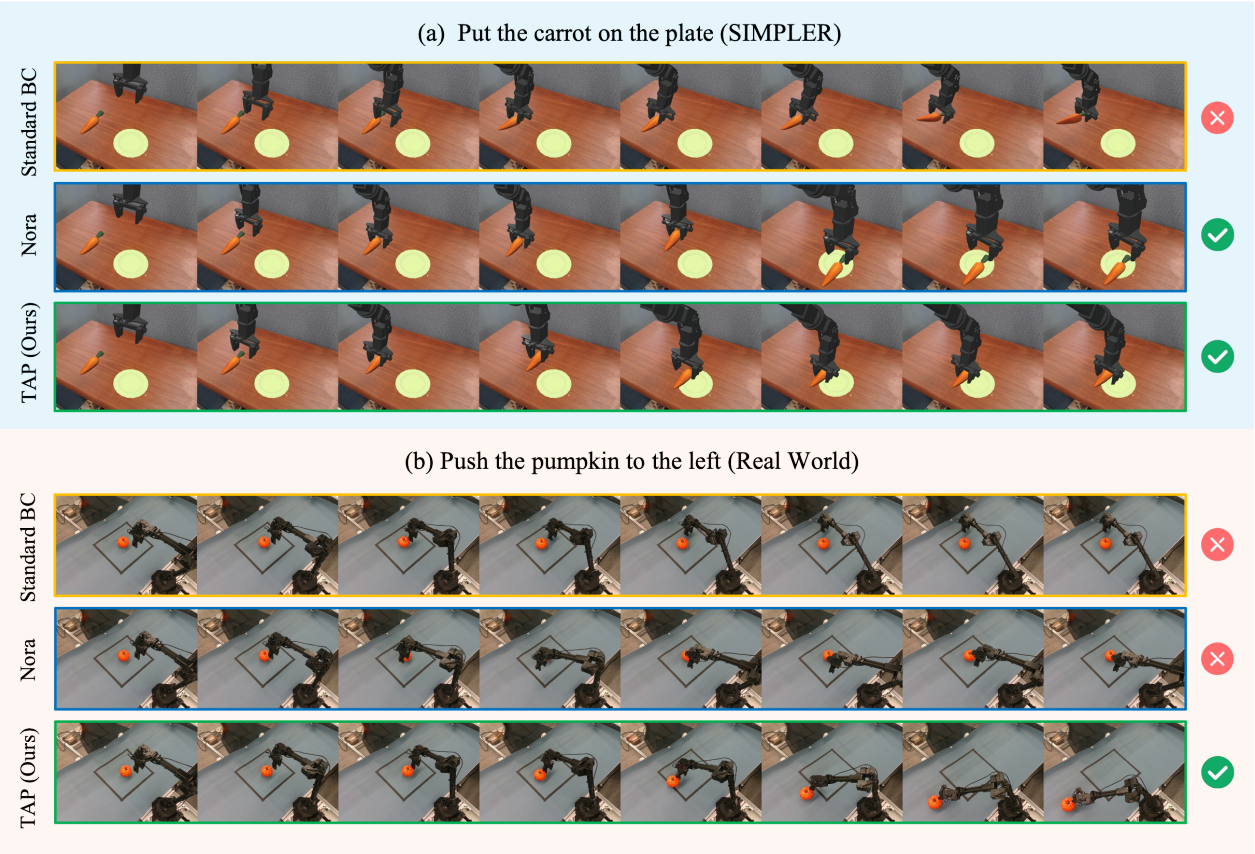

Figure 6 Qualitative Comparison in Simulation and Real-World Environments. (a) On the SIMPLER "Put the carrot on the plate" task, Standard BC fails to execute the final grasp, while Nora and TAP succeed. (b) On the real-world "Push the pumpkin to the left" task under an unseen background, Standard BC exhibits poor contact. The Nora baseline suffers from a visual grounding failure; it attempts a leftward push but misjudges the pumpkin's location due to the novel texture, pushing empty space. TAP accurately isolates the object despite the background shift and successfully completes the task.

D.2 Real-World: Robustness to Unseen Background Shifts

In our real-world WidowX 250 experiments, the distinction between robust physical grounding and brittle visual matching becomes starkly apparent. Figure 6(b) visualizes the "Push the pumpkin to the left" task evaluated on an unseen background texture, a scenario designed to test generalization beyond the training distribution.

The Standard BC model struggles with basic execution, extending the arm but failing to make proper, sustained contact with the object. Interestingly, the Nora baseline exhibits a severe visual grounding failure induced by the out-of-distribution background. While it attempts the semantically correct trajectory (moving its end-effector towards the left), it misjudges the precise 3D spatial coordinates of the pumpkin against the novel table texture. Consequently, it completely misses the object, pushing empty space to the left of the pumpkin. In contrast, TAP successfully leverages its robust, task-agnostic physical priors to accurately isolate the manipulable object from the novel background. It makes solid contact and deliberately pushes the pumpkin to the correct side, demonstrating superior domain-invariant spatial awareness.

E Full Experimental Results

Due to space limitations in the main text, we present the comprehensive breakdown of our experimental results in Table 5, including performance metrics across all individual evaluation checkpoints and sub-tasks.

Table 5 Comprehensive Scaling Results (Origin Method). Detailed breakdown of success rates across all combinations of Stage 1 (pretraining) and Stage 2 (finetuning) steps. Steps are reported in thousands (k).

Training Steps (k)		Spoon on cloth		Carrot on plate		Stack Blocks		Eggplant in Basket		Avg-P	Avg-E	Avg-All
Stage 1	Stage 2	Part.	Ent.	Part.	Ent.	Part.	Ent.	Part.	Ent.			
20	10	8.3%	0.0%	12.5%	0.0%	25.0%	0.0%	0.0%	0.0%	11.45%	0.00%	5.73%
20	20	16.7%	4.2%	37.5%	0.0%	29.2%	0.0%	0.0%	0.0%	20.85%	1.05%	10.95%
20	30	16.7%	4.2%	25.0%	0.0%	29.2%	0.0%	0.0%	0.0%	17.72%	1.05%	9.39%
20	40	33.3%	16.7%	45.8%	12.5%	75.0%	4.2%	0.0%	0.0%	38.52%	8.35%	23.44%
20	50	33.3%	25.0%	33.3%	0.0%	62.5%	16.7%	0.0%	0.0%	32.27%	10.43%	21.35%
20	60	16.7%	12.5%	25.0%	0.0%	62.5%	16.7%	0.0%	0.0%	26.05%	7.30%	16.68%
20	70	20.8%	8.3%	41.7%	4.2%	50.0%	8.3%	0.0%	0.0%	28.12%	5.20%	16.66%
20	80	8.3%	0.0%	33.3%	0.0%	58.3%	12.5%	4.2%	0.0%	26.02%	3.12%	14.57%
20	90	25.0%	4.2%	29.2%	4.2%	66.7%	20.8%	0.0%	0.0%	30.23%	7.30%	18.76%
20	100	25.0%	8.3%	33.3%	4.2%	62.5%	16.7%	0.0%	0.0%	30.20%	7.30%	18.75%
30	10	0.0%	0.0%	0.0%	0.0%	20.8%	0.0%	0.0%	0.0%	5.20%	0.00%	2.60%
30	20	8.3%	0.0%	25.0%	0.0%	41.7%	4.2%	0.0%	0.0%	18.75%	1.05%	9.90%
30	30	29.2%	12.5%	25.0%	4.2%	33.3%	0.0%	4.2%	0.0%	22.93%	4.17%	13.55%
30	40	37.5%	16.7%	29.2%	0.0%	50.0%	8.3%	0.0%	0.0%	29.18%	6.25%	17.71%
30	50	29.2%	16.7%	33.3%	0.0%	45.8%	4.2%	0.0%	0.0%	27.07%	5.23%	16.15%
30	60	29.2%	16.7%	41.7%	8.3%	41.7%	0.0%	4.2%	4.2%	29.20%	7.30%	18.25%
30	70	41.7%	16.7%	37.5%	8.3%	37.5%	0.0%	8.3%	8.3%	31.25%	8.33%	19.79%
30	80	25.0%	20.8%	37.5%	4.2%	45.8%	4.2%	4.2%	4.2%	28.12%	8.35%	18.24%
30	90	25.0%	12.5%	25.0%	0.0%	41.7%	8.3%	0.0%	0.0%	22.93%	5.20%	14.06%
30	100	20.8%	8.3%	33.3%	0.0%	41.7%	4.2%	0.0%	0.0%	23.95%	3.12%	13.54%
40	10	0.0%	0.0%	0.0%	0.0%	16.7%	0.0%	0.0%	0.0%	4.17%	0.00%	2.09%
40	20	25.0%	12.5%	12.5%	0.0%	37.5%	0.0%	12.5%	0.0%	21.88%	3.12%	12.50%
40	30	16.7%	12.5%	29.2%	12.5%	58.3%	8.3%	0.0%	0.0%	26.05%	8.33%	17.19%
40	40	37.5%	29.2%	33.3%	4.2%	45.8%	4.2%	8.3%	0.0%	31.23%	9.40%	20.31%
40	50	33.3%	16.7%	37.5%	0.0%	62.5%	8.3%	0.0%	0.0%	33.32%	6.25%	19.79%
40	60	25.0%	16.7%	29.2%	4.2%	58.3%	8.3%	4.2%	0.0%	29.18%	7.30%	18.24%
40	70	33.3%	20.8%	16.7%	4.2%	50.0%	8.3%	4.2%	0.0%	26.05%	8.33%	17.19%
40	80	45.8%	29.2%	33.3%	0.0%	50.0%	4.2%	0.0%	0.0%	32.27%	8.35%	20.31%
40	90	50.0%	37.5%	37.5%	8.3%	58.3%	4.2%	0.0%	0.0%	36.45%	12.50%	24.47%
40	100	50.0%	37.5%	12.5%	0.0%	50.0%	8.3%	4.2%	0.0%	29.18%	11.45%	20.31%
50	10	0.0%	0.0%	8.3%	0.0%	25.0%	0.0%	0.0%	0.0%	8.33%	0.00%	4.16%
50	20	12.5%	4.2%	8.3%	0.0%	33.3%	4.2%	0.0%	0.0%	13.53%	2.10%	7.81%
50	30	33.3%	16.7%	29.2%	0.0%	37.5%	4.2%	0.0%	0.0%	25.00%	5.23%	15.11%
50	40	33.3%	16.7%	37.5%	4.2%	54.2%	0.0%	4.2%	4.2%	32.30%	6.28%	19.29%
50	50	37.5%	25.0%	33.3%	8.3%	45.8%	12.5%	20.8%	8.3%	34.35%	13.53%	23.94%
50	60	37.5%	20.8%	37.5%	0.0%	50.0%	4.2%	4.2%	4.2%	32.30%	7.30%	19.80%
50	70	29.2%	16.7%	37.5%	4.2%	41.7%	0.0%	25.0%	12.5%	33.35%	8.35%	20.85%
50	80	25.0%	16.7%	33.3%	8.3%	50.0%	0.0%	29.2%	8.3%	34.38%	8.33%	21.35%
50	90	29.2%	25.0%	37.5%	8.3%	45.8%	0.0%	20.8%	4.2%	33.32%	9.38%	21.35%
50	100	29.2%	20.8%	37.5%	0.0%	54.2%	4.2%	16.7%	12.5%	34.40%	9.38%	21.89%
60	10	0.0%	0.0%	0.0%	0.0%	25.0%	0.0%	4.2%	0.0%	7.30%	0.00%	3.65%
60	20	25.0%	12.5%	33.3%	0.0%	45.8%	4.2%	0.0%	0.0%	26.02%	4.17%	15.10%

Continued on next page

Table 5 – continued from previous page

Training Steps (k)		Spoon on cloth		Carrot on plate		Stack Blocks		Eggplant in Basket		Avg-P	Avg-E	Avg-All
Stage1	**Stage2**	Part.	Ent.	Part.	Ent.	Part.	Ent.	Part.	Ent.			
60	30	20.8%	0.0%	8.3%	0.0%	37.5%	0.0%	8.3%	0.0%	18.73%	0.00%	9.36%
60	40	29.2%	25.0%	33.3%	4.2%	58.3%	0.0%	8.3%	0.0%	32.27%	7.30%	19.79%
60	50	20.8%	8.3%	37.5%	0.0%	50.0%	16.7%	0.0%	0.0%	27.07%	6.25%	16.66%
60	60	29.2%	16.7%	29.2%	4.2%	45.8%	4.2%	8.3%	0.0%	28.12%	6.28%	17.20%
60	70	25.0%	12.5%	37.5%	8.3%	58.3%	12.5%	4.2%	0.0%	31.25%	8.33%	19.79%
60	80	41.7%	20.8%	20.8%	4.2%	66.7%	12.5%	8.3%	0.0%	34.38%	9.38%	21.88%
60	90	41.7%	20.8%	29.2%	8.3%	58.3%	12.5%	0.0%	0.0%	32.30%	10.40%	21.35%
60	100	29.2%	12.5%	33.3%	4.2%	62.5%	8.3%	4.2%	0.0%	32.30%	6.25%	19.28%
70	10	20.8%	0.0%	20.8%	0.0%	25.0%	0.0%	4.2%	0.0%	17.70%	0.00%	8.85%
70	20	16.7%	4.2%	29.2%	0.0%	45.8%	0.0%	4.2%	0.0%	23.97%	1.05%	12.51%
70	30	12.5%	12.5%	29.2%	0.0%	45.8%	0.0%	4.2%	0.0%	22.93%	3.12%	13.03%
70	40	54.2%	25.0%	41.7%	0.0%	79.2%	16.7%	8.3%	0.0%	45.85%	10.43%	28.14%
70	50	33.3%	16.7%	41.7%	4.2%	83.3%	16.7%	12.5%	0.0%	42.70%	9.40%	26.05%
70	60	41.7%	37.5%	50.0%	12.5%	62.5%	4.2%	8.3%	0.0%	40.62%	13.55%	27.09%
70	70	41.7%	33.3%	50.0%	16.7%	83.3%	12.5%	4.2%	0.0%	44.80%	15.62%	30.21%
70	80	50.0%	29.2%	45.8%	8.3%	70.8%	8.3%	4.2%	0.0%	42.70%	11.45%	27.07%
70	90	41.7%	25.0%	41.7%	8.3%	70.8%	16.7%	12.5%	4.2%	41.67%	13.55%	27.61%
70	100	50.0%	41.7%	45.8%	8.3%	70.8%	8.3%	0.0%	0.0%	41.65%	14.57%	28.11%
80	10	0.0%	0.0%	12.5%	0.0%	25.0%	0.0%	0.0%	0.0%	9.38%	0.00%	4.69%
80	20	12.5%	0.0%	29.2%	0.0%	41.7%	0.0%	0.0%	0.0%	20.85%	0.00%	10.42%
80	30	33.3%	16.7%	20.8%	0.0%	33.3%	0.0%	4.2%	0.0%	22.90%	4.17%	13.54%
80	40	29.2%	12.5%	37.5%	4.2%	41.7%	0.0%	0.0%	0.0%	27.10%	4.17%	15.64%
80	50	33.3%	12.5%	33.3%	4.2%	41.7%	4.2%	0.0%	0.0%	27.07%	5.23%	16.15%
80	60	4.2%	0.0%	37.5%	4.2%	45.8%	8.3%	12.5%	8.3%	25.00%	5.20%	15.10%
80	70	41.7%	20.8%	41.7%	0.0%	66.7%	4.2%	12.5%	0.0%	40.65%	6.25%	23.45%
80	80	33.3%	25.0%	33.3%	12.5%	54.2%	8.3%	4.2%	4.2%	31.25%	12.50%	21.88%
80	90	20.8%	8.3%	37.5%	12.5%	62.5%	8.3%	0.0%	0.0%	30.20%	7.28%	18.74%
80	100	41.7%	33.3%	29.2%	4.2%	54.2%	4.2%	4.2%	4.2%	32.32%	11.47%	21.90%
90	10	0.0%	0.0%	0.0%	0.0%	16.7%	0.0%	0.0%	0.0%	4.17%	0.00%	2.09%
90	20	0.0%	0.0%	16.7%	0.0%	25.0%	4.2%	0.0%	0.0%	10.43%	1.05%	5.74%
90	30	25.0%	8.3%	29.2%	0.0%	54.2%	4.2%	4.2%	0.0%	28.15%	3.12%	15.64%
90	40	41.7%	20.8%	45.8%	4.2%	58.3%	0.0%	0.0%	0.0%	36.45%	6.25%	21.35%
90	50	20.8%	12.5%	45.8%	0.0%	62.5%	8.3%	4.2%	0.0%	33.32%	5.20%	19.26%
90	60	37.5%	16.7%	58.3%	4.2%	62.5%	4.2%	0.0%	0.0%	39.57%	6.28%	22.93%
90	70	37.5%	25.0%	41.7%	0.0%	58.3%	4.2%	0.0%	0.0%	34.38%	7.30%	20.84%
90	80	37.5%	16.7%	41.7%	0.0%	58.3%	20.8%	0.0%	0.0%	34.38%	9.38%	21.88%
90	90	37.5%	25.0%	37.5%	0.0%	54.2%	8.3%	0.0%	0.0%	32.30%	8.33%	20.31%
90	100	29.2%	16.7%	33.3%	0.0%	50.0%	12.5%	4.2%	0.0%	29.18%	7.30%	18.24%
100	10	0.0%	0.0%	25.0%	0.0%	16.7%	0.0%	0.0%	0.0%	10.43%	0.00%	5.21%
100	20	16.7%	4.2%	12.5%	0.0%	37.5%	8.3%	0.0%	0.0%	16.68%	3.12%	9.90%
100	30	50.0%	16.7%	50.0%	0.0%	70.8%	12.5%	8.3%	0.0%	44.77%	7.30%	26.04%
100	40	41.70%	20.80%	45.80%	4.20%	58.30%	0.00%	0.00%	0.00%	36.45%	6.25%	21.35%
100	50	50.0%	29.2%	50.0%	4.2%	70.8%	8.3%	12.5%	4.2%	45.82%	11.47%	28.65%
100	60	50.00%	25.00%	41.70%	8.30%	58.30%	16.70%	8.30%	0.00%	39.57%	12.50%	26.04%
100	70	62.5%	29.2%	45.8%	12.5%	83.3%	12.5%	4.2%	4.2%	48.95%	14.60%	31.77%
100	80	45.8%	25.0%	25.0%	4.2%	83.3%	8.3%	20.8%	12.5%	43.72%	12.50%	28.11%
100	90	41.7%	29.2%	41.7%	8.3%	66.7%	0.0%	16.7%	16.7%	41.70%	13.55%	27.62%
100	100	66.70%	58.30%	50.00%	0.00%	58.30%	16.70%	8.30%	8.30%	45.82%	20.82%	33.32%