

Visual Verification Enables Inference-time Steering and Autonomous Policy Improvement

Mingtong Zhang, Dhruv Shah
Princeton University
[veritas-improvement.github.io](https://github.com/veritas-improvement)

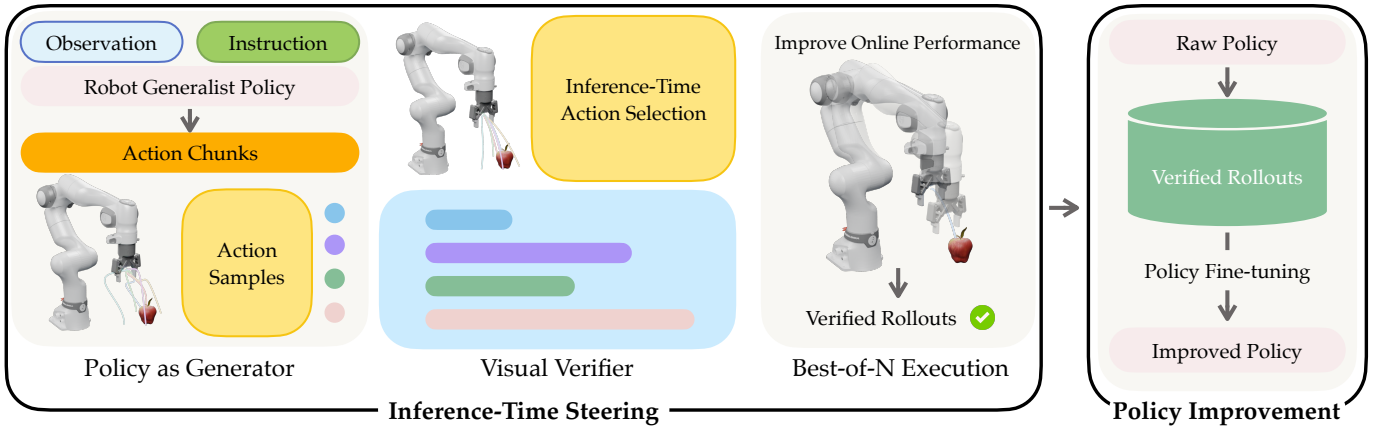


Fig. 1: Overview of VERITAS. A pre-trained generalist policy acts as a stochastic generator, sampling multiple short-horizon action chunks at each decision step. A gradient-free visual verifier scores these candidates based on task alignment and physical plausibility, and the highest-scoring action is executed, yielding immediate performance gains at inference time. Successful verifier-guided rollouts are logged and reused for offline policy improvement, forming a data flywheel that distills verification-time reasoning into the policy and enables continual improvement with minimal human supervision.

Abstract—Robots deployed in the real world should learn from their experience and improve over time. This requires a mechanism of practicing and learning from feedback. In this paper, we propose VERITAS, a generator-verifier framework for generalist robot policies for inference-time policy steering and self-improvement. We use a pre-trained generalist robot policy as a “generator” and pair it with a gradient-free “visual verifier” that evaluates actions at inference time. This framework enables inference-time steering that improves policy performance without additional training. We demonstrate that inference-time verification consistently outperforms vanilla generalists without training on additional demonstration data. Additionally, we demonstrate that the verified rollouts provide effective supervision for offline policy improvement: policies fine-tuned on verified self-generated trajectories achieve consistent performance gains. Notably, we find that post-training with verified rollouts achieves comparable efficiency to expert demonstrations, while requiring no human interventions. Our results highlight inference-time verification as a practical and scalable mechanism for improving robotic policies during deployment.

I. INTRODUCTION

Robotic foundation models trained on large-scale demonstration data [1–7] have shown advanced capabilities in various domains, such as manipulation and navigation [8–12]. These advancements are largely powered by large-scale robotic data collection from different sources. However, unlike texts or images, robotic data collection remains expensive, slow, and limited by the availability of human demonstrators. Thus, robot

foundation models face a fundamental bottleneck: state-of-the-art models are primarily trained on data collected by human experts, which scales linearly in human expert time. This raises a key challenge: how can we improve robotic policies without scaling up the amount of human data? Our key insight is to enable robotic systems to explore the real world and learn from their experience, with minimal supervision and cost.

A promising solution inspired by recent advances in large language modeling is to scale *inference-time computation* for verification [13–15]. Text-based reasoning models can generate multiple solution traces for a problem and select the best one using a “verifier” [16–18]. Analogous to this, perhaps a robot policy can generate and verify multiple action candidates at inference-time before committing to a physical action. Through selecting the *best-of-N* samples using a verifier or value function, a pre-trained model might achieve performance improvements without any additional training. This insight is particularly valuable in robotic applications where data acquisition is significantly more expensive than computation. Rather than collecting additional data to improve coverage or recovery behaviors [19, 20], we can instead use inference-time sampling to enable the policy to “try” multiple times and only execute the best generation. This allows us to trade compute for improved performance through inference-time steering. As the robot experiences successful trials, this autonomously generated data can be used to guide further improvements.

In this work, we introduce Visual **Verification** for **Inference-time Steering** and **Autonomous Policy Improvement**, or **VERITAS**, a robotic self-improvement method built around a *generator-verifier framework*. We treat robot policies as a *generator* capable of sampling diverse candidate actions, and pair it with a VLM-based *verifier* which evaluates these candidates to guide exploration and task execution. This framework enables robots to autonomously explore the physical world with inference-time verification, steering existing generalist policies in an *online* manner out-of-the-box. Additionally, the verified self-generated rollouts serve as high-quality training data for *offline* policy improvement. Notably, this generator-verifier framework enables policy self-improvement without (i) training any additional embodied reasoning models, and (ii) requiring expensive human annotation or data collection. Instead, it externalizes reasoning from the policy itself by introducing an explicit verification stage that evaluates candidate actions at inference time. By creating a flywheel where the robot learns from its own successful executions, we ensure that the post-training data is *on-policy* and physically feasible; we posit that this enables the policy to rapidly improve its behaviors and learn new behaviors, guided by the verifier. Our experiments demonstrate that our framework can improve policy performance while matching the efficiency of learning from human expert demonstrations.

We evaluate the efficacy of VERITAS for inference-time steering and autonomous policy improvement extensively in simulation [21] and in the real world [22]. Across various base policies and manipulation tasks, we find that inference-time steering with VERITAS consistently improves performance over the base policy, without modifying policy parameters. Additionally, we show that verifier-generated rollouts form an effective supervision signal for policy improvement, achieving an average improvement of 10% over the base policy in simulation. We further find that performance gains from “self-improvement” are within 5% of the gains from post-training on human-expert data in the real world, suggesting that verifier-curated data is comparable to human data for policy improvement.

II. RELATED WORKS

Prior work has extensively studied data collection through *human-in-the-loop* learning [23–25] and *shared autonomy* [26–28], where robots improve by incorporating expert interventions, corrections, or demonstrations during deployment. Classic approaches such as DAgger rely on expert relabeling to correct policy mistakes on-policy [19, 29, 30], while shared autonomy frameworks blend human control with autonomous execution to improve task success and collect corrective data. Recent systems refine this paradigm by structuring how corrective data is collected in long-horizon tasks; for example, Hu et al. [20] explicitly decompose robot behavior into nominal execution and recovery phases, triggering human intervention when the policy enters failure states to improve downstream learning [3]. Instead of using human intervention [19, 31] as the source of corrective signal, we leverage *execution-time*

verification to autonomously evaluate candidate actions and select successful behaviors. The verifier replaces the role of the human supervisor by providing a task-aligned preference signal at inference time, allowing the robot to improve performance online and collect success-conditioned trajectories without shared control.

Inspired by advances in language modeling, where scaling test-time compute via verification and search often outperforms scaling model parameters [13, 16, 18], the robotics community has begun exploring mechanisms for embodied control. Recent work integrates explicit reasoning into robot policies by training VLAs with supervision over intermediate visual and semantic representations, enabling multi-step, visually grounded decision making [32–34].

While these works demonstrate impressive embodied reasoning capabilities, these methods rely on expensive annotations during policy pre-training and are not suitable for inference-time steering or improvement of existing policies. Another promising set of directions is sampling and steering online samples, such as Kwok et al. [35] demonstrating that action error scales with the number of actions sampled at inference time using a learned verifier, [36] suggesting scaling verification leads to more effective alignment for policies, Wu et al. [37] introducing a runtime steering method that verifies actions against a reasoning VLA’s self-generated textual plan, and Nakamoto et al. [38], which trains a language-conditioned value function using offline RL and uses it for steering the generalist policy. [39, 40] requires world models’ prediction to select reasonable actions for policy steering. While these methods demonstrate the power of test-time sampling, they typically treat inference-time search as a “momentary boost”, relying on expensive verification, that does not necessarily improve the policy in long-term deployment. Our work takes a novel perspective by treating online verification as a data engine without training task-specific verifiers, and allowing us to distill expensive inference-time compute into the base policy to enhance its capabilities via autonomous verification and post-training. Our experiments further show that our visual verification framework outperforms various other test-time sampling approaches for robot policies.

Long-term deployment of generalist robots requires policies to continually learn and adapt from their experience. However, standard imitation learning suffers from covariate shift, where policy performance degrades as the robot drifts from the expert’s state distribution [29]. Traditional solutions [19, 29] mitigate this by querying human experts for labels on robot-visited states. However, these human-in-the-loop paradigms are difficult to scale. Our approach replaces the human labeler with an automated verifier, ensuring the training data is strictly *on-policy* and physically feasible for the robot’s embodiment. This effectively mitigates the distribution shift problem by aligning the training distribution with the support of the robot’s exploration prior, unlocking scalable self-improvement without human intervention.

III. PRELIMINARIES

We formulate the problem of language-guided robotic manipulation as a Partially Observable Markov Decision Process (POMDP) [41] defined by the tuple $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$.

Problem Formulation. At each sampling timestep t , the robot receives an observation $o_t \in \mathcal{O}$ and a natural language instruction $l \in \mathcal{L}$. Our objective is to train a policy π_θ that maximizes the expected success rate over a distribution of tasks. Unlike standard imitation learning, which assumes a fixed static dataset, our objective is to enable the policy to self-improve its performance post-deployment by collecting and learning from its own experience.

Generalist Robot Policies. In this work, we assume access to a pre-trained “generalist policy” π_θ , that can be conditioned on the current observation and a language instruction, and predicts actions a_t to control a robot. We assume that π_θ can successfully solve a large number of tasks in the target environment, and are agnostic to the specific architecture or data sources the policy was pre-trained on. Following recent advancements in high-frequency control, we employ *action chunking* [42, 43]: the policy predicts a sequence of actions of horizon H :

$$\mathbf{a}_{t:t+H} \doteq (a_t, a_{t+1}, \dots, a_{t+H-1}) \in \mathcal{A}^H. \quad (1)$$

Generative Policies as Priors. Our method relies on the policy’s ability to model complex, multi-modal action distributions (e.g., grasping an object from different angles, solving a multi-step task in different orders, etc.). We focus on generative policies trained with flow-matching or diffusion objectives [5, 43]. Unlike deterministic regression policies that output a single mean action, a generative policy learns the conditional probability distribution $p(\mathbf{a}_{t:t+H} \mid o_t, l)$. At inference time, action chunks are generated via a stochastic sampling process. In this work, we leverage this stochasticity not just for robustness, but as a mechanism for exploring diverse, functional affordances [44] of the scene.

IV. A GENERATOR-VERIFIER FRAMEWORK FOR ROBOT POLICIES

We present a *generator-verifier* inference framework that steers a generalist policy with execution-time selection. The core design principle is to utilize (i) *generation* by a generalist policy π_θ as a stochastic generator, and (ii) *evaluation* by a plug-in verifier V that is policy-agnostic and gradient-free. At every decision step, the robot samples multiple short-horizon action chunks, scores them with V , and executes the best candidate. This yields immediate online performance gains without updating θ , while producing success-conditioned trajectories for offline policy improvement. The complete procedure is detailed in Algorithm 1.

A. The Policy as a Stochastic Generator

Standard policy deployment typically utilizes deterministic or greedy decoding to minimize variance. We instead view

the pre-trained policy π_θ as a *stochastic generator*. We posit that a generalist policy trained on diverse data maintains a broad “prior” over valid affordances, and by sampling from this distribution, we can encourage and recover high-quality behaviors that might have lower initial probability but higher task alignment.

Formally, at discrete sampling timestep t , given observation o_t and language instruction l , the policy generates a set of N candidate action chunks (Algorithm 1, Line 1):

$$\mathbf{a}_{t:t+H}^{(i)} \sim \pi_\theta(\cdot \mid o_t, l), \quad i \in \{1, \dots, N\} \quad (2)$$

where each candidate $\mathbf{a}_{t:t+H}^{(i)} = (a_t, a_{t+1}, \dots, a_{t+H-1})$ represents a sequence of H actions. Utilizing action chunks is critical; it amortizes the cost of verification over the horizon H and provides the verifier with temporal context to evaluate the consequence of a behavior, rather than a micro instantaneous movement.

B. The Verifier Interface

To select the optimal candidate from the generator, we introduce a verifier V . We define the verifier as a plug-and-play, gradient-free function that maps the observation, candidate action chunk, and task instruction to a scalar score:

$$V(o_t, \mathbf{a}_{t:t+H}, l) \in \mathbb{R}. \quad (3)$$

Conceptually, V estimates the utility or alignment of executing $\mathbf{a}_{t:t+H}$ from state o_t under instruction l . Importantly, steering with V operates solely at inference time and does not update π_θ . As a result, the verifier functions as a plug-and-play module that can be instantiated with various mechanisms (e.g., VLMs, geometric constraints, or learned value models) without retraining the base policy.

C. Inference-Time Steering

We combine the generator and verifier via *Best-of- N* selection. This mechanism trades inference-time computation for performance, allowing the robot to “think” (sample and verify) before acting.

As outlined in Algorithm 1, the control loop proceeds as follows:

- 1) **Sample:** The policy generates N diverse action chunks $\{\mathbf{a}_{t:t+H}^{(i)}\}_{i=1}^N$.
- 2) **Verify:** The verifier scores each candidate: $v_i = V(o_t, \mathbf{a}_{t:t+H}^{(i)}, l)$.
- 3) **Select:** The highest-scoring candidate index is selected:

$$i^* = \arg \max_{i \in \{1, \dots, N\}} v_i, \quad \mathbf{a}_{t:t+H}^* = \mathbf{a}_{t:t+H}^{(i^*)} \quad (4)$$

- 4) **Execute:** The robot executes the chunk $\mathbf{a}_{t:t+H}^*$ in the environment.

This steering process serves as a filter, prioritizing physically valid and task-aligned actions. Importantly, because π_θ is fixed, these performance gains come purely from inference-time computation.

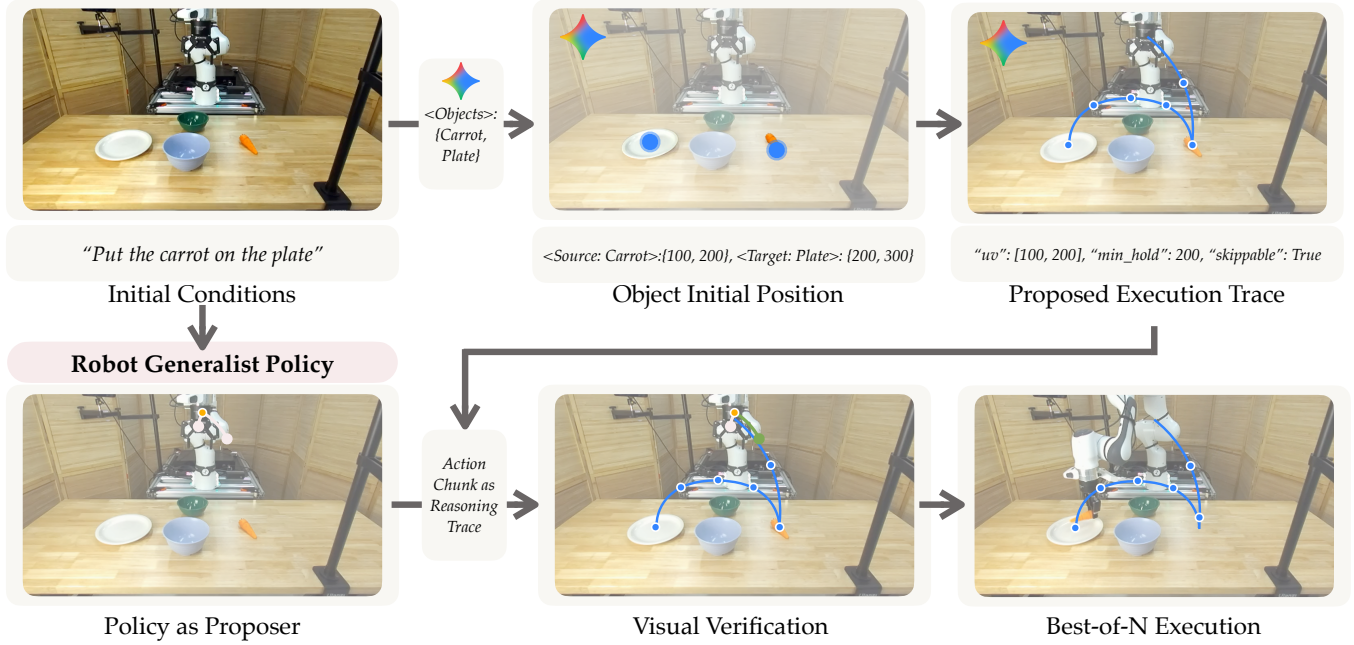


Fig. 2: The overview of VERITAS for inference-time steering. At inference time, a two-call VLM scheme is used to construct a visual verifier. First, given the initial observation and task instruction, the VLM identifies the target object to focus on, a detection model is then used to localize the object in image space. Second, conditioned on the object location, instruction, and initial observation, the VLM proposes a sequence of pixel-space waypoints that define a full visual trace for the task. During execution, the policy samples multiple action candidates, which are scored based on how well their resulting motion follows the proposed visual trace, and the best candidate is selected for execution. Successful verifier-guided rollouts are logged and used as training data for offline policy fine-tuning, resulting in an improved policy with higher task performance.

D. Autonomous Self-Improvement

While inference-time steering improves performance, it incurs a computational cost at every decision step. To amortize this cost and permanently improve the policy, we utilize the verified trajectories as a source of autonomous supervision with no extract human expert efforts.

This process creates a *Data Flywheel*. During policy deployment, we log the successful verified trajectories generated via inference-time steering into a dataset $\mathcal{D}_{\text{auto}}$.

$$\mathcal{D}_{\text{auto}} = \bigcup_k \{(\mathbf{o}_t, \mathbf{a}_{t:t+H}^*, l)\}_{t \in \mathcal{I}_{\text{success}}} \quad (5)$$

We then fine-tune the original policy π_θ on $\mathcal{D}_{\text{auto}}$ using standard supervised behavior cloning to obtain updated parameters θ' :

$$\theta' \leftarrow \arg \min_{\theta} \mathbb{E}_{\mathcal{D}_{\text{auto}}} [-\log \pi_\theta(\mathbf{a}_{t:t+H}^* | \mathbf{o}_t, l)] \quad (6)$$

Mitigating Distribution Shift. A fundamental challenge in imitation learning is the distribution shift between the expert and the learner. Standard human-in-the-loop approaches require expensive expert relabeling. Our framework automates this: because the data in $\mathcal{D}_{\text{auto}}$ was generated by the robot’s own policy (the generator), it is inherently on-policy and kinematically feasible. The verifier acts as the “expert” filter, ensuring only high-quality data is retained. This closes the loop: the reasoning performed by the verifier is distilled into the policy, reducing the need for large N samples in future deployments. Importantly, as the verified data is generated by policy itself, the action distribution is close to the policy’s

action prior, which mitigates the shift and remains efficient for policy improvement.

V. SYSTEM DESIGN FOR INSTANTIATING VERITAS

We now describe a concrete instantiation of the proposed generator-verifier framework for tabletop robotic manipulation. In this work, as an example, we conduct our analysis using the π_0 architecture [5], an open-source Vision-Language-Action (VLA) model trained on thousands of hours of diverse robot trajectories. π_0 utilizes a mixture-of-transformers backbone with a flow-matching action head. Crucially, this flow-matching formulation allows us to effectively sample complex, multi-modal action distributions, providing the necessary diversity to explore valid solutions at test time. We pair this generator with our visual verifier designed to ground high-level semantic instructions into precise geometric constraints.

A. Visual Verifier Design

To ground high-level semantic instructions into precise geometric constraints without requiring expensive 3D annotations, we design a verifier based on absolute pixel-space trajectories proposed by a frontier vision-language model (VLM) [45]. This design leverages the strong spatial and embodied reasoning capabilities of modern VLMs to create a “visual guardrail” for the policy.

The verification process consists of two stages:

- 1) **Trace Generation:** At the onset of a task, the VLM is provided with the initial observation $\mathbf{o}_0$ and the language instruction l . It is prompted to propose a *visual reasoning trace*: a sequence of pixel-space waypoints $\{w_k\}$

Algorithm 1: VERITAS Online Steering and Autonomous Policy Improvement

Input : Policy π_θ , verifier V , instruction l , horizon H , samples N

Output: Executed trajectory, dataset $\mathcal{D}_{\text{auto}}$, updated policy $\pi_{\theta'}$

```
// Phase 1: Inference-time Steering
Initialize  $\mathcal{D}_{\text{auto}} \leftarrow \emptyset$  and observe initial state  $o_0$ ;
for  $t = 0, H, 2H, \dots$  until termination do
    // 1. Sample diverse candidates
    (Eq. 2)
    Sample  $\{\mathbf{a}_{t:t+H}^{(i)}\}_{i=1}^N$  where  $\mathbf{a}_{t:t+H}^{(i)} \sim \pi_\theta(\cdot | o_t, l)$ ;
    // 2. Verify candidates
    Score each candidate  $v_i \leftarrow V(o_t, \mathbf{a}_{t:t+H}^{(i)}, l)$ ;
    // 3. Select Best-of-N (Eq. 4)
    Select  $i^* \leftarrow \arg \max_i v_i$  and set  $\mathbf{a}_{t:t+H}^* \leftarrow \mathbf{a}_{t:t+H}^{(i^*)}$ ;
    // 4. Execute and Log (Eq. 5)
    Execute  $\mathbf{a}_{t:t+H}^*$  and observe  $o_{t+H}$ ;
    Log  $(o_t, \mathbf{a}_{t:t+H}^*, l)$  into  $\mathcal{D}_{\text{auto}}$ ;

// Phase 2: Offline Self-Improvement
if Offline update is enabled then
    Initialize  $\theta' \leftarrow \theta$ ;
    for  $e = 1$  to  $E$  do
        Sample batch  $\mathcal{B} \sim \mathcal{D}_{\text{auto}}$ ;
        Update  $\theta'$  via Behavior Cloning (Eq. 6):
            
$$\theta' \leftarrow \theta' - \eta \nabla_{\theta'} \sum_{(o, \mathbf{a}, l) \in \mathcal{B}} \mathcal{L}_{\text{BC}}(\pi_{\theta'}(\cdot | o, l), \mathbf{a})$$

    return  $\pi_{\theta'}$ ;
```

overlaid on the image that defines the ideal end-effector trajectory for task completion. This trace remains static throughout execution to ensure temporal consistency.

- 2) **Geometric Scoring:** At inference time, for each candidate action chunk $\mathbf{a}_{t:t+H}^{(i)}$ sampled from π_θ , we project the resulting end-effector/gripper positions into the image plane. The verifier score v_i is computed as the negative Euclidean distance between the candidate’s projected trajectory and the nearest segment of the VLM-generated trace.

Crucially, this design allows for efficient verification without requiring online tracking or expensive VLM querying at every verification step. Once the visual trace is generated, verification reduces to a fast geometric consistency check, enabling the evaluation of large N samples with minimal latency overhead. We also have other types of verifier design and ablate them in the experiments. For more details on our verifier and alternate designs, please see Appendix A.

B. Steering at Inference-Time

We implement a simple inference-time steering mechanism using the Best-of- N strategy described in Section IV-C.

Implementation Details. As an example, we deploy the π_0 -DROID policy on a workstation with one NVIDIA GeForce RTX 5090 GPU. At each control frequency step (e.g., 15 Hz), we define an action chunk horizon of H steps. The generator samples N candidates in parallel. The visual verifier evaluates these N chunks against the pre-computed visual trace.

Latency and Throughput. The VLM verifier generates an execution trace only once at the start of episode, and hence, is not run in-the-loop with the policy. The geometric verification incurs negligible overhead (< 1 ms) and runs online. The cost of policy inference scales linearly with the number of sampled actions N , and can be batched during inference. For our experiments, we use $N = 5$ samples and maintain a control frequency of 15 Hz. This trade-off allows us to filter out physically invalid or erratic behaviors that the base policy might occasionally produce.

C. Policy Improvement

To close the loop, we utilize the successful verified trajectories curated by our steerable system for offline policy fine-tuning.

Data Collection. We deploy the steered policy in the real world to solve a diverse set of manipulation tasks (Section VI). We log all executed action chunks $\mathbf{a}_{t:t+H}^*$ that lead to successful task completion after verification, filtering out failures. This results in a high-quality dataset $\mathcal{D}_{\text{auto}}$.

Training Setup. We fine-tune the pre-trained π_0 -DROID checkpoint on $\mathcal{D}_{\text{auto}}$ using the behavior cloning objective (Eq. 6). We use a learning rate of 5×10^{-5} and train for 20,000 steps. As shown in our experiments, even small amounts of this verifier-curated data (e.g., 20 trajectories) are sufficient to significantly shift the policy distribution towards the robust behaviors discovered during steering, effectively distilling the VLM’s reasoning into the policy’s weights.

VI. EXPERIMENTS

In our experiments, we aim to assess the efficacy of the generator-verifier paradigm as a data flywheel for policy improvement. We study the following research questions:

- 1) **Inference-time Steering:** Does the Generator-Verifier architecture improve success rates at test time compared to standard policies?
- 2) **Policy Improvement:** Does training on self-generated, verified data improve the base policy?
- 3) **Data Efficiency:** How does our autonomous data compare to human-expert data?

We ablate the design of verifier in simulation experiments and find the paradigm can improve the policy performance at inference time, agnostic to the design as well. Based on these results, we select the most effective verifier configuration from simulation for deployment in real-world experiments.

A. Experimental Setup

a) *Tasks:* We evaluate our method on a suite of simulated and real-world robotic manipulation tasks specified by natural

language instructions. In simulation, we focus on 4 manipulation tasks from [46], following the evaluation protocol of [21]. These tasks provide a controlled benchmark for assessing policy performance and generalization under standardized conditions. We conduct all real-world experiments on the DROID platform [22]. We design a set of challenging manipulation tasks that stress visual grounding, spatial reasoning, manipulation that requires high-precision, etc. Specifically, we select 2 manipulation tasks in the real world for each policy.

b) Policies: We evaluate our framework using different pretrained robot policies. In simulation, we employ an open-sourced BridgeData policy [47], denoted as π_0 -Bridge, trained on [46] following the architecture of [5]. In real-world experiments, we use various generalist robot policies. Specifically, π_0 -DROID [5], $\pi_{0.5}$ -DROID [48] trained on large-scale real-world demonstration data and fine-tuned on [22]. Unless otherwise specified, all policies are treated as fixed base policies for inference-time steering and are used as action proposers within the generator-verifier framework.

c) Baselines: To show the efficiency of our verifier against several baselines for inference time steering and ablate different verifier designs in the simulation experiments. We compare the VERITAS verifier to several key baselines.

- **V-GPS [38]:** A language-conditioned value function learned via offline RL for policy steering.
- **Heuristic:** A verifier that defines action primitives and use stage transitions for verification.
- **VLM+Constraints:** A verifier that scores action samples by verifying if they follow VLM-proposed reference-based waypoints that resolve to pixel coordinates via tracked objects.

We also compare to RoboMonkey [35], which uses a learned single-step verifier by fitting a Gaussian on the action distribution. While effective for local corrections, we find that the RoboMonkey verifier’s optimization of local single-step action regions limits its application to completely new tasks that require global reasoning. For more details on this baseline, please see Appendix C.

In real-world experiments, we emphasize the comparison between our SOTA verifier design reflected by the simulation results, against baselines:

- **PIVOT [49]:** An approach that uses proposed action primitives and Iterative Visual Optimization for robotic tasks. This baseline combines a naïve action-primitive proposer with an optimization-based action selector, highlighting the importance of a policy proposer.
- **V-GPS-DROID:** We train a V-GPS model on the DROID [22] dataset and deploy it on the same robot hardware setup. This experiment evaluates the efficiency of our combined policy proposer and visual verifier against a learned value function.

B. Inference-Time Steering.

In simulation, we evaluate the performance of different verifiers on top of the base policy π_0 -Bridge and take the base policy as one of the baselines in the simulation. All evaluations

are averaged over 10 trials on top of task variations such as robot or object locations, etc, following [21]. In the real world, we deploy the best-performing verifier in the simulation and evaluate it on 2 different policies using the DROID platform, conducting 50 rollouts for each task across 2 tasks per policy.

As shown in Figure 4 across all 4 simulation evaluated tasks, verifier-guided action selection consistently improves success rates relative to the base policy. While the magnitude of improvement varies with task difficulty and verifier design, performance gains are observed even when using simple heuristic verifiers, indicating that the benefits of the generator-verifier paradigm are not dependent on a specific verifier architecture. On average, all of our verifier designs outperform V-GPS, suggesting that explicit verification of action consequences provides a more reliable signal for policy steering than the learned value estimates.

We observe similar trends in real-world experiments. As shown in Figure 6, inference-time steering significantly improves task success rates across different evaluated real-world tasks and policies.

Across simulation and real-world settings, covering 3 different policies and a total of 1160 evaluation episodes, verifier-guided execution improves policy success rates by an average of 12.6% in simulation and 35% in real-world deployment, without any policy fine-tuning. These gains highlight the effectiveness of inference-time computation alone in improving generalist policy performance. Note that PIVOT’s naïve action primitives can not achieve successful outcomes, whereas a trained policy with a strong action prior performs effectively, highlighting the importance of good action priors for improving performance.

Importantly, the proposed framework is policy-agnostic and verifier-agnostic. We observe consistent improvements across different base policies and across a diverse set of verifier implementations in simulated and real-world environments. This robustness demonstrates that the generator-verifier framework provides a general and reliable mechanism for inference-time policy steering, applicable across tasks, embodiments, and verification strategies.

C. Offline Policy Improvement

We next study whether verified autonomous executions can be used to improve the policy offline. In this setting, we deploy the policy using the generator-verifier framework in the simulation and real world to collect a set of verified successful trajectories for offline policy fine-tuning.

Specifically, in the simulation experiments, we select the verifier that boosts the policy steering performance the most and collect the successful verified rollouts, consisting of 656 demonstrations for the base policy fine-tuning. As shown in Figure 5, fine-tuning on this autonomously collected data improves average policy performance and consistently outperforms the base policy across all 4 tasks.

Notably, inference-time steering provides larger performance gains on tasks that are difficult for the base policy, such as Stack Blocks, than on simpler tasks. One possible

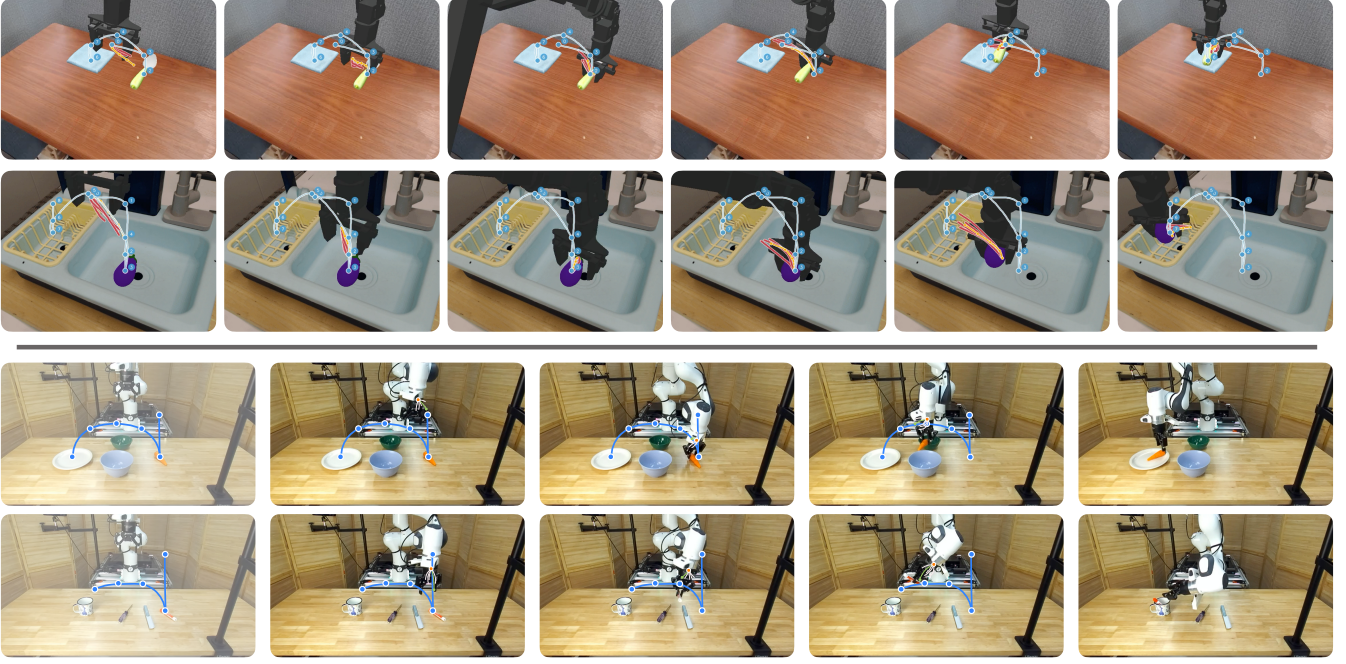


Fig. 3: Inference-time verification improves task performance. Trajectory rollouts illustrating VERITAS steering. The first 2 rows show verification of the π_0 -Bridge policy on a simulated WidowX and the last 2 rows show verification of the π_0 -DROID policy on a real FR3-DROID robot. At each decision step, the generator generates N candidate action chunks $\{\alpha_{t:t+H}^{(i)}\}_{i=1}^N$; the verifier scores each candidate $V(o_t, \alpha_{t:t+H}^{(i)}, l)$ and selects the best action α^* for execution (highlighted). Verifier-guided rollouts achieve successful task completion.

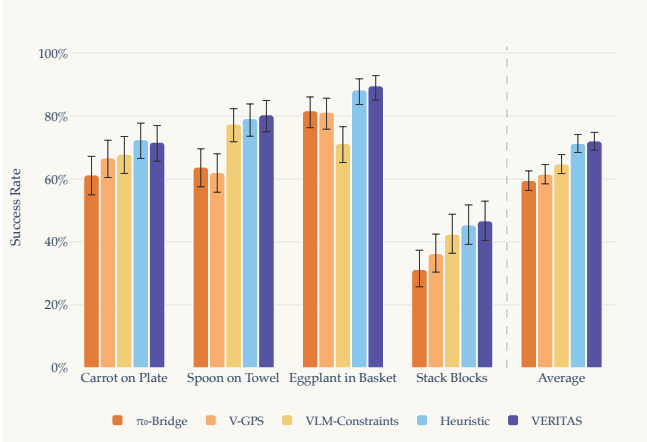


Fig. 4: Inference-time steering in SIMPLER. Simulation success rates comparing standard policy execution and VERITAS execution across 4 tasks and different verifier architectures. VERITAS verifier consistently improves performance over different methods across all tasks. The largest gains are observed in the most challenging task (Stack Blocks). Error bars show 95% finite-sample CI [50].

explanation is that the policy produces more diverse action samples when it is uncertain or underperforms, creating greater opportunity for the verifier to select better candidates. In contrast, for tasks on which the policy already performs well, the sampled actions are less diverse, leaving less room for improvement. These results point to a fundamental trade-off: the effectiveness of inference-time steering depends on the diversity of the proposed actions, and therefore tends to be greatest on more challenging tasks.

In the simulation experiments, VERITAS fine-tuning improves average success rate by 9.7% over the base π_0 -Bridge

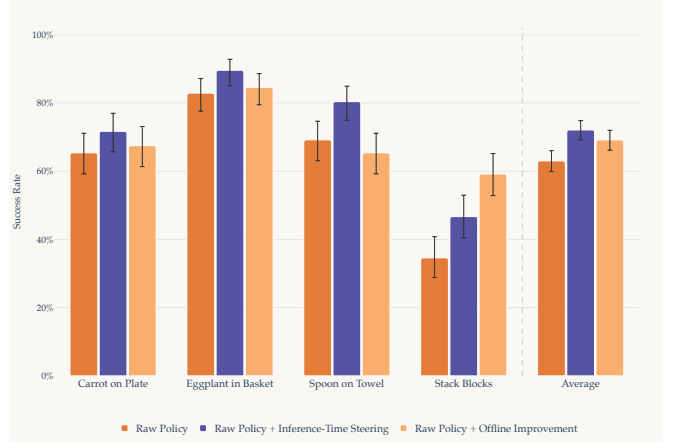


Fig. 5: Simulation success rates of the raw policy, the inference-time steering and the same policy fine-tuned on verifier curated autonomous data. While verification improves performance at inference time through action steering, fine-tuning on the collected verifier curated rollouts successfully distills these gains back into the policy weights. Error bars show 95% finite-sample CI [50].

policy across 4 simulated manipulation tasks in SimplerEnv for 960 total episodes. Success rate gains are consistent across all tasks, with the largest improvement on Stack Blocks, from 31.3% to 59.2%, by 27.9% on the hardest task. In the real world experiments shown in Figure 7, verified autonomous data consistently supports effective policy improvement across 4 challenging real-world tasks for 2 policies, with performance improving reliably as more data is collected, demonstrating that autonomous rollouts alone are sufficient for successful fine-tuning. These results demonstrate that verified autonomous executions can serve as a reliable and effective

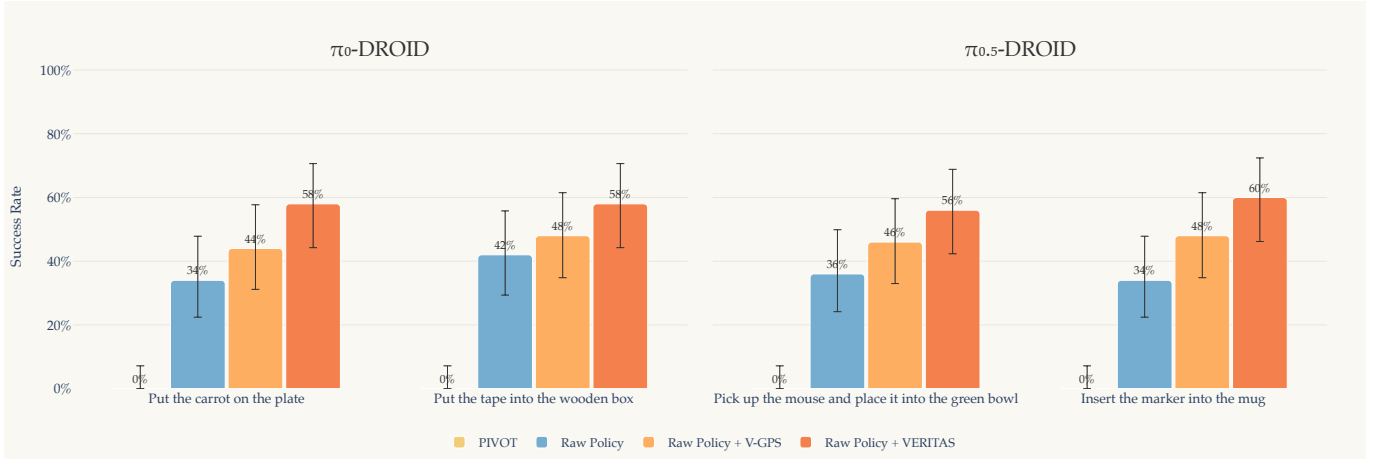


Fig. 6: Inference-time steering in the real world. We evaluate inference-time steering across two real-world policies and two tasks per policy, with 50 trials conducted for each task. VERITAS consistently outperforms the V-GPS baseline across all tasks, demonstrating robust improvements in the real world.

source of supervision for offline policy improvement, without requiring additional human demonstrations. By leveraging the generator-verifier framework to curate high-quality rollouts, VERITAS consistently improves policy performance.

D. Autonomous v/s Human-Collected Data

Finally, we compare the data efficiency of verifier-curated autonomous rollouts to human expert demonstrations as a supervision signal for offline policy improvement. While human demonstrations remain a strong baseline, they are expensive to collect and difficult to scale. An important question is whether autonomously collected, verifier-selected data can provide comparable learning signal.

In our real-world experiments, we fine-tune policies using datasets of 20, 30, 50, and 100 demonstrations collected from either verifier-curated autonomous rollouts or human expert teleoperation. We then evaluate each fine-tuned policy over 20 real-world trials across multiple manipulation tasks, comparing policy performance as a function of the quantity and source of fine-tuning data. This controlled setup allows us to directly compare how policy performance scales with the amount and source of fine-tuning data.

The experimental results are shown in Figure 7. Across all 4 tasks, with 2 tasks evaluated for each policy, verifier curated data demonstrates comparable data efficiency to human expert demonstrations. In several cases, autonomous data achieves stronger performance at moderate data budgets. For example, on *Carrot on plate*, fine-tuning with 50 autonomous demonstrations outperforms fine-tuning with the same number of human demonstrations (0.70 vs. 0.65). Similarly, on *Pick up mouse*, autonomous data consistently achieves higher success rates in the 20 to 50 demonstration regime.

At larger data budgets, verifier curated data remains competitive with expert supervision across all tasks. On *Carrot on plate*, the two data sources converge to the same success rate at 100 demonstrations, while on *Insert marker*, the performance gap remains within 5 percentage points across all data budgets. Overall, these results show that verifier curated autonomous rollouts provide an effective and scalable alternative to costly

human demonstrations for policy improvement.

Overall, these results show that verifier-curated autonomous rollouts can match the data efficiency of costly human teleoperation. By converting deployment-time execution into effective training data, our approach enables scalable policy improvement without requiring continuous expert supervision. These findings suggest that execution-time verification not only improves performance online, but also enables a practical pathway for continual policy improvement by converting deployment experience into effective training data.

More importantly, because the robot itself generates the data, the autonomously collected dataset is inherently on-policy and kinematically feasible. The verifier serves as an expert filter, retaining high-quality executions while discarding failures and suboptimal behaviors. This process effectively closes the learning loop: the reasoning and corrective signals provided by the verifier are distilled back into the policy during offline fine-tuning, progressively improving the generator and reducing the need to sample a large number of candidates in future deployments.

E. Implementation Hyperparameters

We perform a sweep over the action execution horizon, which determines the length of the visual action trace associated with each action chunk (Figure 8). For π_0 -DROID, we find that the open-loop execution horizon does not significantly degrade verification performance in our real-world experiments. Based on this observation, we use the default horizon of 8 steps for all real-world evaluations.

We also sweep over the number of action samples N on task success rate in simulation. As shown in Figure 9, steering performance generally improves as the number of action samples increases, suggesting that larger candidate sets provide better opportunities for verification. However, we find that performance saturates beyond $N = 8$, serving as a practical guide for real-world deployment.

VII. DISCUSSION

In this work, we introduced a generator-verifier framework that transforms inference-time computation into sustainable

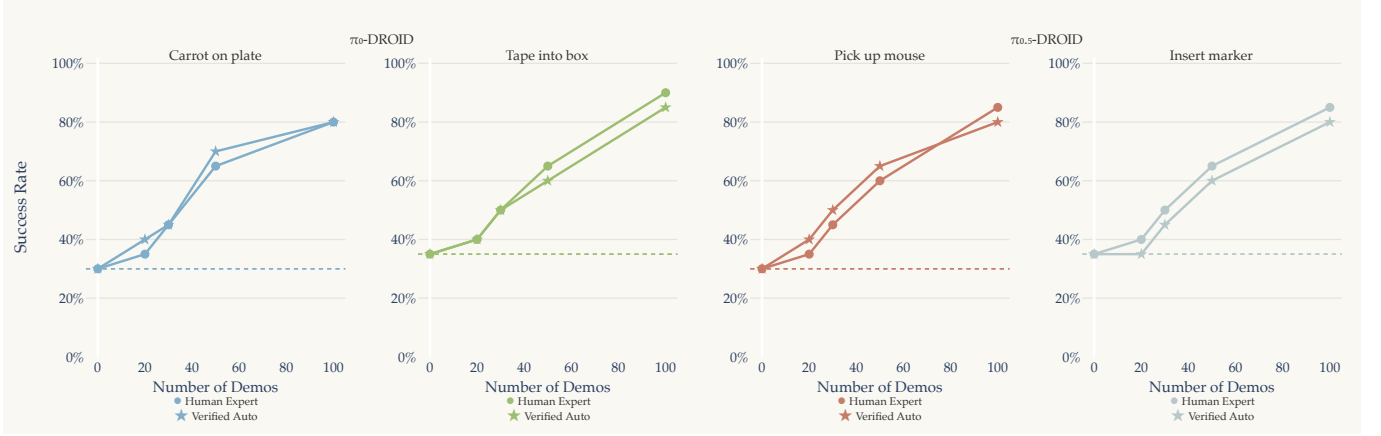


Fig. 7: Comparing data efficiency of autonomous verification v/s human expert. We fine-tune the base policy on datasets of 20, 30, 50, and 100 trajectories collected via either human expert teleoperation or verifier-curated autonomous rollouts, and evaluate each resulting policy over 20 real-world trials. Across all four tasks, verifier-curated data achieves comparable or superior data efficiency to human demonstrations. In multiple cases, autonomous data matches or outperforms human data at the same budget (e.g., 20–50 demos), and remains competitive even at larger scales. For example, on *Pick up mouse*, 50 autonomous demos outperform human demonstrations (0.65 vs. 0.60), while for *Carrot on plate*, performance is identical. These results indicate that verifier-curated autonomous collection can match the data efficiency of expert demonstrations, enabling scalable policy improvement without supervision.

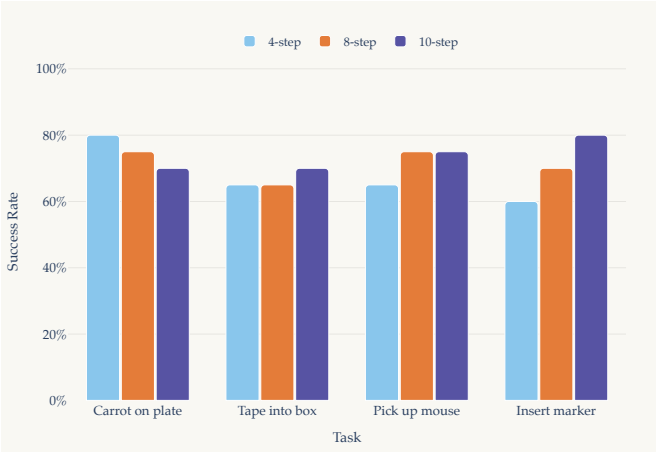


Fig. 8: We find that VERITAS performance does not significantly degrade with open-loop action execution of up to 10 steps.

policy improvement. By decoupling reasoning from action generation, we demonstrated that pre-trained generalist policies can achieve substantial performance gains purely through inference-time steering. We further showed that successfully verified trajectories can be distilled into the base policy via post-training to improve the policy performance in a data-efficient manner. Across a range of simulated and real-world benchmarks, we demonstrate consistent performance improvements with the VERITAS with 20–100 trajectories of experience. Most notably, we find that autonomous verification matches the data efficiency of human-collected expert trajectories. This creates a scalable pathway for robot self-improvement, where deployment experience naturally translates into persistent capability improvements.

While effective, our framework relies on trading inference compute for task performance, utilizing repeated sampling that may be computationally expensive in latency-critical applications. Additionally, our current verifier implementation depends on static visual traces generated at the onset of the task; while this is sufficient for quasi-static manipulation, it

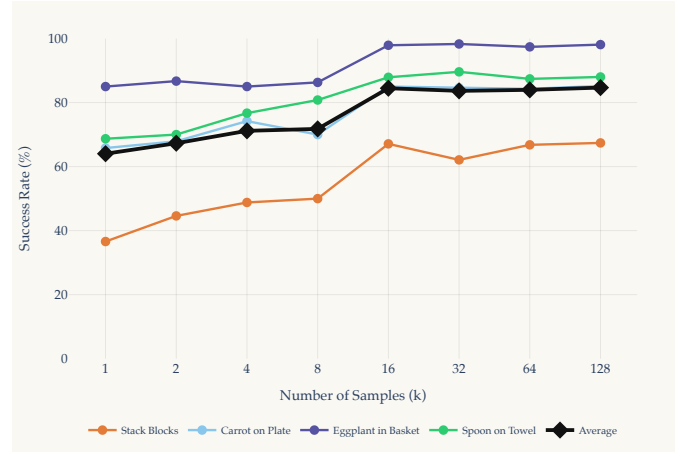


Fig. 9: We find that inference-time verification performance saturates beyond $N = 8$ samples, while the cost of sampling increases linearly with N .

may struggle in highly dynamic environments where the scene changes rapidly during execution. Future work could address these constraints by investigating mechanisms to distill the verifier’s rejection logic into a value function for faster inference, or by exploring joint optimization of the generator and verifier to improve sample efficiency during the search process. Further, it is important to note that our verifier can only improve performance by selecting the best action candidate among those proposed by the policy. This is fundamentally limited by the pre-trained exploration prior learned by the policy, and we expect improvements in policy pre-training to further improve the verification pipeline.

ACKNOWLEDGMENTS

This research was partially supported by the Toyota Research Institute and Microsoft Research, with compute support from Google TPU Research Cloud, NVIDIA Academic Grant Program, and Gemini Academic Program. The authors also thank Mitsuhiro Nakamoto, for help with reproducing V-GPS, and Samuel M. Bateman for helpful discussions.

REFERENCES

- [1] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint*, 2025.
- [2] Abby O’Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903, 2024.
- [3] Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, et al. A careful examination of large behavior models for multitask dexterous manipulation. *Science Robotics*, 2026.
- [4] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint*, 2025.
- [5] Physical Intelligence. π_0 : A vision-language-action flow model for general robot control, 2024.
- [6] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, 2023.
- [7] Yejin Kim, Wilbert Pumacay, Omar Rayyan, Max Argus, Winson Han, Eli VanderBilt, Jordi Salvador, Abhay Deshpande, Rose Hendrix, Snehal Jauhri, et al. MolmoSpaces: A large-scale open ecosystem for robot navigation and manipulation. In *Robotics: Science and Systems (RSS)*, 2026.
- [8] Abhay Deshpande, Maya Guru, Rose Hendrix, Snehal Jauhri, Ainaz Eftekhari, Rohun Tripathi, Max Argus, Jordi Salvador, Haoquan Fang, Matthew Wallingford, et al. MolmoBOT: Large-scale simulation enables zero-shot manipulation, 2026.
- [9] 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*, 2025.
- [10] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, et al. OpenVLA: An open-source vision-language-action model. *Annual Conference on Robot Learning (CoRL)*, 2024.
- [11] Noriaki Hirose, Catherine Glossop, Dhruv Shah, and Sergey Levine. OmniVLA: An omni-modal vision-language-action model for robot navigation. *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.
- [12] Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. GNM: A general navigation model to drive any robot. *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.
- [13] Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [14] Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more LLM calls all you need? towards the scaling properties of compound AI systems. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [15] Dacheng Li, Shiyi Cao, Chengkun Cao, Xiuyu Li, Shangyin Tan, Kurt Keutzer, Jiarong Xing, Joseph E Gonzalez, and Ion Stoica. S*: Test-time scaling for code generation. *arXiv preprint*, 2025.
- [16] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint*, 2021.
- [17] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *International Conference on Learning Representations (ICLR)*, 2023.
- [18] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:11809–11822, 2023.
- [19] Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. HG-DAGGER: Interactive imitation learning with human experts. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 8077–8083, 2019.
- [20] Zheyuan Hu, Robyn Wu, Naveen Enock, Jasmine Jiani Li, Riya Kadakia, Zackory Erickson, and Aviral Kumar. Rac: Robot learning for long-horizon tasks by scaling recovery and correction. In *Workshop on Making Sense of Data in Robotics: Composition, Curation, and Interpretability at Scale at CoRL 2025*, 2025.
- [21] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishika 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 *8th Annual Conference on Robot Learning*, 2024.
- [22] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ash-

- win Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems (RSS)*, 2024.
- [23] Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. *Science Robotics*, 2025.
- [24] Huihan Liu, Soroush Nasiriany, Lance Zhang, Zhiyao Bao, and Yuke Zhu. Robot learning on the job: Human-in-the-loop autonomy and learning during deployment. In *Robotics: Science and Systems (RSS)*, 2023.
- [25] Ruohan Zhang, Faraz Torabi, Lin Guan, Dana H Ballard, and Peter Stone. Leveraging human guidance for deep reinforcement learning tasks. *arXiv preprint*, 2019.
- [26] Anca D. Dragan and Siddhartha S. Srinivasa. A policy blending formalism for shared control. *International Journal of Robotics Research (IJRR)*, 32(7):790–805, June 2013.
- [27] Shervin Javdani, Siddhartha S. Srinivasa, and J. Andrew Bagnell. Shared autonomy via hindsight optimization. *Robotics: Science and Systems (RSS)*, 2015:10–15607, 2015.
- [28] Siddharth Reddy, Anca D Dragan, and Sergey Levine. Shared autonomy via deep reinforcement learning. In *Robotics: Science and Systems (RSS)*, 2014.
- [29] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.
- [30] Xiaomeng Xu, Yifan Hou, Zeyi Liu, and Shuran Song. Compliant residual DAGger: Improving real-world contact-rich manipulation with human corrections. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.
- [31] Jonathan Spencer, Sanjiban Choudhury, Matt Barnes, Matt Schmittle, Mung Chiang, Peter J. Ramadge, and Siddhartha S. Srinivasa. Learning from interventions: Human-robot interaction as both explicit and implicit feedback. *Robotics: Science and Systems (RSS)*, 2020.
- [32] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. In *8th Annual Conference on Robot Learning*, 2024.
- [33] Qingqing Zhao et al. CoT-VLA: Visual chain-of-thought reasoning for vision-language-action models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1702–1713, 2025.
- [34] Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. MolmoAct: Action reasoning models that can reason in space. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026.
- [35] Jacky Kwok, Christopher Agia, Rohan Sinha, Matt Foutter, Shulu Li, Ion Stoica, Azalia Mirhoseini, and Marco Pavone. Robomoney: Scaling test-time sampling and verification for vision-language-action models. In *9th Annual Conference on Robot Learning*, 2025.
- [36] Jacky Kwok, Xilun Zhang, Mengdi Xu, Yuejiang Liu, Azalia Mirhoseini, Chelsea Finn, and Marco Pavone. Scaling verification can be more effective than scaling policy learning for vision-language-action alignment. *arXiv preprint*, 2026.
- [37] Yilin Wu, Anqi Li, Tucker Hermans, Fabio Ramos, Andrea Bajcsy, and Claudia P’erez-D’Arpino. Do what you say: Steering vision-language-action models via runtime reasoning-action alignment verification. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026.
- [38] Mitsuhiko Nakamoto, Oier Mees, Aviral Kumar, and Sergey Levine. Steering your generalists: Improving robotic foundation models via value guidance. In *Conference on Robot Learning (CoRL)*, 2024.
- [39] Yilin Wu, Ran Tian, Gokul Swamy, and Andrea Bajcsy. From foresight to forethought: VLM-in-the-loop policy steering via latent alignment. In *Robotics: Science and Systems (RSS)*, 2025.
- [40] Jessie Yuan, Yilin Wu, and Andrea Bajcsy. When to act, ask, or learn: Uncertainty-aware policy steering. In *Robotics: Science and Systems (RSS)*, 2026.
- [41] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.
- [42] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023.
- [43] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.
- [44] Laura Smith, Alex Irpan, Montserrat Gonzalez Arenas, Sean Kirmani, Dmitry Kalashnikov, Dhruv Shah, and Ted Xiao. STEER: Flexible robotic manipulation via dense language grounding. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.
- [45] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities. *arXiv preprint*, 2025.
- [46] 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.
- [47] Allen Z. Ren. Open- π -zero: Re-implementation of π_0 vision-language-action (VLA) model from Physical Intelligence, 2025.

- [48] Physical Intelligence. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *9th Annual Conference on Robot Learning*, 2025.
- [49] Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, et al. PIVOT: Iterative visual prompting elicits actionable knowledge for VLMs. In *International Conference on Machine Learning (ICML)*, 2024.
- [50] Joseph A. Vincent, Haruki Nishimura, Masha Itkina, Paarth Shah, Mac Schwager, and Thomas Kollar. How generalizable is my behavior cloning policy? a statistical approach to trustworthy performance evaluation. *IEEE Robotics and Automation Letters (RA-L)*, 2024.