# Learning Long-Context Diffusion Policies via Past-Token Prediction

**Anonymous Author(s)**
Affiliation
Address
email

**Abstract:** Reasoning over long sequences of observations and actions is essential for many robotic tasks. Yet, learning effective long-context policies from demonstrations remains challenging. As context length increases, training becomes increasingly expensive due to rising memory demands, and policy performance often degrades as a result of spurious correlations. Recent methods typically sidestep these issues by truncating context length, discarding historical information that may be critical for subsequent decisions. In this paper, we propose an alternative approach that explicitly regularizes the retention of past information. We first revisit the copycat problem in imitation learning and identify an opposite challenge in recent diffusion policies: rather than over-relying on prior actions, they often fail to capture essential dependencies between past and future actions. To address this, we introduce Past-Token Prediction (PTP), an auxiliary task in which the policy learns to predict past action tokens alongside future ones. This regularization significantly improves temporal modeling in the policy head, with minimal reliance on visual representations. Building on this observation, we further introduce a multistage training strategy: pre-train the visual encoder with short contexts, and fine-tune the policy head using cached long-context embeddings. This strategy preserves the benefits of PTP while greatly reducing memory and computational overhead. Finally, we extend PTP into a self-verification mechanism at test time, enabling the policy to score and select candidates consistent with past actions during inference. Experiments across four real-world and six simulated tasks demonstrate that our proposed method improves the performance of long-context diffusion policies by 3× and accelerates policy training by more than 10×. Videos are available at https://ptp-robot.github.io.

## 1 Introduction

Many robotic tasks are inherently non-Markovian: an appropriate choice of action may depend not only on the current observation but also on past observations and actions [1–4]. For example, consider manipulation tasks where the robot arm occludes critical parts of the scene, or multi-stage tasks where early steps inform later strategies [5]. Likewise, past actions can prescribe a style of execution, such as speed, curvature, or strategy, that shapes how future actions should unfold [6, 7].

Despite the importance of historical observations, learning long-context robotic policies through imitation learning remains difficult. First, longer observation histories often introduce features that spuriously correlate with actions in the training data. Policies that latch onto such information may diverge from expert behavior during deployment, leading to performance degradation [8, 9]. Second, conditioning on high-dimensional image sequences imposes a rapidly growing memory and computation burden, making end-to-end training excessively expensive at scale [4, 10].

To cope with these challenges, recent methods typically limit the amount of historical information the policy sees – either by truncating the context length [6, 11] or by engineering past observations into compact representations, such as selecting key frames [12] and summarizing observations [4].

Submitted to the 9th Conference on Robot Learning (CoRL 2025). Do not distribute.

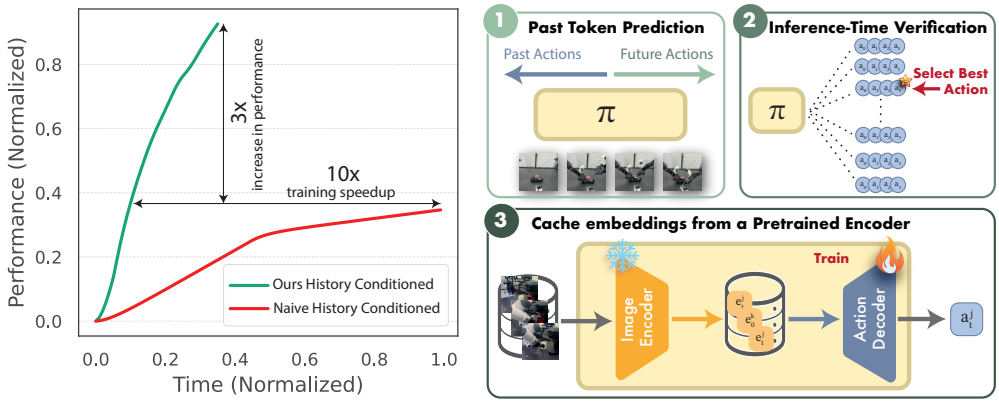


Figure 1: We propose a simple framework for learning long-context diffusion policies from human demonstrations. Our method leads to 3× gains in performance while reducing the training expense by more than 10×.

While these strategies reduce memory requirement, they risk discarding information critical to subsequent decisions.

In this paper, we introduce a simple and effective approach for learning long-context robot policies, illustrated in Fig. 1. At the core of our method is to explicitly regularize the information preserved from past observations. Specifically, we start with an analysis on the discrepancy between recent diffusion policies and their corresponding demonstrations [3, 6]. We observe that action sequences generated by learned policies often exhibit weaker temporal dependencies than those in expert data. To address this, we present past-token prediction (PTP), an auxiliary task where the policy learns to predict past actions alongside future ones. This regularizer encourages the model to attend more effectively to past context, significantly boosting performance. Crucially, we find that the benefits of PTP primarily emerge in the policy head for sequence modeling, rather than the visual encoder.

Building upon this analysis, we introduce a multi-stage training recipe: first, pre-train the visual encoder in a short-context setting, where the policy learns to predict a chunk of future actions from only a few past frames [2, 6], and subsequently fine-tune a long-context decoder that jointly predicts past and future actions from precomputed image embeddings. This design enables the policy to capture long-range temporal dependencies while substantially reducing memory and computational overhead. Beyond training, we further leverage PTP as a self-verification mechanism during inference. At each time step, the policy generates multiple candidate actions and selects the one most consistent with its previously executed actions.

In summary, our main contributions are twofold: (i) identify a critical discrepancy in temporal action dependencies between learned policies and expert demonstrations (§3), (ii) propose a training and inference method for long-context imitation learning via past-token prediction (§4). Empirically, we validate our method on diffusion-based policies [6] across six simulation and four real-world tasks (§5). On average, our method increases the success rate of long-context policies by 3× while reducing training overhead by over 10 times. Notably, it enables policies to achieve 80% success on history-critical tasks where existing methods fail entirely.

## 2 Related Work

**Imitation Learning.** Imitation learning has long served as a simple yet powerful paradigm for robot learning [13–15]. Early approaches typically framed it as a supervised learning problem, where the policy learns to map a given observation to the target action [16]. More recent works have shifted toward modeling the distribution of demonstrations [2, 3, 6, 7, 17–20]. This approach has recently achieved remarkable success towards generalist robot policies [21, 22]. However, imitation learning remains highly susceptible to covariate shift [23–25], e.g. Ross et al. [26] and Spencer et al. [9] characterize compounding errors in a feedback loop once the learned policy diverges from the

demonstration manifold. This problem is exacerbated by high-dimensional visual inputs, where less robust features might be learned due to underspecification [27]. Notably, recent works [4, 6] have empirically found that image-conditioned specialist and generalist policies degrade with history, leading many works to exclude history altogether [2, 22, 28–32]. Our work introduces and analyzes a training recipe that counteracts this degradation.

**Long-Context Policies.** Handling long sequences of high-dimensional observations has been a persistent challenge for robot learning. Many prior works mitigate it by discarding parts of the past, either through regularization strategies such as adversarial objectives [24] or information bottlenecks [33], or by summarizing trajectories using keyframes [12] or motion tracks [34]. Similarly, methods like sketch synthesis [35] and visual trace prompting [4] have been explored for generalist robot policies. These approaches often rely on assumptions about the irrelevance of specific parts of the history, which may not hold in complex tasks. In contrast, we propose a method that directly regularizes diffusion policies to retain information about past actions that would otherwise be lost from historical context.

**Test-Time Scaling.** Recent research in language modeling, image generation, and robotics has shown that inference-time compute may allow models to improve their performance [36–38]. Some seek to build an additional verifier to re-rank the output samples [39–42], while others propose to leverage the internal knowledge to improve reasoning through self-verification [43]. Our method echoes the latter paradigm in the robotic context: our policy is trained to predict accurate past actions before predicting the present action and can self-verify at test-time through past action accuracy. Similarly to how it may be more compute-efficient to use test-time compute on a small LLM [44], we show that checkpoints trained for fewer epochs or at shorter histories can approach the performance of optimal checkpoints by using more test-time compute.

## 3 Preliminaries

**Problem Setting.** We consider the problem of imitation learning, where a robot learns to perform complex tasks from expert demonstrations. At each time step $t$, the robot receives a visual observation $o_t$ and executes an action $a_t$. Crucially, we assume that each observation $o_t$ contains only partial information about the underlying state $s_t$, but the complete information about $s_t$ can be inferred from the history of observations. This setting encapsulates practical challenges commonly encountered in robotic tasks, such as latent strategies in the demonstrations (e.g., expert preference), temporal context (e.g., stage within a task), and perceptual limitations (e.g., visual occlusions).

Given a dataset of $N$ expert demonstrations $\mathcal{D} = \{\tau_i\}_{i=1}^{N}$, where each demonstration trajectory $\tau_i$ consists of a sequence of observation-action pairs, our goal is to learn a long-context policy $\pi_\theta(\mathbf{a}_{t:t+l}|\mathbf{o}_{t-k:t})$ that takes as input the current observation along with the history $\mathbf{o}_{t-k:t} = (o_{t-k}, \ldots, o_t)$ over the past $k$ time steps, and predicts the current and future actions $\mathbf{a}_{t:t+l} = (a_t, \ldots, a_{t+l})$ spanning the next $l$ time steps. While increasing the context length $k$ provides richer historical information, the resulting long-context policies often suffer from substantial performance declines [4, 6].

**Practical Challenges.** One central challenge in long-context imitation learning arises from the prevalence of spurious features in observation history. As context length increases, the model is exposed to a growing set of input features, some of which correlate with but do not causally influence the expert actions. Policies relying on these spurious features in observation history may reach high prediction accuracy within the training distribution but generalize poorly during deployment [8]. One notable manifestation is the copycat behavior [24], where the learned policy simply mimics previous actions as predictions for future ones, ignoring current state observations. Does this phenomenon persist in modern imitation learning methods?

To understand this, we evaluate temporal action dependencies by measuring how predictable the current action is from prior actions alone. Specifically, given a set of demonstrations, we first train long-context policies with varying observation history lengths. We then collect policy rollouts and

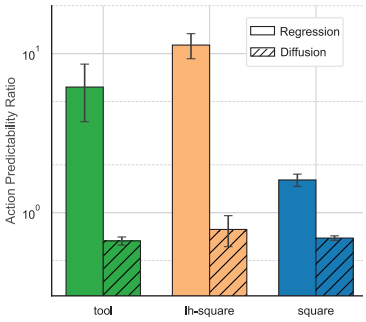


Figure 2: Comparison of regression-based and diffusion-based policies in temporal action dependency, normalized by that in demonstrations.

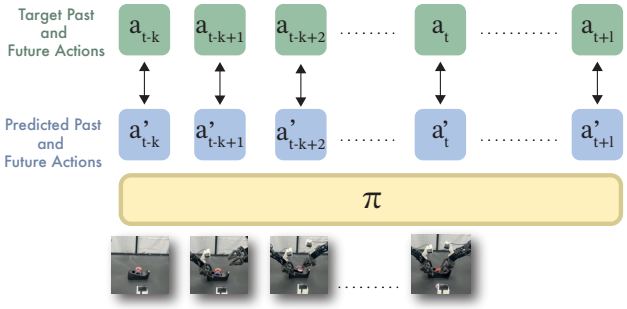


Figure 3: Illustration of past-token prediction. The policy head is trained to jointly predict both past and future action tokens, encouraging the model to capture the temporal dependencies that are otherwise lost between past and future actions.

train a simple two-layer MLP $\phi(a_t|a_{t-1})$ to predict the current action based solely on the previous action. We measure the mean-squared error $\epsilon_\pi$ of the MLP predictor on holdout rollouts and similarly obtain $\epsilon_{\pi^*}$ for expert demonstrations. Following [24, 45], we define the action predictability ratio as $\epsilon_{\pi^*}/\epsilon_\pi$. Intuitively, a ratio greater than 1 indicates an over-reliance on previous actions (i.e., copycat behavior), while a ratio less than 1 indicates weaker-than-expert action dependency.

Fig. 6 shows the action predictability ratios for classical regression-based policies and modern diffusion-based policies [6] across three simulation tasks. Interestingly, the two approaches exhibit opposite failure modes: The regression-based policies indeed exhibit high action predictability, even exceeding that of the expert demonstrations. In contrast, *modern diffusion-based policies yield predictability ratios significantly below 1, indicating a surprising underuse of past action information despite conditioning on long observation histories.* Ideally, an effective imitator should not only learn to accurately predict expert actions in the training set, but also reach a similar level of temporal action dependencies in its rollouts. We will next introduce a method designed to explicitly bridge this gap.

# 4 Method

In this section, we introduce a long-context imitation learning method, aiming to improve both policy performance and training efficiency. We will first describe a simple but crucial auxiliary task to enhance temporal dependencies in sequential decision-making (§4.1). We will then present a multi-stage training recipe that preserves the benefit of this auxiliary task while reducing memory consumption (§4.2). Finally, we will introduce an inference technique that leverages the auxiliary task to effectively self-verify sampled predictions at test time (§4.3).

## 4.1 Past-Token Prediction

One common design choice in imitation learning is next-token prediction, where the policy predicts only the immediate next action token at each time step. To better capture temporal dependencies, recent methods have extended this to predict a chunk of future action tokens [2, 6]. However, as shown in §3, this design alone remains insufficient for modeling the critical dependencies between past and future decisions.

We address this issue through Past-Token Prediction (PTP), an auxiliary objective that tasks the policy to predict past action tokens alongside future ones. Formally, given a sequence of observations $\mathbf{o}_{t-k:t}$, the policy is trained to jointly predict the action tokens from the past time step $t-k$ to the upcoming time step $t+l$:

$$\hat{\mathbf{a}}_{t-k:t+l} = \pi_\theta(\mathbf{o}_{t-k:t}). \tag{1}$$

As illustrated in Fig. 3, this objective expands the prediction window in both temporal directions, explicitly encouraging the policy to preserve information about past actions from the history context.

### 4.2 Memory-Efficient Training with PTP

Recent imitation learning approaches typically train visuomotor policies end-to-end, jointly optimizing both the visual encoder and the policy head. However, this strategy incurs memory costs that grow linearly with context length, making it prohibitively expensive to train long-context policies.

To address this, we propose a multi-stage training recipe that decouples visual representation learning from policy optimization. Our training process consists of three specific stages:

1. **Encoder Training:** We first train the visual encoder with a short observation context but a long prediction horizon, encouraging it to extract representations that retain information critical for predicting $l$ subsequent steps.
2. **Feature Caching:** We then freeze the encoder and precompute embeddings for all frames in the training set. This caching step eliminates redundant computation during policy training.
3. **Policy Training:** Finally, we train the policy head conditioned on long-context observations represented by the cached embeddings. This enables the policy to model long-range dependencies without repeatedly processing visual inputs.

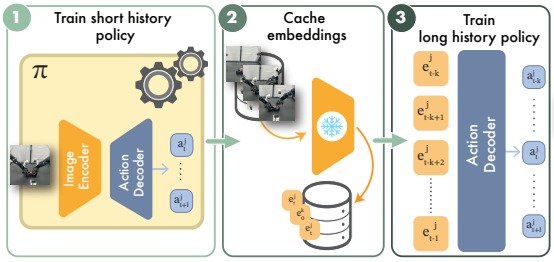


Figure 4: Overview of multistage training with embedding caching. As PTP acts on the decoder, caching embeddings substantially improves inference speed without sacrificing performance. We use a visual encoder from a short-range policy with low validation loss to compute the embeddings of the images in the buffer and cache them in the buffer. With the cached embeddings we can train the long-horizon policy much faster. At test time we take the original encoder.

As shown in Fig. 4, this multistage training approach retains a computational footprint similar to short-context training while enabling efficient scaling to longer observation contexts. In Appendix A.1, we show in more detail how the features of a short-history policy are sufficient to support strong long-context performance.

### 4.3 Test-Time Verification with PTP

Another common challenge in recent diffusion policies lies in the robustness of sampled predictions. Often, not all samples are equally good at capturing the critical temporal dependencies. Recent work has explored re-ranking sampled predictions based on consistency with past predictions [7]. However, when the previous prediction for future actions is suboptimal, e.g. because of unexpected environmental changes, this approach may propagate errors rather than correct them.

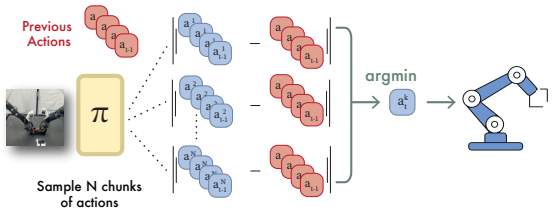


Figure 5: Test-time verification. Multiple action sequences are sampled from the same observation, and the policy selects the sequence that is most consistent compared to ground-truth previous actions.

To address this shortcoming, we cast Past-Token Prediction as a self-verification mechanism during deployment. At each inference step, we sample a batch of $B$ candidate action sequences:

$$\mathcal{A} = \{\hat{\mathbf{a}}^{(1)}, \ldots, \hat{\mathbf{a}}^{(B)}\}, \quad \hat{\mathbf{a}}^{(i)} \sim \pi_\theta(\mathbf{o}_{t-k:t}), \tag{2}$$

where each sampled candidate $\hat{\mathbf{a}}^{(i)} = (a_{t-k}, \ldots, a_{t+l})^{(i)}$ includes both reconstructed past actions and predicted future actions. Since the first $k-1$ actions have already been executed, we use them as a ground-truth reference and select the candidate whose reconstructed past actions best match the executed ones:

$$\hat{\mathbf{a}}^* = \arg\min_{\hat{\mathbf{a}} \in \mathcal{A}} \sum_{\tau=t-k}^{t-1} \|\hat{a}_\tau - a_\tau\|^2 \tag{3}$$

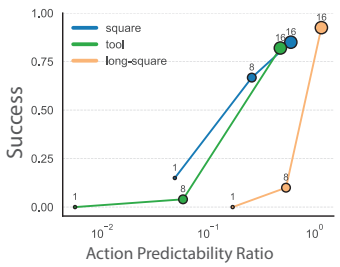


Figure 6: Effect of PTP on temporal action dependency and policy performance. Increasing the amount of past-token supervision aligns the learner more closely with expert action dependencies, resulting in higher success rates.

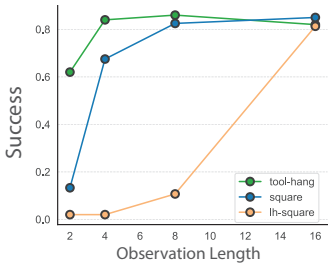


Figure 7: Effect of history observations on PTP-trained diffusion policies. Increasing the context length progressively enhances policy performance, especially in history-critical tasks such as Long-Horizon Square (see Fig. 15).

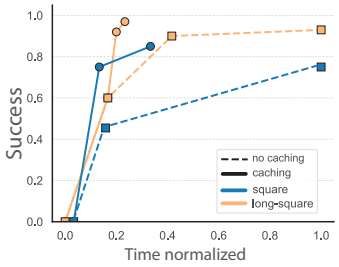


Figure 8: Effect of feature caching. Caching speeds up training by over 5× without hurting performance. On complex tasks like Tool Hang, long-context policies fail to perform even after two days without caching.

As illustrated in Fig. 5, this sample selection procedure is fully parallelizable on GPU devices, enabling self-verification of temporal action dependencies with minimal computational overhead.

# 5 Experiments

In this section, we evaluate the proposed method for learning long-context diffusion policies. We seek to answer the following questions regarding policy performance and training efficiency:

1. How effectively does PTP mitigate the lack of temporal action dependencies shown in §3?
2. How well do the resulting policies perform on tasks that require history-aware decision-making?
3. To what extent does the proposed multi-stage training recipe accelerate policy learning?
4. Could PTP verification further mitigate deficiencies in temporal dependencies at test time?
5. Finally, how do these findings generalize to history-critical tasks in the real world?

To this end, we evaluate our method on the modern diffusion-based policy [6], in comparison with the classical regression-based policy. By default, both policies receive visual and proprioceptive observations from the past 16 time steps as conditional input. We compare policies trained with *PTP* against two baselines: *no-history* policies that take only the current and past single frame as input, and *no-PTP* policies that are trained without PTP. Unless otherwise specified, all policies are trained using the multistage recipe with feature caching and evaluated under a single-sample inference setting. The effect of test-time verification is evaluated separately across multiple checkpoints under varying sample budgets. Additional results are presented in Appendix A, with implementation details provided in Appendix B.

## 5.1 Simulation Experiments

We first evaluate our method across six simulated tasks. Four of these are sourced from existing benchmarks: *square, tool hang, and transport* from RoboMimic [1], each provided by multi-human demonstration datasets, and Push-T from Chi et al. [6]. These tasks feature diverse strategies in demonstrations, requiring the policy to infer and commit to consistent behaviors over time based on historical context. In addition, we introduce two new long-horizon simulation tasks: *long-horizon square*, where the robot must place and remove a square onto the peg twice before finally dropping it in the peg; and *long-horizon aloha*, where one arm must pick up a block, move it to the center of the field of view, and return it precisely to its original location. Success in these new tasks critically depends on the ability to recall and act upon information observed earlier in the episode. Each policy-task pair is evaluated over 100 episodes across three random seeds. We next summarize the key findings from these simulation experiments.

***Takeaway 1: PTP mitigates deficiencies in modeling temporal action dependencies.*** To validate the effect of PTP on modeling temporal action dependencies, we use the same set of tasks as in §3

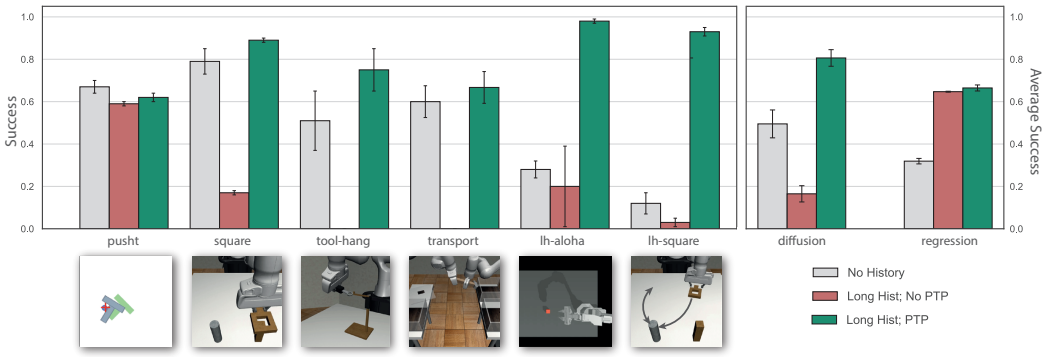


Figure 9: Comparison of different policies across six simulation tasks. Unlike classical regression-based policies, modern diffusion-based policies exhibit a clear drop in performance when conditioned on historical observations. Our method achieves an average improvement of over 30% compared to no-history diffusion policies, and over 60% compared to no-PTP diffusion policies. The gains are especially pronounced on history-critical tasks such as *long-horizon aloha* and *long-horizon square*.

and train policies to predict a variable number of past tokens $\{\hat{a}_{t-c-1}, \ldots, \hat{a}_t\}$, where $c$ denotes the number of actions included in the prediction target. Specifically, we compare three variants: (i) *no-PTP* with $c = 1$, equivalent to the vanilla next-token prediction baseline; (ii) *half-PTP* with $c = 8$, which predicts action tokens corresponding to half the observation window; and (iii) *full-PTP* with $c = 16$. As shown in Fig. 6, PTP consistently increases the action predictability and gets closer to that observed in the expert demonstrations. Notably, the non-PTP baseline exhibits approximately 10× to 100× weaker action predictability ratios compared to expert behavior, whereas full-PTP yields temporal dependencies comparable to demonstrations.

***Takeaway 2: PTP significantly improves the performance of modern policies.*** To assess the impact of PTP on task performance, we compare our method against the no-history and no-PTP baselines on two classes of policies: diffusion-based versus regression-based. All models are evaluated following the protocol from [6], with action chunking set to 8 time steps. As shown in Fig. 9, while the *no-history* baseline already performs competitively on some existing tasks, PTP matches or surpasses its performance. The advantage of PTP is particularly pronounced in long-horizon tasks: both the *no-history* and *no-PTP* baselines struggle with success rates below 30%, whereas our method achieves near-perfect performance on the long-horizon tasks. Averaged across all six simulation tasks, PTP yields an average 50% improvement for diffusion-based policies when conditioned on long contexts, and outperforms the regression counterpart by nearly 20%.

***Takeaway 3: PTP-trained policies benefit from longer contexts.*** To further understand the role of historical contexts, we evaluate PTP-trained diffusion-based policies conditioned on observation histories of varying lengths, ranging from 2 to 16 time steps. As shown in Fig. 7, longer histories generally lead to improved performance. For relatively simple tasks such as *square*, gains tend to saturate beyond 4 steps; however, for more complex tasks, such as *transport*, *long-horizon square*, and *long-horizon aloha*, longer contexts provide substantial performance boosts.

***Takeaway 4: Embedding caching accelerates PTP training without sacrificing performance.*** To assess the effectiveness of the proposed multistage training strategy, we train history-conditioned diffusion policies with and without embedding caching for two days on the three tasks used above (§3), evaluating checkpoints saved every 50 epochs. As shown in Fig. 8, the vanilla training recipe without caching completes only a limited number of epochs within the time budget. In contrast, our caching-based approach matches performance in just 20% of the training time and surpasses it within 40% of the compute budget.

***Takeaway 5: PTP verification boosts performance in challenging settings at test time.*** To validate the potential of self-verification through PTP, we evaluate history-conditioned policies on three challenging tasks, including Tool Hang, Transport, and Long Square, trained under constrained compute budgets and tested with varying sampling budgets $\{1, 3, 5, 10\}$. As shown in Fig. 10, PTP-guided

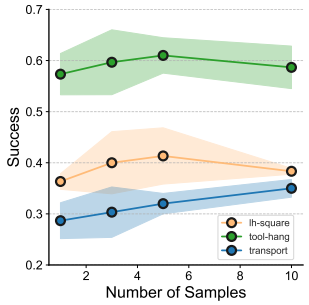


Figure 10: Effect of PTP self-verification. Increasing sampling budgets yields a 5% gain in challenging closed-loop settings.

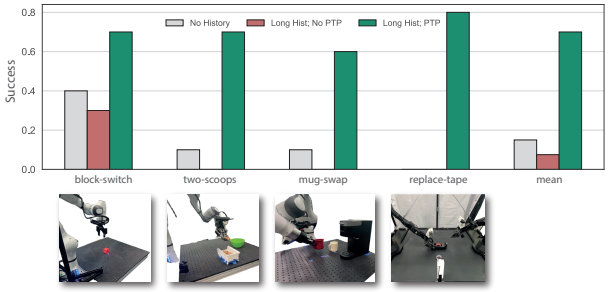


Figure 11: Comparison of different policies on four real-world tasks that critically depend on historical context. Our method yields over a 55% improvement compared to baselines.

sample selection provides notable performance gains. Notably, increasing the number of sampled candidates from 1 to 5 results in approximately 5% improvement in success rate on all these tasks.

### 5.2 Real-World Experiments

We next examine our method on four history-critical tasks across two robot platforms in the real world: *Franka block switch*: move a block from one side to another, where history is needed to correctly infer which side to place the block; *Franka two scoops*, transport two scoops to the target, where history is needed to count scoops; *Franka mug replacement* and *ALOHA tape replacement*: replace one mug or tape by another, where history is needed to distinguish old and new objects. Across all tasks, we use diffusion-based policies with a context length of 16 and a chunk size of 8. Due to different ranges of temporal dependency in these tasks, we apply task-specific subsampling rates detailed in Appendix B.

***Quantitatively, PTP outperforms baselines by over 4× in the real world.*** As shown in Fig. 11, the *no-history* baseline is limited to an average success rate of 15% due to the absence of critical history information. The *no-PTP* baseline, which simply conditions on history without PTP, yields near-zero success on three of four tasks. In contrast, our method achieves an average 70% success rate. Notably, on Tape Replacement, one of the most challenging tasks across the board, our method achieves 80% success, while the two baselines fail entirely.

***Qualitatively, PTP-trained long-context policies excel at both high-level and low-level memory.*** As shown in the videos[1], the two baselines exhibit distinct failure modes: the *no-history* policies often fail at high-level decision-making, such as replacing the wrong object or miscounting scoops, whereas the *no-PTP* baseline struggles with low-level motor control, such as unsuccessful grasps and inaccurate placements. In comparison, policies trained with our method demonstrate improvement in both high-level planning and low-level control, resulting in more coherent and reliable behavior across tasks.

## 6 Conclusion

We have presented Past Token Prediction (PTP), a simple yet effective auxiliary objective for learning history-conditioned diffusion policies from demonstrations. We have shown that PTP can effectively strengthen temporal action dependencies that are often lost in recent diffusion policies. In addition, we have introduced a multistage training strategy and a self-verification mechanism that allow for effective use of PTP during both training and inference. Extensive experiments across ten manipulation tasks in both simulations and the real world demonstrate its advantages in efficiency and effectiveness.

[1] Vidoes at https://ptp-robot.github.io/

## 7 Limitations and Discussion.

Our work has focused on extending context length specifically for diffusion policies, motivated by their growing prevalence in the robot learning community. Nevertheless, the effectiveness of our method may generalize to other classes of modern policies as well. In fact, concurrently with our work, Vuong et al. [46] observes similar challenges in tokenization-based policies. Extending our approach to such settings, and more broadly, designing action tokenizers that better preserve temporal structure, can be an exciting avenue for future research.

Another practical challenge our method faces is inference overhead. While we have shown that caching and reusing visual embeddings can substantially reduce memory consumption and speed up policy training, inference overhead remains a practical bottleneck for closed-loop operations. To make inference time manageable, we followed common practices from recent literature by down-sampling observation history and extending action chunk. However, these adjustments are known to compromise policy reactivity. Designing strategies to further accelerate inference—particularly given the growing scale of VLA models—could be another fruitful direction for future research.

## References


[1] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín. What Matters in Learning from Offline Human Demonstrations for Robot Manipulation. In *Proceedings of the 5th Conference on Robot Learning*, pages 1678–1690. PMLR, Jan. 2022.

[2] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware, Apr. 2023.

[3] S. Lee, Y. Wang, H. Etukuru, H. J. Kim, N. M. M. Shafiullah, and L. Pinto. Behavior Generation with Latent Actions, Mar. 2024.

[4] R. Zheng, Y. Liang, S. Huang, J. Gao, H. D. III, A. Kolobov, F. Huang, and J. Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies, 2024. URL https://arxiv.org/abs/2412.10345.

[5] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu. RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots.

[6] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Robotics: Science and Systems XIX*. Robotics: Science and Systems Foundation, July 2023. ISBN 978-0-9923747-9-2.

[7] Y. Liu, J. I. Hamid, A. Xie, Y. Lee, M. Du, and C. Finn. Bidirectional Decoding: Improving Action Chunking via Closed-Loop Resampling, Dec. 2024.

[8] P. de Haan, D. Jayaraman, and S. Levine. Causal Confusion in Imitation Learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[9] J. Spencer, S. Choudhury, A. Venkatraman, B. Ziebart, and J. A. Bagnell. Feedback in imitation learning: The three regimes of covariate shift, 2021. URL https://arxiv.org/abs/2102.02872.

[10] X. Li, P. Li, M. Liu, D. Wang, J. Liu, B. Kang, X. Ma, T. Kong, H. Zhang, and H. Liu. Towards Generalist Robot Policies: What Matters in Building Vision-Language-Action Models, Dec. 2024.

[11] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky. $\pi_0$: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

[12] C. Wen, J. Lin, J. Qian, Y. Gao, and D. Jayaraman. Keyframe-focused visual imitation learning, 2021. URL https://arxiv.org/abs/2106.06452.

[13] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, May 2009.

[14] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard. Recent Advances in Robot Learning from Demonstration. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(Volume 3, 2020):297–330, May 2020.

[15] M. Zare, P. M. Kebria, A. Khosravi, and S. Nahavandi. A Survey of Imitation Learning: Algorithms, Recent Developments, and Challenges. *IEEE Transactions on Cybernetics*, 54 (12):7173–7186, Dec. 2024.

[16] S. Ross and D. Bagnell. Efficient Reductions for Imitation Learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 661–668. JMLR Workshop and Conference Proceedings, Mar. 2010.

[17] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu. 3D Diffusion Policy, Mar. 2024.

[18] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar. RoboAgent: Generalization and Efficiency in Robot Manipulation via Semantic Augmentations and Action Chunking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4788–4795, May 2024.

[19] D. Wang, S. Hart, D. Surovik, T. Kelestemur, H. Huang, H. Zhao, M. Yeatman, J. Wang, R. Walters, and R. Platt. Equivariant Diffusion Policy. In *8th Annual Conference on Robot Learning*, Sept. 2024.

[20] S. Haldar, Z. Peng, and L. Pinto. BAKU: An Efficient Transformer for Multi-Task Policy Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, Nov. 2024.

[21] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. RT-1: Robotics Transformer for Real-World Control at Scale. In *Robotics: Science and Systems XIX*. Robotics: Science and Systems Foundation, July 2023. ISBN 978-0-9923747-9-2. doi:10.15607/RSS.2023.XIX.025. URL http://www.roboticsproceedings.org/rss19/p025.pdf.

[22] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky. $\pi_0$: A Vision-Language-Action Flow Model for General Robot Control, Nov. 2024.

[23] P. de Haan, D. Jayaraman, and S. Levine. Causal confusion in imitation learning, 2019. URL https://arxiv.org/abs/1905.11979.

[24] C. Wen, J. Lin, T. Darrell, D. Jayaraman, and Y. Gao. Fighting Copycat Agents in Behavioral Cloning from Observation Histories. In *Advances in Neural Information Processing Systems*, volume 33, pages 2564–2575. Curran Associates, Inc., 2020.

[25] D. Shao, T. K. Buening, and M. Kwiatkowska. A Unifying Framework for Causal Imitation Learning with Hidden Confounders, Feb. 2025.

[26] S. Ross, G. J. Gordon, and J. A. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning, 2011. URL https://arxiv.org/abs/1011.0686.

[27] S. Nasiriany, S. Kirmani, T. Ding, L. Smith, Y. Zhu, D. Driess, D. Sadigh, and T. Xiao. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation, 2024. URL https://arxiv.org/abs/2411.02704.

[28] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo, Y. L. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine. Octo: An Open-Source Generalist Robot Policy, May 2024.

[29] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, July 2023.

[30] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn. OpenVLA: An Open-Source Vision-Language-Action Model. In *8th Annual Conference on Robot Learning*, Sept. 2024.

[31] M. Torne, A. Jain, J. Yuan, V. Macha, L. Ankile, A. Simeonov, P. Agrawal, and A. Gupta. Robot Learning with Super-Linear Scaling, Dec. 2024.

[32] Y. Li, Y. Deng, J. Zhang, J. Jang, M. Memmel, C. R. Garrett, F. Ramos, D. Fox, A. Li, A. Gupta, and A. Goyal. HAMSTER: Hierarchical Action Models for Open-World Robot Manipulation. In *The Thirteenth International Conference on Learning Representations*, Oct. 2024.

[33] S. Seo, H. Hwang, H. Yang, and K.-E. Kim. Regularized behavior cloning for blocking the leakage of past action information. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 2128–2153. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/06b71ad997f7e3e4b2e2f2ea12e5a759-Paper-Conference.pdf.

[34] J. Ren, P. Sundaresan, D. Sadigh, S. Choudhury, and J. Bohg. Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning, 2025. URL https://arxiv.org/abs/2501.06994.

[35] P. Sundaresan, Q. Vuong, J. Gu, P. Xu, T. Xiao, S. Kirmani, T. Yu, M. Stark, A. Jain, K. Hausman, D. Sadigh, J. Bohg, and S. Schaal. Rt-sketch: Goal-conditioned imitation learning from hand-drawn sketches, 2024. URL https://arxiv.org/abs/2403.02709.

[36] H. Bansal, A. Hosseini, R. Agarwal, V. Q. Tran, and M. Kazemi. Smaller, Weaker, Yet Better: Training LLM Reasoners via Compute-Optimal Sampling, Aug. 2024.

[37] N. Ma, S. Tong, H. Jia, H. Hu, Y.-C. Su, M. Zhang, X. Yang, Y. Li, T. Jaakkola, X. Jia, and S. Xie. Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps, Jan. 2025.

[38] M. Nakamoto, O. Mees, A. Kumar, and S. Levine. Steering Your Generalists: Improving Robotic Foundation Models via Value Guidance. In *8th Annual Conference on Robot Learning*, Sept. 2024.

[39] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training Verifiers to Solve Math Word Problems, Nov. 2021.

[40] Y. Weng, M. Zhu, F. Xia, B. Li, S. He, S. Liu, B. Sun, K. Liu, and J. Zhao. Large Language Models are Better Reasoners with Self-Verification. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, Dec. 2023.

[41] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's Verify Step by Step. In *The Twelfth International Conference on Learning Representations*, Oct. 2023.

[42] F. Yu, A. Gao, and B. Wang. OVM, Outcome-supervised Value Models for Planning in Mathematical Reasoning. In K. Duh, H. Gomez, and S. Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 858–875, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[43] K. Stechly, K. Valmeekam, and S. Kambhampati. On the Self-Verification Limitations of Large Language Models on Reasoning and Planning Tasks, Aug. 2024.

[44] C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

[45] S. Seo, H. Hwang, H. Yang, and K.-E. Kim. Regularized Behavior Cloning for Blocking the Leakage of Past Action Information. *Advances in Neural Information Processing Systems*, 36: 2128–2153, Dec. 2023.

[46] A. D. Vuong, M. N. Vu, D. An, and I. Reid. Action Tokenizer Matters in In-Context Imitation Learning, Mar. 2025.