Short-to-Long Distillation: Learning Long-Context Policies from Short-Context Supervision

Anonymous Author(s)
Affiliation
Address
email

Abstract

Consistency and reactivity are two essential properties for robotic policies. Yet, recent methods often trade one for the other: using long action chunks and numerous denoising steps to improve intra-chunk consistency, at the cost of lower inference frequency and slower inference speed. In this paper, we first revisit the necessity of these design choices through the lens of data scaling. We find that with sufficient training data, extending history action contexts can substitute for future action chunks, without compromising performance; moreover, conditioning on longer contexts reduces action ambiguity, lessening the need for iterative denoising. Motivated by these observations, we introduce Short-to-Long Distillation, a policy distillation approach that learns a long-context few-step student policy from synthetic data generated by a short-context many-step teacher policy. Central to our approach are two data curation strategies: (i) on-policy noise injection to broaden the coverage of action contexts, and (ii) mode-seeking chunk optimization to sharpen the distribution of action labels. Empirically, our method achieves strong results on diffusion policies across Push-T and RoboMimic tasks. Notably, using only 1k distilled sequences, the student policies match their teachers in static settings and surpass them by up to 40% in stochastic environments. Our results suggest the promise of synthetic data as a scalable alternative to inductive biases for robot learning.

1 Introduction

Humans solve physical tasks with both consistency and reactivity—producing coherent long-horizon behaviors while adapting rapidly to unexpected changes. Robots should possess the same capabilities, yet learning such policies from human demonstrations remains challenging. Often, collected demonstration datasets exhibit three properties: long temporal dependencies, large behavioral variations, and limited state coverage, resulting in a highly multimodal yet sparsely covered distribution over valid action sequences [7, 10]. To address these challenges, recent policies are typically built with: (i) action chunking - predict and execute a chunk of future actions over multiple steps [6, 20], (ii) action denoising – refine an action chunk from random noise through multiple steps [3, 6]. While these designs improve consistency within a chunk, they inherently compromise reactivity: larger chunks lower inference frequency and multi-step denoising slows inference speed.

To enhance reactivity, recent studies have explored three main approaches. First, guided inference improves cross-chunk consistency during inference [1, 4, 11], reducing the reliance on large chunks. Second, hierarchical policies introduce low-level controllers beneath high-level chunk predictors [2, 12, 19], enabling rapid local adaptation. Third, accelerated denoising reduces the number of refinement steps [9, 15, 17]. Despite these advances, existing approaches remain tied to multi-step chunking and denoising, imposing an inherent trade-off between long-term consistency and short-term reactivity.

In this work, we first revisit the necessity of action chunking and action denoising. We hypothesize that these design choices are not intrinsic requirements of robotic policies, but rather artifacts of

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

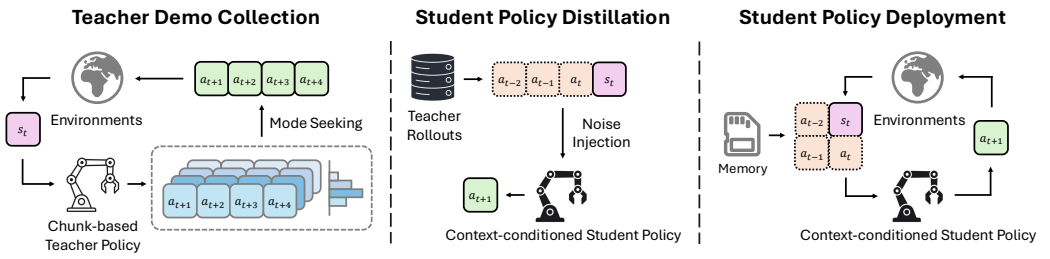


Figure 1: Illustration of Short-to-Long Distillation: A chunk-based teacher policy first interacts with the environment to generate demonstrations using mode seeking. The context-conditioned student policy is then trained on stored teacher rollouts, augmented with context noise injection to expand action coverage. At deployment, the student conditions on longer context actions while predicting shorter chunks, achieving higher reactivity without compromising performance or consistency.

limited data. We thus conduct a set of controlled experiments by scaling demonstration data up to $10\times$ larger per task than typical academic datasets. Interestingly, we find that with sufficient training data, extending history action context can replace long action chunks, without sacrificing performance. Moreover, conditioning on longer contexts reduces action ambiguity, substantially decreasing the need for iterative denoising. Despite these promising outcomes, acquiring human demonstrations at such a scale is costly and even impractical. This raises a natural question: Can we leverage existing policies, rather than human supervision, as data generators to reach this critical scale?

To this end, we introduce Short-to-Long Distillation (S2L), a policy distillation framework for learning performant long-context few-step policies. S2L proceeds in two stages: we first train a short-context, long-chunk, many-step teacher policy on human demonstrations to capture consistency, and then distill its knowledge into a long-context, short-chunk, few-step student policy using synthetic rollouts generated by the teacher. Unlike conventional distillation methods [8, 16, 18], S2L operates in a heterogeneous setting where teacher and student differ in both input and output spaces, precluding direct one-to-one supervision. To bridge this structural mismatch, we propose two data curation strategies: (i) on-policy noise injection to expand the coverage of action contexts, and (ii) mode-seeking chunk optimization to sharpen the distribution of action labels. Together, these strategies enable the learned student to inherit the consistency strength of the teacher policy while discarding its reliance on chunking and denoising, thereby achieving substantially higher reactivity.

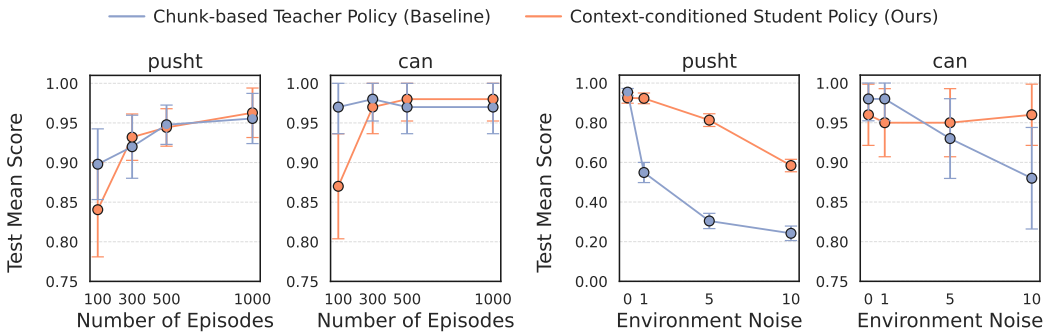
The main contributions of this paper are twofold: an analysis revealing how data scale influences policy design (§2) and a distillation method for learning long-context, few-step policies from existing short-context multi-step ones (§3). Empirically, we evaluate our method on diffusion policies across Push-T and RoboMimic tasks. Using only 1k distilled sequences, the student policies achieve performance comparable to their teachers in static environments and exceed them by up to 40% in stochastic environments. Our results suggest the promise of synthetic data curation as a scalable alternative to inductive biases for robot learning.

2 Analysis: Effect of Data Scaling on Policy Design

Problem Setup. We study imitation learning from a finite set of demonstrations $\mathcal{D} = \{(s_t, a_t)\}_{t=1}^{T}$. At each time step t, the policy conditions on the current observation s_t and the previous c actions to predict a *chunk* of h future actions: $\pi_\theta\big(a_{t:t+h-1} \,\big|\, s_t,\, a_{t-c:t-1}\big)$. We assume the chunk distribution is modeled via an *iterative refinement* process (e.g., diffusion or flow-matching), requiring k denoising steps per chunk prediction.

Recent policies often favor *short* contexts (c), *long* chunks (h), and *many* refinement steps (k) to improve intra-chunk consistency. However, this design reduces reactivity in two ways: (i) the decision frequency decreases as h grows, and (ii) inference latency increases with k. In principle, an ideal closed-loop regime would invert this design—*long* context, *short* chunks (up to token-by-token), and *few* refinement steps—as commonly seen in language modeling. Yet, applying this strategy in robotics often hurts performance, arguably due to the comparatively limited data scale.

How does the optimal design of robot policies evolve as demonstration data increases?



(a) Data scaling (b) Stochastic enviroment

Figure 2: Effect of dataset size and context conditioning. (a) Increasing the demonstration dataset shifts the optimal design toward longer context lengths ($c \uparrow$) and shorter prediction horizons ($h \downarrow$). (b) Context-conditioned student policies surpass chunk-based teacher policies as the environment noise level increases.

Scaling Analysis. Existing robot learning benchmarks typically provide only a few hundred demonstrations per task. We hypothesize that this data scale is limitating, constraiing the optimal policy design and biasing current methods toward short-context, chunk-heavy, many-step designs. To understand how the data scale influences the design space, we conduct a controlled scaling analysis on two widely used tasks, *Push-T* and *Can*, varying the amount of demonstration data up to $\mathbf{10}\times$ the original scale. For each data scale, we keep the total modeling length $l = c + h$ fixed, but vary the context length $c \in \{0, 2, 4, 6, 8\}$ and the number of refinement steps $k \in \{1, 3, 8, 15, 100\}$.

Interestingly, we observe that both *action chunking* and *iterative denoising*, which dominate recent policy designs [4, 5, 14, 17], become less critical as data increases. As shown in Fig. 2, policies with *longer context* and *fewer refinement steps* gradually match—and eventually surpass—the performance of chunk-based, many-step counterparts. This result suggests that *learning long-context, few-step policies is not infeasible—the bottleneck lies in data*.

3 Method: Short-to-Long Distillation for Policy Learning

Motivated by the analysis in §2, we next introduce a policy distillation method for learning performant policies conditioned on longer action contexts and operating with few denoising steps. While learning such a policy directly from limited human demonstrations is generally challenging, we seek to achieve this goal by progressively expanding the data in two stages: first, we train a teacher policy π_{chunk} on human demonstrations, using a long chunk horizon and many denoising steps to prioritize intra-chunk consistency; subsequently, we train a student policy π_{context} on synthetic rollouts generated by π_{chunk}, conditioning on longer contexts and using fewer refinement steps to improve reactivity.

The core challenge lies in enabling the student not only to inherit the teacher's consistency but also to *outperform* it in reactivity. This may seem out of reach at first glance. However, we notice two structural differences between the teacher and student policies that offer valuable guidance for effective distillation:

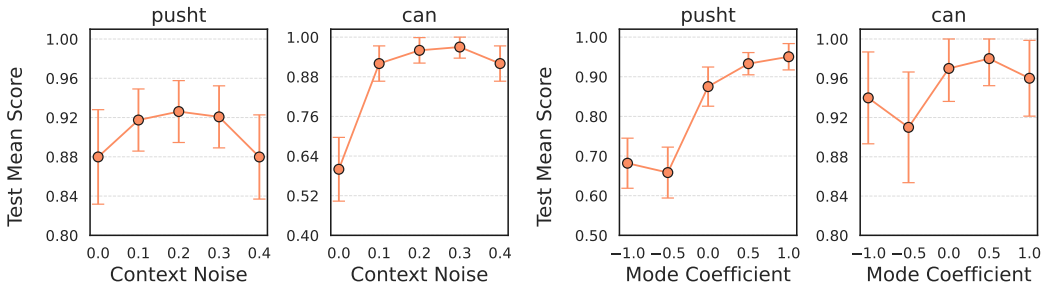
- *Smooth chunks vs. noisy contexts.* The teacher generates temporally consistent action chunks, whereas the student conditions on histories of individual actions, which can be much noisier.
- *Ambiguous samples vs. specified labels.* The teacher may yield diverse chunk samples, whereas the student operates under extended histories that make the next actions more specified.

These differences imply that, even if not all raw teacher rollouts are directly useful, they can potentially be turned into effective training signals for the student. Specifically, we consider two curation strategies tailored to history action contexts and future action labels, respectively.

Noise-Corrupted Context Expansion. Given a teacher rollout, we corrupt the action context with additive Gaussian noise and train the student on the perturbed history:

$$\tilde{a}_{t-c:t-1} \;=\; a_{t-c:t-1} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

Random noise is sampled at every use so each iteration presents a distinct history; the student predicts $a_{t:t+h-1}$ from $(s_t, \tilde{a}_{t-c:t-1})$ with the teacher chunk as the target. This stochastic corruption expands



(a) Effect of noise injection

(b) Effect of mode seeking

Figure 3: Effect of chunk curation. (a) Adding Gaussian noise to context actions improves student policy performance by expanding action coverage, though excessive noise eventually degrades results. (b) Increasing the degree of mode seeking during teacher demonstration enhances student performance, as clustering on denser modes provides sharper supervision.

context coverage and reduces overfitting to deterministic teacher behavior. The noise level σ is a hyperparameter and can be tuned based on a teacher–student discrepancy.

Mode-Seeking Chunk Optimization. Given the state and context at time t, let the teacher produce $N = 16$ candidate chunks $\mathcal{C}_t = \{\hat{a}^{(i)}_{t:t+h-1}\}_{i=1}^{N}$. We assign each candidate a density score $z_i = \phi(\hat{a}^{(i)}_{t:t+h-1})$ based on average distance to 5 nearest neighbors. We then use $d \in [-1, 1]$ controls *mode seeking* vs. *mode covering* with coupled truncation: for $d > 0$ keep only the densest top fraction; for $d < 0$ keep only the least-dense bottom fraction; for $d = 0$ keep all. We then select the supervision chunk from the kept set, favoring higher z_i when $d > 0$ and thereby sharpening the label distribution in the mode-seeking regime.

4 Experiments

In this section, we evaluate our method on two robotic manipulation benchmarks: Push-T [6] and RoboMimic [13]. Specifically, we use the public checkpoints from prior work [6, 11] as the teacher policies, and assess the effectiveness of the proposed curation strategies on the student performance.

4.1 Effect of Noise Injection

Setup. We study the effect of Gaussian noise injection on student policy performance by varying the context noise level $\sigma \in \{0.0, 0.1, 0.2, 0.3, 0.4\}$ in *Push-T* and *RoboMimic (Can)*. Student policies are trained on 500 teacher demonstrations collected with mode coefficient $d = 0.5$, using context length $c = 8$ and prediction horizon $h = 1$. All other factors (architecture, optimizer, learning rate, training steps) are held fixed.

Results. As shown in Fig. 3a, performance follows a non-monotonic trend with respect to σ. Moderate noise ($\sigma \leq 0.2$) improves generalization by expanding action coverage, while larger values degrade performance as injected noise dominates the context signal.

4.2 Effect of Mode Seeking

Setup. We then examine how the mode-seeking coefficient $d \in \{-0.5, 0.0, 0.5, 1.0\}$ in teacher demonstration collection affects student policy performance on *Push-T* and *RoboMimic (Can)*. Each teacher chunk is sampled $N = 16$ times, with $d < 0$ promoting diversity, $d = 0$ sampling uniformly, and $d > 0$ emphasizing high-density modes. Student policies are trained under the same conditions as above (500 demos, $c = 8$, $h = 1$), with all other variables controlled.

Results. As shown in Fig. 3b, *Push-T* shows consistent gains as d increases, highlighting the benefits of clustering on dense modes. *Can* exhibits a similar upward trend but with greater variance across runs. We attribute this noisiness to the binary nature of the task's binary success metric, which may obscure finer-grained differences in policy quality.

References

[1] C. Agia, R. Sinha, J. Yang, Z.-a. Cao, R. Antonova, M. Pavone, and J. Bohg. Unpacking Failure Modes of Generative Policies: Runtime Monitoring of Consistency and Progress. *arXiv preprint arXiv:2410.04640*, Oct. 2024.

[2] L. Ankile, A. Simeonov, I. Shenfeld, M. Torne, and P. Agrawal. From Imitation to Refinement – Residual RL for Precise Assembly. *arXiv preprint arXiv:2407.16677*, Dec. 2024.

[3] 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. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*, Nov. 2024.

[4] K. Black, M. Y. Galliker, and S. Levine. Real-Time Execution of Action Chunking Flow Policies, June 2025. URL `http://arxiv.org/abs/2506.07339`. arXiv:2506.07339 [cs].

[5] C. Cheang, S. Chen, Z. Cui, Y. Hu, L. Huang, T. Kong, H. Li, Y. Li, Y. Liu, X. Ma, H. Niu, W. Ou, W. Peng, Z. Ren, H. Shi, J. Tian, H. Wu, X. Xiao, Y. Xiao, J. Xu, and Y. Yang. GR-3 Technical Report. *arXiv preprint arXiv:2507.15493*, July 2025.

[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] O. X.-E. Collaboration, A. Padalkar, A. Pooley, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Garg, A. Brohan, A. Raffin, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Ichter, C. Lu, C. Xu, C. Finn, C. Xu, C. Chi, C. Huang, C. Chan, C. Pan, C. Fu, C. Devin, D. Driess, D. Pathak, D. Shah, D. Büchler, D. Kalashnikov, D. Sadigh, E. Johns, F. Ceola, F. Xia, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Schiavi, G. Kahn, H. Su, H.-S. Fang, H. Shi, H. B. Amor, H. I. Christensen, H. Furuta, H. Walke, H. Fang, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Peters, J. Schneider, J. Hsu, J. Bohg, J. Bingham, J. Wu, J. Wu, J. Luo, J. Gu, J. Tan, J. Oh, J. Malik, J. Booher, J. Tompson, J. Yang, J. J. Lim, J. Silvério, J. Han, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Zhang, K. Rana, K. Srinivasan, L. Y. Chen, L. Pinto, L. Fei-Fei, L. Tan, L. Ott, L. Lee, M. Tomizuka, M. Spero, M. Du, M. Ahn, M. Zhang, M. Ding, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. Di Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, P. R. Sanketi, P. Wohlhart, P. Xu, P. Sermanet, P. Sundaresan, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Martín-Martín, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Moore, S. Bahl, S. Dass, S. Sonawani, S. Song, S. Xu, S. Haldar, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Dasari, S. Belkhale, T. Osa, T. Harada, T. Matsushima, T. Xiao, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, V. Jain, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Wang, X. Zhu, X. Li, Y. Lu, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Xu, Y. Wang, Y. Bisk, Y. Cho, Y. Lee, Y. Cui, Y.-H. Wu, Y. Tang, Y. Zhu, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Xu, and Z. J. Cui. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. *arXiv preprint arXiv:2310.08864*, Dec. 2023.

[8] G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, Mar. 2015.

[9] S. H. Høeg, Y. Du, and O. Egeland. Streaming Diffusion Policy: Fast Policy Synthesis with Variable Noise Diffusion Models. *arXiv preprint arXiv:2406.04806*, Oct. 2024.

[10] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, A. Jain, A. Lee, Y. Lee, M. Memmel, S. Park, I. Radosavovic, K. Wang, A. Zhan, K. Black, C. Chi, K. B. Hatch, S. Lin, J. Lu, J. Mercat, A. Rehman, P. R. Sanketi, A. Sharma, C. Simpson, Q. Vuong, H. R. Walke, B. Wulfe, T. Xiao, J. H. Yang, A. Yavary, T. Z. Zhao, C. Agia, R. Baijal, M. G. Castro, D. Chen, Q. Chen, T. Chung,

J. Drake, E. P. Foster, J. Gao, D. A. Herrera, M. Heo, K. Hsu, J. Hu, D. Jackson, C. Le, Y. Li, K. Lin, R. Lin, Z. Ma, A. Maddukuri, S. Mirchandani, D. Morton, T. Nguyen, A. O'Neill, R. Scalise, D. Seale, V. Son, S. Tian, E. Tran, A. E. Wang, Y. Wu, A. Xie, J. Yang, P. Yin, Y. Zhang, O. Bastani, G. Berseth, J. Bohg, K. Goldberg, A. Gupta, A. Gupta, D. Jayaraman, J. J. Lim, J. Malik, R. Martín-Martín, S. Ramamoorthy, D. Sadigh, S. Song, J. Wu, M. C. Yip, Y. Zhu, T. Kollar, S. Levine, and C. Finn. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. *arXiv preprint arXiv:2403.12945*, Mar. 2024.

[11] 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. URL `http://arxiv.org/abs/2408.17355`. arXiv:2408.17355 [cs].

[12] X. Ma, S. Patidar, I. Haughton, and S. James. Hierarchical Diffusion Policy for Kinematics-Aware Multi-Task Robotic Manipulation. *arXiv preprint arXiv:2403.03890*, Mar. 2024.

[13] 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.

[14] NVIDIA, J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. J. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, J. Jang, Z. Jiang, J. Kautz, K. Kundalia, L. Lao, Z. Li, Z. Lin, K. Lin, G. Liu, E. Llontop, L. Magne, A. Mandlekar, A. Narayan, S. Nasiriany, S. Reed, Y. L. Tan, G. Wang, Z. Wang, J. Wang, Q. Wang, J. Xiang, Y. Xie, Y. Xu, Z. Xu, S. Ye, Z. Yu, A. Zhang, H. Zhang, Y. Zhao, R. Zheng, and Y. Zhu. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots. *arXiv preprint arXiv:2503.14734*, Mar. 2025.

[15] A. Prasad, K. Lin, J. Wu, L. Zhou, and J. Bohg. Consistency Policy: Accelerated Visuomotor Policies via Consistency Distillation. *arXiv preprint arXiv:2405.07503*, May 2024.

[16] A. A. Rusu, S. G. Colmenarejo, C. Gulcehre, G. Desjardins, J. Kirkpatrick, R. Pascanu, V. Mnih, K. Kavukcuoglu, and R. Hadsell. Policy Distillation. *arXiv preprint arXiv:1511.06295*, Jan. 2016.

[17] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, S. Alibert, M. Cord, T. Wolf, and R. Cadene. SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics. *arXiv preprint arXiv:2506.01844*, June 2025.

[18] L. Wang and K.-J. Yoon. Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3048–3068, June 2022.

[19] H. Xue, J. Ren, W. Chen, G. Zhang, Y. Fang, G. Gu, H. Xu, and C. Lu. Reactive Diffusion Policy: Slow-Fast Visual-Tactile Policy Learning for Contact-Rich Manipulation. *arXiv preprint arXiv:2503.02881*, Apr. 2025.

[20] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. *arXiv preprint arXiv:2304.13705*, Apr. 2023.

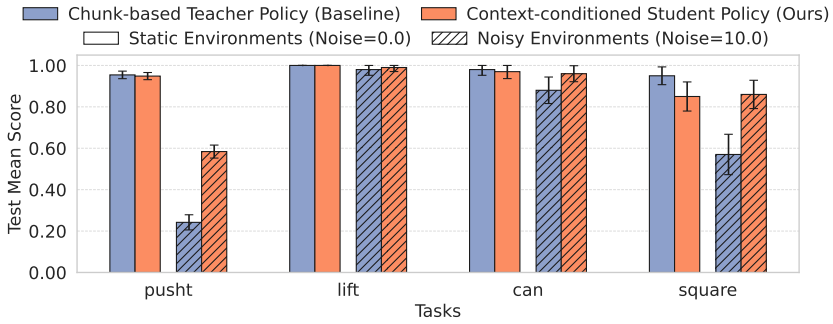


Figure 4: Context-conditioned student policies match the performance of chunk-based teacher policies in static environments, and surpass them in stochastic environments across four Push-T and RoboMimic tasks.

A Implementation Details

We parameterize the student policy with a U-Net backbone and augment it with FiLM layers for conditioning on observations, diffusion timesteps, and context actions following standard diffusion policy[6] architectures.

Teacher demo collection. During chunk-based teacher demonstration collection, we employ early stopping to mitigate mode collapse, which can otherwise reduce performance in mode-seeking settings. The chunk-based teacher policy has context length $c = 0$ and action horizon $h = 8$. We use 16 samples and 3 modes, which is sufficient to capture the multimodal behavior of teacher policies in both Push-T and Robomimic tasks.

Student policy distillation. For context-conditioned student policy distillation, we set context length $c = 8$ and action horizon $h = 1$. Training uses a batch size of 256 with 100 diffusion steps for both training and inference. Rollouts are generated with a DDPM sampler.

Unless otherwise noted, all experiments share the same architecture, optimizer, diffusion scheduler, and saved teacher demo data. To ensure convergence and fair comparison, we train each model for up to 500 epochs and report results from the checkpoint achieving the best validation performance.

Environment Setup. We evaluate student policy performance on four robot manipulation benchmarks:

Push-T: We use the Push-T environment from Chi et al. [6], where the goal is to push a T-shaped block to a designated target location. The action space consists of two-dimensional end-effector velocity commands, and policies are conditioned on keypoint-based observations.

Robomimic: We adopt three tasks from the Robomimic suite [13] — Lift, Can, and Square. Policies are conditioned on state-based observations, with continuous action spaces following the suite's default setup.

Name	Value
Context length	0
Action horizon	8
Number of samples	16
Number of modes	3
Mode coefficient	0.5

Table 1: Default hyperparameters for teacher demonstration collection.

Name	Value
Context length	8
Action horizon	1
Context noise	0.2
Number of teacher episodes	500
Batch size	256
Epochs	300
Learning rate	1×10^{-4}
Weight decay	1×10^{-6}
Optimizer	AdamW
Training diffusion steps	100
Inference diffusion steps	100
Diffusion scheduler	DDPM
Number of test environments	100
Environment noise	0.0

Table 2: Default hyperparameters for student policy distillation.