Agile Imitation: Efficient Robot Skill Learning via Iterative Human Feedback and Data Aggregation

Chenjun Liu, Haojie Luo, Wei Li
School of Intelligent Robotics and Advanced Manufacturing
Fudan University, Shanghai, China
Email: 24210860058@m.fudan.edu.cn

***Abstract*—Interactive Imitation Learning (IIL) is a promising paradigm for the online optimization of robotic skills, enabling agents to learn from sparse expert demonstrations. However, in few-shot scenarios, effectively combining human feedback with data aggregation to mitigate distributional shift remains a significant challenge. While previous works explored this integration, systematic research into internal mechanisms—such as dynamic feedback weighting, delay compensation, and lightweight trajectory resampling—remains insufficient under extreme data constraints. This paper proposes the Agile Iterative Aggregation (AIA) framework, which systematically optimizes these core iterative feedback and data processing mechanisms using a human-in-the-loop approach. We evaluate AIA on six simulated manipulation tasks and further validate it on a real-world platform. The results show that our optimized processing and aggregation strategies significantly enhance policy performance and robustness. Compared to baselines, AIA achieves superior success rates and finer sub-skill acquisition, offering a cost-effective solution for high-efficiency robotic skill learning.**

***Index Terms*—interactive imitation learning, human-in-the-loop, few-shot learning, data aggregation.**

I. Introduction

Imitation learning (IL) [1]–[4] is a widely used paradigm in robotics for enabling intelligent systems to learn skills by observing expert demonstrations. Compared to reinforcement learning (RL) [5], imitation learning avoids the complexity of reward function design and can guide policy formation directly from human experience. However, traditional IL methods, such as behavioral cloning (BC) [6], face an inherent distribution shift problem. Due to small differences between the learned and expert policies, the agent gradually enters state regions not covered by the expert's demonstrations during execution, leading to a dramatic degradation in performance. To address this challenge, Interactive Imitation Learning (IIL) has been developed [7], which allows human experts to intervene in the learning process by providing online feedback or supplementary demonstrations. This human-in-the-loop (HITL) [8]–[10] mechanism effectively corrects the agent's errors and guides it to explore the correct behavioral space, proving particularly crucial in low-data scenarios where data acquisition is costly or expert resources are limited.

Despite the promise of IIL, achieving efficient and robust robot skill learning with very few initial samples remains a significant challenge. Much of the existing research [11], [12] focuses on high-level algorithmic combinations, while often overlooking the subtle yet critical mechanisms within the interactive loop that profoundly impact learning efficiency. Specifically, several key questions lack systematic investigation: How can we precisely handle the inherent delay between an agent's mistake and the expert's corrective feedback to avoid learning from corrupted data? How can we design lightweight trajectory standardization strategies that are effective for the sparse and non-uniform data collected in few-shot settings? Furthermore, how can the influence of corrective feedback be dynamically weighted to balance expert guidance and agent autonomy?

Compounding these challenges is the reliance on coarse evaluation metrics, such as binary task success rates. For complex manipulation tasks, such metrics fail to capture the policy's nuanced understanding of the task, making it difficult to assess the true effectiveness of iterative human teaching. To address these gaps, this paper introduces the Agile Iterative Aggregation (AIA) analytical framework. Rather than proposing a single new algorithm, AIA provides a systematic approach to investigating and optimizing the core components of the interactive learning process under severe data constraints. Our main contributions are threefold:

1. We conduct a systematic study of several critical, yet often overlooked, interaction mechanisms within the AIA framework. This includes proposing and validating strategies for feedback delay compensation, developing lightweight methods for trajectory completion and standardization, and designing adaptive strategies for the dynamic adjustment of feedback factors.

2. We propose a novel, fine-grained evaluation metric based on task decomposition. By breaking down a complex task into key sub-steps (e.g., gripper alignment, trajectory logic, grasping position), this metric allows for a more precise measurement of a policy's understanding and progress. It provides deeper insights into the effectiveness of each round of human interaction, revealing which specific sub-skills are being successfully acquired.

3. Through extensive experiments on six simulated robot manipulation tasks, we demonstrate that optimizing these key mechanisms within the AIA framework leads to significant improvements in learning efficiency and policy robustness compared to baseline methods.

Ultimately, this work aims to provide both practical guidance and theoretical support for achieving truly low-data, high-efficiency robot imitation learning, moving beyond high-level concepts to focus on the impactful details of the human-robot

learning partnership.

II. Related Works

A. Imitation Learning and Distribution Shift

Imitation Learning (IL) has emerged as a cornerstone for enabling robots to acquire complex manipulation skills by observing and mimicking expert demonstrations. This paradigm effectively bypasses the "reward engineering" bottleneck often encountered in Reinforcement Learning (RL) [13]–[15]. The most direct implementation is Behavioral Cloning (BC), which formulates policy learning as a supervised regression problem that maps observed states $\mathbf{s}$ to expert actions $\mathbf{a}$. In continuous action spaces, the policy π is typically trained to minimize the discrepancy between predicted actions and expert demonstrations using standard regression objectives [16].

Despite its simplicity, BC [17], [18] is fundamentally plagued by the *distribution shift* problem. Because the learned policy is inherently imperfect, small execution errors lead the agent into state regions not covered by the expert's training distribution. Without guidance in these unfamiliar states, initial deviations compound over time, resulting in catastrophic performance collapse. While extensions like Weighted BC attempt to prioritize high-quality samples, and Inverse Reinforcement Learning (IRL) [19], [20] seeks to infer underlying reward functions, these methods often remain computationally prohibitive or fail to converge in few-shot scenarios. These challenges underscore the necessity for interactive learning paradigms that can correct policy drift through real-time feedback.

B. Interactive Learning and DAgger Variants

Interactive Imitation Learning (IIL) addresses distribution shift by incorporating human experts directly into the training loop. A landmark contribution is the Dataset Aggregation (DAgger) algorithm, which introduces an iterative "learning-on-the-job" process. In DAgger, the agent executes its current policy while an expert provides corrective labels for the states encountered; this new data is then aggregated with the original dataset to retrain the policy. Subsequent research has produced numerous variants to improve interaction efficiency: HG-DAgger [21] allows for flexible human intervention and takeover, while EnsembleDAgger [22] utilizes policy ensembles to identify states with high uncertainty for targeted queries.

Beyond the DAgger family [23]–[28], broader Human-in-the-Loop (HITL) frameworks like Reinforcement Learning from Human Feedback (RLHF) [29], [30] and CEILing [31] have gained prominence. However, existing methods often focus on high-level interaction protocols while overlooking the fine-grained execution mechanics. Specifically, the impact of feedback delay compensation, dynamic weighting of expert signals, and lightweight trajectory processing remains insufficiently explored under severe data constraints. Our AIA framework identifies these critical research gaps, systematically optimizing these internal mechanisms to ensure robust learning when expert resources are extremely limited.

C. Towards Data-Efficient Skill Acquisition

Achieving high performance with minimal human interaction [32]–[34] is a central goal in modern robotics. Current research pursues this through two primary paths. The first involves data-level optimizations such as procedural data augmentation and residual policy learning. However, training residual networks remains challenging in few-shot settings where corrective data is too sparse to model effectively. The second path leverages large-scale Vision-Language-Action (VLA) models [35], [35], such as RT-2 [36] or Octo [37], which are pre-trained on massive datasets like Open X-Embodiment [2]. Building upon these foundations, a promising direction involves deploying pre-trained VLA models in real-world environments and performing fine-tuning via human-in-the-loop interaction, utilizing Reinforcement Learning (RL) or Imitation Learning (IL) to bridge the gap between generalized representations and task-specific execution [38]–[40].

While these foundation models exhibit impressive generalization, fine-tuning them on new tasks via DAgger-like loops still typically requires hundreds of demonstration episodes to achieve proficiency. This stands in sharp contrast to our Agile Iterative Aggregation (AIA) framework, which focuses on maximizing the utility of an extremely small set of 10–20 interactions using a lightweight policy. By intelligently managing the entire data lifecycle—from feedback delay compensation to standardized resampling—AIA ensures that each piece of expert intervention provides maximal performance gains. This provides a practical path for low-cost, high-efficiency robotic skill acquisition without the need for massive pre-training or large-scale data collection.

III. Method

In this section, we present the Agile Iterative Aggregation (AIA) framework, a systematic approach for improving interactive imitation learning under extreme data constraints. Rather than introducing a new learning paradigm, AIA focuses on optimizing several fine-grained yet critical mechanisms within the human-in-the-loop learning process, which are often overlooked in prior work.

A. Overview of the AIA Framework

The AIA framework follows an iterative data aggregation paradigm inspired by interactive imitation learning. Given a small set of initial expert demonstrations, an initial policy is trained via behavioral cloning. The agent then executes the learned policy in the environment while a human expert monitors its behavior and provides corrective interventions when necessary. These corrective trajectories are processed and aggregated into the training dataset, and the policy is subsequently updated. This process is repeated over multiple interaction rounds, progressively refining the policy. Figure 1 illustrates the overall workflow of AIA. A key design principle of AIA is to maximize the utility of each human intervention in few-shot settings. To this end, AIA introduces several

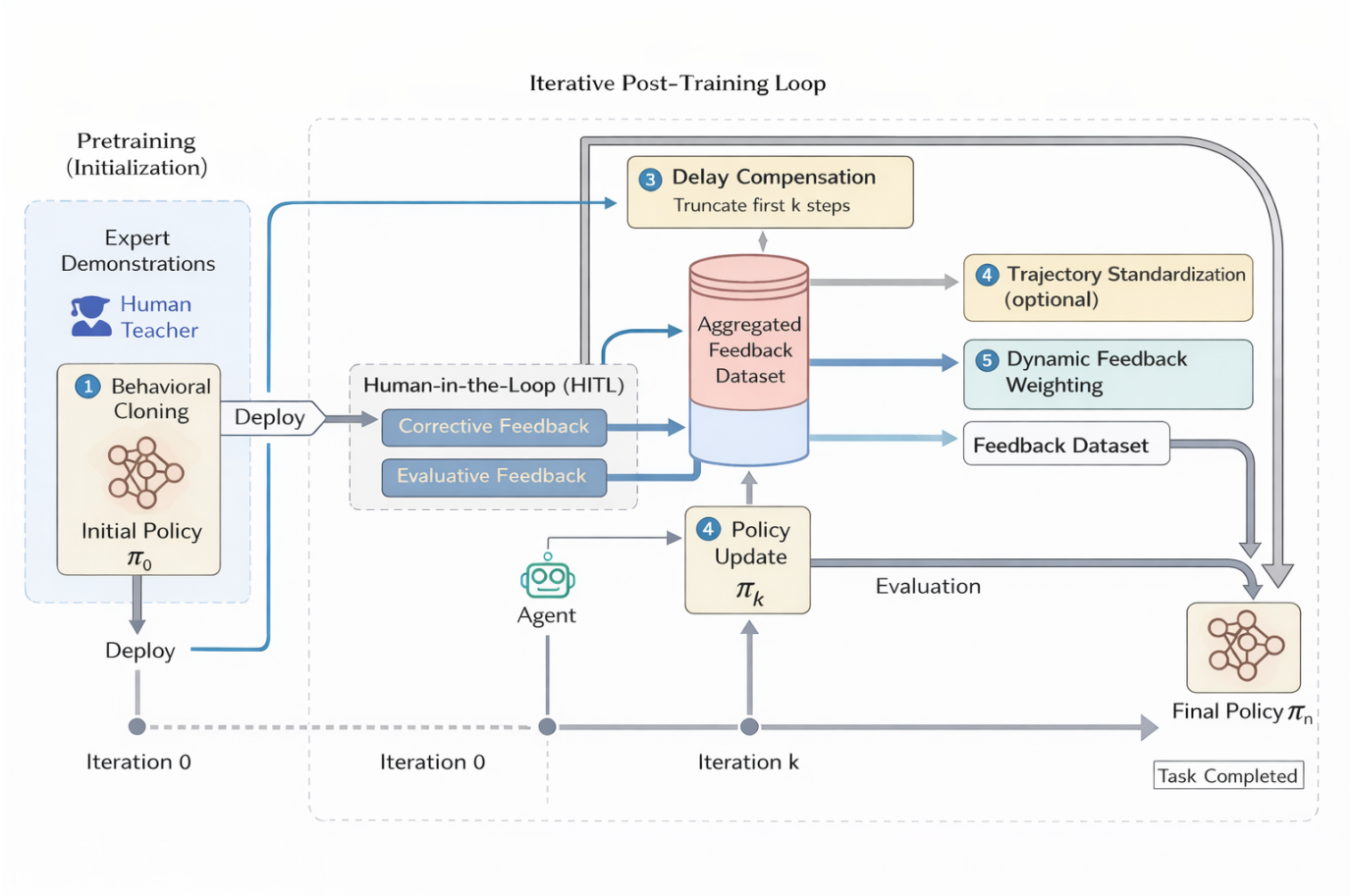


Fig. 1: The workflow of Agile Iterative Aggregation (AIA) framework.

lightweight yet effective mechanisms that regulate how corrective feedback is processed, weighted, and integrated into the training data.

B. Trajectory Standardization

In interactive learning, collected demonstrations and corrective trajectories often exhibit varying temporal lengths due to task diversity and inconsistent human intervention timing. While recurrent policies can theoretically handle variable-length sequences, such variability may introduce instability during multi-round data aggregation, particularly when data is scarce. To address this issue, AIA applies a lightweight trajectory standardization strategy. Specifically, all trajectories are converted to a fixed length through resampling while preserving task-relevant terminal states. This design ensures consistent temporal alignment across training samples and stabilizes policy optimization without introducing heavy preprocessing or additional supervision.

C. Feedback Delay Compensation

Human corrective feedback is inherently delayed relative to the agent's erroneous actions due to perception, decision-making, and physical reaction latency. Without proper handling, this delay can corrupt the training data by associating failure-inducing states with recovery actions, thereby misleading the policy during learning. AIA explicitly compensates for feedback delay by truncating the initial segment of each corrective trajectory. By removing the ambiguous portion corresponding to the agent's erroneous behavior prior to expert takeover, this strategy aligns corrective actions with the relevant state distribution and mitigates the adverse effects of delayed intervention.

D. Dynamic Feedback Weighting

Not all corrective feedback carries equal informational value. Early interventions often address fundamental deficiencies in the policy, whereas later corrections tend to refine fine-grained behaviors. Assigning a fixed importance to all corrective data may therefore either underutilize critical feedback or overemphasize noisy corrections. To balance these effects, AIA introduces a dynamic feedback weighting mechanism. A feedback factor α is used to modulate the contribution of corrective trajectories during data aggregation. The value of α is adaptively adjusted based on the frequency of human interventions relative to autonomous executions, allowing the framework to emphasize rare, informative corrections while down-weighting frequent, potentially noisy feedback.

E. Human Guidance Protocol

A fundamental design choice in human-in-the-loop systems is determining how and when expert guidance should be provided. AIA adopts a reactive correction protocol, in which the agent is allowed to act autonomously until the expert intervenes to correct erroneous behavior. This protocol ensures that corrective feedback is collected directly on the agent-induced state distribution, enabling more targeted learning and improved distributional alignment. Compared to proactive full demonstrations, reactive correction provides higher information density per intervention, making it particularly suitable for few-shot interactive learning scenarios.

IV. Experiments

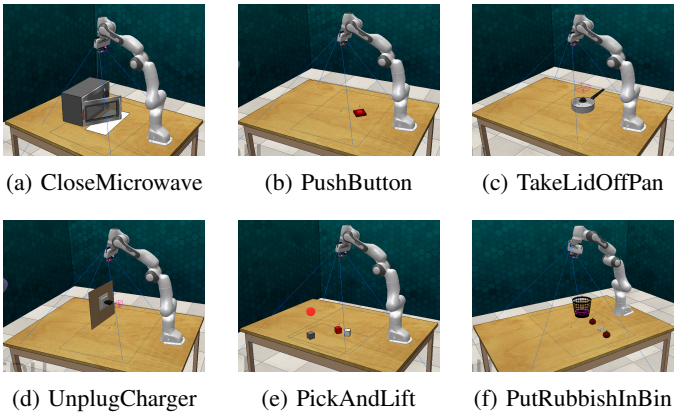
This section presents a comprehensive experimental evaluation of the proposed Agile Iterative Aggregation (AIA) framework. We first describe the experimental settings, including environments, tasks, the real-world validation setup, and our policy network architecture. We then establish benchmarks using standard Behavioral Cloning (BC) and interactive CEIL baselines, followed by systematic ablations of key AIA components. Finally, we report overall performance and provide a fine-grained analysis using the proposed task decomposition metric.

A. Experimental Settings

We evaluate our framework in the CoppeliaSim simulation environment, utilizing a 7-DOF Franka Emika Panda robotic arm equipped with a parallel gripper and a wrist-mounted RGB camera (128×128) for visual input.

a) Simulation Tasks: We evaluate our agent on six simulated manipulation tasks from RLBench [41], as shown in Fig. 2, selected to cover a diverse set of core manipulation primitives commonly found in household and laboratory settings. Specifically, the suite includes: PickAndLift (stable grasping and lifting), PushButton (precise end-effector positioning and force control), CloseMicrowave (interaction with an articulated object under hinge constraints), TakeLidOffSaucepan (rim grasping and constrained vertical lifting), UnplugCharger (high-precision alignment and force-controlled extraction), and PutRubbishInBin (a long-horizon task integrating grasping, transport, and placement). Together, these tasks span diverse contact dynamics, physical constraints, and success criteria, providing a compact yet comprehensive benchmark for evaluating robustness and generalization from limited demonstrations.

b) Real-World Validation: To evaluate sim-to-real transfer capability, we conducted preliminary real-world experiments on a UR5 robotic manipulator equipped with a two-finger parallel gripper. As illustrated in Fig. 3, the task is formulated as a vision-based pick-and-place operation, in which the robot identifies a target object, executes a stable grasp, and places it into a designated container.



(a) CloseMicrowave (b) PushButton (c) TakeLidOffPan

(d) UnplugCharger (e) PickAndLift (f) PutRubbishInBin

Fig. 2: An overview of the six simulated robot manipulation tasks used for evaluation.

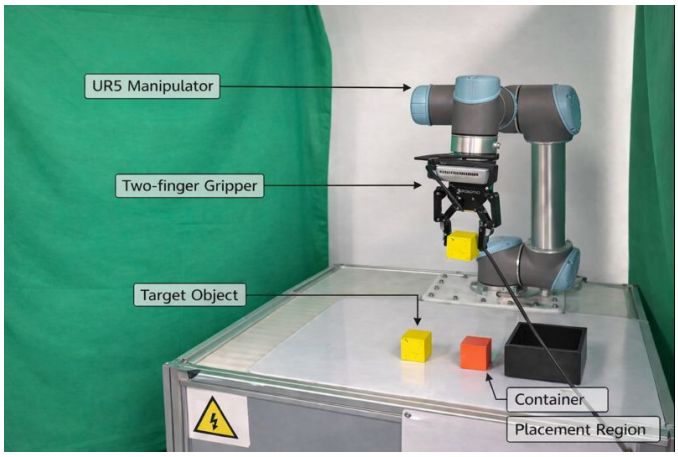


Fig. 3: Real-world experimental setup for sim-to-real validation. A UR5 robotic arm equipped with a two-finger gripper performs a pick-and-place task by grasping a target object from the workspace and placing it into a designated container.

c) Policy Network Architecture: Our policy, denoted as $\pi(\mathbf{a}|\mathbf{s})$, is parameterized by an end-to-end convolutional LSTM [42] network, as illustrated in Fig. 4. The network takes the robot state $\mathbf{s}$ as input, including a 128×128 RGB image from the wrist-mounted camera and the robot's proprioceptive state (joint angles and gripper opening). The visual input is processed by a three-layer convolutional backbone to produce a compact embedding, which is concatenated with the proprioceptive state and fed into an LSTM with 256 hidden units. The network outputs the final action $\mathbf{a}$ representing desired changes in joint angles and gripper state.

B. Baselines and Evaluation Metrics

We compare AIA against baselines that represent different learning paradigms. Our non-interactive baseline is Behavioral Cloning (BC), trained on datasets of varying sizes (10, 20, and 100 expert demonstrations). To ensure consistency across tasks, demonstrations are generated using a unified scripted collection pipeline provided by the simulator APIs under the

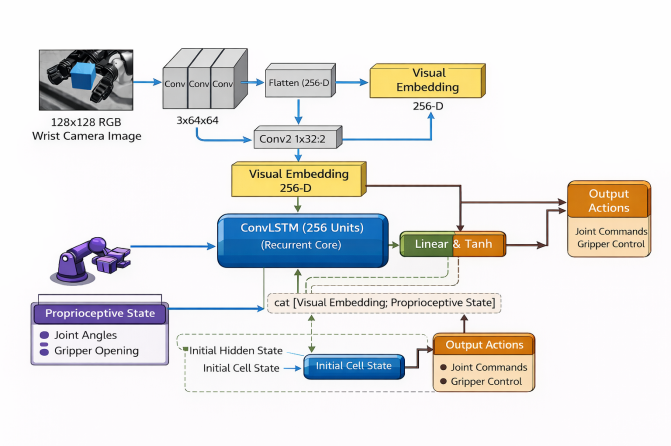


Fig. 4: Network architecture.

same observation and control interface. Our primary interactive baseline is CEIL (Correct Me If I'm Wrong) [31]. Motivated by iterative data aggregation, we further evaluate multi-round variants (DoubleCEIL and TripleCEIL). For evaluation, our primary metric is the *Task Success Rate*, computed as the percentage of successful trials over a fixed number of evaluation runs. This is supplemented by a fine-grained *task decomposition metric* (introduced in Sec. IV-E) to analyze sub-skill acquisition.

Table I and Fig. 5 show that BC performance improves with more demonstrations, but remains limited on complex long-horizon tasks, highlighting the distribution shift issue in non-interactive imitation learning. CEIL improves performance in few-shot settings; however, naively increasing interaction rounds does not consistently yield gains. On certain tasks (e.g., TakeLidOffSaucepan and UnplugCharger), performance peaks after one or two rounds and subsequently declines, motivating a systematic optimization of internal interactive learning mechanisms.

C. Ablation Study of Key AIA Components

We conduct systematic ablations to analyze how each component in AIA affects few-shot interactive imitation learning performance. Specifically, we examine: (1) trajectory standardization strategies, (2) feedback delay compensation, (3) dynamic feedback weighting, and (4) guidance protocols.

1) Trajectory Standardization: To evaluate our lightweight standardization strategy, we set a standard trajectory length of 150 steps and conduct a comparative study on the TakeLidOffSaucepan (TLS) task. We evaluate five methods: (1) End Extension + Linear Truncation; (2) Linear Interpolation + Linear Truncation; (3) Mirror Extension + Linear Truncation; (4) Linear Interpolation + Direct Truncation; and (5) Linear Interpolation + Uniform Sampling Truncation.

2) Feedback Delay Compensation: We evaluate the impact of our delay compensation mechanism by comparing a no-compensation baseline against various truncation lengths (3, 5, 7, 10, 12, and 15 steps). As shown in Fig. 6(b), truncation is crucial for performance, with 10 steps yielding the best results. Consequently, we apply a 10-step truncation to all corrective trajectories in our main experiments.

3) Dynamic Feedback Factor: To assess the effectiveness of the dynamic feedback factor α, we first compare against a commonly used baseline strategy that adjusts α based on the relative frequency of autonomous executions and human corrections:

$$\alpha = \begin{cases} 1, & N_{cor} < 10 \\ \dfrac{N_{pos} + N_{neg}}{N_{cor}}, & \text{otherwise} \end{cases} \tag{1}$$

where N_{cor} denotes the cumulative number of human corrections, and N_{pos} and N_{neg} denote the numbers of autonomous successes and failures, respectively. We further evaluate adaptive strategies based on online performance and correction statistics. Due to space constraints, we report the best-performing strategy, termed *Correction Frequency Weighting*:

$$\alpha = \begin{cases} 1.5, & F_{cor} < 0.1 \\ 1.2, & 0.1 \leq F_{cor} < 0.3 \\ 0.8, & F_{cor} \geq 0.3 \end{cases} \tag{2}$$

where $F_{cor} = \dfrac{N_{cor}}{N_{pos} + N_{neg}}$ represents the frequency of corrective feedback relative to autonomous executions.

4) Guidance Strategy: We compare the *Proactive Demonstration* protocol against the *Reactive Correction* (or Delegation) protocol [43]. As shown in Fig. 6(d), reactive correction leads to more effective and robust learning. Therefore, AIA primarily employs the reactive correction protocol to maximize the efficiency of expert feedback.

D. Overall Performance of AIA

Having identified effective designs for each component, we evaluate the overall performance of the fully integrated AIA agent. We compare AIA against a non-interactive BC policy and a standard interactive agent based on the original CEIL framework. All methods are trained under the same few-shot conditions, starting with an initial dataset of 20 demonstrations. As summarized in Fig. 7, AIA demonstrates substantial and consistent improvements over both baselines across all six tasks. The benefits are particularly pronounced on precision-critical tasks. For example, on PushButton (PB), AIA achieves an 88% success rate, improving over CEIL (59%) and substantially outperforming BC. On UnplugCharger (UC), CEIL fails to robustly acquire the skill, while AIA reaches 70% success by optimizing delay processing, feedback weighting, and the human guidance protocol. Similar improvement trends are also observed on PickAndLift and PutRubbishInBin, confirming the robustness of the proposed framework.

E. Fine-grained Analysis with Task Decomposition

Binary success rates can be insufficient to reveal how a policy improves through interaction on complex manipulation tasks. We therefore introduce a fine-grained evaluation protocol by decomposing each task into key sub-stages. This

TABLE I: Comprehensive performance comparison of baselines. The table shows the evolution from non-interactive Behavioral Cloning (BC) with varying data amounts to interactive learning with CEIL. All success rates are percentages (%).

Task	**Behavioral Cloning (BC)**			**Interactive Learning (CEIL Iterations)**		
	10 Demos	**20 Demos**	**100 Demos**	**1 Iter (CEIL)**	**2 Iters (DoubleCEIL)**	**3 Iters (TripleCEIL)**
CloseMicrowave (CM)	51	59	70	79	82	84
PushButton (PB)	4	25	50	59	77	80
TakeLidOffSaucepan (TLS)	0	15	30	50	72	70
UnplugCharger (UC)	0	6	15	23	42	20
PickAndLift (PL)	0	4	20	18	33	50
PutRubbishInBin (PRB)	0	0	0	2	13	16

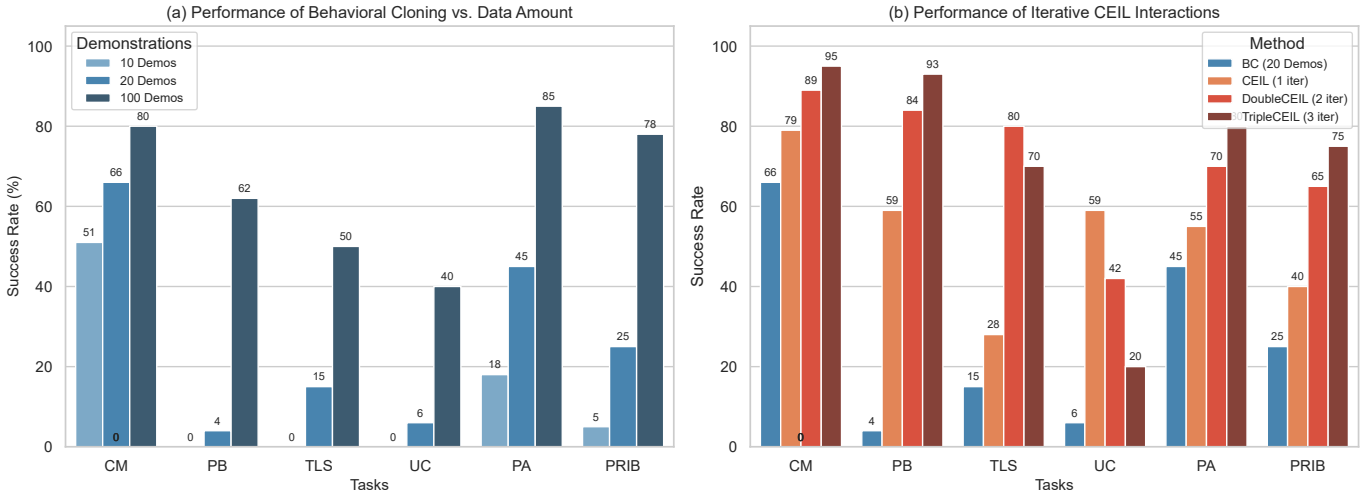


Fig. 5: Baseline comparison: Behavioral Cloning (BC) with different numbers of demonstrations and CEIL with increasing interaction rounds.

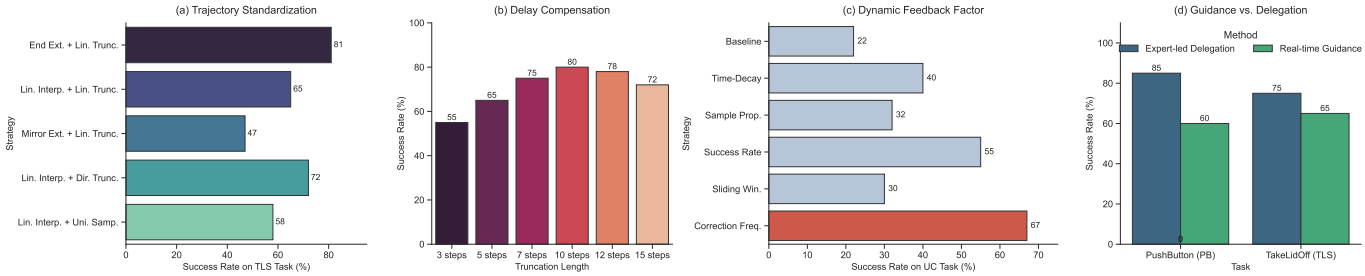


Fig. 6: Ablations of AIA components: (a) trajectory processing, (b) delay compensation, (c) dynamic feedback weighting, (d) guidance protocol.

metric allows us to localize performance gains to specific sub-skills and analyze which stages are most improved by iterative human teaching.

Contrasting the agent's performance with and without human intervention provides deeper insight. As shown in Table II, the CEIL baseline can often acquire early coarse stages but fails at high-precision stages (e.g., grasping), resulting in overall task failure. In contrast, AIA achieves more consistent performance across sub-stages, indicating that performance gains are concentrated on critical failure points. These results suggest that human feedback in AIA functions as targeted fine-tuning rather than holistic re-demonstration. Corrective interventions selectively reinforce specific sub-skills where systematic errors occur, while leaving already learned behaviors largely unaffected.

V. Conclusion

In this paper, we addressed the challenge of efficient robot skill acquisition from limited data by proposing the Agile Iterative Aggregation (AIA) framework. Rather than claiming the isolated mechanisms—such as trajectory standardization or dynamic feedback weighting—as entirely standalone novelties, our primary contribution lies in their systematic integration and targeted adaptation for extreme few-shot scenarios. We demonstrated that the synergistic optimization of trajectory

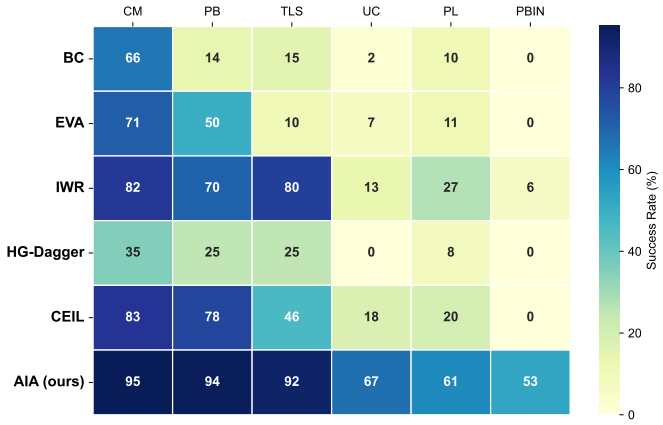


Fig. 7: Overall performance comparison across six manipulation tasks. The heatmap illustrates the success rates of various imitation and feedback-based learning methods, including BC, EVA, IWR, HG-Dagger, CEIL, and the proposed AIA approach. Cell values and color gradients represent the task success rate (%).

TABLE II: Detailed performance comparison via task decomposition. Values represent cumulative success rate (%) at each sub-task stage across tasks.

Task	Stage	BC	CEIL	AIA
CloseMicrowave	1. Correct Trajectory	80	90	**100**
	2. Approach Door	60	80	**96**
	3. Grasp & Push	70	90	**95**
PushButton	1. Correct Trajectory	60	70	**100**
	2. Approach Button	30	52	**100**
	3. Press Correctly	10	50	**96**
TakeLidOff	1. Correct Trajectory	80	85	**100**
	2. Approach Lid	60	75	**100**
	3. Grasp Lid	6	50	**92**
UnplugCharger	1. Correct Trajectory	40	50	**90**
	2. Align Gripper	10	26	**67**
	3. Approach & Enclose	30	40	**85**
	4. Grasp	4	23	**70**
	5. Unplug	2	10	**70**
PickAndLift	1. Correct Trajectory	20	70	**80**
	2. Approach Correct Block	14	35	**75**
	3. Grasp Block	10	18	**61**
	4. Lift	43	61	**61**
PutRubbishInBin	1. Correct Trajectory	40	50	**80**
	2. Approach Rubbish	20	80	**100**
	3. Grasp Rubbish	0	2	**64**
	4. Return to Observe	6	20	**76**
	5. Move to Bin	2	4	**53**
	6. Release	0	0	**60**

processing, feedback delay compensation, and dynamic feedback weighting within the cohesive AIA framework leads to significant performance gains and strong robustness across different human experts. Furthermore, we introduced a novel task decomposition metric that offers a more nuanced evaluation of policy learning beyond simple success rates. Collectively, our findings provide a practical and lightweight guide for developing more effective and reliable interactive robot learning systems.

Looking forward, we anticipate a paradigm shift towards migrating interactive learning from simulations to physical platforms by integrating Reinforcement Learning with Vision-Language-Action (VLA) models for large-scale fine-tuning. While Large Language Models (LLMs) excel as "robot teachers" in simulation, the transition to the real world is hindered by the absence of an omniscient "ground truth" mentor. Consequently, a critical future direction involves alleviating the dependency on perfect expert models and developing frameworks capable of learning robustly from imperfect, heterogeneous real-world feedback. Such cross-disciplinary efforts will be essential to bridge high-level reasoning with low-level control, ultimately capturing a more practical understanding of robot intelligence.

Looking forward, we anticipate a paradigm shift towards migrating interactive learning from simulations to physical platforms by integrating Reinforcement Learning with Vision-Language-Action (VLA) models for large-scale fine-tuning. While Large Language Models (LLMs) excel as "robot teachers" in simulation, the transition to the real world is hindered by the absence of an omniscient "ground truth" mentor. Consequently, a critical future direction involves alleviating the dependency on perfect expert models and developing frameworks capable of learning robustly from imperfect, heterogeneous real-world feedback. Such cross-disciplinary efforts will be essential to bridge high-level reasoning with low-level control, ultimately capturing a more practical understanding of robot intelligence.

ACKNOWLEDGMENT

This work was supported by the Shanghai Institute for Mathematics and Interdisciplinary Sciences (SIMIS) under Grant SIMIS-ID-2025-RB. The authors thank SIMIS for the financial support and for providing the computational resources and facilities essential to this research. This work was also supported by the Shanghai Science and Technology Innovation Action Plan (Grant No. 24YL1900900).

REFERENCES

[1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[2] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6892–6903.

[3] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg *et al.*, "A generalist agent," *arXiv preprint arXiv:2205.06175*, 2022.

[4] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters *et al.*, "An algorithmic perspective on imitation learning," *Foundations and Trends® in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.

[5] C. Szepesvári, *Algorithms for reinforcement learning*. Springer nature, 2022.

[6] S. Ambhore, "A comprehensive study on robot learning from demonstration," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. IEEE, 2020, pp. 291–299.

[7] C. Celemin, R. Pérez-Dattari, E. Chisari, G. Franzese, L. de Souza Rosa, R. Prakash, Z. Ajanović, M. Ferraz, A. Valada, J. Kober *et al.*, "Interactive imitation learning in robotics: A survey," *Foundations and Trends® in Robotics*, vol. 10, no. 1-2, pp. 1–197, 2022.

[8] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.

[9] Z. Xue, Z. Peng, Q. Li, Z. Liu, and B. Zhou, "Guarded policy optimization with imperfect online demonstrations," *arXiv preprint arXiv:2303.01728*, 2023.

[10] Z. M. Peng, W. Mo, C. Duan, Q. Li, and B. Zhou, "Learning from active human involvement through proxy value propagation," *Advances in neural information processing systems*, vol. 36, pp. 77 969–77 992, 2023.

[11] X. Xu, Y. Hou, Z. Liu, and S. Song, "Compliant residual dagger: Improving real-world contact-rich manipulation with human corrections," *arXiv preprint arXiv:2506.16685*, 2025.

[12] Y. Chen, S. Tian, S. Liu, Y. Zhou, H. Li, and D. Zhao, "Conrft: A reinforced fine-tuning method for vla models via consistency policy," *arXiv preprint arXiv:2502.05450*, 2025.

[13] I. Kostrikov, A. Nair, and S. Levine, "Offline reinforcement learning with implicit q-learning," *arXiv preprint arXiv:2110.06169*, 2021.

[14] A. Nair, A. Gupta, M. Dalal, and S. Levine, "Awac: Accelerating online reinforcement learning with offline datasets," *arXiv preprint arXiv:2006.09359*, 2020.

[15] Z. Wang, A. Novikov, K. Zolna, J. S. Merel, J. T. Springenberg, S. E. Reed, B. Shahriari, N. Siegel, C. Gulcehre, N. Heess *et al.*, "Critic regularized regression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7768–7778, 2020.

[16] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[17] F. Sasaki and R. Yamashina, "Behavioral cloning from noisy demonstrations," in *International Conference on Learning Representations*, 2020.

[18] K. Zolna, A. Novikov, K. Konyushkova, C. Gulcehre, Z. Wang, Y. Aytar, M. Denil, N. De Freitas, and S. Reed, "Offline learning from demonstrations and unlabeled experience," *arXiv preprint arXiv:2011.13885*, 2020.

[19] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: Challenges, methods and progress," *Artificial Intelligence*, vol. 297, p. 103500, 2021.

[20] A. Y. Ng, S. Russell *et al.*, "Algorithms for inverse reinforcement learning." in *Icml*, vol. 1, no. 2, 2000, p. 2.

[21] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer, "Hg-dagger: Interactive imitation learning with human experts," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8077–8083.

[22] K. Menda, K. Driggs-Campbell, and M. J. Kochenderfer, "Ensembledagger: A bayesian approach to safe imitation learning," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5041–5048.

[23] R. Hoque, L. Y. Chen, S. Sharma, K. Dharmarajan, B. Thananjeyan, P. Abbeel, and K. Goldberg, "Fleet-dagger: Interactive robot fleet learning with scalable human supervision," in *Conference on Robot Learning*. PMLR, 2023, pp. 368–380.

[24] S.-W. Lee, X. Kang, and Y.-L. Kuo, "Diff-dagger: Uncertainty estimation with diffusion policy for robotic manipulation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 4845–4852.

[25] M. S. Nazeer, C. Laschi, and E. Falotico, "Soft dagger: Sample-efficient imitation learning for control of soft robots," *Sensors*, vol. 23, no. 19, p. 8278, 2023.

[26] X. Sun, S. Yang, M. Zhou, K. Liu, and R. Mangharam, "Mega-dagger: Imitation learning with multiple imperfect experts," *arXiv preprint arXiv:2303.00638*, 2023.

[27] J. Luijkx, Z. Ajanović, L. Ferranti, and J. Kober, "Askdagger: Active skill-level data aggregation for interactive imitation learning," *arXiv preprint arXiv:2508.05310*, 2025.

[28] T. Kobayashi, "Cubedagger: Improved robustness of interactive imitation learning without violation of dynamic stability," *arXiv preprint arXiv:2505.04897*, 2025.

[29] T. Kaufmann, P. Weng, V. Bengs, and E. Hüllermeier, "A survey of reinforcement learning from human feedback," 2024.

[30] J. Dai, X. Pan, R. Sun, J. Ji, X. Xu, M. Liu, Y. Wang, and Y. Yang, "Safe rlhf: Safe reinforcement learning from human feedback," *arXiv preprint arXiv:2310.12773*, 2023.

[31] E. Chisari, T. Welschehold, J. Boedecker, W. Burgard, and A. Valada, "Correct me if i am wrong: Interactive learning for robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3695–3702, 2022.

[32] H. Cai, Z. Peng, and B. Zhou, "Predictive preference learning from human interventions," *arXiv preprint arXiv:2510.01545*, 2025.

[33] ——, "Robot-gated interactive imitation learning with adaptive intervention mechanism," *arXiv preprint arXiv:2506.09176*, 2025.

[34] Z. Peng, Z. Liu, and B. Zhou, "Data-efficient learning from human interventions for mobile robots," *arXiv preprint arXiv:2503.04969*, 2025.

[35] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[36] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[37] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[38] P. Intelligence, A. Amin, R. Aniceto, A. Balakrishna, K. Black, K. Conley, G. Connors, J. Darpinian, K. Dhabalia, J. DiCarlo *et al.*, "$\pi^*_{0.6}$: A VLA that learns from experience," *arXiv preprint arXiv:2511.14759*, 2025.

[39] M. Pan, S. Feng, Q. Zhang, X. Li, J. Song, C. Qu, Y. Wang, C. Li, Z. Xiong, Z. Chen *et al.*, "Sop: A scalable online post-training system for vision-language-action models," *arXiv preprint arXiv:2601.03044*, 2026.

[40] J. Luo, C. Xu, J. Wu, and S. Levine, "Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning," *Science Robotics*, vol. 10, no. 105, p. eads5033, 2025.

[41] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.

[42] A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[43] J. MacGlashan, M. K. Ho, R. Loftin, B. Peng, G. Wang, D. L. Roberts, M. E. Taylor, and M. L. Littman, "Interactive learning from policy-dependent human feedback," in *International conference on machine learning*. PMLR, 2017, pp. 2285–2294.