Residual RL with Uncertainty for Dexterous Manipulation under Imperfect Imitation (CS234 Final Project Report)

Kuo-Han Hung* [1] **Max Liu*** [1] **Ava Kouhana*** [1]

Abstract

Residual reinforcement learning (Residual RL) integrates imitation learning (IL) with reinforcement learning (RL) by training a residual policy to refine an imperfect base imitation policy. While effective for dexterous manipulation (Li et al., 2025a), prior work typically assumes access to high-quality demonstrations , which are often noisy, limited, or inconsistent in real-world settings. In this study, we systematically investigate how imitation quality impacts residual RL performance and show that smaller, more consistent datasets can yield better results than larger, noisier ones. Building on this insight, we propose an uncertainty-aware residual RL framework that dynamically adjusts residual magnitudes based on the confidence of the imitation policy. Empirical evaluation on the *write on paper* task in OakInk v2 (Zhan et al., 2024) demonstrates a 6.3% relative improvement over the baseline, highlighting the effectiveness of our approach.

1. Introduction

In this work, we aim to systematically study residual reinforcement learning (Residual RL) (Johannink et al., 2018b; Ankile et al., 2024b) for dexterous manipulation (Andrychowicz et al., 2020; OpenAI et al., 2019b). Dexterous manipulation demands precise, contact-rich control of high-degree-of-freedom robotic hands interacting with articulated objects. It is widely regarded as one of the most challenging problems in robotics due to complex and discontinuous contact dynamics, partial observability, long horizons, and high-dimensional action spaces.

Learning dexterous manipulation policies purely with reinforcement learning remains extremely difficult. The large control space, sparse and delayed rewards, and costly exploration make end-to-end RL both data inefficient and unstable. Residual reinforcement learning has recently emerged as a promising alternative (Li et al., 2025a). As shown in Figure 1, residual policy is trained to produce corrective actions on top of a pre-existing base policy, typically obtained via imitation learning from human demonstrations. The base policy provides a structured prior and reasonable initial behavior, while the residual policy focuses on refinement and error correction. This decomposition can substantially improve sample efficiency and training stability, especially in dexterous manipulation settings.

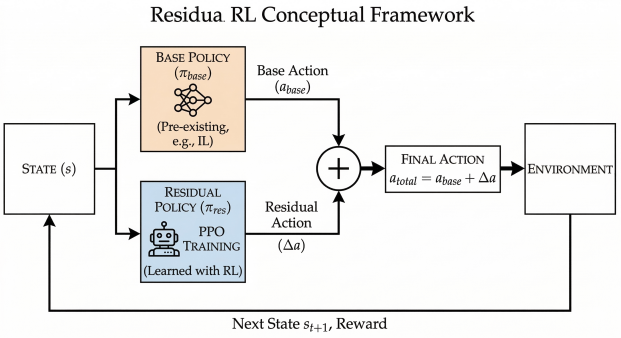


Figure 1. Residual RL framework.

Despite these advantages, most prior work (Johannink et al., 2018b; Ankile et al., 2024b) assumes that the imitation policy is reasonably accurate and reliable. In practice, however, demonstrations are often noisy, limited, or inconsistent (Ross et al., 2011), especially for challenging dexterous manipulation tasks. When the imitation policy is imperfect, residual learning may struggle because the residual policy must compensate for large or unpredictable errors in the base policy. Understanding how the quality of imitation affects residual RL therefore becomes an important but underexplored question.

In this work, we study the interaction between imitation learning and residual RL in dexterous manipulation. Our goal is to understand what characteristics of an imitation policy make it effective for residual learning, and how residual RL behaves when the imitation prior is imperfect. Through

Stanford University [1]. Correspondence to: All authors are enrolled in CS234. Kuo-Han Hung,Max Liu,Ava Kouhana <khhung,maxliu01,akouhana@stanford.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

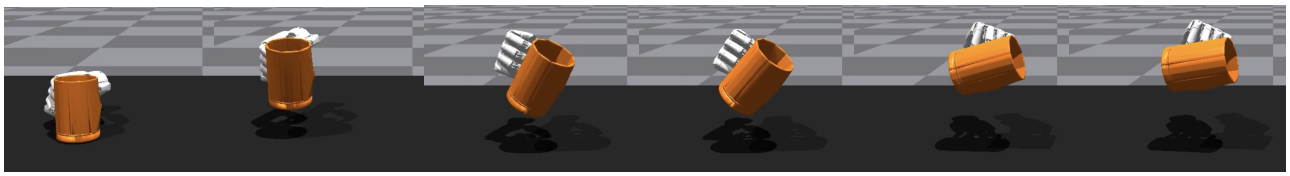


Figure 2. Dexterous manipulation on a pour cup task in OakInk v2 dataset (Zhan et al., 2024).

a series of empirical analyses on the OakInk v2 benchmark (Zhan et al., 2024), we investigate how factors such as demonstration noise, dataset size, and trajectory consistency influence residual RL performance.

Our experiments reveal an unexpected finding: increasing the amount of demonstration data does not necessarily improve residual learning. Instead, we observe that training imitation policies on smaller but more consistent datasets can lead to significantly better residual RL performance. This suggests that residual policies benefit from stable and structured imitation priors rather than highly diverse or multimodal demonstrations.

Motivated by this observation, we further propose an uncertainty-aware residual learning approach that adapts the magnitude of residual actions based on the confidence of the imitation policy. Intuitively, the residual policy should make small corrections when the imitation policy is confident, and larger adjustments when the imitation policy is uncertain or encounters out-of-distribution states. By incorporating an uncertainty-aware regularization into the PPO objective, we encourage adaptive exploration that focuses on states where the imitation policy is less reliable.

Empirical results show that the proposed uncertainty-aware residual RL improves performance on the *write on paper* task in the OakInk v2 dataset (Zhan et al., 2024) with a relative success rate improvement of 6.3% over the baseline residual RL method. These findings highlight the importance of accounting for imperfect imitation when designing residual RL systems for dexterous manipulation.

In summary, our contributions are threefold:

- **Showed importance of the Imitation Prior:** Residual RL is highly sensitive to the quality of the imitation prior, with performance degrading sharply as demonstration noise increases.
- **Discovered Consistency Outweighs Quantity:** Increasing the quantity of demonstration data does not necessarily improve performance; instead, training on smaller, more consistent datasets leads to significantly better residual RL outcomes.
- **Invented Uncertainty-Aware Adaptation:** We introduce an uncertainty-aware mechanism that modulates residual action magnitudes based on the confidence of the imitation policy, enabling larger corrections in out-of-distribution states.

2. Related Work

RL for Dexterous Manipulation Dexterous manipulation remains challenging due to high-dimensional control and contact-rich dynamics. Early successes relied on large-scale reinforcement learning (RL) in simulation, demonstrating in-hand manipulation with multi-fingered hands (OpenAI et al., 2019a; Li et al., 2025b). However, due to the large action space in dexterous manipulation, RL from scratch is highly inefficient and requires a long time to learn a good policy. Therefore, a more recent trend is to leverage human demonstrations to guide RL training. For example, some approaches use human motion to generate wrist trajectories while using RL only to learn finger movements (Chen et al., 2024). There are also works that utilize affordance information (Mandikal & Grauman, 2021; Zhang et al., 2025) to facilitate RL learning. More recently, some methods (Mandi et al., 2025) propose learning policies through curriculum strategies that gradually increase task difficulty during training. In this work, we focus on the residual RL formulation (Li et al., 2025a; Hsieh et al., 2025), where an imitation policy is provided and RL is applied on top of it.

RL for Imitation Learning Imitation learning (IL) is a compelling alternative to RL, as the use of demonstrations can mitigate or even eliminate the burden of exploration. However, it often requires accurate on-robot action data, which is challenging to capture for dexterous hands. Most existing approaches (Li et al., 2023; Arunachalam et al., 2022; Fang et al., 2025; Qin et al., 2023) require setting up a teleoperation system customized for a particular robot hand embodiment. Human hand data (such as videos) provide another source of information. Prior work has used human hand data for learning rough grasp affordances and improving retargeting (Antotsiou et al., 2018), but these approaches have been limited to short-horizon manipulation (mainly grasping). Even with perfect demonstrations, IL still easily suffers from out-of-distribution scenarios, resulting in lower success rates. In this work, we instead use residual RL, which improves upon the imperfect imitation prior.

Residual Learning Residual RL combines IL and RL by learning corrective actions on top of a base policy (Johannink et al., 2018a). This decomposition improves sample efficiency and stabilizes training by leveraging a structured prior. Recent work has applied residual learning to manipulation and assembly tasks (Ankile et al., 2024a; Li et al., 2025a; Hsieh et al., 2025; Su et al., 2026), typically assuming a strong imitation policy. However, this assumption is often unrealistic in dexterous manipulation, where demonstrations are imperfect. Our work builds on this line by explicitly studying how imitation quality affects residual learning and proposing uncertainty-aware residual updates.

3. Preliminaries

3.1. Problem Settings

Following (Li et al., 2025a), we model the dexterous manipulation task as a Markov Decision Process (MDP):

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma) \tag{1}$$

where $s \in \mathcal{S}$ is the state (e.g., robot joint states and object pose), $a \in \mathcal{A}$ is the action (e.g., joint torques), P denotes the transition dynamics, r the reward function, and $\gamma \in [0, 1]$ the discount factor.

We learn a policy $\pi(a \mid s)$ using proximal policy optimization (PPO) (Schulman et al., 2017), maximizing the discounted return:

$$E\left[\sum_{t=1}^{T} \gamma^{t-1} r_t^{\text{stage}}\right]. \tag{2}$$

In residual reinforcement learning, the policy is decomposed as:

$$a = \pi_{\text{IL}}(s) + \pi_\theta(s), \tag{3}$$

where π_{IL} is a fixed (or pre-trained) imitation policy and π_θ is a residual policy trained with RL.

The residual policy is optimized with PPO using the clipped objective:

$$\mathcal{L}_{\text{PPO}} = E\left[\min\left(r_t(\theta)\hat{A}_t,\ \text{clip}\big(r_t(\theta), 1-\epsilon, 1+\epsilon\big)\hat{A}_t\right)\right]. \tag{4}$$

3.2. Experiment Settings

All experiments are conducted on the OakInk v2 benchmark (Zhan et al., 2024), a large-scale simulation dataset for dexterous bimanual hand–object manipulation. Policies are trained in simulation with the OakInk environment, where observations consist of robot joint states and object poses, and actions correspond to joint position commands. We adopt a Gaussian policy architecture, where actions are sampled from a Gaussian distribution parameterized by a neural network policy, and trained with proximal policy optimization (Schulman et al., 2017). To evaluate performance, we report success rate by initializing each episode from a standard initial state distribution and averaging results over the last three training checkpoints. Success detection follows the protocol in ManipTrans (Li et al., 2025a) by using the object trajectory: a rollout is deemed successful if the object remains within the ground-truth demonstration trajectory bounds throughout the episode. The reward function is task-specific and provided by the environment, extending the base imitation reward with object-tracking and interaction terms as in ManipTrans (Li et al., 2025a), where rewards encourage the robot to follow object pose progression and maintain stable contact, including additional angular and translational penalties for articulated objects (e.g., difference between actual and demonstration object poses and velocities) to promote faithful trajectory following. An example visualization of a task setup is shown in Figure 2.

Task	Noise 0.005	Noise 0.01	Noise 0.02
Trigger lever	36.1%	22.4%	12.4%
Place on test tube rack	24.8%	18.3%	3.8%

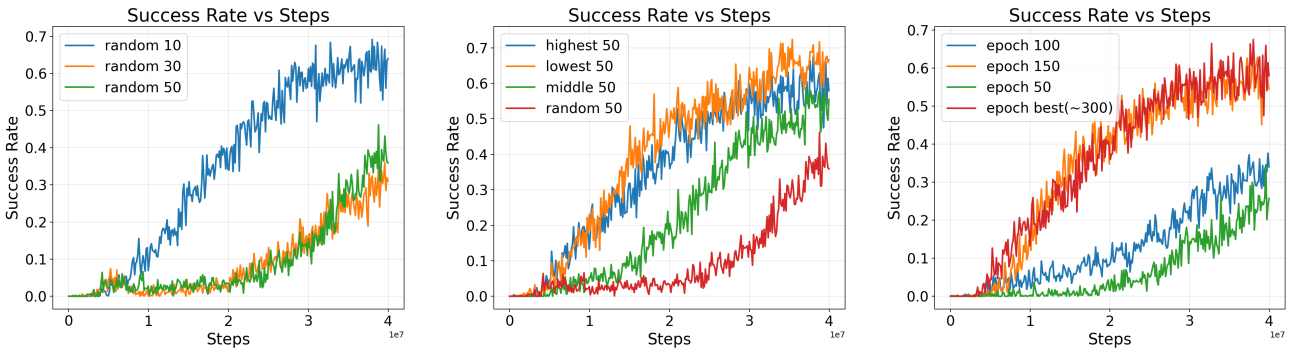
Table 1. Task success rate of residual RL for different tasks under varying noise levels in imitation prior.

4. Analysis

In this section, we analyze the relationship between imitation learning and residual RL, with a focus on the role of the imitation policy. We first show that residual RL is highly sensitive to the quality of the imitation prior; even small amounts of noise can significantly degrade performance. To investigate this further, we train several imitation policies varying in dataset size, data fitting, and data quality, and evaluate their impact on residual RL performance. Surprisingly, we find that increasing the amount of demonstration data does not necessarily improve residual RL outcomes. Instead, smaller but more consistent datasets yield better results. Motivated by this observation, we propose a data filtering strategy that selects consistent trajectories to enhance the quality of the imitation policy, ultimately leading to improved residual RL performance by 8.37%.

4.1. How important is imitation to residual RL?

To evaluate robustness to imperfect imitation, we injected Gaussian noise into the demonstration datasets and measured performance under increasing corruption levels. As shown in Table 1, as noise increases, performance degrades substantially, especially on the *place on test tube rack* task, where success rates collapse to near zero even at low noise



(a) Number of demonstrations (b) Quality of demonstrations (c) Fitting of demonstrations

Figure 3. Success rate of residual RL trained with varying imitation policies. **(a)** Leveraging imitation priors trained on uniformly sampled subsets of 10, 30, and 50 episodes. **(b)** Policies trained on different quality subsets of 50 episodes (highest, middle, and lowest returns versus random). **(c)** Base policies trained for different numbers of epochs, where the fully converged checkpoint (∼300 epochs) provides the most stable prior.

levels. This suggests that the policy heavily relies on clean imitation priors for this task and fails to generalize when demonstrations are corrupted. In contrast, the *trigger lever* task shows more resilience at low noise (notably retaining substantial success rates at 0.005 and 0.01), but performance still deteriorates sharply at higher noise (0.02). Overall, the results indicate limited robustness to degraded imitation data, with task-dependent sensitivity. This highlights the importance of the imitation prior to the residual RL training.

4.2. What kind of imitation policy is good for residual RL?

To understand the characteristics of an imitation learning policy that facilitate effective residual reinforcement learning, we analyze the imitation prior from three distinct perspectives: (1) the number of demonstrations, (2) the quality of the demonstrations, and (3) how closely the policy fits the demonstration data.

In our data collection phase, we deployed a pre-trained imitation policy coupled with a residual policy in the *write on paper* environment to collect 2,500 rollout episodes. Among these, 2,085 episodes were marked as successful. We utilize this successful subset as our core dataset for the following evaluations.

4.2.1. Number of demonstrations

To investigate the impact of demonstration quantity on residual learning, we uniformly sampled subsets of 10, 30, and 50 episodes from our successful dataset to train the base imitation policies for an identical number of steps.

Figure 3a illustrates the learning curves for the subsequent residual RL training. Surprisingly, the results demonstrate that the residual policy utilizing the imitation prior trained on just 10 episodes significantly outperforms those trained on 30 and 50 episodes. This observation indicates that simply increasing the amount of demonstration data used to train the imitation policy does not necessarily improve the performance of residual RL.

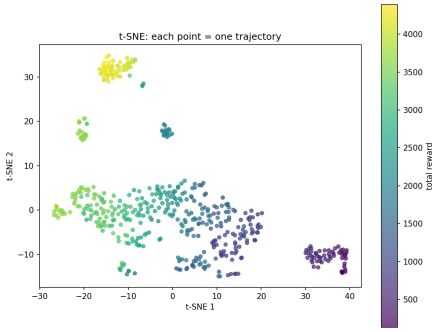


Figure 4. t-SNE visualization of the uniformly sampled 500 successful demonstration trajectories. Each point represents a single trajectory, colored by its total return. High-reward and low-reward trajectories form dense, isolated clusters, indicating high behavioral consistency. In contrast, median-reward trajectories are broadly scattered across the center, revealing high multimodality.

4.2.2. Quality of demonstrations

To further investigate how the quality of demonstrations affect residual learning, we ranked the 2,085 successful episodes based on their ground truth rewards. From this sorted dataset, we extracted three distinct subsets of 50 episodes each: the highest 50, the middle 50, and the lowest 50. We trained separate imitation policies on these subsets and compared them against the uniformly random 50-episode baseline from the previous experiment, keeping all other training configurations identical.

As illustrated in Figure 3b, the learning curves reveal an intriguing dynamic. The residual policies utilizing imitation priors trained on the highest and lowest subsets perform comparably well, with the lowest subset even demonstrating a marginal performance edge. Both of these subsets significantly outperform the middle subset, which in turn yields better results than the randomly sampled baseline.

4.2.3. Fitting of Demonstrations

Finally, to evaluate the impact of the degree of imitation fit, we selected the highest 50 subset from the previous experiment as our training data. We extracted intermediate imitation policy checkpoints trained for 50, 100, and 150 epochs, alongside the fully converged checkpoint that achieved the best validation performance (at approximately 300 epochs). These checkpoints were then independently deployed as the base policies for residual RL training.

As depicted in Figure 3c, the success rate of the residual policy consistently improves as the number of imitation training epochs increases and converges around 150 epochs and remains stable up to 300 epochs. This positive correlation clearly indicates that the imitation prior must sufficiently fit the demonstration data to provide a reliable structural foundation.

4.2.4. Formulating a Hypothesis: The Role of Consistency

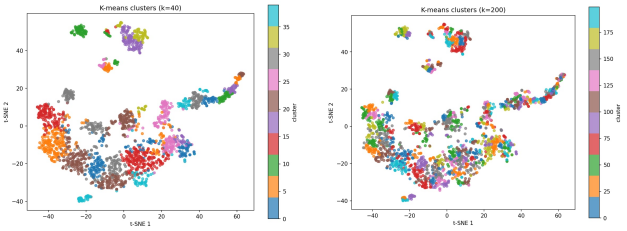
The results from the previous ablations present a seemingly counterintuitive picture: how can an imitation policy trained on significantly less data, or even on the lowest-performing demonstrations, yield a superior foundation for residual RL?

Synthesizing these observations, we formulate a core hypothesis: the effectiveness of an imitation prior for residual RL relies fundamentally on its *behavior consistency* and predictability, rather than the raw volume or the optimality of the demonstration data. We conjecture that datasets with strictly limited episodes, or those naturally grouped by extreme returns might have a higher consistency.

4.3. Is Behavioral Consistency the Key Factor?

To empirically verify our hypothesis regarding the critical role of trajectory consistency, we applied t-SNE dimensionality reduction to visualize the behavioral distribution of a subset of 500 successful trajectories, uniformly sampled from our core 2,085 episode dataset.

As illustrated in Figure 4, the visualization strongly corroborates our intuition. Each point represents a single trajectory, colored by its total reward. We observe that the trajectories with the highest total rewards form a dense, isolated cluster in the top-left region. Similarly, the lowest-reward successful trajectories cluster tightly in the bottom-right.



(a) $k = 40$ (b) $k = 200$

Figure 5. K-means clustering of successful trajectories in the t-SNE latent space. (a) At $k = 40$, the cluster with the highest consistency perfectly overlaps with the highest-return cluster. (b) At the finer $k = 200$ resolution, consistency and return decouple into distinct clusters.

Conversely, trajectories with median rewards are widely dispersed across the center of the latent space, exhibiting significant behavioral variance and multimodality.

4.4. Revisiting Data Quantity and Quality

With the critical role of behavioral consistency established, we can now provide a cohesive explanation for the counterintuitive results in Figure 3. An imitation prior trained on just 10 random episodes outperformed larger subsets because smaller samples are statistically less likely to capture the full multimodality of the dataset, forcing the policy to learn a more deterministic distribution. This same principle explains why extreme-reward subsets (the ”highest 50” and ”lowest 50”) dramatically outperformed the ”middle 50”. As revealed by our t-SNE analysis, extreme-return trajectories naturally form dense, isolated clusters representing highly consistent physical behaviors. In contrast, median-reward trajectories are widely scattered, creating a high-variance, multimodal prior that severely hinders downstream residual exploration.

4.5. How can we get a better imitation policy for residual RL?

To systematically construct a training dataset optimized for behavioral consistency, we applied K-means clustering to the latent representations obtained from our previous analysis. We evaluated two distinct clustering resolutions: a broader grouping with $k = 40$ and a more fine-grained grouping with $k = 200$, as visualized in Figure 5.

Our objective was to isolate and compare subsets based on two criteria: the highest consistency and the highest average return. For the $k = 40$ resolution, we originally intended to extract 50 trajectories from the most consistent cluster and 50 from the highest-return cluster. Interestingly, we observed that at this resolution, the most consistent cluster completely coincided with the highest-return cluster, yielding a unified 50-episode subset. For the finer $k = 200$

Figure 6. Success rate of residual RL trained with imitation priors from different K-means subsets. The "Highest Consistency 10" subset, despite having the lowest mean reward, achieves the highest success rate, demonstrating that trajectory consistency is the most critical factor for residual RL optimization.

Table 2. Quantitative characteristics of the subsets extracted via K-means clustering. Similarity represents the degree of trajectory consistency.

Subset Configuration	Mean Reward	Similarity
Consistency (Reward) 50	4226.86	0.9152
Reward 10	**4290.52**	0.9406
Consistency 10	363.23	**0.9639**

resolution, the attributes successfully decoupled. We extracted 10 trajectories from the most consistent cluster and 10 trajectories from the highest-return cluster.

Since all selected subsets are derived from K-means clusters, they inherently possess a strong baseline of behavioral consistency. By evaluating these specific settings, our goal is to answer a more nuanced question: once a high level of consistency is already established, what becomes the dominant factor for downstream residual RL performance? Specifically, we aim to determine whether it is more critical to further maximize consistency, or if trajectory quality takes precedence as the primary driver of success.

To evaluate these three settings, we present the quantitative characteristics of each extracted subset in Table 2 and the learning curves in Figure 6. Strikingly, the "Highest Consistency 10" subset achieves the highest success rate despite a drastically lower mean reward (363.23 vs. >4200). This confirms our hypothesis: strict behavioral consistency is far more critical for residual RL than absolute demonstration quality. A predictable prior, even from sub-optimal trajectories, provides a stable foundation, allowing the residual policy to focus on refinement rather than correcting erratic base actions.

4.6. Result Analysis

Building upon the insights from our previous analyses, we evaluated the overall impact of prioritizing trajectory consistency by comparing the residual RL performance of our best-performing imitation policy against the originally pre-trained imitation baseline. The quantitative results demonstrate a clear advantage for the consistency-driven approach: the residual policy leveraging the highly consistent prior achieved a task success rate of 73.8%, compared to the 68.1% success rate of the original pre-trained baseline. This represents a relative performance improvement of 8.37%.

To further investigate the physical behaviors driving this statistical improvement, we conducted a qualitative examination of the rollout trajectories generated by both policies, as visualized in Figure 7. Our observations revealed that the residual policy guided by the highly consistent imitation prior produced noticeably smoother and continuous manipulation sequences (Figure 7a). In contrast, the residual policy relying on the original pre-trained prior frequently exhibited erratic, localized movements, as seen in Figure 7b, with the robotic hand prone to vibrating or getting stuck in place without making meaningful progress.

5. Uncertainty-Aware Residual RL

5.1. Motivation

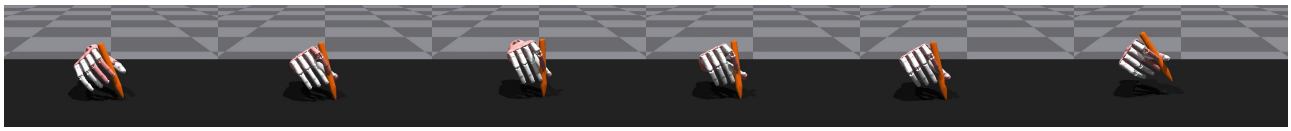
From the previous analysis, we observe that Residual RL benefits most from a highly consistent, predictable imitation prior, even if it is sub-optimal in terms of reward. Because the base policy is trained on a narrowly consistent distribution, it may exhibit high uncertainty when encountering out-of-distribution states during RL exploration. Therefore, we propose *uncertainty-aware residual RL*, which adapts residual magnitude to local imitation uncertainty by penalizing large residuals in high-certainty regions and encouraging larger residuals in high-uncertainty regions. We introduce two independent methods to achieve this.

5.2. Uncertainty Function

Let $\sigma(s) \in R_{\geq 0}$ denote the uncertainty of the imitation policy at state s. In general, $\sigma(s)$ can be any measure of epistemic or distributional uncertainty. For the experiments in this work, we use the minimum Euclidean distance from the current wrist position to any timestep in the demonstration trajectory:

$$\sigma(s) = \min_{t \in [1,T]} \left\| p_{\text{wrist}}(s) - p_{\text{demo}}^{(e,t)} \right\|_2 \tag{5}$$

where $p_{\text{wrist}}(s) \in R^3$ is the wrist position in state s, $p_{\text{demo}}^{(e,t)} \in R^3$ is the wrist position at timestep t in the demonstration trajectory for environment e, and T is the trajectory length. High $\sigma(s)$ indicates the agent is far from the demonstrated



(a) Highly consistent prior: Fluid and successful reorientation.

(b) Original pre-trained prior: Static manipulation and poor control.

Figure 7. Qualitative comparison of physical behaviors between policies. (a) The trajectory generated using the highly consistent imitation prior shows a clear, fluid strategy for reorienting the pen. (b) The trajectory relying on the original pre-trained prior reflects hesitant, static manipulation where the hand is prone to vibrating in place without meaningful progress.

distribution.

5.3. Method 1: Loss-Based Penalization

We add an auxiliary loss term to the PPO objective:

$$\mathcal{L}_{\text{uncertainty}} = -\lambda \cdot E_{(s,a)\sim\mathcal{B}} \left[\sigma(s) \cdot \|\mu_\theta(s)\|_2\right] \tag{6}$$

where $\mu_\theta(s)$ is the mean residual action, $\mathcal{B}$ is the mini-batch, and $\lambda > 0$ is a hyperparameter. The negative sign ensures that minimizing this term increases $\|\mu_\theta(s)\|$ when $\sigma(s)$ is large. The total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{PPO}} + \mathcal{L}_{\text{uncertainty}} \tag{7}$$

where $\mathcal{L}_{\text{PPO}}$ denotes the standard PPO loss (actor, critic, entropy, and bound terms).

5.4. Method 2: Reward-Based Penalization

Alternatively, we shape the reward at each timestep:

$$r_{\text{uncertainty}}(s, a_{\text{res}}) = \alpha \cdot \sigma(s) \cdot \|a_{\text{res}}\|_2 \tag{8}$$

where a_{res} is the executed residual action and $\alpha > 0$ is a scaling coefficient. This term is added to the environment reward: $r_{\text{total}} = r_{\text{task}} + r_{\text{uncertainty}}$. PPO maximizes cumulative reward, so this incentivizes larger residuals in high-uncertainty states without modifying the loss function.

5.5. Results

Table 3 compares the baseline residual RL with the two proposed uncertainty-aware methods on the *write on paper* task. Both approaches improve performance over the original residual RL policy. The loss-based method achieves the largest improvement, increasing the success rate from 78.6% to 83.6% (+6.3%). In contrast, the reward-based method provides a smaller gain, reaching 80.1% (+1.9%).

These results suggest that explicitly regularizing the residual magnitude through the optimization objective is more effective than reward shaping for adapting residual behavior to imitation uncertainty. One possible reason is that the loss-based formulation directly influences the policy update at every gradient step, whereas the reward-based signal must propagate through value estimation and may therefore be weaker or noisier. Overall, the improvement demonstrates that encouraging larger residual corrections in high-uncertainty states can help the agent recover from imitation errors and improve task success.

Method	Success Rate (%)	Improvement
Residual RL (Original)	78.6%	+0.0
Uncertainty-Aware (Loss)	83.6%	+6.3%
Uncertainty-Aware (Reward)	80.1%	+1.9%

Table 3. Comparison of residual RL methods on task *write on paper* with uncertainty awareness. The report success rate is average across last 3 checkpoints for fairness.

6. Conclusion and Discussion

Most prior work on residual RL assumes access to a nearly optimal imitation policy and focuses on learning small corrective residuals on top of this strong prior. However, this assumption is often unrealistic for challenging real-world manipulation problems, such as dexterous manipulation, where obtaining high-quality demonstrations is difficult and imitation policies are frequently imperfect. In such settings, additional reinforcement learning exploration is necessary for the agent to fully understand the task and achieve reliable performance.

In this project, we investigated the role of the imitation policy in residual learning by examining two key questions: (1) what constitutes a "good" imitation policy for effective residual learning, and (2) what happens when the imitation policy is substantially imperfect. Our analysis suggests that the quality and consistency of the imitation policy significantly affect residual learning dynamics, and that simply increas-

ing the amount of demonstration data does not necessarily lead to better performance.

Furthermore, we show that uniformly applying residual corrections across all states can be suboptimal. Instead, adapting the magnitude of residual exploration based on the uncertainty of the imitation policy can improve performance, particularly in states that are out-of-distribution relative to the demonstrations. This observation highlights the importance of selectively allocating exploration where the imitation policy is less reliable.

7. Future Direction

The findings in our work highlight an important direction for future works: developing reinforcement learning methods that can effectively leverage imperfect or weak priors. Rather than assuming access to high-quality imitation policies, future work can focus on adaptive mechanisms that determine when to rely on the prior and when to deviate from it through exploration. Such approaches are particularly valuable for complex real-world robotic tasks, where demonstrations are often limited and imperfect.

Furthermore, our results show that more consistent imitation policies can lead to improved performance. This suggests a promising avenue for studying the learning dynamics of multimodal policies, such as diffusion-based or flow matching policies, within the residual RL framework. In particular, it is important to understand how residual learning interacts with multimodal imitation priors, and to identify the most effective architectures and training strategies for leveraging these priors.

8. Contribution

Here we list the contributions of each members:

- Kuo-Han Hung: report writing, brainstorming, design and experiment uncertainty aware residual RL (sec 5).
- Max Liu: report writing, brainstorming, train different imitation learning policies for analysis (sec 4).
- Ava Kouhana: report writing, brainstorming, design and running - error analysis (sec 4.1).

9. Additional Materials

code link: Dex-RL

References

Andrychowicz, M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

Ankile, L., Simeonov, A., Shenfeld, I., Torne, M., and Agrawal, P. From imitation to refinement – residual rl for precise assembly, 2024a. URL https://arxiv.org/abs/2407.16677.

Ankile, L., Simeonov, A., Shenfeld, I., Torne, M., and Agrawal, P. From imitation to refinement – residual rl for precise assembly, 2024b. URL https://arxiv.org/abs/2407.16677.

Antotsiou, D., Garcia-Hernando, G., and Kim, T.-K. Task-oriented hand motion retargeting for dexterous manipulation imitation, 2018. URL https://arxiv.org/abs/1810.01845.

Arunachalam, S. P., Silwal, S., Evans, B., and Pinto, L. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation, 2022. URL https://arxiv.org/abs/2203.13251.

Chen, Y., Wang, C., Yang, Y., and Liu, C. K. Object-centric dexterous manipulation from human motion data, 2024. URL https://arxiv.org/abs/2411.04005.

Fang, H.-S., Romero, B., Xie, Y., Hu, A., Huang, B.-R., Alvarez, J., Kim, M., Margolis, G., Anbarasu, K., Tomizuka, M., Adelson, E., and Agrawal, P. Dexop: A device for robotic transfer of dexterous human manipulation, 2025. URL https://arxiv.org/abs/2509.04441.

Hsieh, J., Tu, K.-H., Hung, K.-H., and Ke, T.-W. Dexman: Learning bimanual dexterous manipulation from human and generated videos, 2025. URL https://arxiv.org/abs/2510.08475.

Johannink, T., Bahl, S., Nair, A., Luo, J., Kumar, A., Loskyll, M., Ojea, J. A., Solowjow, E., and Levine, S. Residual reinforcement learning for robot control, 2018a. URL https://arxiv.org/abs/1812.03201.

Johannink, T., Bahl, S., Nair, A., Luo, J., Kumar, A., Loskyll, M., Ojea, J. A., Solowjow, E., and Levine, S. Residual reinforcement learning for robot control, 2018b. URL https://arxiv.org/abs/1812.03201.

Li, K., Li, P., Liu, T., Li, Y., and Huang, S. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning, 2025a. URL https://arxiv.org/abs/2503.21860.

Li, S., Huang, Z., Chen, T., Du, T., Su, H., Tenenbaum, J. B., and Gan, C. Dexdeform: Dexterous deformable object manipulation with human demonstrations and differentiable physics, 2023. URL https://arxiv.org/abs/2304.03223.

Li, Y., Ma, X., Xu, J., Cui, Y., Cui, Z., Han, Z., Huang, L., Kong, T., Liu, Y., Niu, H., Peng, W., Qiao, J., Ren, Z., Shi, H., Su, Z., Tian, J., Xiao, Y., Zhang, S., Zheng, L., Li, H., and Wu, Y. Gr-rl: Going dexterous and precise for long-horizon robotic manipulation, 2025b. URL https://arxiv.org/abs/2512.01801.

Mandi, Z., Hou, Y., Fox, D., Narang, Y., Mandlekar, A., and Song, S. Dexmachina: Functional retargeting for bimanual dexterous manipulation, 2025. URL https://arxiv.org/abs/2505.24853.

Mandikal, P. and Grauman, K. Learning dexterous grasping with object-centric visual affordances, 2021. URL https://arxiv.org/abs/2009.01439.

OpenAI, Andrychowicz, M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., Schneider, J., Sidor, S., Tobin, J., Welinder, P., Weng, L., and Zaremba, W. Learning dexterous in-hand manipulation, 2019a. URL https://arxiv.org/abs/1808.00177.

OpenAI, Andrychowicz, M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., Schneider, J., Sidor, S., Tobin, J., Welinder, P., Weng, L., and Zaremba, W. Learning dexterous in-hand manipulation, 2019b. URL https://arxiv.org/abs/1808.00177.

Qin, Y., Su, H., and Wang, X. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation, 2023. URL https://arxiv.org/abs/2204.12490.

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

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Su, E., Westenbroek, T., Nagabandi, A., and Gupta, A. Rfs: Reinforcement learning with residual flow steering for dexterous manipulation, 2026. URL https://arxiv.org/abs/2602.01789.

Zhan, X., Yang, L., Zhao, Y., Mao, K., Xu, H., Lin, Z., Li, K., and Lu, C. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion, 2024. URL https://arxiv.org/abs/2403.19417.

Zhang, L., Mondal, S., Bing, Z., Bai, K., Zheng, D., Chen, Z., Knoll, A. C., and Zhang, J. Dora: Object affordance-guided reinforcement learning for dexterous robotic manipulation, 2025. URL https://arxiv.org/abs/2505.14819.