Extended Abstract

Multimodal RLPD for Industrial Robotic Cable Insertion

Jehan Shah, Stanford University

Motivation. This project studies robotic cable insertion on a randomized industrial task board. A UR5e must insert an SFP or SC plug into a specified port while the board pose, port pose, target identity, and populated board configuration vary across trials. The task combines visually identical ports, sub-millimeter and degree-level contact precision, near-contact occlusion, cable snagging, and domain randomization. The central question is how to combine model-based perception, demonstrations, and online deep RL so learning is concentrated on contact-rich insertion rather than target selection or long-horizon approach.

Method. The final system is staged. A model-based tracker localizes the correct target port and a perception-based geometric controller moves the plug to a clean, mouth-relative handoff. A learned policy then owns last-mile motion from deployment-realistic inputs: three wrist RGB views, tared force/torque, TCP pose and velocity, and plug/port type. The learner follows Reinforcement Learning with Prior Data (RLPD): a frozen perception ResNet18-GN trunk feeds a tanh-squashed SAC actor and an ensemble of LayerNorm critics with random subset-min targets. Each update samples half from a fixed demonstration prior and half from online rollouts, with plug-type stratification in the offline half. Rewards are labeled from privileged simulator state during training only, using funnel-shaped progress shaping plus a terminal task-success score.

Findings. The work proceeded through three diagnostic experiments. First, an end-to-end flow-matching imitation policy trained on 571 successful privileged-teacher demos fit the demonstrations well: replayed 16-step action chunks matched the teacher to about 2.4 mm and 1.2 degrees after 50k training steps. However, rollouts revealed that cable snagging was not the dominant failure. Instead, goal disambiguation failed: a target token did not reliably bind to the correct physical port among visually identical alternatives. Second, explicit state tracking solved this target-binding problem. On a 100-configuration validation set at a 2 cm plug-tip-to-port-mouth diagnostic standoff, the tracker achieved 0.10 mm median target-position error, 0.096 degree median yaw error, and zero wrong-port nearest-target selections. But the perception-based geometric controller still reached only partial insertion, about 38/75 on the Tier-3 task-success score. Third, a force/torque- and proprioception-only SAC residual policy tested whether contact cues alone were sufficient after partial insertion. It typically scored around 25/75, below the geometric baseline, with rare full insertions but no durable improvement trend.

Finally, the multimodal RLPD last-mile policy was trained from the 4 cm handoff as the last experiment in the diagnostic sequence. Over a run of 88 scored rollouts it trained stably—no divergence, no collapse—but never seated: zero full insertions, zero partials, a Tier-3 score that plateaued at the 25/75 proximity cap, and a critic value that rose to the demonstrated seated value (≈ 82) and then decayed monotonically to the true on-policy value (≈ 44) as the actor accumulated only non-seating experience. A controlled ablation that replaces wrist vision with the ground-truth port pose at the same handoff reproduces this plateau point-for-point, ruling out localization accuracy as the dominant bottleneck. A separate 1 cm relative-frame behavior-cloning diagnostic fits the same deploy stack to sub-millimeter open-loop action error yet still drifts off-axis in closed loop. Together these isolate the bottleneck as unmet success seeding in prior-data RL—no online episode ever reaches the rewarding seated state—rather than a simple perception, open-loop capacity, or SAC-loss explanation.

Status and Contributions. The contribution is a grounded diagnosis of where learning is needed and a code-complete online RLPD pipeline for that regime: mouth-relative filtering, actor-started rollout logging, bag conversion, privileged reward labeling, prior-online replay updates, and deploy-actor export. The completed online run is reported as a characterized negative result, and the ground-truth-pose and behavior-cloning probes localize the cause to a missing online success signal. For this task, the useful learning problem is not broad end-to-end imitation but visually guided, contact-rich correction from a controlled handoff, and solving it requires actively seeding successful contact transitions.

Multimodal RLPD for Industrial Robotic Cable Insertion

Jehan Shah
Stanford University
jehan8@stanford.edu

Abstract

Industrial cable insertion combines visual ambiguity, precise contact dynamics, and flexible-object nuisance effects. I study this problem in the AIC robotic cable-insertion environment, where a UR5e must insert SFP or SC plugs into a specified port on a randomized task board. The project began with end-to-end flow-matching imitation from privileged-teacher demonstrations, but rollout diagnostics showed that the main failure was not cable snagging; it was binding a target token to the correct physical port among identical ports. This motivated a staged architecture: explicit model-based tracking and geometric control for target selection and hand-off, followed by multimodal RLPD for last-mile insertion. A force/torque-only SAC residual baseline was a negative result, suggesting that contact cues without wrist vision do not provide enough directional information near the port mouth. The system implements an online RLPD loop with a frozen perception backbone, 50/50 prior-online replay, plug-type-balanced offline sampling, funnel-shaped privileged reward labels, and a deployment actor that observes only wrist RGB and proprioceptive/contact state. The completed RLPD run is a characterized negative result: mechanically faithful components trained stably but never discovered a successful seating transition. A controlled ground-truth-pose ablation rules out localization accuracy as the dominant bottleneck, and a behavior-cloning probe rules out gross open-loop capacity limitations. The remaining failure is the unmet online success-seeding condition of prior-data RL. I identify automated, privileged-teacher intervention as the most direct next step.

1 Introduction

Many industrial assembly tasks require placing a compliant or flexible part into a tight fixture. Cable insertion is a useful representative case: the robot must reason visually about which connector is the target, approach without snagging the cable, and then complete a short but contact-rich seating motion. The AIC cable-insertion task instantiates this challenge with a UR5e robot, a randomized board, visually similar SFP and SC ports, and a task score that distinguishes proximity, partial insertion, and full seating.

The natural first attempt is end-to-end visuomotor imitation. A sufficiently large action-chunking policy could, in principle, map images, proprioception, and a target token directly to a sequence of Cartesian commands. However, this conflates three problems with very different structure: target disambiguation, free-space approach, and contact-rich seating. My experiments show that this conflation matters. Offline action matching looked strong, but rollouts failed because the policy did not reliably bind the target token to the correct physical port. Conversely, model-based tracking was very accurate at the pre-insertion standoff, but the perception-based geometric controller could not reliably seat the connector. The final approach therefore decomposes the task: use explicit geometry for the part it solves well, then use deep RL only for the last-mile regime where contact makes hand coding brittle.

This paper makes four contributions. First, it reports a diagnostic sequence of negative and positive results that localize the learning problem. Second, it shows that explicit tracking can remove the identical-port ambiguity and define a clean mouth-relative handoff. Third, it evaluates a force/torque-only SAC residual and finds that contact cues alone do not reliably solve seating. Fourth, it implements an online multimodal RLPD pipeline that mixes prior demonstrations with online actor-started rollouts while keeping privileged state out of the deployed policy observation.

2 Related Work

Visuomotor imitation and action chunks. Recent robot imitation policies often predict short action sequences rather than single actions. ACT uses action chunks to reduce effective horizon in fine-grained manipulation (Zhao et al., 2023), while Diffusion Policy represents action generation as a conditional denoising process (Chi et al., 2023). Flow matching provides another continuous generative modeling objective (Lipman et al., 2023); in this project it was used to learn 16-step Cartesian action chunks. These methods provide strong open-loop action priors, but they do not by themselves resolve whether a goal token has been grounded to the correct instance among visually identical objects.

Interactive correction. DAgger reduces compounding error by collecting expert labels on states visited by the learned policy (Ross et al., 2011); HG-DAgger adapts this idea to human-guided robot learning (Kelly et al., 2019). My initial plan was similar: if cable snagging dominated failures, human VR corrections could add examples of recovery behavior. The experiments changed that diagnosis. The first bottleneck was target binding, a state-estimation problem better addressed by explicit tracking than by more corrective action labels.

Offline and prior-data reinforcement learning. Offline RL methods such as IQL can extract policies from fixed datasets without online interaction (Kostrikov et al., 2022), but they cannot create new experience in the exact contact states where the demonstrator or geometric controller is weak. SAC is a maximum-entropy off-policy actor-critic method for continuous control (Haarnoja et al., 2018). RLPD asks how to use offline data during online RL with minimal changes to off-policy learning; its key recommendations include symmetric sampling from prior and online data, LayerNorm critics, high update-to-data ratios, and critic ensembles with pessimistic target backups (Ball et al., 2023). Those ideas directly motivate the current insertion-RL implementation.

Robot RL systems. HIL-SERL demonstrates that human-in-the-loop RL can train vision-based policies for precise manipulation by combining demonstrations, online RL, reward models, and human interventions (Luo et al., 2024). This project uses a related systems view but replaces human interventions with a tracked geometric handoff and privileged simulator reward labels. The gap addressed here is a simulated industrial insertion task with visually ambiguous targets and deployment observations that exclude ground-truth pose.

3 Task, Data, and Evaluation

Task. Each trial specifies a plug type, target module, and port name. SFP targets occupy slots 0–9 and SC targets slots 10–11. The board pose, port pose, and population of distractor connectors vary across trials. The robot's deployed policy may use wrist images and proprioceptive/contact state, but not simulator ground truth. The official Tier-3 score ranges from proximity credit through partial insertion to 75 points for correct full insertion; a perception-based geometric baseline reaches roughly 38/75, corresponding to partial insertion but not seating.

Demonstrations. The initial dataset contains 571 successful privileged-teacher demonstrations. The completed RLPD experiments in this report use a 4 cm mouth-relative handoff: the segment begins before the port mouth and continues through seating. This mouth-relative cut is important because SFP and SC mouth-to-seated depths differ; a seated-relative cut would start the two plug types at inconsistent physical phases. The 4 cm prior used for the reported vision and ground-truth-pose RLPD runs contains 106,989 mouth-relative transitions. A later diagnostic tightened the handoff to 1 cm and re-expressed observations and actions in a relative frame, but it was used only for the

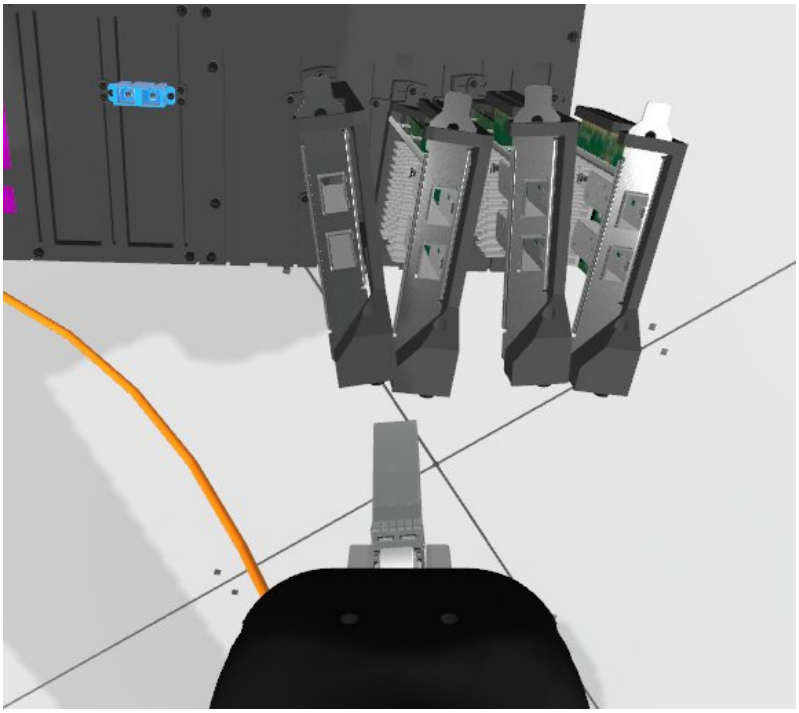

Figure 1: Randomized cable-insertion task board in simulation. The policy must insert the specified plug into the correct target port among visually similar alternatives while board pose and populated configuration vary.

behavior-cloning probe reported in Experiment 5, not as a completed RLPD result. That diagnostic contains 80,860 filtered transitions.

Evaluation metrics. I report the official Tier-3 task-success score, plug-type-specific summaries when available, and diagnostic geometry metrics. For perception, the validation set contains 100 configurations evaluated at a 2 cm plug-tip-to-port-mouth diagnostic standoff. For online RLPD, the intended final ladder compares geometric control, the initial deploy actor, and the trained RLPD actor on fresh randomized targets.

4 Method

4.1 Flow-Matching Imitation Baseline

The first policy was an end-to-end flow-matching behavior-cloning model. Its inputs were three RGB views, tared force/torque and proprioceptive history, and a target token. These streams were encoded and fused into a conditional flow model that predicted a 16-step Cartesian action chunk. At rollout time, the policy executed the first four actions from the chunk at 20 Hz, then recomputed a fresh chunk in a receding-horizon loop. This design was meant to cover the whole task and, if needed, be improved later through DAgger-style intervention.

4.2 Model-Based Tracking and Handoff

After the flow-matching rollout failures, I moved target selection out of the policy. The tracking stack detects board and port keypoints across views, triangulates them, fits CAD-constrained board and port geometry, and aggregates estimates over time. A geometric controller then moves to a safe standoff above the selected port and descends along the estimated port axis. This tracked pose is used for target selection and handoff only; it is not part of the learned insertion policy's observation.

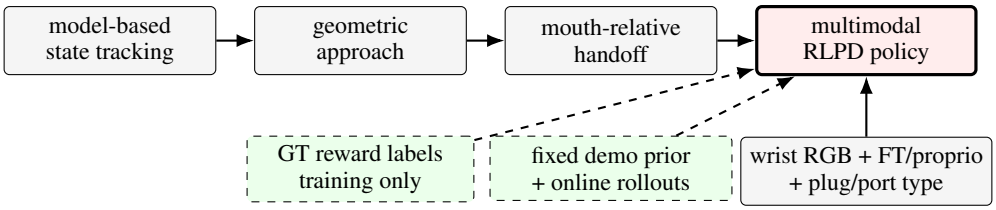

Figure 2: Staged architecture. Tracking and geometric control standardize target selection and the handoff. The learned policy owns only last-mile insertion and does not observe the tracked port pose or simulator ground truth.

4.3 FT-Only Residual SAC

The second learned baseline kept the perception-based insertion controller and trained a SAC residual on top of it. The residual policy observed force/torque and proprioception, but no wrist vision. This tested a narrow hypothesis: once the controller had reached partial insertion, perhaps contact sensing alone was enough to learn a local wiggle that seats the plug. The result was negative, which motivated restoring local vision to the learned policy.

4.4 Multimodal RLPD

The current method trains the last-mile policy with RLPD. The deployed actor observes three wrist RGB images, tared force/torque, TCP pose, TCP velocity, and plug type. It outputs a bounded 6-DoF Cartesian delta, which is converted to a pose target for the existing impedance/control stack. The policy does not observe ground truth, the target port pose, or any privileged reward features.

The visual encoder is a frozen perception ResNet18-GN trunk initialized from the trained perception checkpoint. Each view is processed by the shared trunk, then a trainable 1×1 convolution, GroupNorm, Mish nonlinearity, and pooling produce per-view embeddings. The three view embeddings are concatenated and projected to a visual latent. A state encoder processes proprioceptive/contact features and plug type. The actor is a tanh-squashed SAC policy over Cartesian deltas.

The critic follows the RLPD recipe. A shared encoder feeds an ensemble of LayerNorm Q-functions with $N = 10$ critics. Bellman targets use a random subset-min backup with $Z = 2$, which provides pessimism without taking the minimum over the entire ensemble. The update-to-data ratio is 8 by default. To avoid the memory cost of one full ResNet per critic, the perception trunk is shared and frozen; only the projection, state encoder, actor head, and critic heads train. The actor consumes detached embeddings in the first phase, so the critic loss, rather than the policy gradient, shapes the trainable representation.

Replay is symmetric: each batch contains half offline demonstration transitions and half online rollout transitions. The offline prior is fixed and sized to hold all demonstrations; it is sampled throughout training rather than rotated away. Because the raw demonstration transitions are skewed toward SFP due to slot count and longer insertion depth, the offline half is stratified to sample SFP and SC equally. Online rollouts are collected from the current actor, sliced from the recorded actor-start time to the terminal tick, labeled with the same schema and reward, added to the online replay buffer, and then used for the next RLPD update. Each online round therefore follows:

$$\text{rollout} \rightarrow \text{bag conversion} \rightarrow \text{actor-start labeling} \rightarrow \text{RLPD update} \rightarrow \text{deploy actor export}.$$

Reward. The reward uses privileged simulator geometry during training only. Dense per-step reward is a funnel-shaped progress potential:

$$r_t^{\text{shape}} = \Phi(s_{t+1}) - \Phi(s_t) - c,$$

where Φ is defined in the ground-truth port frame. It contains an always-on lateral-alignment term and an axial-progress term gated by a smooth funnel: descending in free space remains rewarded, but descending past the mouth while far off-axis receives little or negative dense progress. A terminal task-success estimate is added on the final transition. This separates smooth local guidance from the piecewise Tier-3 geometry, whose proximity and partial-insertion discontinuities produced sharp reward cliffs when used directly at every step.

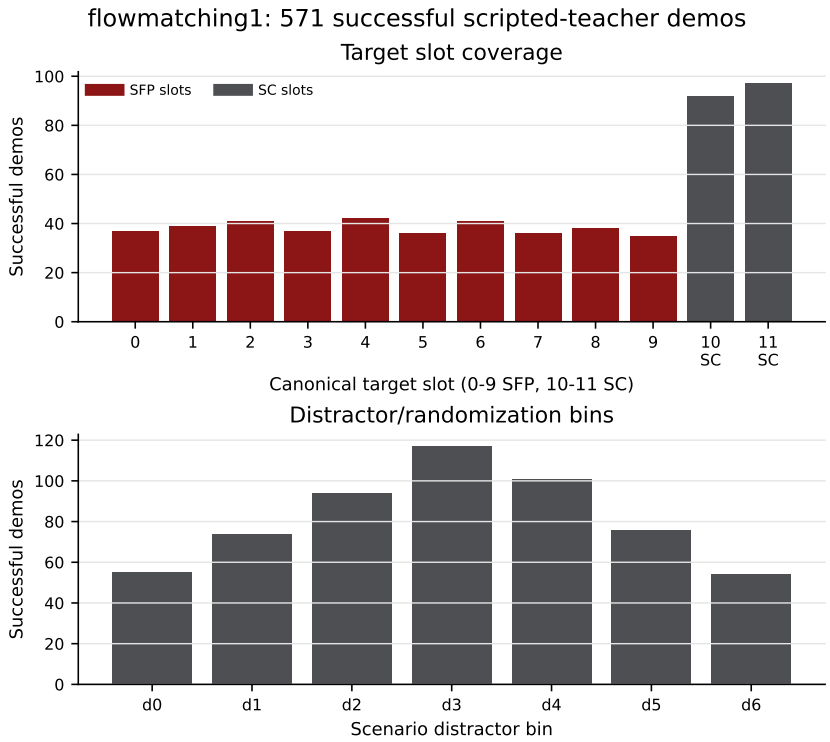

Figure 3: Teacher demonstration coverage used for the initial flow-matching imitation baseline. The dataset spans target slots, plug types, and randomized board configurations.

5 Experiments and Results

5.1 Experiment 1: Flow-Matching Imitation

The flow-matching policy was trained for 50k steps on the broad privileged-teacher demonstration set. Offline action replay looked plausible: predicted chunks matched the privileged teacher to approximately 2.4 mm and 1.2 degrees. Figure 3 shows the broad coverage of the teacher demonstrations.

Despite the offline fit, rollouts failed in a way that changed the diagnosis. Cable snagging occurred, but was not the primary bottleneck. The more severe failure was goal disambiguation: the learned policy had to infer which physical port corresponded to the target token while also generating precise motion. When several ports looked identical, this implicit binding was unreliable. This is an important distinction. More demonstrations or DAgger corrections might help closed-loop recovery, but they do not directly give the policy an explicit selected-port state.

5.2 Tracking Pivot

The model-based tracker was evaluated at the 2 cm diagnostic standoff on 100 validation configurations. The results in Table 1 show that explicit tracking solved the target-binding part of the problem: the nearest target check made no wrong-port selections. However, tracking did not solve seating. The perception-based geometric controller reached partial insertion but stalled at the lip, with a typical Tier-3 score around 38/75.

5.3 Experiment 2: FT-Only SAC Residual

The FT-only residual SAC baseline tested whether local contact sensing was sufficient once geometry had moved the plug near the mouth. Figure 4 shows the residual RL scores. The typical score was about 25/75, below the geometric baseline of about 38/75. Rare full insertions occurred, but there was no durable upward trend. The interpretation is that force/torque can reveal that contact has occurred,

Table 1: Tracker diagnostic at 2 cm plug-tip-to-port-mouth standoff on 100 validation configurations.

Metric	Median	p90 / count
Target position error	0.10 mm	0.49 mm
Target yaw error	0.096°	0.49°
Wrong-port nearest-target selections	0	0 / 100

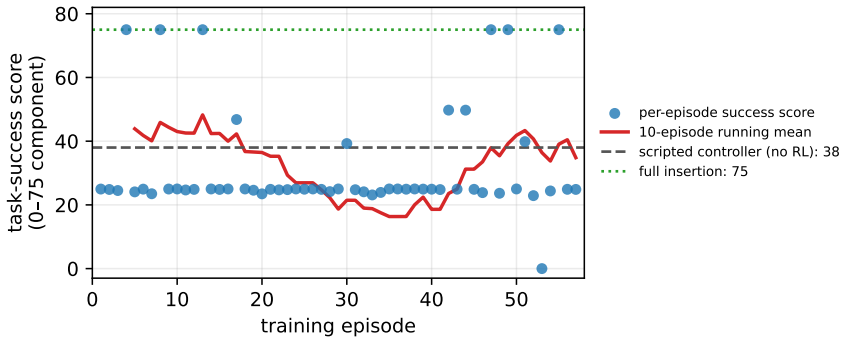

Figure 4: FT/proprio-only SAC residual scores. The blind residual policy did not reliably improve over the perception-based geometric baseline.

but without wrist vision the policy lacks a reliable directional cue for whether it is off-axis, on the lip, or entering the mouth at a bad angle. The final policy therefore needs both local vision for alignment and force/proprioception for contact.

5.4 Experiment 3: 4 cm Multimodal RLPD on the Last Mile

The last-mile RLPD policy was trained to completion overnight: a frozen ImageNet-pretrained ResNet18–GroupNorm trunk shared across actor and a ten-critic ensemble, a two-critic min target, update-to-data ratio eight, 50/50 symmetric sampling from the offline demonstration prior and the online buffer, the potential-based funnel reward, and no behavior cloning—the canonical RLPD recipe. Online episodes used the perception/tracking stack to reach the mouth-relative handoff, then handed the insertion segment to the learned actor under a ten-second cap.

Training was stable throughout with no divergence, NaNs, or collapse, but did not learn to insert. Across 88 scored rollouts there were zero full seats and zero partial insertions; the Tier-3 score plateaued at the 25/75 proximity cap (median 16.5), with best-per-pass scores flat across five passes (21.5, 23.8, 23.1, 24.4, 25.0). The Tier-2 contact-quality score was positive in the large majority of rollouts (median $+16.2$, negative in 14/88), so the failure is timid and imprecise, not over-forceful. The most informative signal is the critic: the bootstrapped target value climbed from the offline prior toward the demonstrated seated value (peak ≈ 82 near step 1650), then decayed *monotonically* to ≈ 44 by step 4400 as the online buffer filled with non-seating experience, while the temperature α annealed from 0.96 to 0.29. The policy converges to parking at the port mouth and the critic honestly revises the value of the seated state downward because no online episode ever reaches it. The $+75$ terminal reward exists only in the offline replay and is never reinforced on-policy. Stable RLPD is not the same as learned insertion.

(This interpretation depends on trustworthy value statistics. An earlier version of the loop carried a mis-scaled action log-probability; only after correcting the SAC log-probability to normalized tanh space did the value curves become meaningful. I therefore treat learner-health metrics—per-critic loss, mean and target Q, α, online buffer size, reward and terminal statistics—as part of the experimental result, not incidental logging.)

5.5 Experiment 4: 4 cm Ground-Truth-Pose Control

To separate a perception bottleneck from a reinforcement-learning bottleneck, I changed exactly one variable: port localization. In the control arm the policy reads the ground-truth port pose (a seven-vector in `base_link`, from the ground-truth TF tree) through a small MLP instead of reading the wrist images through the frozen ResNet. Everything else is identical.

The ground-truth arm reproduces the vision plateau point-for-point. Over 77 online rounds the critic value peaks and then decays on the same schedule ($\approx 67 \rightarrow 62$ across steps 2850–3150), Tier-3 stays capped in the 22–24 band, and there are zero seats (last-ten-round mean Tier-3 15.8, full seats 0/10). Perfect localization does not unlock seating. The failure therefore lies in exploration and value learning and is independent of how the port is localized.

5.6 Experiment 5: 1 cm Relative-Frame Behavior-Cloning Probe

The remaining hypothesis is that the actor, action representation, or controller simply cannot express a seating motion. I test this directly by behavior-cloning the deploy stack on a late diagnostic version of the data: a 1 cm handoff, start-relative TCP pose, body-frame velocity, and end-effector-frame Cartesian actions. The objective is Gaussian negative log likelihood plus mean-squared-error on the six-dimensional Cartesian delta. Open-loop, the fit is excellent: the validation action error converges to sub-millimeter (mean-squared error $\approx 3 \times 10^{-7}\,\mathrm{m}^2$, i.e. ≈ 0.5 mm RMS) against 5–7 mm demonstration actions. This rules out a simple open-loop expressivity failure: the stack can represent teacher-like seating actions when supervised on the demonstration distribution.

Closed-loop, the same cloned policy drifts off-axis from the handoff rather than seating. A sub-millimeter open-loop fit that fails in closed loop is the signature of compounding error: small per-step biases accumulate over the rollout, and at a one-centimeter handoff the policy is close enough to two visually similar ports that the unimodal action head regresses toward their average. This both confirms that the architecture is capable and explains why open-loop accuracy alone does not transfer—closing the loop needs either an explicit success signal during learning or on-policy correction, which motivates the intervention scheme in Section 7.

6 Discussion

The central lesson is that the failures were architectural, not merely hyperparameter-level. End-to-end imitation failed because it combined visual goal binding with precise motion generation. Model-based tracking fixed the binding problem but could not adapt to small contact states at the lip. FT/proprio-only residual RL failed because contact magnitude alone did not identify a reliable correction direction. These results justify the final decomposition: tracking and geometric control define a clean last-mile problem, and multimodal RL is used only where visual alignment and contact response must interact.

This also clarifies the role of prior demonstrations. The demonstrations are not treated as a complete solution. They anchor the critic and provide a behavior prior over reasonable insertion motions, but online interaction is needed to generate data in the failure states created by the current actor. Conversely, the online buffer should not replace the offline prior: keeping the full prior available through 50/50 replay stabilizes learning and prevents the value function from forgetting successful insertion behavior while collecting noisy early rollouts.

6.1 Root-Cause Analysis of the RLPD Failure

Experiments 3–5 form a diagnostic ladder that rules hypotheses in or out. The ground-truth-pose control (Experiment 4) removes localization error as the primary explanation: perfect localization reproduces the same plateau. The behavior-cloning probe (Experiment 5) removes the simplest capacity explanation: the same stack fits seating actions to sub-millimeter open-loop. To reduce the chance that the result is an implementation artifact, I audited the update against the RLPD reference and against the relevant SERL/HIL-SERL robot-RL design patterns; the Bellman target $r + \gamma\,\mathrm{mask}\,(\min_Z \bar{Q} - \alpha \log \pi')$, the mean-$Q$ actor, autotuned temperature, LayerNorm critic, 50/50 symmetric sampling, EMA targets, and update-to-data cadence all match. What remains is the training signal itself.

Table 2: Deviations from the RLPD-pixels and HIL-SERL RAM-insertion references, ordered by our estimated severity for this task.

Dimension	Reference	Ours
Online success seeding	warm-up / human intervention	none in Exp. 3–4
Controller compliance	tuned impedance + clips	not tuned/exposed
Action frame	local end-effector action	base-frame in Exp. 3–4; rel-frame probe later
Force handling	physical compliance	terminal reward penalty
Network capacity	ResNet + 256×256	gross capacity not limiting (Exp. 5)
Perception	task dependent	localization accuracy not limiting (Exp. 4)

The mechanism is a starvation of online success. The demonstrations teach the critic that seated states are valuable, but the actor, trained only by maximizing Q, with no behavior-cloning term, never visits the narrow bore-entry region, so every online transition is proximity or failure. The critic correctly revises the value of the *reachable* states down to the proximity ceiling, and the actor optimizes toward parking at the mouth. The $+75$ terminal lives only in offline replay and is never reinforced on-policy. This is exactly the success-seeding condition that RLPD satisfies with random warm-up and that HIL-SERL satisfies with human interventions; our last-mile-only, success-free online buffer satisfies neither (Table 2). A secondary interaction compounds it: with proximity capped near $+25$ and a force penalty of -12, a failed push (roughly $+13$) scores worse than a gentle hover (roughly $+25$), so the reward mildly reinforces not-pushing. HIL-SERL instead bounds force physically through compliance rather than penalizing it.

7 Limitations and Future Work

The central limitation is the one the analysis exposes: prior-data RL on this task needs successful online transitions, and our last-mile-only loop never produces them on its own. The learned policy is also intentionally not given the tracked pose, which keeps deployment observations realistic but makes it sensitive to handoff quality; diagnostic rollouts confirm that a poor handoff can place the actor several millimeters off-axis before it begins, and the actor can then drift farther while commanding descent. Handoff lateral error should be reported separately from closed-loop actor drift in future evaluations. The system is also simulated only and relies on CAD-specific perception and task geometry.

Automated privileged-teacher intervention (immediate next step). The direct remedy for the success-seeding deficit, and the work I identified but did not have time to land, is to inject successful and corrective transitions into the online buffer automatically, in the spirit of HIL-SERL but with the privileged teacher standing in for a human. During an online rollout a monitor would watch the ground-truth lateral offset of the plug from the port axis and, when the actor leaves a tight funnel, hand control to a teacher that re-centers the plug onto the axis (lifting clear of the lip first when it is at or below the mouth so a lateral move cannot jam), then return control once the plug is back inside the funnel. The corrective motion would be recorded as the commanded action and flow into the replay buffer; if the teacher carries recoveries through seating, the agent would observe seated and near-seated states its own exploration cannot reach. The deployed observation is unchanged, since the monitor consults ground truth only to decide when to take over. I prototyped this intervention but did not reach a version that produced stable corrections within the project window—an early implementation drove the plug away from the port rather than onto its axis—so I report it as the next step rather than as a result, and the working insertion policy remains future work. Beyond it, the ranked follow-ups are: tune controller compliance so contact yields into the bore instead of jamming; complete the relative-frame RLPD evaluation rather than only the BC probe; replace the force penalty with physical compliance; and add a behavior-cloning warm-start so the first online rollouts begin near seating. Longer-term work includes hardware transfer and ablations on frozen versus trainable visual features and the update-to-data ratio.

8 Contributions and Changes from Proposal

This was a solo project. I collected and processed the demonstration data, integrated the perception/tracking stack into the insertion handoff, trained and evaluated the flow-matching imitation baseline, implemented and diagnosed the FT-only SAC residual baseline, and built the multimodal insertion-RLPD pipeline: data filtering, reward labeling, replay buffers, stratified prior sampling, critic ensemble, env-free update loop, deploy actor export, online rollout runner, and evaluation diagnostics. I also ran and analyzed the two diagnostics that localize the RLPD failure—the ground-truth-pose control arm and the behavior-cloning probe—and prototyped the automated privileged-teacher intervention proposed as future work.

The main change from the proposal is the learning target. The proposal expected cable snagging to dominate and therefore emphasized DAgger-style interactive correction. The experiments showed that target disambiguation and last-mile contact adaptation were more important. The final system moved away from a DAgger-first and blind-residual-RL framing toward explicit tracking plus multimodal prior-data RL.

9 AI tools disclosure

AI tools were used as coding aids for this project inlcuding Claude Code and ChatGPT Codex for the following:

- Infrastructure and boilerplate: I used it to assist with the implementation of the domain-randomized configuration generation, the trial runner, ROS 2 bag $\rightarrow$ dataset conversion, dataloaders, checkpoint export/loading, plotting and reporting scripts, and test scaffolding. - Debugging: AI assisted in diagnosing failures, including the quaternion-sign bug and interpreting the critic value statistics.

The rest of the work including the research directionm, problem decomposition (perception/tracking $\rightarrow$ scripted handoff $\rightarrow$ learned last mile), the experimental design and the diagnostic ladder that isolates the RLPD failure (the ground-truth-pose control and the behavior-cloning probe), the interpretation of results, and all decisions about what to build, run, keep, and report were my own.

References

Philip J. Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. 2023. Efficient Online Reinforcement Learning with Offline Data. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*. PMLR. `https://arxiv.org/abs/2302.02948`

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. 2023. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Robotics: Science and Systems*.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 1861–1870. `https://proceedings.mlr.press/v80/haarnoja18b.html`

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. 2019. HG-DAgger: Interactive Imitation Learning with Human Experts. In *IEEE International Conference on Robotics and Automation*.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2022. Offline Reinforcement Learning with Implicit Q-Learning. In *International Conference on Learning Representations*.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*.

Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. 2024. Precise and Dexterous Robotic Manipulation via Human-in-the-Loop Reinforcement Learning. arXiv:2410.21845 [cs.RO] `https://hil-serl.github.io/`

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. 2011. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *International Conference on Artificial Intelligence and Statistics*.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. 2023. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. arXiv:2304.13705 [cs.RO]

A Code-Grounded Implementation Map

The mouth-relative filtering and online actor-start slicing are implemented in `insertion_rl/data/filter_demos.py`. The smooth shaping and terminal Tier-3 reward labels are in `insertion_rl/env/reward.py`. Replay, 50/50 symmetric sampling, and plug-type-stratified offline batches are in `insertion_rl/training/replay_buffer.py`. The RLPD agent, frozen perception trunk, critic ensemble, SAC losses, and update cadence are in `insertion_rl/training/train_rlpd.py` and `insertion_rl/training/rlpd_update.py`. The deploy node and tracked-handoff boundary are in `insertion_rl/ros.py`. The online runner is `pipeline/generation/batch_train_insertion_rl.sh`, and scoring summaries are produced by `insertion_rl/eval/scoring_report.py`. The ground-truth-pose control arm (Experiment 4) is in `insertion_rl/model/gt_pose_encoder.py`, and the behavior-cloning probe (Experiment 5) is in `insertion_rl/training/bc_pretrain_deploy.py`.