INFORMATION-GUIDED REWARD LEARNING FOR CONTACT-RICH ROBOTIC MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Contact-rich robotic manipulation requires distinguishing valid progress from ineffective motion. Yet correct insertion, misalignment, and jamming can appear visually similar, making image-based rewards unreliable. We propose an information-guided reward learning framework, which combines a physically supervised visual reward model with an action-history-based risk estimator. Privilege physical information, such as pose and contact signals, provide physical labels for reward learning, and do not appear during deployment. A dual-view visual model learns to predict positive, negative, or unclear physical progress, assigning unclear targets to visually ambiguous segments. The risk estimator learns whether visual positives are physically negative, based on executed-action history and visual class probabilities. At deployment, the risk estimator calibrate the visual reward model by changing high-risk positives to unclear. In Isaac Factory PegInsert, the framework suppresses 83.3% of negative-to-positive errors while retaining 82.9% of correct positives. Adding the learned reward to PPO increases aggregate success from 9.72% to 13.49%. On real Franka PegInsert data, Bayes-derived supervision increases mean episode-level student–teacher agreement from 43.70% to 59.32%, indicating improved supervision learnability rather than demonstrated gains in task success. These results support combining physical supervision and action history for more reliable contact-rich rewards.

1 INTRODUCTION

Contact-rich robotic manipulation requires reliable feedback on whether a motion produces physically valid progress. In assembly tasks such as peg insertion, a small pose error can turn correct insertion into misaligned loading, while both appear as similar downward motion in images Narang et al. (2022); Tang et al. (2023). If a reward model classifies both motions as positive, it may encourage a policy to persist in an invalid interaction. Reliable rewards must therefore reflect the physical outcome of motion.

Visual and vision-language reward models evaluate task progress from images, reducing reliance on manually designed state features. Existing methods improve reward prediction through demonstration similarity, temporal representations, visual value estimation, and failure analysis Sermanet et al. (2016; 2018); Nair et al. (2022); Ma et al. (2022; 2023; 2025); Duan et al. (2025); Lee et al. (2026). However, contact outcomes can depend on geometric relationships that are difficult to distinguish visually. Even with multiple frames and views, a visual model may confidently assign positive rewards to incorrect contact.

Partial observability in contact-rich manipulation makes it difficult for visual observations to fully capture contact outcomes Kaelbling et al. (1998); Koval (2016). We therefore introduce executed-action history, available at deployment, as additional evidence for visual reward prediction. Images capture visible motion changes, whereas action history records the control applied during those changes. Under occlusion or contact constraints, this control information may not be recoverable from images but may help distinguish valid progress from invalid interactions. Let Y_t denote the physical reward label, Z_t the visual history available at deployment, and H_t the executed interaction history. We focus on settings satisfying $I(Y_t; H_t \mid Z_t) > 0$, where interaction history retains information about the physical reward label conditional on the visual history.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

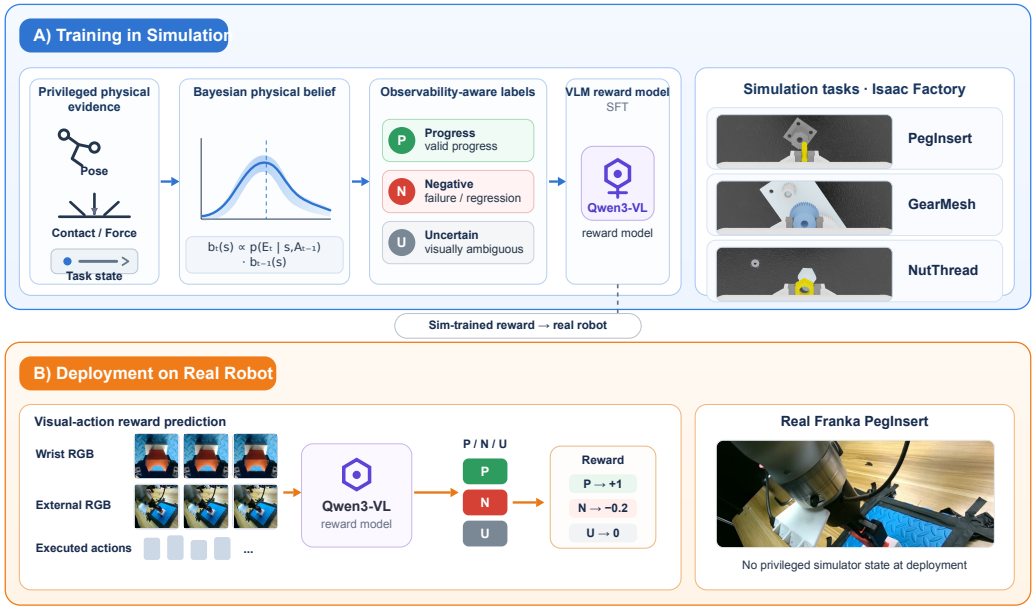

Figure 1: Overview of our Bayesian-privileged reward learning framework for contact-rich manipulation. We evaluate the method on three Isaac Lab Factory benchmark tasks—PegInsert, GearMesh, and NutThread—together with a real-robot Franka PegInsert setting. During training, privileged physical evidence is used to construct observability-aware P/N/U supervision. At deployment, the reward model predicts local progress from dual-view visual observations and executed interaction history, without access to privileged physical state. The same reward formulation is shared across the simulated Factory tasks and the real-robot setting.

Following this perspective, we use privileged physical information to establish local physical outcomes during training and action history to supplement visual judgments at deployment. Simulator pose, insertion depth, and contact state provide supervision Vapnik & Izmailov (2015); Lopez-Paz et al. (2015); Chen et al. (2020); Pinto et al. (2017). At deployment, the model evaluates rewards from images and executed actions without reading these physical states.

When constructing visual supervision, we consider both whether an interaction is valid and whether images can distinguish that validity. For segments with known physical states but insufficient visual evidence, such as a confirmed jam that still resembles valid contact in images, we retain the physical label and assign an unclear visual target. A dual-view Qwen3-VL model Bai et al. (2025) learns positive, negative, and unclear local trends from wrist and external RGB sequences, avoiding forced judgments about contact states that cannot be distinguished visually.

For the positive predictions that the visual model still produces, we introduce a history-conditioned risk estimator. It takes action history and visual class probabilities as inputs and uses the retained physical labels to learn which positive predictions actually correspond to invalid interactions. At deployment, a prediction is changed to unclear when its estimated risk reaches a threshold selected on a calibration set; otherwise, the positive prediction is retained. Abstention indicates insufficient evidence to support a positive reward and does not directly classify the segment as a physical failure El-Yaniv et al. (2010); Geifman & El-Yaniv (2019). The visual model thus estimates local progress, while the risk estimator reduces incorrect positive feedback.

Our contributions are as follows:

- **Physically grounded, observability-aware reward supervision.** We use privileged pose and contact evidence to construct local positive, negative, and unclear supervision, while explicitly assigning unclear targets to physical distinctions that are not reliably observable from deployment-time RGB.

- **History-conditioned selective rewards.** We use executed interaction history, available at deployment, to estimate the physical risk of visually positive rewards and selectively map unreliable positives to unclear, without privileged physical inputs.
- **Evaluation from physical evidence to downstream deployment.** Across three Factory assembly tasks and real Franka PegInsert, we evaluate physical belief quality, visual reward learnability, history-conditioned disambiguation, and downstream policy learning. On fresh PegInsert trajectories, action history retains 82.9% of valid positives at 83.3% false-positive suppression, compared with 52.9% retention for RGB-only selection; the learned reward further increases matched PPO aggregate success from 9.72% to 13.49%, while Bayes-derived real-robot supervision improves mean student–teacher agreement from 43.70% to 59.32%.

2 Related Work

2.1 Contact-Rich State Estimation and Active Physical Information

Contact and tactile measurements provide physical constraints that complement vision during manipulation. Classical work on Bayesian estimation and touch-based localization uses contact observations to maintain uncertainty over object pose Petrovskaya et al. (2006); Petrovskaya & Khatib (2011); Hebert et al. (2011). Later work represents pose distributions from tactile observations or formulates contact-rich manipulation directly in belief space Bauza et al. (2023); Koval (2016); Platt et al. (2010); Kaelbling et al. (1998). Active information gathering further selects interactions that reduce uncertainty about task-relevant physical states Hebert et al. (2013); Hsiao et al. (2010); Gordon et al. (2025). BayesContact combines visual proposals with simulation-based physical likelihoods for contact-rich state estimation Kamireddypalli et al. (2026).

In our method, the Bayesian physical belief is used as privileged physical evidence for constructing and auditing reward supervision. It is not used as the deployed visual reward itself.

2.2 Visual and Vision-Language Reward Learning

Traditional reward learning includes inverse reinforcement learning, ranked demonstrations, and preference-based feedback Fu et al. (2017); Brown et al. (2019; 2020); Christiano et al. (2017); Lee et al. (2021). More recently, visual reward learning has replaced hand-designed state features with perceptual and task-progress representations Sermanet et al. (2016; 2018); Nair et al. (2022); Ma et al. (2022; 2023). Vision-language models further enable language-conditioned and zero-shot reward inference Radford et al. (2021); Xu et al. (2021); Sontakke et al. (2023); Rocamonde et al. (2024); Wang et al. (2024).

Recent methods have explored several directions. GVL uses in-context temporal structure for value estimation, AHA reasons about manipulation failures, and RoboReward uses large vision-language models for robot-specific reward modeling Ma et al. (2025); Duan et al. (2025); Lee et al. (2026); Wu et al. (2026); Zhang et al. (2025).

These methods mainly improve reward estimation from visual information. Our focus is complementary: even when a visual predictor is strong, the physical validity of a visually plausible transition can remain ambiguous in contact-rich interaction.

2.3 Selective Prediction and Privileged Supervision

Selective prediction allows a model to abstain from unreliable decisions, trading coverage for lower predictive risk El-Yaniv et al. (2010); Geifman & El-Yaniv (2017; 2019); Liu et al. (2019). Related work has studied confidence-based rejection, failure prediction, and explicit confidence estimation Guo et al. (2017); DeVries & Taylor (2018); Corbière et al. (2019). Learning to defer and explicit risk control provide closely related mechanisms for handling uncertain predictions Mozannar & Sontag (2020); Angelopoulos et al. (2024).

Privileged-information learning and knowledge distillation use richer information during training than is available to the deployed model Vapnik & Izmailov (2015); Hinton et al. (2015); Lopez-Paz et al. (2015). This idea has also been widely used in robotics, where simulator state or other

privileged observations can supervise policies that later operate with limited sensing Chen et al. (2020); Pinto et al. (2017); Kumar et al. (2021).

Our setting combines these two ideas in a specific way. Privileged physical information is used to define reward supervision during training, while deployment-time interaction history is used to selectively reject unreliable visual positives.

3 Problem Formulation

3.1 Reward Prediction in Contact-Rich Manipulation

We study local reward prediction in contact-rich manipulation. Let S_t denote the latent physical state, including relative object–target pose, insertion depth, and contact mode. We denote dual-view visual history by Z_t, executed action history by H_t, and training-time privileged physical evidence by E_t. Both visual and action histories end at the current evaluation time. The prediction target is the physical trend within a local interaction window rather than the terminal outcome of an entire episode.

3.2 Reward Labels

A privileged physical teacher assigns a reward label from local state changes:

$$Y_t \in \{P, N, U\}.$$

Here, P denotes physically valid progress, such as improved lateral alignment or insertion progress under correct alignment. N denotes local regression or invalid motion, including retreat, lateral divergence, and clearly misaligned advancement. U denotes an unresolved trend or insufficient physical evidence.

Physical distinctions available to the privileged teacher may not be visible in the RGB window. Visual training must therefore account for the deployment observation boundary. We denote the visual training target by Y_t^{vis}: positive and negative physical labels are retained when supported by visual evidence, while visually unresolved windows receive U. For example, a physically confirmed jam with insufficient visual evidence can have $Y_t = N$ and $Y_t^{\text{vis}} = U$. Physical labels supervise risk estimation and evaluation; visual targets train the image-based predictor. Both are preserved independently in the data.

4 Physically Grounded Reward Learning

Our method produces local rewards through physical supervision, visual prediction, and history-conditioned risk checking. During training, a privileged physical teacher constructs physical labels Y_t and visual targets Y_t^{vis} to supervise the risk estimator and dual-view visual model, respectively. At deployment, the visual model first predicts local trends. The risk estimator then combines visual class probabilities with executed action history to check positive predictions, mapping high-risk positives to U to obtain the final label $\widetilde{Y}_t$. Prediction uses only visual history, a task description, and executed actions; privileged pose, contact truth, and the Bayesian posterior are not deployment inputs. The following subsections describe supervision construction, the visual model, risk estimation, and the output rule in turn.

4.1 Supervision from Physical Evidence

Training-time pose and contact evidence ground local rewards in physical interaction. For latent states that must be inferred from historical evidence, we maintain a Bayesian belief

$$b_t(s) = p(S_t = s \mid E_{0:t}, A_{0:t-1}),$$

where A_j denotes the action applied between states S_j and S_{j+1}. Under standard Markov assumptions, the belief update is

$$b_t(s) \propto p(E_t \mid S_t = s) \int p(S_t = s \mid S_{t-1} = s', A_{t-1}) b_{t-1}(s')\, ds'.$$

This update first predicts state changes using the executed action, then revises the probabilities of candidate states using new contact evidence. Directly available simulator states establish physical references, while the belief retains uncertainty in contact inference Petrovskaya et al. (2006); Petrovskaya & Khatib (2011); Kamireddypalli et al. (2026). For directly observed state quantities, label construction uses the corresponding evidence directly. The Bayesian belief is an auxiliary training-time physical representation and is not provided to the deployed reward model.

We construct trend supervision from local state changes. For peg insertion, a decrease in lateral alignment error d_t^{xy} or an increase in insertion depth d_t^z under correct alignment indicates positive progress. Lateral divergence, retreat, or advancement under clear misalignment indicates a local negative trend; weak changes or insufficient evidence receive U. Depth increase must be interpreted together with alignment to avoid labeling misaligned loading as valid progress. Detailed threshold criteria are given in Appendix A.2.

After establishing the physical label, we construct Y_t^{vis} using the visual window. When the visual observation supports the physical distinction, $Y_t^{\text{vis}} = Y_t$; when the images lack sufficient evidence, or the physical label itself is U, we set $Y_t^{\text{vis}} = U$. This adapts visual supervision to deployment observations while retaining the original physical label for risk learning and physical-correctness evaluation.

4.2 Dual-View Visual Reward Model

We use Qwen3-VL-2B Bai et al. (2025) as the visual reward model and fine-tune it on synchronized five-frame wrist and external RGB windows. The visual input is

$$Z_t = \left(I^{\text{wrist}}_{t-K+1:t}, I^{\text{ext}}_{t-K+1:t}\right), \qquad K = 5.$$

Given a task description ℓ, the model predicts the conditional distribution of the visual target:

$$p_\theta(Y_t^{\text{vis}} \mid Z_t, \ell), \qquad Y_t^{\text{vis}} \in \{P, N, U\}.$$

The class-probability vector and discrete prediction are

$$\hat{\mathbf{p}}_t^{\text{vis}} = [p_\theta(P \mid Z_t, \ell), p_\theta(N \mid Z_t, \ell), p_\theta(U \mid Z_t, \ell)],$$

and

$$\hat{Y}_t^{\text{vis}} = \arg\max_{y \in \{P,N,U\}} p_\theta(y \mid Z_t, \ell).$$

Privileged physical information influences learning through supervision labels, while inference uses only images and the task description. The class U allows the model to express uncertainty for visually unresolved interactions, but does not guarantee that all physical failures are identified. Since the probabilities refer to visual supervision categories, a confident positive can still conflict with the physical outcome. Such predictions are further checked by the history-conditioned risk estimator.

4.3 History-Conditioned Physical Risk Estimation

To identify physically negative windows among positive visual predictions, we combine executed action history H_t with visual class probabilities to estimate

$$\rho_t = p_\psi\left(Y_t = N \mid H_t, \hat{\mathbf{p}}_t^{\text{vis}}\right).$$

Here, ρ_t is the estimated probability of a physically negative outcome; a larger value indicates greater risk in accepting the current positive prediction. Action history provides the control context for visual changes and ends at the same evaluation time as the visual window. The estimator is supervised by physical labels Y_t, so confirmed physical negatives remain available for risk learning even when their visual targets are U.

4.4 Selective Reward Output

At deployment, risk checking applies only to positive visual predictions. We define the acceptance gate as $g_t = \mathbf{1}[\rho_t < \tau]$, where τ is chosen on an independent calibration set and fixed at test time.

Positive predictions are retained when their risk is below the threshold and mapped to U otherwise; existing N and U predictions remain unchanged:

$$\widetilde{Y}_t = \begin{cases} P, & \hat{Y}_t^{\text{vis}} = P \text{ and } g_t = 1, \\ U, & \hat{Y}_t^{\text{vis}} = P \text{ and } g_t = 0, \\ \hat{Y}_t^{\text{vis}}, & \hat{Y}_t^{\text{vis}} \in \{N, U\}. \end{cases}$$

Rejection indicates insufficient evidence to support a positive reward, so the output is U rather than a confirmed physical failure El-Yaniv et al. (2010); Geifman & El-Yaniv (2019). The threshold controls the trade-off between suppressing erroneous positive feedback and retaining valid positive feedback.

5 Experiments

Our experiments follow the full reward-learning pipeline from privileged physical supervision to deployment-time reward use. We first verify that contact evidence improves the privileged physical belief used to construct reward supervision. We then evaluate how well the resulting P/N/U reward labels can be learned from deployment-visible observations across PegInsert, GearMesh, and NutThread, highlighting task-dependent failure modes under contact ambiguity. Next, we isolate the role of executed interaction history by measuring how much it suppresses physically invalid visual positives while retaining valid ones, and further analyze this effect under controlled partial observability and threshold variation. We then test whether the learned reward provides a useful optimization signal in downstream PegInsert PPO. Finally, we evaluate the supervision formulation on real Franka PegInsert data, measuring student–teacher agreement and examining how the learned local reward transfers to held-out real-robot interactions. Together, these experiments evaluate the pipeline at five levels: physical evidence, reward learnability, history-based disambiguation, downstream policy learning, and real-robot transfer.

5.1 Experimental Setup

Tasks. We use PegInsert, GearMesh, and NutThread from Isaac Factory. The three tasks contain different contact failure modes. PegInsert mainly requires lateral alignment followed by insertion, GearMesh requires correct engagement before further motion, and NutThread additionally involves rotational alignment and sustained threading progress.

For all tasks, the visual reward model receives synchronized wrist and external RGB observations. Privileged quantities, including relative pose, insertion depth, contact state, and task outcome, are available only when constructing physical supervision and evaluation labels. They are not given to the deployed reward model.

Reward model. We fine-tune Qwen3-VL-2B to predict the local reward label

$$Y_t \in \{P, N, U\}$$

from a short dual-view visual history. For history-conditioned disambiguation, the selector additionally receives the executed action history H_t. Training, calibration, and test trajectories are separated at the episode level.

Metrics. For reward prediction, we report class-wise P/N/U recall, macro-F1, and the negative-to-positive false-positive rate. We refer to

$$Y_t = N, \qquad \hat{Y}_t = P$$

as a dangerous false positive, since it assigns a positive reward to physically invalid progress.

For selective prediction, we additionally report false-positive suppression, positive retention, and coverage. They are defined as

$$\text{FPS} = 1 - \frac{N_{\text{FP}}^{\text{after}}}{N_{\text{FP}}^{\text{before}}},$$

$$\text{Retention} = \frac{N_{\text{valid positive}}^{\text{after}}}{N_{\text{valid positive}}^{\text{before}}},$$

where coverage is the fraction of visual-positive samples retained by the selector.

5.2 PRIVILEGED PHYSICAL EVIDENCE

We first evaluate the privileged Bayesian belief used to construct physical supervision. For PegInsert, we compare the initial physical prior with the posterior obtained after incorporating contact evidence. Table 1 summarizes the resulting localization quality.

Table 1: Evaluation of privileged physical belief on PegInsert.

Method	Pose Error (mm) ↓	Grid Acc. ↑	Coverage@90 ↑
Initial prior	2.24	11.1%	–
Bayesian posterior	**0.57**	**86.7%**	**93.3%**

The posterior reduces uncertainty about the local geometric state after physical interaction. We use this estimate as privileged evidence for reward supervision rather than as an input to the deployed reward model. Additional matched probing results are reported in Appendix A.6.

5.3 VISUAL REWARD UNDER CONTACT AMBIGUITY

We next examine held-out P/N/U reward prediction across the three Factory tasks. Rather than assuming a uniform failure pattern across tasks, we report the class-wise behavior of the analyzed reward models in Table 2.

Table 2: Held-out P/N/U reward diagnostics across the three Factory tasks. $N \to P$ denotes physically negative samples predicted as positive.

Task	P Recall	N Recall	U Recall	Macro-F1	$N \to P$
PegInsert	32.33%	73.77%	42.88%	41.44%	20.49%
GearMesh	83.45%	1.72%	87.06%	52.89%	1.72%
NutThread	73.08%	23.08%	50.00%	45.72%	23.08%

The three tasks exhibit different failure patterns. PegInsert retains substantial negative recall but still produces a non-negligible $N \to P$ rate. GearMesh instead shows very low negative recall: its low $N \to P$ rate therefore reflects a tendency to map negative samples toward U, rather than reliable negative recognition. NutThread shows an intermediate but similarly imbalanced three-way boundary.

These results indicate that contact ambiguity is task dependent and is not well characterized by a single accuracy or false-positive metric. We therefore retain the full P/N/U behavior when analyzing the effect of physical supervision and history-conditioned disambiguation.

5.4 HISTORY-CONDITIONED PHYSICAL DISAMBIGUATION

We next study visual-positive predictions on fresh PegInsert trajectories. At matched 83.3% false-positive suppression, the RGB-only selector retains 52.9% of valid positives, whereas the action-history selector retains 82.9%. The resulting positive-retention gain is 30.0 percentage points, with an episode-level bootstrap 95% confidence interval of $[16.2, 43.8]$ percentage points. The corresponding coverage gain is 26.1 percentage points.

Figure 2(a) visualizes this matched operating point. The complete selector ablation, including RGB, Qwen, and action-history combinations, is reported in Appendix A.7.

This result matches the partial-observability setting in Section A.5: the current RGB observation does not always contain enough information to determine whether a positive-looking transition is physically valid, while preceding executed interactions can provide additional evidence.

5.5 PARTIAL OBSERVABILITY AND THRESHOLD SENSITIVITY

We further vary visual separability while keeping the interaction-history signal fixed, and evaluate threshold sensitivity of the history selector. The results are summarized in Figure 2.

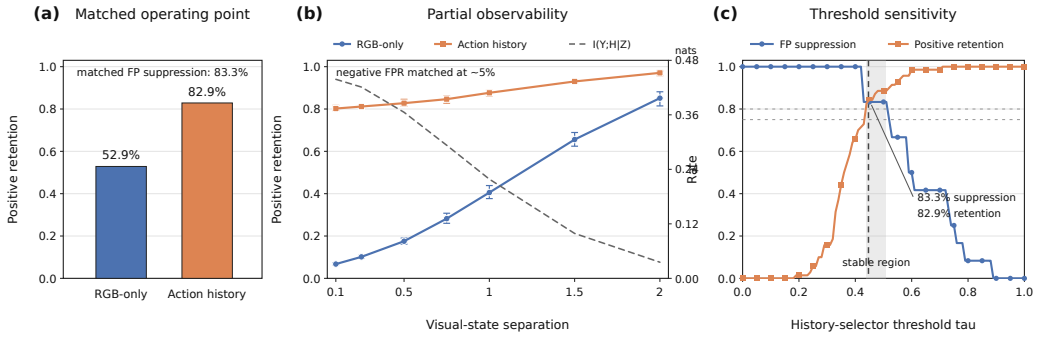

Figure 2: History-conditioned reward disambiguation under partial observability. (a) At matched 83.3% false-positive suppression, action history retains 82.9% of valid positives compared with 52.9% for RGB-only selection. (b) History-conditioned retention remains substantially more stable as visual separability decreases. (c) Threshold sensitivity shows a stable operating region.

As visual separation decreases, visual-only positive retention drops from 85.1% to 6.8%, whereas the history-conditioned selector retains 97.1–80.2%. Over the same range, the estimated conditional information $I(Y; H \mid Z)$ increases from 0.036 to 0.438 nats.

For thresholds $\tau \in [0.44, 0.51]$, the history selector maintains at least 80% false-positive suppression while retaining at least 75% of valid positives.

5.6 Downstream Policy Learning

We finally evaluate whether the learned reward provides a useful downstream training signal. Under the matched PegInsert PPO setting, we compare the learned reward with the native sparse reward. Results are reported in Table 3.

Table 3: Downstream PPO performance on PegInsert.

Reward	Aggregate Success (%) ↑	Absolute Gain
Sparse / native reward	9.72	–
Learned reward	**13.49**	**+3.77 pp**

Using the learned reward increases aggregate success from 9.72% to 13.49%, an absolute improvement of 3.77 percentage points. This provides downstream evidence that the learned reward can serve as a useful optimization signal in the PegInsert setting.

5.7 Real-Robot Evaluation

We further evaluate the supervision formulation on real Franka PegInsert. The deployment-visible student receives synchronized wrist and external RGB histories together with executed robot actions, while privileged physical signals are used only to construct training-time supervision. The frozen dataset contains 192 training episodes, 48 held-out evaluation episodes, and 80 independent test episodes.

Table 4 reports supervision learnability on the 48 held-out episodes.

Bayes-derived supervision improves mean episode-level teacher agreement from 43.70% to 59.32%, a paired gain of 15.62 percentage points. This measures *supervision learnability*, not human P/N/U ground-truth accuracy or real-robot task-success improvement.

Among 56 nonterminal teacher-disagreement windows, 55 involve U and only one is a direct $P \leftrightarrow N$ disagreement, indicating that the two teachers differ mainly in their uncertainty boundary. On the

Table 4: Real-robot supervision learnability on 48 held-out Franka PegInsert episodes. The interval uses 10,000 paired bootstrap resamples over episode ID.

Method	Mean episode agreement	Paired comparison
Bayes-derived supervision	**59.32%**	–
Heuristic supervision	43.70%	–
Δ (Bayes − Heuristic)	**+15.62 pp**	**95% CI [9.03, 22.17] pp**

independent 80-episode terminal-outcome test, the local learnability improvement does not clearly transfer to better outcome discrimination.

Figure 3 further breaks down supervision learnability across representative real-robot interaction types.

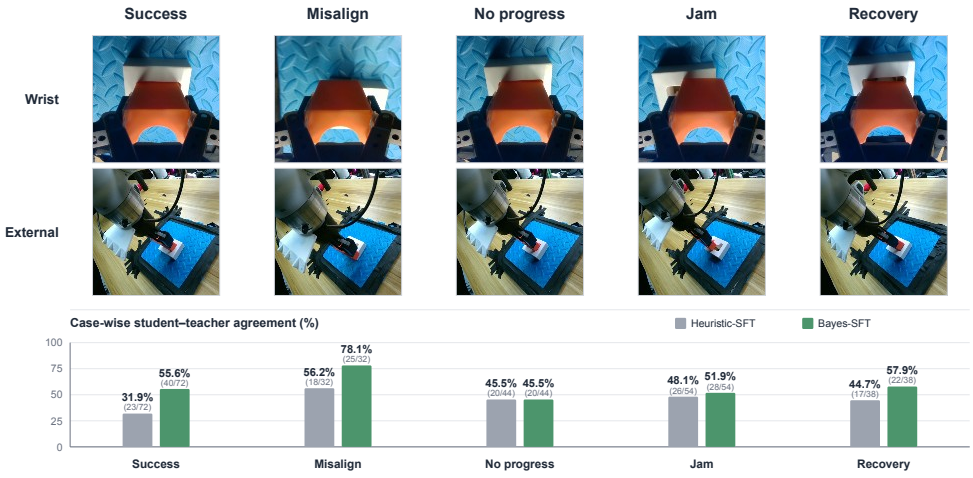

Figure 3: Real-robot Franka PegInsert examples and case-wise supervision learnability. Top: synchronized wrist and external observations from five representative interaction types. Bottom: case-wise student–teacher agreement for heuristic-supervised and Bayes-privileged SFT students, each evaluated against its own frozen privileged teacher.

6 Limitations and Conclusion

Limitations. The physical supervision used in this work requires privileged information during training. This information is directly available in simulation, while real-robot data requires calibrated geometry, additional sensing, or offline outcome annotation. Our current P/N/U representation describes local physical progress and does not cover every form of long-horizon task progress. The history selector also uses a finite action window; contact failures that depend on substantially longer interaction histories may require a richer temporal representation. Finally, the experiments focus on rigid assembly tasks, and we do not study deformable contact or tool-use settings.

Conclusion. We presented a reward-learning framework for contact-rich manipulation under partial observability. Privileged physical evidence is used to construct P/N/U supervision, while the deployed visual reward model predicts local task progress from dual-view RGB observations. When a visual prediction is positive, the executed interaction history provides an additional physical consistency signal and allows unreliable positives to be abstained to U.

The experiments cover PegInsert, GearMesh, and NutThread, together with a real-robot peg-insertion setting. They evaluate the physical belief used for supervision, reward prediction under contact ambiguity, and selective disambiguation of visual positives. The results show when interaction history provides information beyond the current visual observation and how this information can be used to make contact-rich visual rewards more reliable.

Reproducibility Statement

Implementation details and additional experimental settings are provided in the appendix. Anonymous source code for constructing the P/N/U supervision, training the visual reward model, and reproducing the selective reward evaluation is available at: `https://anonymous.4open.science/r/Bayesian_contact-0F0F/README.md`.

AI Use Statement

Generative AI tools were used during this work to assist with manuscript drafting and language editing, literature organization, code development and debugging, experiment planning, and the analysis and presentation of intermediate experimental results. AI tools were also used to help organize figures, tables, and supplementary material.

All AI-assisted text, code, experimental configurations, mathematical statements, and reported results were reviewed by the authors. Numerical results reported in the paper were obtained from executed experiments and verified against the corresponding logs and evaluation outputs. The authors take responsibility for the final content of the paper.

References

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Maria Bauza, Antonia Bronars, and Alberto Rodriguez. Tac2pose: Tactile object pose estimation from the first touch. *The International Journal of Robotics Research*, 42(13):1185–1209, September 2023. ISSN 1741-3176. doi: 10.1177/02783649231196925. URL `http://dx.doi.org/10.1177/02783649231196925`.

Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pp. 783–792. PMLR, 2019.

Daniel S Brown, Wonjoon Goo, and Scott Niekum. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Conference on robot learning*, pp. 330–359. PMLR, 2020.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Conference on robot learning*, pp. 66–75. PMLR, 2020.

P Christiano, J Leike, T Brown, M Martic, S Legg, and D Amodei. Deep reinforcement learning from human preferences. neurips. *arXiv preprint arXiv:1706.03741*, 2017.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. *Advances in neural information processing systems*, 32, 2019.

Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

Jiafei Duan, Wilbert Pumacay, Nishanth Kumar, Yi Ru Wang, Shulin Tian, Wentao Yuan, Ranjay Krishna, Dieter Fox, Ajay Mandlekar, and Yijie Guo. AHA: A vision-language-model for detecting and reasoning over failures in robotic manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=JVkdSi7Ekg`.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Ethan K Gordon, Bruke Baraki, Hien Bui, and Michael Posa. Active tactile exploration for rigid body pose and shape estimation. *arXiv preprint arXiv:2510.13595*, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Paul Hebert, Nicolas Hudson, Jeremy Ma, and Joel Burdick. Fusion of stereo vision, force-torque, and joint sensors for estimation of in-hand object location. In *2011 IEEE International Conference on Robotics and Automation*, pp. 5935–5941. IEEE, 2011.

Paul Hebert, Thomas Howard, Nicolas Hudson, Jeremy Ma, and Joel W Burdick. The next best touch for model-based localization. In *2013 IEEE International Conference on Robotics and Automation*, pp. 99–106. IEEE, 2013.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. arxiv 2015. *arXiv preprint arXiv:1503.02531*, 2, 2015.

Kaijen Hsiao, L Kaelbling, and Tomás Lozano-Pérez. Task-driven tactile exploration. *Robotics: Science and Systems VI*, 2010.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Aditya Kamireddypalli, Matias Mattamala, Joao Moura, Russell Buchanan, Sethu Vijayakumar, and Subramanian Ramamoorthy. Bayescontact: Uncertain pose estimation via visuo-tactile proposals and simulation-based inference. *arXiv preprint arXiv:2607.16123*, 2026.

Michael C Koval. *Robust manipulation via contact sensing*. PhD thesis, Carnegie Mellon University, 2016.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots. *arXiv preprint arXiv:2107.04034*, 2021.

Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training, 2021. URL `https://arxiv.org/abs/2106.05091`.

Tony Lee, Andrew Wagenmaker, Karl Pertsch, Percy Liang, Sergey Levine, and Chelsea Finn. Roboreward: General-purpose vision-language reward models for robotics. *arXiv preprint arXiv:2601.00675*, 2026.

Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. *Advances in Neural Information Processing Systems*, 32, 2019.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. Liv: Language-image representations and rewards for robotic control. In *International Conference on Machine Learning*, pp. 23301–23320. PMLR, 2023.

Yecheng Jason Ma, Joey Hejna, Chuyuan Fu, Dhruv Shah, Jacky Liang, Zhuo Xu, Sean Kirmani, Peng Xu, Danny Driess, Ted Xiao, et al. Vision language models are in-context value learners. In *International Conference on Learning Representations*, volume 2025, pp. 33984–34009, 2025.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pp. 7076–7087. PMLR, 2020.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

Yashraj Narang, Kier Storey, Iretiayo Akinola, Miles Macklin, Philipp Reist, Lukasz Wawrzyniak, Yunrong Guo, Adam Moravanszky, Gavriel State, Michelle Lu, et al. Factory: Fast contact for robotic assembly. *arXiv preprint arXiv:2205.03532*, 2022.

Anna Petrovskaya and Oussama Khatib. Global localization of objects via touch. *IEEE Transactions on Robotics*, 27(3):569–585, 2011.

Anna Petrovskaya, Oussama Khatib, Sebastian Thrun, and Andrew Y Ng. Bayesian estimation for autonomous object manipulation based on tactile sensors. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pp. 707–714. IEEE, 2006.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. *arXiv preprint arXiv:1710.06542*, 2017.

Robert Platt, Russell Louis Tedrake, Leslie P Kaelbling, and Tomas Lozano-Perez. Belief space planning assuming maximum likelihood observations. 2010.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-language models are zero-shot reward models for reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 28446–28463, 2024.

Pierre Sermanet, Kelvin Xu, and Sergey Levine. Unsupervised perceptual rewards for imitation learning. *arXiv preprint arXiv:1612.06699*, 2016.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 1134–1141. IEEE, 2018.

Sumedh Sontakke, Jesse Zhang, Séb Arnold, Karl Pertsch, Erdem Bıyık, Dorsa Sadigh, Chelsea Finn, and Laurent Itti. Roboclip: One demonstration is enough to learn robot policies. *Advances in Neural Information Processing Systems*, 36:55681–55693, 2023.

Bingjie Tang, Michael A Lin, Iretiayo Akinola, Ankur Handa, Gaurav S Sukhatme, Fabio Ramos, Dieter Fox, and Yashraj Narang. Industreal: Transferring contact-rich assembly tasks from simulation to reality. *arXiv preprint arXiv:2305.17110*, 2023.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *The Journal of Machine Learning Research*, 16(1):2023–2049, 2015.

Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. Rl-vlm-f: Reinforcement learning from vision language foundation model feedback. *arXiv preprint arXiv:2402.03681*, 2024.

Yanru Wu, Weiduo Yuan, Ang Qi, Vitor Guizilini, Jiageng Mao, and Yue Wang. Large reward models: Generalizable online robot reward generation with vision-language models. *arXiv preprint arXiv:2603.16065*, 2026.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6787–6800, 2021.

Jiahui Zhang, Yusen Luo, Abrar Anwar, Sumedh Anand Sontakke, Joseph J Lim, Jesse Thomason, Erdem Biyik, and Jesse Zhang. Rewind: Learning new tasks from language without new demonstrations. In *Inductive Biases in Reinforcement Learning Workshop@ RLC 2025*, 2025.

A Additional Methodological Details

This appendix provides additional details for the formulation introduced in Sections 3 and 4. We focus on the physical belief, the construction of observability-aware supervision, and the role of interaction history under partial observability.

A.1 Privileged Physical Belief

Let S_t denote the latent physical state at time t, including the local geometric relation between the manipulated object and the target as well as contact-related variables that are not directly observable from RGB images. During training, we additionally have access to privileged physical evidence E_t, such as relative pose, insertion depth, contact state, and task outcome. We maintain a belief over the latent physical state,

$$b_t(s) = p(S_t = s \mid E_{0:t}, A_{0:t-1}), \tag{1}$$

where $A_{0:t-1}$ denotes the executed action history.

The belief can be updated recursively through a standard Bayesian filtering step,

$$b_t(s) \propto p(E_t \mid S_t = s) \int p(S_t = s \mid S_{t-1} = s', A_{t-1}) b_{t-1}(s')\, ds'. \tag{2}$$

In our setting, this belief is used only as a training-time representation of physical evidence and uncertainty. It is not provided to the deployed reward model.

For a candidate probing action a, the expected information gain can be written as

$$\mathrm{IG}_t(a) = \mathcal{H}(b_t) - \mathbb{E}_{E_{t+1}|a,b_t}\left[\mathcal{H}(b_{t+1})\right], \tag{3}$$

where $\mathcal{H}(\cdot)$ denotes entropy. This quantity measures the expected reduction in physical uncertainty after an interaction. Similar formulations have been used in active contact sensing and Bayesian localization (Hebert et al., 2013; Hsiao et al., 2010; Kamireddypalli et al., 2026).

We do not use Eq. equation 3 as the deployed reward. Its role is to characterize the physical information available to the privileged teacher and to distinguish confident physical evidence from uncertain contact states.

A.2 Observability-Aware Physical Supervision

The privileged teacher constructs a three-way local reward label

$$Y_t \in \{P, N, U\}, \tag{4}$$

where P denotes physically valid progress, N denotes physically invalid local motion or regression, and U denotes an unresolved physical trend or insufficient physical evidence. The visual training target is separately denoted by Y_t^{vis} to account for visual observability.

Let

$$d_t^{xy} = \left\| \mathbf{p}_{\mathrm{obj},t}^{xy} - \mathbf{p}_{\mathrm{target},t}^{xy} \right\|_2 \tag{5}$$

denote the lateral misalignment, and define its local change over a temporal window ending at t, with start time $t_0 = t - K + 1$, as

$$\Delta d_t^{xy} = d_t^{xy} - d_{t_0}^{xy}. \tag{6}$$

Before stable insertion or engagement, decreasing lateral error is treated as positive progress,

$$\Delta d_t^{xy} < -\epsilon_{xy} \quad \Rightarrow \quad Y_t = P, \tag{7}$$

while a clear increase in lateral error is treated as negative motion,

$$\Delta d_t^{xy} > \epsilon_{xy} \quad \Rightarrow \quad Y_t = N. \tag{8}$$

During insertion, downward progress is considered valid only when the object is sufficiently aligned with the target. Let Δd_t^z denote the local change in insertion depth over the same window, $\Delta d_t^z = d_t^z - d_{t_0}^z$. We use

$$\Delta d_t^z > \epsilon_z, \qquad d_t^{xy} < \epsilon_c \quad \Rightarrow \quad Y_t = P, \tag{9}$$

whereas downward motion under clear lateral misalignment is labeled negative,

$$\Delta d_t^z > \epsilon_z, \qquad d_t^{xy} > \epsilon_f \quad \Rightarrow \quad Y_t = N, \tag{10}$$

with $\epsilon_f > \epsilon_c$.

The visual unclear label is important in contact-rich manipulation. Privileged state may distinguish valid contact from a jam while the corresponding RGB windows remain nearly indistinguishable. We then assign

$$Y_t^{\text{vis}} = U \tag{11}$$

without overwriting the physical label Y_t. Thus, a confirmed physical negative can remain available for risk supervision even when its visual target is unclear.

The visual target can be written as

$$Y_t^{\text{vis}} = \Phi(Y_t, Z_t), \tag{12}$$

where Φ retains a positive or negative physical label when supported by the visual window and returns U otherwise. A physical label of U also yields an unclear visual target. This separates physical correctness from the distinctions supported by deployment-time observations, following privileged-information learning (Vapnik & Izmailov, 2015; Lopez-Paz et al., 2015).

A.3 Deployable Visual Reward

Let

$$Z_t = \left(I_{t-K+1:t}^{\text{wrist}}, I_{t-K+1:t}^{\text{ext}}\right) \tag{13}$$

denote the synchronized visual history from the wrist and external cameras. The deployed visual reward model estimates

$$p_\theta(Y_t^{\text{vis}} \mid Z_t, \ell), \qquad Y_t^{\text{vis}} \in \{P, N, U\}, \tag{14}$$

where ℓ denotes the task description.

Importantly, the deployed model does not observe E_t or the belief b_t. Consequently, its prediction is restricted to physical trends that are identifiable from visual observations. The discrete visual prediction is

$$\hat{Y}_t^{\text{vis}} = \arg \max_{y \in \{P,N,U\}} p_\theta(y \mid Z_t, \ell). \tag{15}$$

This separation is intentional. The privileged teacher defines physically meaningful supervision, while the visual model learns only the part of that supervision supported by deployment-time observations.

A.4 History-Conditioned Physical Disambiguation

Current visual observations may remain ambiguous after contact. We therefore consider the executed interaction history

$$H_t = (A_{t-L}, \ldots, A_{t-1}), \tag{16}$$

which is available at deployment without privileged sensing. Here A_j acts between S_j and S_{j+1}, so the history contains no future actions.

We are mainly concerned with visually positive predictions because a false positive can provide an incorrect reward for a physically invalid state. For samples satisfying

$$\hat{Y}_t^{\text{vis}} = P, \tag{17}$$

we estimate the probability that the visual positive is physically inconsistent,

$$\rho_t = p_\psi \left(Y_t = N \mid H_t, \hat{\mathbf{p}}_t^{\text{vis}} \right), \tag{18}$$

where

$$\hat{\mathbf{p}}_t^{\text{vis}} = \left[p_\theta(P), p_\theta(N), p_\theta(U) \right]. \tag{19}$$

A visual positive is retained only when

$$\rho_t < \tau, \tag{20}$$

where τ is selected on a separate calibration set. The final label is therefore

$$\widetilde{Y}_t = \begin{cases} P, & \hat{Y}_t^{\text{vis}} = P \text{ and } \rho_t < \tau, \\ U, & \hat{Y}_t^{\text{vis}} = P \text{ and } \rho_t \geq \tau, \\ \hat{Y}_t^{\text{vis}}, & \hat{Y}_t^{\text{vis}} \in \{N, U\}. \end{cases} \tag{21}$$

Rejected positives are mapped to U rather than N. The history module is therefore not required to prove that a state is negative; it only determines whether the available evidence is sufficient to trust a positive visual reward. This is related to selective prediction and abstention (El-Yaniv et al., 2010; Geifman & El-Yaniv, 2019), but the additional evidence used here comes from the interaction history rather than from visual confidence alone.

A.5 Why Interaction History Can Help

The role of interaction history can be stated directly in information-theoretic terms. Let Y denote the physical reward label, Z the current visual observation, and H the interaction history. If

$$I(Y; H \mid Z) > 0, \tag{22}$$

then H contains information about the physical reward label that is not already contained in Z. In this case, the current visual observation is not a sufficient statistic for reward prediction.

This observation also has a simple decision-theoretic consequence. Let $\mathcal{R}^*(X)$ denote the minimum achievable classification risk when the predictor has access to observation X. Since a predictor using (Z, H) may always ignore H, its optimal hypothesis class contains the visual-only predictor as a special case. Therefore,

$$\mathcal{R}^*(Z, H) \leq \mathcal{R}^*(Z). \tag{23}$$

For zero–one loss, a change in posterior probabilities alone need not make this inequality strict. Strict improvement requires history to change optimal classification decisions in a way that reduces expected error.

Equation equation 23 does not imply that arbitrary temporal context must improve reward prediction. It only states that improvement is possible when the interaction history carries label-relevant information that is absent from the visual observation. This is the regime targeted by our history-conditioned disambiguation module. The inequality concerns optimal predictors with access to (Z, H); it is not a guarantee for the implemented module, which receives visual class probabilities rather than the full Z.

A.6 Additional Physical-Probing Analysis

We additionally compare several probe-selection strategies under a matched PegInsert probing protocol. All methods share the same first probe and differ only in the subsequent probe-selection objective. The Bayesian belief and contact evidence used in this analysis are privileged and are not provided to the deployed reward model.

Table 5: Matched physical-probing comparison on PegInsert. Each method follows the same probing protocol.

Probe Objective	Pose Error (mm) ↓	Posterior Entropy ↓	Peak Probe Force (N) ↓
Random	0.653	0.411	6.09
Information gain	0.690	0.376	6.27
MAP-directed	**0.245**	0.375	5.69
Task-aware	0.527	**0.323**	**5.32**

As shown in Table 5, task-aware probing gives the lowest posterior entropy and peak probing force, whereas MAP-directed probing gives the lowest final pose error. The comparison therefore illustrates the trade-off between geometric localization accuracy and physical interaction cost rather than establishing a single uniformly superior probing objective.

A.7 Additional Selective-Reward Ablations

We report the complete selective-reward ablation corresponding to Section 5.4. All variants operate on the same set of visual-positive predictions from fresh PegInsert trajectories.

Table 6: Full selective-reward ablation on fresh PegInsert trajectories. All methods operate on the same visual-positive prediction set.

Method	FP Supp. ↑	Pos. Ret. ↑	Coverage ↑
Visual reward only	0.0%	100.0%	100.0%
RGB-only selector	83.3%	52.9%	44.3%
Action-history selector	**83.3%**	**82.9%**	**70.4%**
Qwen + Action	100.0%	41.4%	30.4%
RGB + Action	83.3%	57.1%	48.7%
RGB + Qwen + Action	83.3%	55.7%	46.1%

At the matched 83.3% false-positive suppression operating point, the action-history selector retains 82.9% of valid positives compared with 52.9% for the RGB-only selector. This corresponds to a positive-retention difference of 30.0 percentage points, with an episode-level paired-bootstrap 95% confidence interval of $[16.2,\ 43.8]$ percentage points. The corresponding coverage difference is 26.1 percentage points.