PRICE THE ACTION CHUNKS: PHYSICAL RELATIONAL CREDIT ASSIGNMENT FOR EMBODIED REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Outcome-based reinforcement learning (RL) post-trains vision–language–action policies using terminal success signals, but assigns the same trajectory-level advantage to every action chunk. A failed episode can thus penalize useful early actions as if they caused the failure. Existing approaches seek finer-grained feedback through learned evaluators, adding task-specific supervision or additional model training. We explore, for the first time to our knowledge, whether physical relations across trajectories can provide action-chunk credit in embodied RL from terminal outcomes alone, without an auxiliary evaluator. The key insight is that rollouts reaching corresponding physical situations can serve as references for one another: their terminal outcomes provide evidence for assessing local progress. We introduce *Physical Relations for Inferring Credit from Episodes* (PRICE), with two components: (i) a physical relational graph that pools current and historical outcomes at corresponding chunk boundaries to estimate success potentials; and (ii) confidence-gated credit assignment that uses changes in these potentials to refine trajectory-level supervision. Our analysis connects oracle potential changes to the terminal-success objective and provides a finite-sample directional bound for outcome-independent evidence pools. Independent continuation tests show that PRICE's retained credits align with local progress, while experiments on LIBERO, RoboTwin 2.0, and real robots demonstrate improved task success over outcome-based baselines and faster learning.

1 INTRODUCTION

Vision–language–action (VLA) models learn general-purpose manipulation policies from large-scale imitation data. Reinforcement learning (RL) provides a way to refine these initial capabilities through interaction (Kim et al., 2024; Tan et al., 2025; Li et al., 2026a). Outcome-based approaches make this post-training practical by using task success as the reward: a completed rollout supplies the feedback needed for learning, without requiring annotations or reward design for each intermediate manipulation stage.

Long-horizon manipulation, however, contains action chunks with different effects on the final outcome. A robot may grasp an object securely and transport it to the target, then fail only during placement. The failed episode contains useful behavior as well as a decision that needs correction, but a trajectory-level objective assigns the same advantage to every chunk. Its terminal label tells the learner that the attempt failed, while leaving unresolved where progress was made or lost. This makes local credit assignment a central challenge for learning from sparse outcomes.

Existing approaches address this missing feedback through designed process rewards, learned reward or value models, and world-model-based progress estimation (Lu et al., 2025; Shu et al., 2025; Xiao et al., 2025; Zhu et al., 2026a). These mechanisms supply finer-grained supervision, but require task-specific reward design or additional learned evaluators. This motivates our question: *Can embodied RL assign fine-grained credit to action chunks from terminal outcomes alone, without introducing an additional evaluator?*

To investigate this question, we revisit the experience generated by group-based RL. Rollouts for the same task often pass through corresponding physical situations before diverging in their later

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.

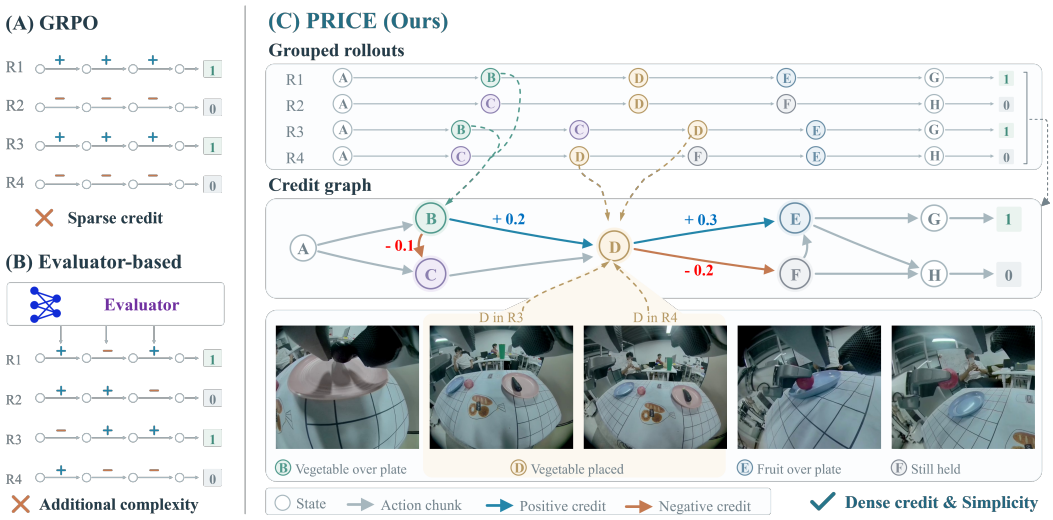

Figure 1: Motivation for PRICE. PRICE derives local action-chunk credit by comparing physically corresponding states across rollouts, using only terminal outcomes and no auxiliary evaluator.

actions and outcomes. For example, different attempts may establish a similar grasp, yet differ in whether the object remains stable during transport. Such shared situations allow trajectories to serve as references for one another. Their terminal outcomes provide evidence about which intermediate situations are more conducive to eventual success, making local progress visible even within a failed episode. Figure 1 illustrates how shared and diverging manipulation attempts connect sparse outcomes to local decisions.

The key challenge is to establish which physical situations can be compared across trajectories. Cross-trajectory credit assignment has been explored in agentic RL (Feng et al., 2026; Wang et al., 2026b;a; Yan et al., 2026); in embodied control, however, the relevant correspondence must be inferred from continuous visual and proprioceptive observations. States that look similar can differ in object pose or contact, with different implications for subsequent actions. Useful relations must therefore capture physical conditions that matter for task progress.

We therefore introduce *Physical Relations for Inferring Credit from Episodes* (PRICE) to derive action-chunk credit from physically grounded cross-trajectory relations. PRICE identifies corresponding physical situations from visual and proprioceptive observations and estimates their success potentials using outcomes from current and historical rollouts. Changes in these potentials across an action chunk provide local credit, which is selected through confidence gating and added to the GRPO advantage. This supervision comes directly from collected experience, without an auxiliary evaluator. Our analysis connects oracle potential changes to the terminal-success objective and provides a finite-sample directional bound for outcome-independent evidence pools. To our knowledge, this is the first study to investigate action-chunk-level credit assignment from physically grounded cross-trajectory relations in embodied RL.

Empirically, PRICE narrows the performance gap between evaluator-free and evaluator-based RL. It outperforms evaluator-free baselines and learns faster than SimpleVLA-RL, while remaining competitive with SRPO, a state-of-the-art evaluator-based method (Fei et al., 2026). Independent continuation evaluation shows that PRICE credit aligns with local progress measured by changes in node-averaged success. Controlled ablations establish the contributions of historical evidence and confidence gating to policy improvement. These benefits extend to real-robot manipulation, where PRICE-AWR improves offline flow-matching adaptation over trajectory-level GRPO-AWR.

Our contributions are threefold: (i) We introduce PRICE, a method for inferring action-chunk credit from physical relations and terminal outcomes without auxiliary evaluators. (ii) We establish the connection between oracle endpoint credit and the terminal-success objective, and analyze the directional reliability of pooled credit estimates under explicit evidence-selection assumptions. (iii)

We demonstrate improved simulated and real-robot manipulation, with independent credit evaluation and controlled component ablations.

2 PRELIMINARIES

Problem setup. For task context κ, let H_t denote the complete interaction history at action-chunk boundary t, and let h_t denote the policy input constructed from this history. The policy samples an action chunk $\mathbf{a}_t \sim \pi_\theta(\cdot \mid h_t)$, whose execution produces H_{t+1}. A trajectory $\tau = (H_1, \mathbf{a}_1, \ldots, \mathbf{a}_T, H_{T+1})$ receives only a terminal success indicator $Y \in \{0, 1\}$. The post-training objective is $J(\theta) = \mathbb{E}_{\kappa,\tau\sim\pi_\theta}[Y]$. Let $T_i \leq T$ be the number of chunks executed in trajectory i. For analysis, early termination is padded with absorbing histories and policy-independent dummy actions.

Outcome-based GRPO. Outcome-based GRPO (Shao et al., 2024) samples a group g of K trajectories from the same initial context under a frozen collection policy $\pi_{\theta_{\text{old}}}$. It assigns trajectory i the group-normalized terminal advantage

$$\widehat{A}_i^{\text{GRPO}} = \frac{Y_i - \bar{Y}_g}{s_{Y,g} + \varepsilon_{\text{num}}}, \qquad i \in g, \tag{1}$$

where $\bar{Y}_g$ and $s_{Y,g}$ are the group mean and sample standard deviation, and $\varepsilon_{\text{num}} > 0$ stabilizes normalization. If all outcomes in a group agree, every advantage is zero.

For tokenized action chunks, GRPO maximizes the clipped surrogate

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{N_{\text{tok}}}\sum_{i,t,\ell}\min\Big(&\rho_{i,t,\ell}(\theta)\widehat{A}_i^{\text{GRPO}},\\ &\text{clip}(\rho_{i,t,\ell}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})\,\widehat{A}_i^{\text{GRPO}}\Big) - \beta D_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\Bigg].\end{aligned} \tag{2}$$

Here $\rho_{i,t,\ell}$ is the current-to-collection policy likelihood ratio for a valid action token, and $N_{\text{tok}} = \sum_{i,t} L_{i,t}$ counts valid tokens, with $L_{i,t}$ the number in chunk t of trajectory i. The clipping bounds are ϵ_{low} and ϵ_{high}; β weights an optional reference-policy KL penalty. Every chunk in a trajectory receives the same advantage, leaving its local contribution unresolved.

3 METHOD

Given grouped rollouts labeled only by terminal success, PRICE maps action-chunk boundaries to task-specific physical nodes, estimates endpoint success potentials from cross-trajectory evidence, and converts changes in graph potential into confidence-gated action-chunk credit that augments the trajectory-level GRPO advantage, as shown in Figure 2.

We seek to credit action chunks that increase the probability of eventual success. This motivates measuring local progress by the change in success probability across chunk boundaries. For a fixed policy π, let $V_t^\pi(H_t) = \mathbb{E}_\pi[Y \mid H_t]$ denote this probability given history H_t, and define the chunk advantage $A_t^\pi(H_t, \mathbf{a}_t) = \mathbb{E}_\pi[Y \mid H_t, \mathbf{a}_t] - V_t^\pi(H_t)$. We define the oracle chunk credit as $\Delta_t^{\text{ora}} = V_{t+1}^\pi(H_{t+1}) - V_t^\pi(H_t)$, with $V_{T+1}^\pi(H_{T+1}) = Y$.

Theorem 1 (Oracle alignment of additive credit)**.** *Assume a finite-horizon process whose task and initial-context distribution, environment transitions, and terminal-outcome law do not depend on θ, together with the standard regularity conditions for likelihood-ratio policy gradients. For $\pi = \pi_\theta$,*

$$\mathbb{E}_\pi[\Delta_t^{\text{ora}} \mid H_t, \mathbf{a}_t] = A_t^\pi(H_t, \mathbf{a}_t). \tag{3}$$

For on-policy grouped rollouts, let $S_t = \nabla_\theta \log \pi_\theta(\mathbf{a}_t \mid h_t)$ and $g_{\text{base}} = \mathbb{E}_\pi[\sum_t S_t \widehat{A}^{\text{GRPO}}]$. With stop-gradient weights and $\lambda \geq 0$, the oracle score update satisfies

$$g_{\text{aug}}^{\text{ora}} := \mathbb{E}_\pi\left[\sum_t S_t\big(\widehat{A}^{\text{GRPO}} + \lambda\Delta_t^{\text{ora}}\big)\right] = g_{\text{base}} + \lambda\nabla_\theta J(\theta). \tag{4}$$

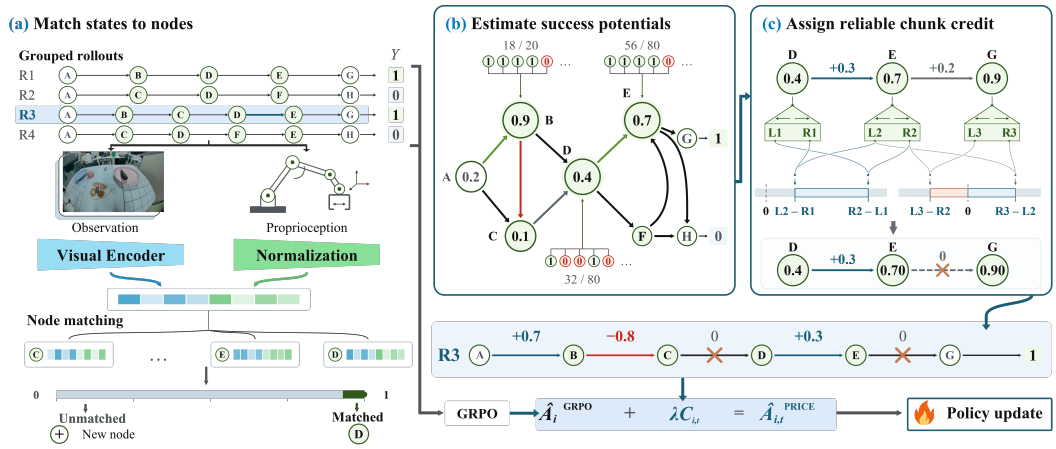

Figure 2: Overview of PRICE. (a) Visual–proprioceptive descriptors match physically corresponding states across rollouts. (b) Cross-trajectory terminal outcomes determine node success potentials. (c) Confidence-gated endpoint differences provide action-chunk credit that augments the GRPO advantage for policy optimization.

The theorem shows that weighting policy scores by oracle potential differences recovers the gradient of the terminal-success objective in expectation. Motivated by this connection, PRICE constructs local credit from differences in pooled success potentials at physically matched boundaries and uses confidence gating to select differences relative to their estimated uncertainty. Appendix B gives the proof.

3.1 Physical Relational Graph Construction

Physical Interaction Representation. Let i index a trajectory and t an action-chunk boundary. PRICE combines the visual observation $o_{i,t}$ and robot proprioception $p_{i,t}$ at that boundary into a physical interaction descriptor

$$z_{i,t} = \left[\sqrt{w_{\text{vis}}}\,\mathcal{N}(\phi_{\text{vis}}(o_{i,t}))\,;\,\sqrt{w_{\text{prop}}}\,\mathcal{N}(\phi_{\text{prop}}(p_{i,t}))\right], \quad w_{\text{vis}}, w_{\text{prop}} \geq 0, \quad w_{\text{vis}} + w_{\text{prop}} = 1, \tag{5}$$

where $\mathcal{N}$ denotes ℓ_2 normalization, ϕ_{vis} is a frozen visual encoder, and ϕ_{prop} applies fixed normalization to proprioception. This fixed representation keeps node matching comparable across RL updates. The resulting nodes form an observation-based abstraction of interaction histories, over which continuation evidence is pooled.

Node matching and relation edges. For each task κ, PRICE maintains a directed graph shared across its initial-condition cases. Nodes cluster similar boundary descriptors, and edges represent action chunks. The prototype c_v of node v is the normalized sum of its assigned descriptors. In round q, PRICE compares $z_{i,t}$ with every node prototype frozen at the start of the round using cosine similarity $\langle z_{i,t}, c_v \rangle$. If the highest similarity reaches the matching threshold η, the boundary is assigned to the corresponding node. Unmatched descriptors are jointly clustered within each rollout group using the same threshold η, forming temporary nodes for that group's queries. Let $v_{i,t}$ denote a boundary's node assignment. Once two consecutive boundaries are assigned, $\mathbf{a}_{i,t}$ defines the directed relation $(v_{i,t}, \mathbf{a}_{i,t}, v_{i,t+1})$.

Cross-trajectory node potentials. PRICE estimates each node's success potential by pooling terminal outcomes across the histories represented by the node. Let $g(i)$ denote the rollout group containing query trajectory i, whose trajectories share the same initial context. Let $I_j(v) = \mathbb{I}[\exists\, t : v_{j,t} = v]$ indicate whether trajectory j visits node v. Each trajectory contributes once per node, using its first visit as the representative history. For query trajectory i of task $\kappa(i)$, the current counts

at node v are

$$n_{-i}^{\mathrm{cur}}(v) = \sum_{j \in g(i),\, j \neq i} I_j(v),$$
$$r_{-i}^{\mathrm{cur}}(v) = \sum_{j \in g(i),\, j \neq i} I_j(v) Y_j. \tag{6}$$

These count contributing trajectories and successful outcomes, respectively. When $n_{-i}^{\mathrm{cur}}(v) > 0$, the current estimate is $r_{-i}^{\mathrm{cur}}(v)/n_{-i}^{\mathrm{cur}}(v)$. Only trajectories matched to the same node share evidence, and the query trajectory is excluded. GRPO normalization remains within each rollout group.

Cross-round node potentials. Current-group evidence may be sparse or have little outcome variation. PRICE therefore augments node estimates with historical evidence. For each node, a summary s stores per-round counts (n_s, r_s) of distinct visiting trajectories and successful outcomes, aggregated across groups of the same task after each update. To limit mismatch from policy changes, it uses only summaries satisfying $D_{s,q} \leq d_{\max}$, where $D_{s,q}$ is the cumulative adjacent-policy KL from the summary's collection round u_s to the current round q, and $d_{\max}$ is the maximum allowed cumulative KL for reusing historical evidence.

PRICE combines current-group and eligible historical counts for the same node. For a boundary x matched to node v, the pooled success potential is

$$\widehat{V}_x = \frac{r_{-i}^{\mathrm{cur}}(v) + \sum_s r_s}{n_{-i}^{\mathrm{cur}}(v) + \sum_s n_s}, \tag{7}$$

where the sums run over eligible historical summaries for v. The denominator, denoted by N_x, counts all contributing trajectories, giving each equal weight. No estimate is formed when $N_x = 0$.

3.2 Confidence-Gated Action-Chunk Credit

Finite-sample noise can reverse the sign of an estimated potential change. PRICE uses a Hoeffding-inspired gate to compare this change with a threshold determined by endpoint support. For chunk t in trajectory i, let x and x^+ denote its source and destination boundaries. When both estimates are available, the candidate credit is $\widehat{\Delta}_{i,t} = \widehat{V}_{x^+} - \widehat{V}_x$.

An endpoint is supported when it is assigned to a node and $N_x > 0$. For a chosen gate parameter $\delta_{\mathrm{edge}} \in (0, 1)$, its Hoeffding radius and interval are

$$\epsilon_x = \sqrt{\frac{\log(4/\delta_{\mathrm{edge}})}{2N_x}}, \qquad L_x = \max(0, \widehat{V}_x - \epsilon_x), \qquad R_x = \min(1, \widehat{V}_x + \epsilon_x). \tag{8}$$

Appendix C provides the derivation and sampling assumptions.

Propagating the two endpoint intervals gives an interval for the potential change:

$$\mathcal{I}_{i,t} = [L_{x^+} - R_x,\ R_{x^+} - L_x]\,. \tag{9}$$

PRICE accepts the candidate credit only if this interval excludes zero:

$$C_{i,t} = \begin{cases} \widehat{\Delta}_{i,t}, & \text{if } t < T_i,\ x, x^+ \text{ are supported, and } 0 \notin \mathcal{I}_{i,t}, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Here T_i is the number of chunks in trajectory i; its final chunk retains the base GRPO advantage.

Let $\overline{V}_x$ denote the mean conditional continuation-success probability of the pooled records under their respective collection policies, and define $\overline{\Delta}_{i,t} = \overline{V}_{x^+} - \overline{V}_x$.

Gate interpretation. For edge $e = (i, t)$, write $r_e = \epsilon_x + \epsilon_{x^+}$ and $E_e = |\widehat{V}_x - \overline{V}_x| + |\widehat{V}_{x^+} - \overline{V}_{x^+}|$. An accepted credit satisfies $|\widehat{\Delta}_e| > r_e$. Since $|\widehat{\Delta}_e - \overline{\Delta}_e| \leq E_e$, the triangle inequality gives

$$C_e \neq 0, \quad E_e \leq r_e \quad \Longrightarrow \quad \mathrm{sign}(C_e) = \mathrm{sign}(\overline{\Delta}_e). \tag{11}$$

Thus, a wrong nonzero credit requires $E_e > r_e$, for any archive selection. Appendix D gives a finite-sample bound under outcome-independent evidence selection, and Appendix D.1 relates these errors to collection-policy mismatch and state abstraction.

3.3 Optimization Objective

PRICE augments the trajectory-level GRPO advantage with confidence-gated action-chunk credit:

$$\widehat{A}_{i,t}^{\mathrm{PRICE}} = \widehat{A}_{i}^{\mathrm{GRPO}} + \lambda C_{i,t}, \qquad \lambda \geq 0. \tag{12}$$

Here λ controls the strength of local credit. We optimize the clipped objective in Equation 2, assigning $\widehat{A}_{i,t}^{\mathrm{PRICE}}$ to every valid action token in chunk t of trajectory i. Implementation details are provided in Appendix E.1.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate PRICE on two simulated benchmarks and two real-robot tasks. For LIBERO (Liu et al., 2023), we use the Spatial, Object, Goal, and Long suites. From RoboTwin 2.0 (Chen et al., 2025a), we select six manipulation tasks on a dual-arm platform covering grasping, coordinated lifting, object transport, placement, and handover. For evaluation, we follow the RoboTwin protocol used by RLinf-VLA (Zang et al., 2025). Real-robot experiments use Remote Insertion and Produce Sorting from HiFi-UMI (AI et al., 2026), covering precise placement and category-conditioned sorting (Section 4.3).

Implementation details. For LIBERO, we use OpenVLA-OFT initialized by one-shot SFT. For RoboTwin 2.0, we use OpenVLA-OFT with task-specific SFT checkpoints through RLinf (Zang et al., 2025). We fine-tune all parameters on eight NVIDIA A800 GPUs with 80 GB memory each, using a learning rate of 5×10^{-6} and an optimization mini-batch size of 128. Each raw rollout batch contains 64 initial-condition cases with eight rollouts per case (512 trajectories), sampled at a temperature of 1.6. For our main PRICE experiments on LIBERO and RoboTwin 2.0, we report mean success rates over three independent training runs. The corresponding standard deviations are reported in Appendix G.1. PRICE-specific settings are provided in Appendix E.2. Real-robot experiments use the offline flow-matching adaptation described in Section 4.3.

Metrics. Following the official evaluation protocols, we report task success rate (%) and measure training efficiency by the steps needed to reach target success rates. We assess local credit using independent Monte Carlo (MC) estimates of the change in node-averaged continuation success across each chunk. Direction accuracy (DirAcc, %) measures sign agreement between credit and MC changes on non-tied comparisons. Spearman rank correlation (RankCorr) measures whether signed credits and MC changes rank chunks consistently. Aligned MC gap (AlignedGap, pp) measures progress in the credit's direction by averaging the MC change multiplied by the credit sign. Higher is better for all three. The MC protocol and metric definitions are given in Appendix E.3.

4.2 Main Results

Task performance. Tables 1 and 2 summarize task performance on LIBERO and RoboTwin 2.0. On LIBERO, PRICE achieves 99.0% average success, outperforming all compared evaluator-free methods and closely matching SRPO (99.2%). On LIBERO-Long, PRICE reaches 97.6%, exceeding the strongest evaluator-free baseline by 3.6 percentage points. Across the six RoboTwin 2.0 tasks reported in Table 2, PRICE achieves an average success rate of 90.0%, compared with 31.9% for the OpenVLA-OFT (SFT) initialization, an improvement of 58.1 percentage points. This average is numerically higher than Feat2Go (88.8%), RLinf-VLA (84.5%), and SimpleVLA-RL (71.3%) by 1.2, 5.5, and 18.7 percentage points, respectively.

Training efficiency. We compare PRICE with SimpleVLA-RL under the same settings. PRICE reaches both 90% and 95% success with fewer training steps and lower GPU-hour costs than SimpleVLA-RL (Figure 3(a–b)). At 90% success, PRICE reduces training steps by 34.5% on Goal and 37.1% on Spatial, with corresponding GPU-hour savings of 38.3% and 32.8%. Timing details are provided in Appendix G.2. PRICE also achieves average success comparable to evaluator-based SRPO under similar training budgets (Table 1; Figure 3(c)).

Table 1: Performance comparison on the LIBERO benchmark. We report task success rates (%). Rows below the dashed line are our results; other values follow the unified comparison of Fei et al. (2026). T, W, P, and I denote third-person images, wrist images, proprioception, and language instructions, respectively. Full and One denote full-shot and one-trajectory-per-task SFT, respectively.

Method	Policy input	Spatial	Object	Goal	Long	Avg.
VLA baseline						
OpenVLA (Kim et al., 2024)	T+I	84.7	88.4	79.2	53.7	76.5
OpenVLA*-Full (Fei et al., 2026)	T+I	91.6	95.3	90.6	86.5	91.0
Evaluator-based RL						
TGRPO (Chen et al., 2025b)	T+I	90.4	92.2	81.0	59.2	80.7
GRAPE (Zhang et al., 2024)	T+I	88.5	92.1	83.1	57.2	80.2
VLA-RL (Lu et al., 2025)	T+I	90.2	91.8	82.2	59.8	81.0
World-Env (Xiao et al., 2025)	T+I	87.6	86.6	86.4	57.8	79.6
SRPO (Fei et al., 2026)	T+I	98.8	100.0	99.4	98.6	99.2
Evaluator-free RL						
SimpleVLA-RL (Li et al., 2026a)	T+I	98.2	98.7	98.8	91.7	96.9
RIPT-VLA (Tan et al., 2025)	T+W+P+I	99.0	98.6	98.6	93.8	97.5
RLinf-VLA (Zang et al., 2025)	T+W+P+I	99.4	99.8	98.8	94.0	98.0
OpenVLA-OFT-One	T+I	63.6	54.9	59.6	17.3	48.9
+ **PRICE (Ours)**	T+I	99.2	99.6	99.4	97.6	99.0
		↑35.6	↑44.7	↑39.8	↑80.3	↑50.1

Table 2: Performance comparison on six selected RoboTwin 2.0 manipulation tasks. We report task success rates (%) and the average across the six tasks. The OpenVLA-OFT (SFT) row is the initialization used for PRICE; results for the other compared methods are from the cited works.

Method	Handover Block	Lift Pot	Move Can Pot	Pick Dual Bottles	Place Container Plate	Place Empty Cup	Avg.
SimpleVLA-RL (Li et al., 2026a)	57.8	64.1	61.2	68.3	82.1	94.2	71.3
RLinf-VLA (Zang et al., 2025)	70.31	70.31	83.59	92.96	95.31	94.53	84.5
Feat2Go (Shu et al., 2026)	76.6	78.9	**91.4**	93.8	96.9	95.3	88.8
OpenVLA-OFT (SFT)	28.1	3.1	9.4	20.3	54.7	75.8	31.9
+ **PRICE (Ours)**	**79.7**	**80.5**	90.6	**94.5**	**97.7**	**96.9**	**90.0**
	↑51.6	↑77.4	↑81.2	↑74.2	↑43.0	↑21.1	↑58.1

4.3 Real-World Experiments

Experimental setup. We evaluate on two HiFi-UMI tasks (AI et al., 2026): Remote Insertion places a remote control into a storage box, while Produce Sorting requires placing both a fruit and a vegetable into their category-specific trays (Figure 4(b)). All variants start from the same OpenPI-$\pi_{0.5}$ checkpoint post-trained on 400 HiFi-UMI trajectories. The initialization collects 60 rollouts per task for offline adaptation, with no additional online collection during optimization. We compare the initialization with two offline adaptation methods based on advantage-weighted regression (AWR) (Peng et al., 2019). GRPO-AWR weights the flow-matching loss using trajectory-level advantages, while PRICE-AWR adds action-chunk relational credit before computing these weights. Both adaptation methods share the same data, random seed, and optimization budget. We follow the HiFi-UMI evaluation protocol, reporting task success and the macro-average across tasks. Appendix F details the adaptation objectives, observation configuration, and training hyperparameters.

Results. PRICE-AWR achieves 55.0% success on Remote Insertion and 67.5% on Produce Sorting, exceeding GRPO-AWR by 12.5 and 15.0 percentage points, respectively (Figure 4(a)). Macro-average success reaches 61.25%, compared with 47.5% for GRPO-AWR and 41.25% for the initialization. These gains on both tasks support the benefit of action-chunk relational credit under the same offline adaptation budget.

4.4 Analysis of PRICE

Alignment with local progress. We independently assess local progress on LIBERO-Spatial by continuing the frozen OpenVLA-OFT policy from saved states in each chunk's source and destination evidence pools. The difference in node-averaged Monte Carlo success rates provides a reference

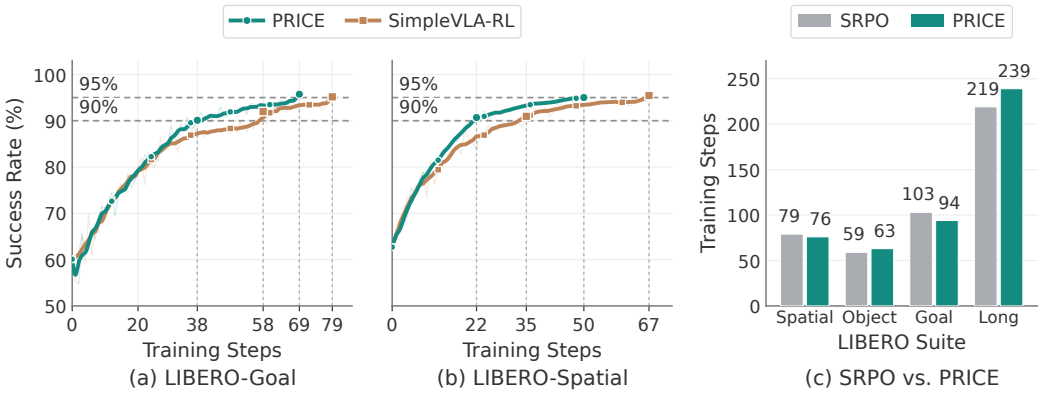

Figure 3: Training efficiency on LIBERO. (a–b) Rollout success during training on Goal and Spatial. Solid curves show seven-step moving averages and faint lines show raw observations; dashed guides and x-axis ticks mark the steps for 90% and 95% success. (c) Training steps of SRPO and PRICE across four suites. SRPO counts are from Fei et al. (2026); corresponding success rates are reported in Table 1.

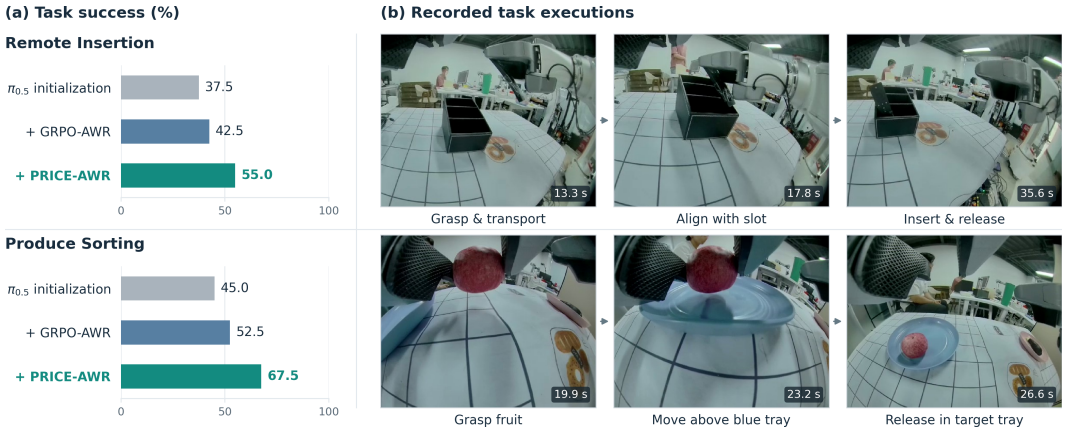

Figure 4: Real-robot performance and representative executions. (a) Success rates of the initialization, GRPO-AWR, and PRICE-AWR on Remote Insertion and Produce Sorting under the HiFi-UMI evaluation protocol. (b) Chronological keyframes show remote transport, alignment, and insertion into the storage box (top), and fruit grasping, transport, and release into its designated tray during Produce Sorting (bottom).

for each chunk's progress (Appendix E.3). Figure 5(a–c) shows that PRICE distinguishes chunks even when their trajectories share the same terminal outcome. Retained credits agree in sign with 97.5% of non-tied reference differences, with a RankCorr of 0.709 and an AlignedGap of +28.47 pp (Figure 5(f)).

Roles of historical evidence and confidence gating. Historical evidence enables local contrasts even when all trajectories in a rollout group share the same outcome. Pooling outcomes across rounds yields distinct endpoint potentials in these groups (Figure 5(d,e)). Confidence gating then selects a subset with stronger agreement with independent MC progress, increasing DirAcc from 77.2% before gating to 97.5% among retained credits (Figure 5(b,e,f)).

Component ablations. Removing historical evidence lowers average success from 98.3% to 94.0%, close to GRPO's 93.9% (Table 3). Under our settings, current-group evidence alone is

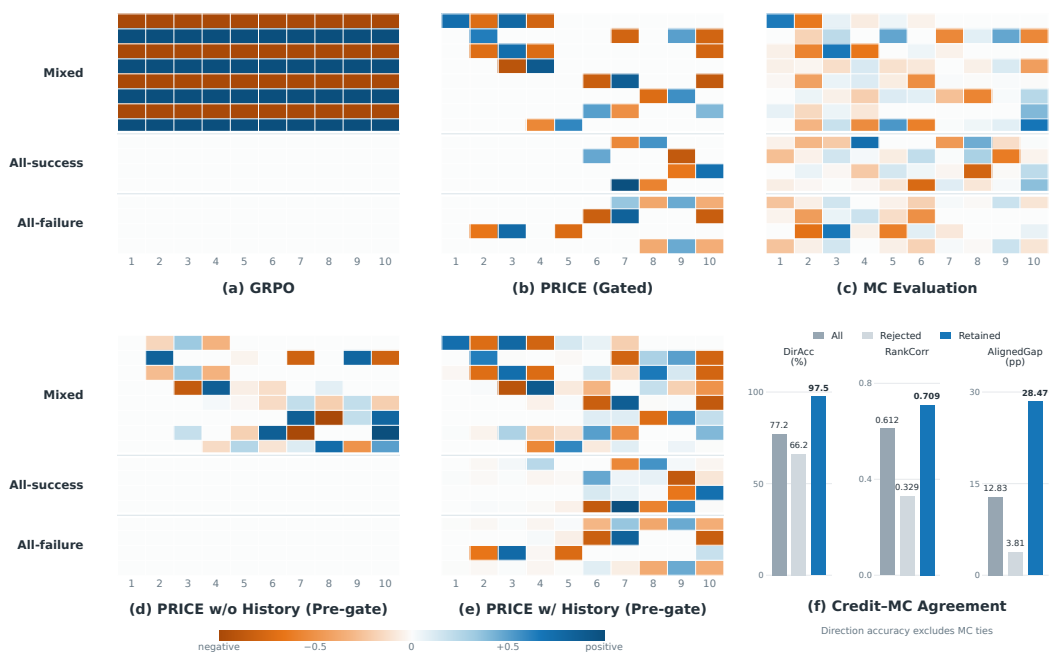

Figure 5: Credit analysis on LIBERO-Spatial. Rows denote trajectory windows and columns action chunks. (a) GRPO advantage sign. (b) Gated PRICE credit from (e). (c) Independent MC progress over the pooled evidence in (e). (d–e) Pre-gate contrasts without/with history at fixed nodes. (f) MC-based credit evaluation for all nonzero candidates in (e) and their rejected/retained subsets.

Table 3: Component ablations on LIBERO-Spatial and LIBERO-Object. All variants use 63 training steps with the same initialization, interaction budget, and optimization settings within each suite. We report task success (%) and the average across the two suites.

Variant	History	Gate	Spatial	Object	Avg.
GRPO	×	×	93.2	94.6	93.9
PRICE w/o Historical Evidence	×	✓	93.4	94.6	94.0
PRICE w/o Gate	✓	×	95.8	97.9	96.9
Full PRICE	✓	✓	**97.0**	**99.6**	**98.3**

insufficient for any credit to pass the gate, so the update reduces to GRPO (Appendix C). Removing the gate also lowers average success to 96.9%. Although it provides denser local feedback, it admits uncertain credit directions into policy updates: more credit does not necessarily provide better supervision.

5 Conclusion

We introduced PRICE, which derives action-chunk credit from terminal outcomes through physical relations across trajectories. By pooling current and historical evidence and applying confidence gating, PRICE refines trajectory-level supervision without an auxiliary evaluator. Our analysis connects oracle potential changes to the terminal-success objective and provides a finite-sample directional bound for outcome-independent evidence pools. Experiments on LIBERO, RoboTwin 2.0, and real robots demonstrate gains over outcome-based baselines, while controlled comparisons in simulation show improved training efficiency. Independent continuation evaluation further shows that retained credits align with local progress. Together, these results show that physical correspondence across collected experience can turn sparse terminal feedback into useful local supervision for embodied policy learning.

AI USE STATEMENT

We used generative AI tools to assist with drafting, translation, and language polishing, and to support literature retrieval and discovery. AI assistance also included feedback on theoretical statements, proofs, experimental design, and result interpretation, as well as illustrative numerical examples for planning. The authors are responsible for verifying AI-assisted material and for the final claims, references, and reported results.

ETHICS STATEMENT

This work studies robotic manipulation in simulation and on real hardware. Learned policies may behave unpredictably outside the evaluated conditions. Deployment should therefore include task-specific safety assessment, appropriate safeguards, and human oversight. Use of the underlying datasets and pretrained models should respect their licenses and intended scope.

REPRODUCIBILITY STATEMENT

Section 3 specifies the graph construction, credit estimation, and optimization objective. Appendices B–D provide the theoretical assumptions and proofs. Appendix E details the training procedure, hyperparameters, and credit-evaluation protocol; Appendices F and H describe the real-robot and world-action-model experiments.

REFERENCES

David Abel, David Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In *International Conference on Machine Learning*, pp. 2915–2923. PMLR, 2016.

Simple AI, Yuteng Wei, Jinming Ma, Jiawei Wang, Weitao Zhou, Yushen Zuo, Ke Rui, Minglei Li, Jinhao Zhang, Zhikang Pan, et al. Hifi-umi: Learning deployable manipulation policies from high-fidelity umi data alone. *arXiv preprint arXiv:2607.25895*, 2026.

Jose A Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards. *Advances in Neural Information Processing Systems*, 32, 2019.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025a.

Zengjue Chen, Runliang Niu, He Kong, Qi Wang, Qianli Xing, and Zipei Fan. Tgrpo :fine-tuning vision-language-action model via trajectory-wise group relative policy optimization, 2025b. URL https://arxiv.org/abs/2506.08440.

Xin Cheng, Shuo He, Lang Feng, HaiYang Xu, Ming Yan, Lei Feng, and Bo An. Beyond trajectory-level attribution: Graph-based credit assignment for agentic reinforcement learning. *arXiv preprint arXiv:2605.26684*, 2026.

Senyu Fei, Siyin Wang, Li Ji, Ao Li, Shiduo Zhang, Liming Liu, Jinlong Hou, Jingjing Gong, Xianzhong Zhao, and Xipeng Qiu. Srpo: Self-referential policy optimization for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6718–6728, 2026.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In *International conference on machine learning*, pp. 2170–2179. PMLR, 2019.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Zhennan Jiang, Shangqing Zhou, Yutong Jiang, Zefang Huang, Mingjie Wei, Yuhui Chen, Tianxing Zhou, Zhen Guo, Hao Lin, Quanlu Zhang, et al. Wovr: World models as reliable simulators for post-training vla policies with rl. *arXiv preprint arXiv:2602.13977*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Yang Zhaohui, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, et al. Simplevla-rl: Scaling vla training via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 125846–125869, 2026a.

Jiazheng Li, Yawei Wang, Qiaojing Yan, Yijun Tian, Zhichao Xu, Huan Song, Panpan Xu, and Lin Lee Cheong. Salt: Step-level advantage assignment for long-horizon agents via trajectory graph. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 4709–4725, 2026b.

Qiyang Li, Zhiyuan Paul Zhou, and Sergey Levine. Reinforcement learning with action chunking. *Advances in Neural Information Processing Systems*, 38:55518–55553, 2026c.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Guanxing Lu, Wenkai Guo, Chubin Zhang, Yuheng Zhou, Haonan Jiang, Zifeng Gao, Yansong Tang, and Ziwei Wang. Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning. *arXiv preprint arXiv:2505.18719*, 2025.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287. Citeseer, 1999.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Zezhong Qian, Xiaowei Chi, Yu Qi, Haozhan Li, Zhi Yang Chen, and Shanghang Zhang. Wam-rl: World-action model reinforcement learning with reconstruction rewards and online video sft. *arXiv preprint arXiv:2606.17906*, 2026.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Junyang Shu, Zhiwei Lin, and Yongtao Wang. Rftf: Reinforcement fine-tuning for embodied agents with temporal feedback. *arXiv preprint arXiv:2505.19767*, 2025.

Junyang Shu, Zhiwei Lin, Bingqing Wei, and Yongtao Wang. Feat2go: Visual feature-grounded value estimation for embodied reinforcement learning, 2026. URL https://arxiv.org/abs/2605.30795.

Chengye Song, Jiawei Zhang, Rui Song, Shengqi Wang, Xiangrong Zhang, Ziyi Wang, Huanbin Zhou, and Hongzhou Wang. Pace: Phase-progress-aware credit for long-horizon embodied manipulation. *arXiv preprint arXiv:2608.15026*, 2026.

Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive post-training for vision-language-action models. *arXiv preprint arXiv:2505.17016*, 2025.

Hanyang Wang, Weijieying Ren, Yuxiang Zhang, Ding Cao, Zhizhao Zeng, Ke Zeng, and Tianxiang Zhao. Bipace: Bisimulation-guided policy optimization with action counterfactual estimation for llm agents. *arXiv preprint arXiv:2606.25556*, 2026a.

Yunan Wang, Minghui Song, Zihan Zhang, Shaohan Huang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Group-graph policy optimization for long-horizon agentic reinforcement learning. *arXiv preprint arXiv:2606.22995*, 2026b.

Junjin Xiao, Yandan Yang, Xinyuan Chang, Ronghan Chen, Feng Xiong, Mu Xu, Wei-Shi Zheng, and Qing Zhang. World-env: Leveraging world model as a virtual environment for vla post-training. *arXiv preprint arXiv:2509.24948*, 2025.

Hongxi Yan, Ziyue Huang, Shichao Fan, and Qingjie Liu. Gated-bepo: Confidence-gated bellman credit assignment for large language model agents. *arXiv preprint arXiv:2608.06861*, 2026.

Jiazhi Yang, Kunyang Lin, Jinwei Li, Wencong Zhang, Tianwei Lin, Longyan Wu, Zhizhong Su, Hao Zhao, Ya-Qin Zhang, Li Chen, et al. Rise: Self-improving robot policy with compositional world model. *arXiv preprint arXiv:2602.11075*, 2026.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

Hongzhi Zang, Mingjie Wei, Si Xu, Yongji Wu, Zhen Guo, Yuanqing Wang, Hao Lin, Peihong Wang, Liangzhi Shi, Yuqing Xie, et al. Rlinf-vla: A unified and efficient framework for reinforcement learning of vision-language-action models. *arXiv preprint arXiv:2510.06710*, 2025.

Haoyu Zhang, Zecui Zeng, Bin Wang, Lusong Li, Liang Lin, and Long Cheng. Dream2reward: Transition-alignment reward models from positive demonstrations for robotic manipulation. *arXiv preprint arXiv:2608.18787*, 2026.

Zijian Zhang, Kaiyuan Zheng, Zhaorun Chen, Joel Jang, Yi Li, Siwei Han, Chaoqi Wang, Mingyu Ding, Dieter Fox, and Huaxiu Yao. Grape: Generalizing robot policy via preference alignment. *arXiv preprint arXiv:2411.19309*, 2024.

Yao Zhou, Hang Gao, Fengge Wu, Changwen Zheng, and Wenwen Qiang. Temporal grpo: Beyond trajectory-level credit in vision-language-action reinforcement learning. *arXiv preprint arXiv:2608.13026*, 2026.

Fangqi Zhu, Zhengyang Yan, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. Wmpo: World model-based policy optimization for vision-language-action models. In *International Conference on Learning Representations*, volume 2026, pp. 62486–62502, 2026a.

Siyuan Zhu, Chao Yu, Rongxin Yang, Zongkai Liu, Jinjun Hu, Qiwen Chen, and Yibo Zhang. Gagpo: Generalized advantage grouped policy optimization. *arXiv preprint arXiv:2605.13217*, 2026b.

Appendix

A Related Work

A.1 RL Post-Training for Embodied Policies

Vision–language–action policies acquire broad manipulation capabilities from imitation data, and reinforcement learning can further correct imitation errors through interaction (Kim et al., 2024). Recent work extends post-training across policy architectures and optimization frameworks: RIPT-VLA constructs leave-one-out advantages directly from binary success, SimpleVLA-RL adopts outcome-based GRPO, RLinf-VLA provides unified infrastructure for embodied RL, and WAM-RL extends online adaptation to world–action models (Tan et al., 2025; Li et al., 2026a; Zang et al., 2025; Qian et al., 2026). Together, these studies establish the value of interactive post-training. Outside VLA post-training, Q-Chunking treats an action chunk as the decision unit and uses multi-step temporal-difference backups for sparse-reward manipulation (Li et al., 2026c). Outcome-based variants, however, typically assign the same episode result to every decision in a trajectory, revealing which rollout failed but not where its probability of success changed.

A.2 Fine-Grained Feedback in Embodied Reinforcement Learning

To obtain finer feedback, one family introduces task-dependent process supervision: TGRPO constructs dense rewards from intermediate task conditions, VLA-RL learns a process reward model from automatically extracted manipulation stages, and RFTF learns value changes by temporally ordering successful demonstrations (Chen et al., 2025b; Lu et al., 2025; Shu et al., 2025). Feat2Go derives visual progress targets from a pretrained world model and trains a value model to reshape terminal rewards; PACE uses a phase-aware critic to assign step-level credit; and Dream2Reward learns a successful-transition model from positive demonstrations to score observed motion (Shu et al., 2026; Song et al., 2026; Zhang et al., 2026). Other world-model approaches generate intermediate supervision through imagined future interaction: World-Env combines an action-conditioned visual simulator with a vision–language reflector for continuous reward and termination prediction; WMPO couples pixel-space imagined rollouts with a learned reward model for on-policy optimization; RISE decomposes its world model into controllable dynamics and a progress-value model, scoring imagined actions by predicted value improvement; and WoVR uses keyframe-started rollouts and policy–simulator co-evolution to limit hallucination and accumulated model error (Xiao et al., 2025; Zhu et al., 2026a; Yang et al., 2026; Jiang et al., 2026). SRPO uses the policy's own successful rollouts as references and scores failures by their distance to the success set, avoiding expert reference trajectories (Fei et al., 2026). Temporal GRPO instead compares rollouts within detectable task stages and assigns stage-relative advantages to their action intervals; its stage construction and alignment use task-stage predicates during training (Zhou et al., 2026). These approaches rely on task-specific progress conditions or learned evaluators, while SRPO assigns progress at the rollout level. This leaves open whether final task outcomes from grouped rollouts can provide action-chunk credit without task-stage predicates or an auxiliary evaluator.

A.3 Credit Assignment from Cross-Trajectory Relations

Credit assignment from delayed outcomes has broader precedents. RUDDER redistributes delayed returns through a learned return decomposition (Arjona-Medina et al., 2019). One-step value changes also underlie generalized advantage estimation and potential-based reward shaping (Schulman et al., 2015; Ng et al., 1999). Recent agentic RL methods propagate outcome signals through repeated states, shared edges, or non-parametric node values, and by using latent-space clustering to estimate local $Q-V$ signals from neighboring rollouts (Li et al., 2026b; Feng et al., 2026; Zhu et al., 2026b; Cheng et al., 2026; Wang et al., 2026b; Yan et al., 2026; Wang et al., 2026a). These studies show that cross-trajectory relations can provide local credit without predefined progress labels. PRICE estimates value changes from pooled terminal outcomes, and the policy-invariance result for potential-based reward shaping does not directly apply to its estimated, gated score weights. Relational credit, however, presupposes knowing when different trajectories have reached the "same state." In text or reasoning tasks, this can often be approximated using discrete states, repeated anchors, or semantic matching. In embodied control, two semantically similar frames can differ in

gripper–object contact or relative pose, while the same physical situation can appear different under viewpoint or appearance changes. Incorrectly merging or splitting such states directly distorts node values and their differences (Abel et al., 2016; Gelada et al., 2019). The central problem for embodied relational credit is therefore to establish reliable cross-trajectory physical correspondence from continuous visual and proprioceptive signals, and thereby determine whether an action chunk makes genuine progress. PRICE constructs physical relations at chunk boundaries, estimates local progress by comparing the terminal-success potentials of preceding and succeeding endpoints across trajectories, and selects credits using endpoint differences and evidence counts with a Hoeffding-style uncertainty threshold (Hoeffding, 1963).

B Endpoint Value Changes and the Terminal Objective

Under the assumptions of Theorem 1, H_t contains the task context, boundary observations, proprioception, time, and previous chunks. It excludes other trajectories' outcomes and graph estimates. The actor input h_t is a fixed function of H_t. The terminal history includes Y, so $V^{\pi}_{T+1}(H_{T+1}) = Y$.

Proof of Theorem 1. By the tower property,

$$\begin{aligned}\mathbb{E}_\pi[V^{\pi}_{t+1}(H_{t+1}) \mid H_t, \mathbf{a}_t] &= \mathbb{E}_\pi[\mathbb{E}_\pi[Y \mid H_{t+1}] \mid H_t, \mathbf{a}_t] \\ &= \mathbb{E}_\pi[Y \mid H_t, \mathbf{a}_t]. \end{aligned} \tag{B.1}$$

Subtracting $V^{\pi}_t(H_t)$ proves Equation 3.

For the policy-gradient identity, fix θ_0, set $\pi = \pi_{\theta_0}$, and define $S_t(\theta_0) = \nabla_\theta \log \pi_\theta(\mathbf{a}_t \mid h_t)|_{\theta=\theta_0}$. Because the policy samples from $\pi_{\theta_0}(\cdot \mid h_t)$, $\mathbb{E}_{\pi_{\theta_0}}[S_t(\theta_0) \mid H_t] = 0$, so $V_t^{\pi_{\theta_0}}(H_t)$ is a valid baseline. The likelihood-ratio identity and iterated expectation give

$$\begin{aligned}\nabla_\theta J(\theta)|_{\theta=\theta_0} &= \mathbb{E}_{\pi_{\theta_0}}\left[\sum_{t=1}^{T} S_t(\theta_0) A_t^{\pi_{\theta_0}}(H_t, \mathbf{a}_t)\right] \\ &= \mathbb{E}_{\pi_{\theta_0}}\left[\sum_{t=1}^{T} S_t(\theta_0) \Delta_t^{\mathrm{ora}}\right]. \end{aligned} \tag{B.2}$$

Stop-gradient implements this score-function estimator by differentiating only the action probability.

Additive oracle credit. Let the expectation also include the on-policy rollout group used to compute $\widehat{A}^{\mathrm{GRPO}}$ for the query trajectory. Define $g_{\mathrm{base}}(\theta_0) = \mathbb{E}_{\pi_{\theta_0}}[\sum_t S_t(\theta_0)\widehat{A}^{\mathrm{GRPO}}]$. Linearity and Equation B.2 give

$$\begin{aligned} g^{\mathrm{ora}}_{\mathrm{aug}}(\theta_0) &= g_{\mathrm{base}}(\theta_0) + \lambda \mathbb{E}_{\pi_{\theta_0}}\left[\sum_t S_t(\theta_0)\Delta_t^{\mathrm{ora}}\right] \\ &= g_{\mathrm{base}}(\theta_0) + \lambda\, \nabla_\theta J(\theta)|_{\theta=\theta_0}\,, \end{aligned} \tag{B.3}$$

which proves Equation 4.

Trajectory-wise identity. Consecutive endpoint values also cancel pathwise:

$$\begin{aligned}\sum_{t=1}^{T} \Delta_t^{\mathrm{ora}} &= \sum_{t=1}^{T} \left(V_{t+1}^{\pi_{\theta_0}}(H_{t+1}) - V_t^{\pi_{\theta_0}}(H_t)\right) \\ &= Y - V_1^{\pi_{\theta_0}}(H_1). \end{aligned} \tag{B.4}$$

C Pooled Endpoint Concentration

Fix a supported endpoint x and $N_x > 0$ distinct trajectory records, with sample size and membership fixed independently of their terminal labels. Reveal the records in collection order. Let $Y_{x,k} \in \{0, 1\}$

be the outcome of record k, and let $\mathcal{F}_{x,k}$ be an increasing pre-outcome filtration containing its first-visit history and all earlier revealed outcomes. Assume $Y_{x,k} \mid \mathcal{F}_{x,k} \sim \mathrm{Bernoulli}(\mu_{x,k})$, where

$$\mu_{x,k} = \mathbb{E}[Y_{x,k} \mid \mathcal{F}_{x,k}], \qquad \overline{V}_x = \frac{1}{N_x}\sum_{k=1}^{N_x} \mu_{x,k}, \qquad \widehat{V}_x = \frac{1}{N_x}\sum_{k=1}^{N_x} Y_{x,k}. \tag{C.1}$$

For any $\nu \in \mathbb{R}$, conditional Hoeffding's lemma gives

$$\mathbb{E}[\exp(\nu(Y_{x,k} - \mu_{x,k}))|\mathcal{F}_{x,k}] \le \exp\left(\frac{\nu^2}{8}\right). \tag{C.2}$$

Iterating this bound along the filtration and applying the Chernoff method to both tails yields

$$\Pr\left(\left|\widehat{V}_x - \overline{V}_x\right| \ge \epsilon\right) \le 2\exp(-2N_x\epsilon^2). \tag{C.3}$$

Substituting $\epsilon = \sqrt{\log(4/\delta_{\mathrm{edge}})/(2N_x)}$ makes the right-hand side $\delta_{\mathrm{edge}}/2$. Clipping the interval to $[0,1]$ preserves coverage because both $\widehat{V}_x$ and $\overline{V}_x$ lie in $[0,1]$. Therefore,

$$\Pr\big(\overline{V}_x \notin [L_x, R_x]\big) \le \frac{\delta_{\mathrm{edge}}}{2}. \tag{C.4}$$

For supported endpoints, the interval excludes zero exactly when $|\widehat{\Delta}_{i,t}| > \epsilon_x + \epsilon_{x^+}$; clipping endpoint intervals to $[0,1]$ preserves this criterion. With equal support $N_x = N_{x^+} = N$ and observed contrast magnitude $d > 0$, this requires

$$N > \frac{2\log(4/\delta_{\mathrm{edge}})}{d^2}. \tag{C.5}$$

Under our parameter settings, $\delta_{\mathrm{edge}} = 0.15$, so contrast magnitudes 1.0, 0.4, and 0.3 require at least 7, 42, and 73 supporting trajectories per endpoint, respectively. This motivates accumulating same-task evidence across rounds. If evidence is restricted to a single group of $K = 8$ trajectories, excluding the query leaves at most seven peers. An endpoint with at most six peers has radius at least 0.5231, while seven peers give radius 0.4843. Thus, a gate could pass only if both endpoints have seven peers; these endpoints then share the same peer set and have identical means. Such within-group evidence alone yields no nonzero credit under the default gate. Without historical outcome summaries, the augmented advantage therefore reduces to the GRPO advantage.

D Pooled Finite-Sample Edge Direction

Corollary 2 (Direction bound for fixed evidence pools)**.** *Fix a supported non-terminal edge. Suppose each endpoint pool has sample size and membership fixed independently of its terminal labels and satisfies the conditional Bernoulli assumptions in Appendix C. Then $\overline{\Delta}_{i,t} \in \mathcal{I}_{i,t}$ with probability at least $1 - \delta_{\mathrm{edge}}$, and*

$$\Pr\big(C_{i,t} \ne 0,\ \mathrm{sign}(C_{i,t}) \ne \mathrm{sign}(\overline{\Delta}_{i,t})\big) \le \delta_{\mathrm{edge}}. \tag{D.1}$$

Proof of Corollary 2. Apply Equation C.4 to the source endpoint x and destination endpoint x^+. A union bound gives

$$\Pr\big(\overline{V}_x \in [L_x, R_x] \text{ and } \overline{V}_{x^+} \in [L_{x^+}, R_{x^+}]\big) \ge 1 - \delta_{\mathrm{edge}}. \tag{D.2}$$

No independence between the two endpoint estimates is needed. On this joint event,

$$\overline{\Delta}_{i,t} = \overline{V}_{x^+} - \overline{V}_x \in [L_{x^+} - R_x,\ R_{x^+} - L_x] = \mathcal{I}_{i,t}. \tag{D.3}$$

The empirical means lie in their own intervals by construction, so $\widehat{\Delta}_{i,t} \in \mathcal{I}_{i,t}$ as well. If $0 \notin \mathcal{I}_{i,t}$, the interval lies strictly on one side of zero and both contrasts have that sign. Since an accepted credit equals $\widehat{\Delta}_{i,t}$, an incorrect nonzero credit implies failure of the joint endpoint coverage event. Its probability is at most δ_{edge}, proving Equation D.1 and Corollary 2.

This is a marginal bound for a fixed edge under the stated selection assumptions, not a bound conditional on emission or a simultaneous training-wide guarantee. Historical KL screening and archive retention can depend on terminal outcomes through policy updates, so the corollary does not directly cover the adaptive training archive. Retained-credit direction accuracy is evaluated under frozen-policy replay in Section 4.4. The endpoint union bound accommodates shared trajectories without requiring independent estimates. Paired or variance-adaptive concentration could exploit this structure to obtain tighter intervals.

D.1 From Pooled Evidence to Current-Policy Progress

Fix collection round q with policy π_q. For each endpoint x, let $(H_{x,k}, t_{x,k})$ denote the first-visit history and boundary index of pooled record k. Keeping these histories and their weights fixed, define the current-policy reference potential and contrast as

$$\begin{aligned} V_x^{q,\mathrm{pool}} &= \frac{1}{N_x} \sum_{k=1}^{N_x} V_{t_{x,k}}^{\pi_q}(H_{x,k}), \\ \Delta_e^{q,\mathrm{pool}} &= V_{x^+}^{q,\mathrm{pool}} - V_x^{q,\mathrm{pool}}. \end{aligned} \tag{D.4}$$

For query edge $e = (i, t)$, the oracle contrast is $\Delta_e^{\mathrm{ora},q} = V_{t+1}^{\pi_q}(H_{i,t+1}) - V_t^{\pi_q}(H_{i,t})$. The estimation error decomposes as

$$\begin{aligned} \widehat{\Delta}_e - \Delta_e^{\mathrm{ora},q} = & \underbrace{\widehat{\Delta}_e - \overline{\Delta}_e}_{\text{sampling error}} \\ & + \underbrace{\overline{\Delta}_e - \Delta_e^{q,\mathrm{pool}}}_{\text{collection-policy mismatch}} \\ & + \underbrace{\Delta_e^{q,\mathrm{pool}} - \Delta_e^{\mathrm{ora},q}}_{\text{state-abstraction error}}. \end{aligned} \tag{D.5}$$

Corollary 2 controls the sampling term under its evidence assumptions. The second term compares continuation policies at the same recorded histories. The third measures the discrepancy between the pooled histories and the query, including differences in physical state, interaction history, and remaining horizon.

Direction transfer. Let $B_e = |\overline{\Delta}_e - \Delta_e^{q,\mathrm{pool}}| + |\Delta_e^{q,\mathrm{pool}} - \Delta_e^{\mathrm{ora},q}|$. The triangle inequality gives $|\widehat{\Delta}_e - \Delta_e^{\mathrm{ora},q}| \le E_e + B_e$. For an accepted edge,

$$E_e + B_e < |\widehat{\Delta}_e| \quad \Longrightarrow \quad \mathrm{sign}(C_e) = \mathrm{sign}(\Delta_e^{\mathrm{ora},q}). \tag{D.6}$$

In particular, $E_e \le r_e$ and $B_e < |\widehat{\Delta}_e| - r_e$ suffice. The historical KL screen restricts the contributing collection rounds, and physical matching determines which histories share evidence. Together with the sampling interval, these choices govern the reliability of local credit.

E Implementation and Evaluation Details

E.1 PRICE Training Algorithm

Visual descriptors use an independent frozen copy of the initial SFT encoder; proprioception uses fixed checkpoint normalization statistics. Each round computes credit from a frozen archive, updates the policy, and then commits evidence from all raw rollout groups for the next round (Algorithm 1).

Algorithm 1 PRICE training and history updates.

Input: Initial policy π_0, fixed physical descriptor, and parameters η, δ_{edge}, λ, $d_{\max}$, C_{hist}, C_{node}

Initialize task graphs $\mathcal{G}_0$, summaries $\mathcal{S}_0$, and cumulative policy KL $\xi_0 = 0$
for each training round q **do**
  Freeze $\mathcal{G}_q$ and $\mathcal{S}_q$ as the query snapshot
  Collect grouped rollouts and terminal outcomes under fixed π_q
  for each rollout group g **do**
    Match frozen prototypes; cluster unmatched boundaries within g
    Normalize outcomes within g to obtain $\widehat{A}_i^{\text{GRPO}}$
    for each query trajectory $i \in g$ **do**
      Pool current-group leave-one-out counts with history satisfying $D_{s,q} \le d_{\max}$
      Compute gated credit $C_{i,t}$ using Equations 7–10
      $\widehat{A}_{i,t}^{\text{PRICE}} \leftarrow \widehat{A}_i^{\text{GRPO}} + \lambda C_{i,t}$
    end for
  end for
  Update π_q to π_{q+1} using $\widehat{A}^{\text{PRICE}}$ in Equation 2
  Update prototypes; cluster unmatched descriptors across groups of the same task
  Append one summary (q, v, n_v, r_v, ξ_q) per visited node from all raw rollouts
  Apply the summary and node capacity limits C_{hist} and C_{node}
  $\xi_{q+1} \leftarrow \xi_q + \max\{0, \widehat{\text{KL}}(\pi_q \| \pi_{q+1})\}$
  Publish $\mathcal{G}_{q+1}$ and $\mathcal{S}_{q+1}$ for queries in the next round
end for

Node matching does not explicitly use terminal labels. Each trajectory contributes once per node, using its first visit; temporary nodes use current-group evidence only. The final chunk receives zero local credit, and advantages are not renormalized after credit addition. Matching follows fixed trajectory–time order within each group; archive updates follow task–group–trajectory–time order.

Each summary stores its collection-round cumulative KL ξ_{u_s}, giving $D_{s,q} = \xi_q - \xi_{u_s}$. Appendix E.2 specifies the parameters and KL estimator. Nodes retain their most recent C_{hist} summaries; task graphs retain at most C_{node} nodes, evicting the least recently matched. Archive changes are committed only after a successful policy update and saved with actor checkpoints; failed updates discard staged records.

E.2 Training Settings

Table 4 summarizes simulation training settings, PRICE credit parameters, and real-robot AWR weights. General training settings are given in Section 4.1, and real-robot adaptation details are provided in Appendix F.

For historical-evidence eligibility, adjacent-policy KL is estimated from the old-minus-post-update log probabilities on valid action tokens in the update batch, averaged over microbatches and workers. Negative estimates are clipped to zero before accumulation. The separate reference-policy KL penalty is disabled ($\beta = 0$). Real-robot AWR weights follow Equation F.3.

For ablations, PRICE w/o Historical Evidence sets $C_{\text{hist}} = 0$ while retaining graph construction and gating; PRICE w/o Gate uses raw potential differences for all supported non-terminal chunks; GRPO sets $\lambda = 0$.

E.3 Credit Evaluation Protocol

Main credit audit. Figure 5 uses a frozen OpenVLA-OFT policy on LIBERO-Spatial. The replay corpus contains five rollout rounds, 320 groups, and 20,480 valid non-terminal action chunks. The audit uses $\eta = 0.93$, $\delta_{\text{edge}} = 0.15$, and up to four historical summaries per node, covering 160 chunks in 16 windows selected before observing MC outcomes. Table 8 and Figure 6 reuse this audit.

Controlled history and gate comparisons. The heatmaps share locked windows and endpoint node assignments. Panel (d) uses terminal outcomes from current-group peers, excluding the query trajectory and retaining each peer's first node visit. It computes $\widehat{V}_x^{\text{cur}} = r_{-i}^{\text{cur}}(v)/n_{-i}^{\text{cur}}(v)$ and

Table 4: Core training hyperparameters. Optimization batch sizes count trajectories and are global across eight GPUs. Simulation and real-robot credit weights are reported separately.

Parameter	Setting	Parameter	Setting
Simulation training and GRPO		*PRICE in simulation*	
Optimizer	AdamW	Credit weight λ	0.2
Learning rate	5×10^{-6}	Matching threshold η	0.93
Learning-rate schedule	Constant	Gate parameter δ_{edge}	0.15
Cases per raw rollout batch	64	Summaries per node C_{hist}	4
Rollouts per case	8	Nodes per task C_{node}	1,024
Trajectories per raw batch	512	Visual weight w_{vis}	0.5
Optimization mini-batch size	128	Proprioceptive weight w_{prop}	0.5
Optimization micro-batch size	8	Cumulative KL limit $d_{\max}$	0.2
GRPO epochs per update batch	1	*Real-robot AWR adaptation*	
Rollout temperature	1.6	Credit weight λ_{AWR}	0.2
Gradient clipping norm	1.0	Temperature τ_{AWR}	1.0
Actions per chunk	8	Minimum weight $w_{\min}$	0.1
Lower GRPO clip ϵ_{low}	0.20	Maximum weight $w_{\max}$	5.0
Upper GRPO clip ϵ_{high}	0.28	Time-sampling shape α_{time}	1.5
Reference KL coefficient β	0	Time-sampling shape β_{time}	1.0
		Minimum flow time $u_{\min}$	0.001

displays the destination-minus-source difference before gating; unsupported endpoints yield zero. Panel (e) uses the pooled potentials in Equation 7; applying the gate in Equation 10 yields (b). The training history ablation in Table 3 retains the gate.

Independent continuation evaluation. For an evaluated edge $e = (x, \mathbf{a}, x^+)$, we restore physical states from the exact current-plus-historical evidence pools used in Figure 5(e) and continue the same frozen policy. The main credit audit uses 12,695 continuations across 9,878 distinct evidence states in 159 endpoint contexts. We first estimate each state's continuation success rate, then average these rates with equal weight within each endpoint pool, regardless of the number of continuations per state. Their difference defines the independent MC estimate in panel (c):

$$\widehat{\Delta}_e^{\text{MC}} = \widehat{V}_{x^+}^{\text{MC}} - \widehat{V}_x^{\text{MC}}. \tag{E.1}$$

Credit-quality metrics. Let $\mathcal{S}$ be an evaluated set of chunks with nonzero signed credits $\widetilde{C}_e$ and MC differences $\widehat{\Delta}_e^{\text{MC}}$. Direction accuracy (DirAcc) is the percentage of sign matches on $\mathcal{S}_{\neq} = \{e \in \mathcal{S} : \widehat{\Delta}_e^{\text{MC}} \neq 0\}$:

$$\text{DirAcc}(\mathcal{S}) = \frac{100}{|\mathcal{S}_{\neq}|} \sum_{e \in \mathcal{S}_{\neq}} \mathbb{I}\Big[\text{sign}(\widetilde{C}_e) = \text{sign}(\widehat{\Delta}_e^{\text{MC}})\Big]. \tag{E.2}$$

Spearman rank correlation (RankCorr) compares the ordering of signed credits and MC differences:

$$\text{RankCorr}(\mathcal{S}) = \text{Spearman}\Big((\widetilde{C}_e)_{e \in \mathcal{S}}, (\widehat{\Delta}_e^{\text{MC}})_{e \in \mathcal{S}}\Big). \tag{E.3}$$

Aligned MC gap (AlignedGap) averages the per-chunk aligned differences $g_e = 100\ \text{sign}(\widetilde{C}_e)\widehat{\Delta}_e^{\text{MC}}$, in percentage points:

$$\text{AlignedGap}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{e \in \mathcal{S}} g_e. \tag{E.4}$$

Positive g_e indicates directional agreement and negative g_e indicates disagreement. RankCorr and AlignedGap include MC ties. Confidence intervals use rollout-group-cluster bootstrap resampling.

Evaluation subsets. Figure 5(f) evaluates the 123 nonzero candidates in (e), their 78 rejected candidates, and the 45 retained credits in (b). All subsets compare pre-gate credit from (e) with the MC reference in (c). For Table 8, the correct/non-tied counts are 39/40 for confidence gating and 29/44 for endpoint-count selection.

Online archive audit. We audit three PRICE runs on LIBERO-Spatial at rounds 16, 32, and 63, with $\delta_{\text{edge}} = 0.15$. At each stage, we uniformly sample 400 supported non-terminal chunks per run, irrespective of gate acceptance. For retained credits, we fix the selected evidence pools and estimate $\overline{\Delta}_e$ using 512 fresh continuations per endpoint pool from uniformly sampled evidence histories under their collection policies. DirAcc, RankCorr, and AlignedGap follow the definitions above, including the treatment of MC ties. Results are reported in Table 7.

F Real-Robot Experimental Details

F.1 Robot Observations and Task Definitions

Both tasks use the real-robot platform, task definitions, and deployment configuration of HiFi-UMI (AI et al., 2026). In Remote Insertion, the robot grasps a remote control and inserts it into a target storage box. In Produce Sorting, it places a designated fruit and vegetable into their respective category-specific trays; an episode succeeds only if both placements are completed. The policy receives camera images and a 20-dimensional proprioceptive state. The recorded executions illustrated in Figure 4(b) contain four camera streams, with upper and lower views from each wrist. The action prediction horizons are 50 for Remote Insertion and 20 for Produce Sorting.

F.2 Offline Adaptation Protocol

We initialize every variant from an OpenPI-$\pi_{0.5}$ checkpoint post-trained with 400 HiFi-UMI trajectories. Before adaptation, we run this initialization policy on each task to collect 60 interaction trajectories. These fixed task-specific datasets provide both the terminal outcomes and the physical relations used for offline adaptation; no additional online rollouts are introduced during optimization.

GRPO-AWR forms a trajectory-level advantage from task-normalized terminal outcomes and assigns it to all action chunks in that trajectory. PRICE-AWR adds the action-chunk relation credit to the same terminal advantage before mapping the combined signal to a bounded AWR weight for the flow-matching loss. The two methods share the same single-epoch schedule, random seed, global batch size of 4, and learning rate of 10^{-5}.

F.3 Offline Relational Credit and Weighted Flow Matching

Fixed-buffer relational credit. Let $\mathcal{D}$ denote the training action chunks in a task-specific offline buffer. We construct physical nodes once from frozen visual–proprioceptive descriptors. Each distinct episode contributes its binary terminal outcome once per visited node. For a query chunk (i, t), episode i is excluded from both endpoint outcome estimates. The offline contrast $c_{i,t}$ is the destination-minus-source potential difference for supported non-terminal chunks, and zero for unsupported or final chunks. This fixed-buffer adapter uses the raw contrasts without the online confidence gate; graph assignments and credits remain fixed during adaptation.

Advantage construction. For task context κ, let $\bar{Y}_\kappa$ and σ_κ be the mean and population standard deviation of outcomes across distinct training episodes. Let $\mathcal{D}_\kappa^+$ contain the training chunks whose contrast is nonzero, and define their root-mean-square contrast by

$$s_\kappa = \sqrt{\frac{1}{|\mathcal{D}_\kappa^+|} \sum_{(i,t)\in\mathcal{D}_\kappa^+} c_{i,t}^2}. \tag{F.1}$$

We set $s_\kappa = 1$ when this set is empty and use $\varepsilon = 10^{-8}$ for numerical stability. The offline advantages are

$$\begin{aligned} A_i^{\text{out}} &= \frac{Y_i - \bar{Y}_{\kappa(i)}}{\max(\sigma_{\kappa(i)}, \varepsilon)}, \\ A_{i,t}^{\text{AWR}} &= A_i^{\text{out}} + \lambda_{\text{AWR}} \frac{c_{i,t}}{\max(s_{\kappa(i)}, \varepsilon)}. \end{aligned} \tag{F.2}$$

This scaling preserves zero credit. PRICE-AWR uses $\lambda_{\text{AWR}} = 0.2$; GRPO-AWR uses only A_i^{out}.

Table 5: Main PRICE results (%, mean ± SD) across three independent training runs.

LIBERO Method	Spatial	Object	Goal	Long	Avg.
PRICE	99.2 ± 0.2	99.6 ± 0.2	99.4 ± 0.3	97.6 ± 0.5	99.0 ± 0.2

RoboTwin 2.0 Method	Handover Block	Lift Pot	Move Can Pot	Pick Dual Bottles	Place Container Plate	Place Empty Cup	Avg.
PRICE	79.7 ± 2.3	80.5 ± 1.6	90.6 ± 1.6	94.5 ± 1.4	97.7 ± 0.8	96.9 ± 0.8	90.0 ± 0.8

Bounded AWR weights. Index training chunks by j, and let A_j denote the corresponding method's advantage. We exponentiate, normalize by the training-buffer mean, and clip the weights:

$$r_j = \exp(A_j/\tau_{\mathrm{AWR}}), \qquad \bar{r} = \frac{1}{|\mathcal{D}|}\sum_{j\in\mathcal{D}} r_j,$$
$$w_j = \mathrm{clip}(r_j/\bar{r}, w_{\min}, w_{\max}), \quad \bar{w} = \frac{1}{|\mathcal{D}|}\sum_{j\in\mathcal{D}} w_j. \tag{F.3}$$

We use $\tau_{\mathrm{AWR}} = 1$, $w_{\min} = 0.1$, and $w_{\max} = 5$. Each method computes its own $\bar{r}$ and $\bar{w}$, which remain fixed during training. The loss uses $w_j/\bar{w}$ so that its scale is normalized by the dataset mean after clipping.

Flow-matching objective. Both adaptation methods retain the OpenPI flow-matching time sampler, with parameters reported in Table 4. For normalized action chunk $\mathbf{a}_j$, sample $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ and $b \sim \mathrm{Beta}(\alpha_{\mathrm{time}}, \beta_{\mathrm{time}})$, and set $u = u_{\min} + (1 - u_{\min})b$. The interpolated action is $\mathbf{x}_u = (1-u)\mathbf{a}_j + u\boldsymbol{\epsilon}$, with target velocity $\boldsymbol{\epsilon} - \mathbf{a}_j$. Let $m_{j,r}$ mark valid action timesteps, and let $d = 32$ be the padded model action dimension. The per-chunk loss and weighted minibatch objective are

$$\ell_j(\theta) = \mathbb{E}_{u,\boldsymbol{\epsilon}}\left[\frac{\sum_r m_{j,r}\,\|v_\theta(h_j, \mathbf{x}_u, u)_r - (\boldsymbol{\epsilon} - \mathbf{a}_j)_r\|_2^2}{d\sum_r m_{j,r}}\right],$$
$$\mathcal{L}_{\mathrm{AWR}}(\theta) = \frac{1}{|\mathcal{B}|\bar{w}}\sum_{j\in\mathcal{B}} w_j \ell_j(\theta). \tag{F.4}$$

Here $\mathcal{B}$ contains chunks with at least one valid action timestep; padded timesteps are masked out. All graph statistics, advantages, and weights are fixed supervision signals. Both adaptation methods optimize this same flow-matching objective, differing in whether relational credit enters the AWR advantage.

G Additional Results and Analyses

G.1 Run-to-Run Variation in Main Results

Table 5 reports mean success rates and sample standard deviations across three independent PRICE training runs. The Avg. standard deviation (SD) is computed from the average success rate within each run.

G.2 Training Efficiency

GPU hours (GPUh) are computed as the number of GPUs multiplied by elapsed training time in hours. Table 6 reports cumulative cost at the 90% and 95% rollout-success milestones in Figure 3(a–b). At the 90% milestone, PRICE uses 272.8 GPU hours on Goal and 221.2 on Spatial, reducing cost by 38.3% and 32.8%, respectively, relative to SimpleVLA-RL. At 95%, the reductions are 17.3% and 23.8%. These results show that PRICE's gains in training-step efficiency also translate into lower cumulative GPU-hour costs.

Table 6: Cumulative GPU hours at the marked milestones in Figure 3(a–b). Reductions are relative to SimpleVLA-RL.

Suite	Success (%)	SimpleVLA-RL	PRICE	Reduction (%)
Goal	90	441.9	**272.8**	38.3
	95	625.2	**517.2**	17.3
Spatial	90	329.0	**221.2**	32.8
	95	714.6	**544.3**	23.8

G.3 Mechanism Analysis

Credit alignment during online training. Retained credits achieve mean DirAcc above 94%, positive rank correlations, and positive aligned MC gaps at all three training stages (Table 7). These results support agreement between online PRICE credit and the independent collection-policy MC reference as the policy and archive evolve. The audit protocol is detailed in Appendix E.3.

Table 7: Online credit alignment on LIBERO-Spatial. Of 1,200 audited chunks per stage across three PRICE runs, 24, 48, and 36 receive nonzero credit at rounds 16, 32, and 63, respectively.

Training round	DirAcc (%) ↑	RankCorr ↑	AlignedGap (pp) ↑
16	95.8	0.67	+24.6
32	95.8	0.71	+27.8
63	94.3	0.68	+26.2

Physical context of credited chunks. We analyze the physical context of action chunks assigned nonzero credit by PRICE in LIBERO-Spatial. Each chunk is assigned to one of four mutually exclusive categories, in order: (i) gripper–target contact, detected at a chunk boundary, or target-object displacement greater than 1 cm; (ii) an arm-induced reduction of more than 1 cm in gripper-to-object surface distance, provided category (i) does not apply; (iii) a source or destination configuration within 5 cm of the target-object surface, provided neither of the first two categories applies; and (iv) other. The first three categories constitute interaction-related intervals. Of the 777 credited chunks, 572 (73.6%) fall within these intervals, accounting for 72.2% of the total absolute credit magnitude. Contact or manipulation, approach, and proximity alone account for 33.7%, 19.4%, and 20.5% of all credited chunks, respectively. Thus, most credited chunks occur during approach to, contact with, or positioning near the target object, as well as object movement.

Credit selection at equal coverage. We examine whether confidence gating selects more reliable credits than random subsampling or selection based on evidence quantity. All three selectors retain 45 of the same 123 nonzero candidates in Figure 5. Endpoint-count selection keeps candidates with the largest minimum endpoint support, $\min(N_x, N_{x^+})$, while random selection samples uniformly without replacement. Confidence gating achieves 97.5% DirAcc and $+28.5$ pp AlignedGap, outperforming both controls (Table 8). These results support assessing potential changes relative to their uncertainty when selecting local credit.

Alignment of retained and rejected credits. We examine whether gating retains credits that better reflect local progress. Figure 6(a) compares pre-gate credit with independent MC estimates for the same candidates as Figure 5. Retained credits exhibit a higher rate of directional agreement with MC progress than rejected candidates. Figure 6(b) additionally accounts for the magnitude of MC changes by aligning them with each candidate's credit direction. Retained credits achieve an AlignedGap of $+28.47$ pp, compared with $+3.81$ pp for rejected candidates. Their aligned-gap distribution has an interquartile range above zero, whereas the rejected subset's spans zero. These results indicate that gating selects credits with more consistent directional agreement and larger average MC changes in the credited direction.

Table 8: Credit selection at equal coverage on LIBERO-Spatial. Each selector retains 45 of the same 123 candidates. MC ties are excluded from DirAcc; the random row reports means over 100,000 subsets.

Selector	DirAcc (%) ↑	AlignedGap (pp) ↑
Confidence gate	**97.5**	**+28.5**
Endpoint-count selection	65.9	+4.6
Random selection	77.2	+12.8



Figure 6: Credit selection and local progress on LIBERO-Spatial. (a) Raw PRICE credit versus MC progress for the same 123 candidates as Figure 5; blue and gray denote the 45 retained and 78 rejected candidates, respectively. Shaded quadrants indicate agreement in sign. (b) Per-chunk aligned MC gaps g_e (pp), whose means give AlignedGap. Vertical jitter separates points; thick segments show interquartile ranges, dark ticks mark medians, and annotations report means.

Table 9: Independent descriptor comparison. Coverage is the percentage of the 20,480 corpus chunks receiving nonzero credit; quality metrics use descriptor-specific MC samples and the definitions in Appendix E.3. Brackets give 95% rollout-group-cluster bootstrap intervals.

Descriptor	Coverage (%)	DirAcc ↑ (%)	RankCorr ↑	AlignedGap ↑ (pp)
Proprio	3.6	88.2 [75.7, 97.4]	0.650	+23.4 [13.6, 33.0]
Vision	4.2	87.2 [71.4, 100.0]	0.595	+29.0 [19.6, 38.5]
Fusion	3.8	**93.9** [83.9, 100.0]	**0.758**	**+33.7** [23.7, 43.5]

G.4 Descriptor Robustness

PRICE remains reliable across descriptor choices. In an independent experiment on 20,480 action chunks, we rebuild the graph using proprioceptive, fused visual–proprioceptive, and visual descriptors with the remaining settings fixed. All three yield DirAcc above 87% and positive AlignedGap (Table 9). Fusion achieves the highest credit-quality scores: 93.9% DirAcc, 0.758 RankCorr, and +33.7 pp AlignedGap. Vision provides the highest coverage at 4.2%, compared with 3.8% for Fusion and 3.6% for Proprio.

Table 10: Credit-weight sensitivity on LIBERO-Spatial. Task success is evaluated after 63 training steps; bold marks the default weight.

Credit weight λ	0.1	**0.2**	0.3	0.4
Success (%)	95.8	**97.0**	96.5	96.9

Table 11: Gate-parameter sensitivity on 160 locked chunks. Retained counts nonzero credits; bold marks the default setting. Metric definitions are given in Appendix E.3.

δ_{edge}	Retained	DirAcc (%) ↑	AlignedGap (pp) ↑
0.05	25	100.0	+30.8
0.10	30	100.0	+29.3
0.15	45	97.5	+28.5
0.30	46	95.1	+27.3

Table 12: Credit coverage and archive size under frozen-policy replay of the same 20,480 chunks. Credited chunks counts nonzero gated credits; Nodes and Summaries are final archive totals across tasks. Evictions counts cumulative node removals due to the 1,024-node limit per task.

η	Credited chunks	Nodes	Summaries	Evictions
0.90	628	287	873	0
0.93	**777**	484	1,405	0
0.96	604	1,234	3,064	0
0.99	302	7,458	12,325	4,678

G.5 Hyperparameter Sensitivity

Sensitivity to the credit weight. After 63 training steps on LIBERO-Spatial, $\lambda = 0.2$ and 0.4 achieve similar success (97.0% and 96.9%, respectively; Table 10). However, $\lambda = 0.2$ reaches 90% success in 22 steps versus 29 for $\lambda = 0.4$, supporting our default choice for its faster learning at comparable final success.

Credit selection across gate settings. Table 11 varies δ_{edge} on 160 locked chunks while holding node assignments and evidence pools fixed. The stricter settings 0.05 and 0.10 retain 25 and 30 credits, respectively, with 100% DirAcc on non-tied MC comparisons. At the default setting of 0.15, the gate retains 45 credits with 97.5% DirAcc and $+28.5$ pp AlignedGap. Relaxing the gate to 0.30 retains 46 credits, while DirAcc decreases to 95.1% and AlignedGap to $+27.3$ pp.

Choice of the matching threshold. Table 12 compares credit coverage and archive size across matching thresholds, with other parameters fixed to the defaults in Table 4. We use $\eta = 0.93$ to obtain the highest observed coverage of gated credit with a relatively small archive: 777 credited chunks and 484 nodes. Higher tested thresholds increase archive size while reducing coverage.

H Extension to World Action Models

To examine applicability beyond VLA policies, we study PRICE on Fast-WAM (Yuan et al., 2026), a representative world action model that combines video modeling with action generation. We compare GRPO and PRICE on LIBERO-10 to assess the benefit of relational credit with a WAM backbone.

Initialization. We follow the official Fast-WAM SFT pipeline and select the `step_006000.pt` checkpoint, which achieves 53.4% success on LIBERO-10, as the base model for both GRPO and PRICE.

Policy adaptation. Fast-WAM generates continuous action means through flow matching. For each rollout chunk, we cache the features $z_{i,t}$ and action latent $x_{i,t}$ entering the final flow step,

Table 13: Success rates (%) on LIBERO-10 with Fast-WAM, evaluated over 50 episodes per task. The official full-SFT result is included for reference.

Method	LIBERO-10 success (%)
Fast-WAM (official full SFT)	95.2
SFT initialization (step 6,000)	53.4
+ GRPO	92.8
+ PRICE	**97.2**

together with its integration increment Δs. Writing $\xi_{i,t} = (z_{i,t}, x_{i,t}, \Delta s)$, the update uses

$$\begin{aligned} \mu_\theta(\xi_{i,t}) &= \mathcal{P}(x_{i,t} + \Delta s\, f_\theta(z_{i,t}))\,, \\ q_\theta(\mathbf{a} \mid \xi_{i,t}) &= \mathcal{N}\big(\mu_\theta(\xi_{i,t}), \mathrm{diag}(\sigma_\theta^2)\big)\,, \end{aligned} \tag{H.1}$$

where f_θ is the action output head and $\mathcal{P}$ applies action denormalization, environment-specific post-processing, and selection of the executed action horizon. Rollouts sample from this diagonal Gaussian; updates hold $\xi_{i,t}$ fixed and optimize the output head and log standard deviations, keeping the remaining backbone frozen. For action coordinate k within a chunk, the likelihood ratio is

$$\rho_{i,t,k}(\theta) = \frac{\mathcal{N}\Big(a_{i,t,k}; \mu_{\theta,k}(\xi_{i,t}), \sigma^2_{\theta,k}\Big)}{\mathcal{N}\Big(a_{i,t,k}; \mu_{\theta_{\mathrm{old}},k}(\xi_{i,t}), \sigma^2_{\theta_{\mathrm{old}},k}\Big)}. \tag{H.2}$$

We use these ratios in the clipped surrogate of Equation 2, averaging over valid action coordinates and sharing each chunk's advantage across its coordinates.

Controlled comparison. Both methods share the initial checkpoint, rollout grouping, optimizer settings, training budget, and evaluation protocol. GRPO uses the trajectory-level advantage, while PRICE applies the augmentation in Equation 12. PRICE uses visual–proprioceptive boundary descriptors, cross-trajectory outcome pooling, and confidence-gated credit with the settings in Table 4; credit addition is not followed by advantage renormalization. Table 13 reports aggregate success rates and includes the official Fast-WAM full-SFT result of 95.2% for reference. PRICE achieves 97.2% success, compared with 92.8% for GRPO, an improvement of 4.4 percentage points. This result supports the applicability of relational credit beyond VLA policies to a WAM-based control policy.

I Additional Visualizations

We show chronological execution sequences from simulation and the real robot.

I.1 Simulated Manipulation

Figure 7 shows four recorded OpenVLA-OFT executions on LIBERO-Spatial. Different bowl starting locations require different extraction and transport motions before placement on the plate. The unsuccessful attempt reaches the plate region but does not complete the task.

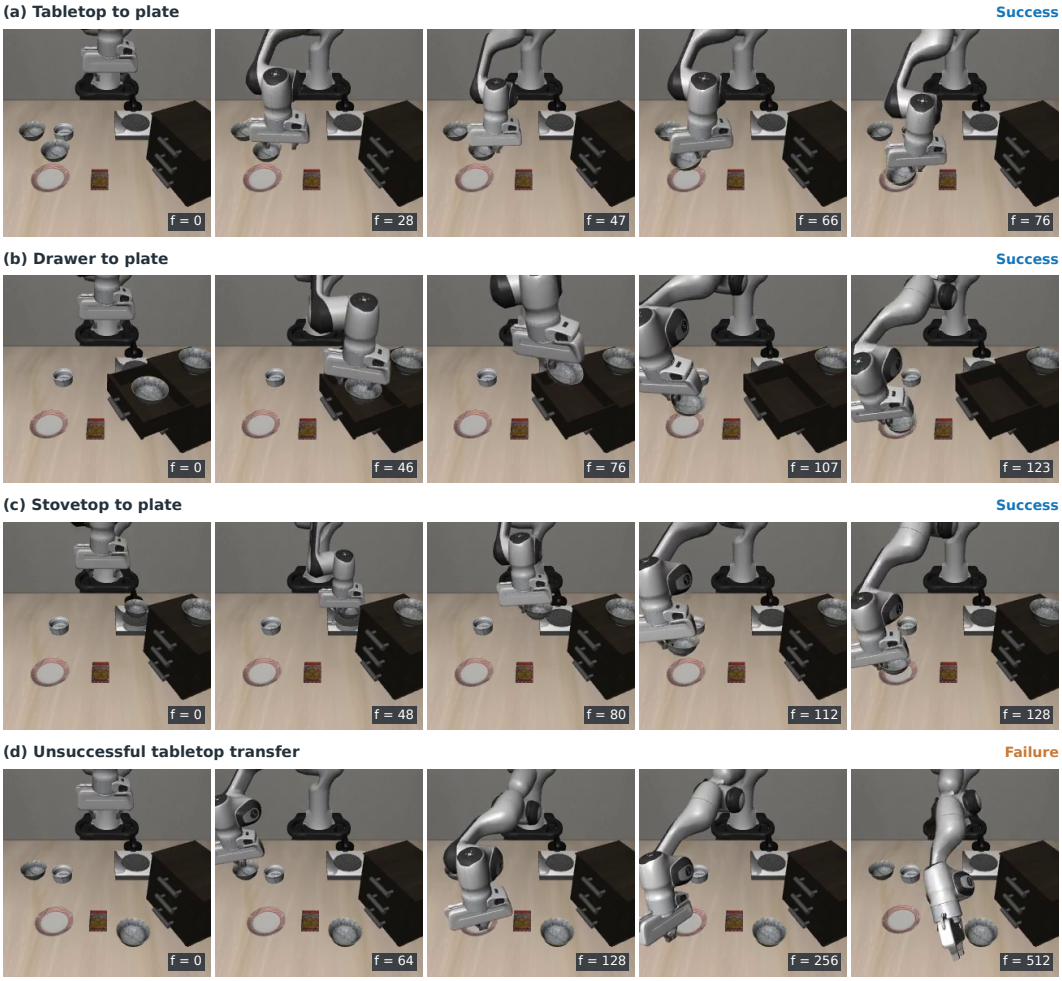

Figure 7: Additional simulated execution examples on LIBERO-Spatial. Each row shows five chronological frames from one recorded rollout: successful bowl transfers from (a) the tabletop, (b) a drawer, and (c) the stovetop, followed by (d) an unsuccessful tabletop transfer. Labels f are zero-based video frame indices.

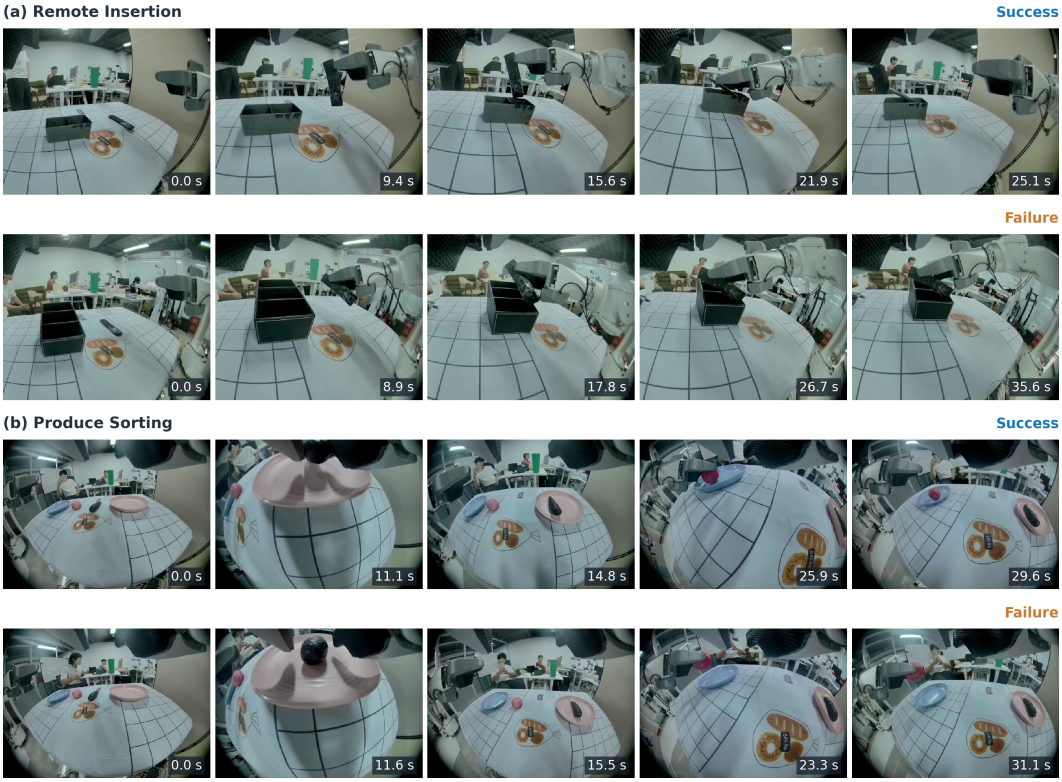

Figure 8: Additional real-robot execution examples. (a) Remote Insertion and (b) Produce Sorting each show one successful episode (upper row) and one unsuccessful episode (lower row), with five chronological frames per episode. Timestamps are relative to episode start. Remote Insertion uses the lower left-wrist camera; Produce Sorting uses the lower right-wrist camera to show both trays. Outcomes are operator-provided episode labels.

I.2 Real-Robot Manipulation

Figure 8 presents four additional episodes, distinct from those in Figure 4(b). The unsuccessful insertion attempt keeps the remote near the box, while the unsuccessful sorting attempt completes the vegetable placement but leaves the fruit held near its tray. The successful sorting sequence shows both placements, extending the fruit-only view in the main figure.

J LIMITATIONS

PRICE trades off evidence reuse against collection-policy mismatch and state-abstraction error (Appendix D.1). Credit quality depends on how well the descriptor preserves distinctions relevant to success and how representative historical outcomes remain under the current policy. Short observation histories could improve physical correspondence, while more targeted history selection could reduce policy mismatch. Local credit also requires repeated visits to corresponding states, limiting its coverage during early training or in tasks with little overlap across trajectories. Improved exploration and evidence reuse could expand this coverage. Finally, training becomes less efficient as success rates approach saturation. Further gains beyond 90–95% require additional interaction, while increasingly uniform terminal outcomes reduce the diversity of new evidence for credit assignment. Sampling difficult tasks or initial conditions more frequently could concentrate the interaction budget on remaining failures and improve learning efficiency in this regime.