E2HiL: Entropy-Guided Sample Selection for Efficient Real-World Human-in-the-Loop Reinforcement Learning

Haoyuan Deng, Yudong Lin, Yuanjiang Xue, Haoyang Du, Qianzhun Wang, Boyang Zhou, Zhenyu Wu, and Ziwei Wang, *Member, IEEE*

***Abstract*—Human-in-the-loop guidance has emerged as an effective approach for accelerating online reinforcement learning (RL) in real-world manipulation. However, existing human-in-the-loop RL (HiL-RL) frameworks often suffer from low sample efficiency, requiring substantial human interventions to achieve convergence and thereby leading to high labor costs. To address this, we propose a sample-efficient real-world human-in-the-loop RL framework named `E2HiL`, which requires fewer human interventions by actively selecting informative samples. Specifically, stable reduction of policy entropy enables improved trade-off between exploration and exploitation with higher sample efficiency. We first build influence functions of different samples on the policy entropy, which is efficiently estimated by the covariance of action probabilities and soft advantages of policies. Then we select samples with moderate values of influence functions, where shortcut samples that induce sharp entropy drops and noisy samples with negligible effect are pruned. Extensive experiments across 10 real-world manipulation tasks, spanning multiple embodiments and learning frameworks, demonstrate that `E2HiL` improves success rates by 24.9% while reducing human interventions by 9.3% compared to state-of-the-art HiL-RL baselines. These results validate its effectiveness as a policy- and embodiment-agnostic plug-and-play module for efficient real-world RL.**

***Index Terms*—Human-in-the-loop reinforcement learning, robotic manipulation, sample efficiency.**

I. Introduction

ROBOTIC manipulation is a fundamental capability for robots deployed in both industrial assembly and household services [1]. Despite decades of progress, achieving human-level manipulation skills in dynamic and unstructured environments remains a challenge [2]. Meanwhile, conventional manually designed controllers struggle to generalize across diverse tasks and high-dimensional policy spaces, motivating the transformation toward learning methods.

Although pre-trained imitation policies show strong performance, they still struggle with out-of-distribution (OOD) states [3], [4], [5], [6]. To improve robustness, recent work leverages RL's trial-and-error exploration to acquire more reliable task-specific skills. However, RL trained purely in simulation often fails to transfer to real robots due to gaps in dynamics and perception [7], prompting the adoption of real-world reinforcement learning. Nevertheless, directly training on real-world robots intensifies challenges such as low sample efficiency [8] and sparse rewards [9], [10]. Human-in-the-loop RL (HiL-RL) [2] leverages human corrective takeovers to accelerate online policy learning. However, existing human-in-the-loop RL methods still require extensive human intervention samples to converge, as the complexity of policy learning grows exponentially with the state–action dimensionality and task horizon, ultimately leading to high human labor costs [11]. This naturally raises the question: *How can we achieve more stable and sample-efficient human-in-the-loop RL in real-world robotic manipulation?*

In this letter, we propose an efficient real-world human-in-the-loop RL framework named **`E2HiL`** with active entropy-guided sample selection. Specifically, we identify a key limitation of existing HiL-RL methods: they cannot distinguish which intervention samples are truly informative, leading to inefficient use of human effort and degraded exploration. Fig. 1 reveals a clear trade-off between entropy dynamics and policy improvement during HiL-RL training. This pattern mirrors recent LLM-RL findings [12], [13], where rapid entropy reduction leads to premature convergence to suboptimal policies and stable reduction of policy entropy enables improved trade-off between exploration and exploitation with higher sample efficiency. To address this limitation, we derive influence functions that estimate effect of each sample on policy entropy using the covariance between action probabilities and soft advantages of policies. Building on this, **`E2HiL`** identifies uninformative samples with extreme values of influence functions, including shortcut samples that induce sharp entropy drops and noisy samples with negligible influence. Rather than treating all samples equally, **`E2HiL`** further proposes an entropy-bounded sample selection mechanism to prune uninformative samples, which stabilizes policy entropy dynamics and prevents premature collapse. We validate **`E2HiL`** across 10 real-world manipulation tasks

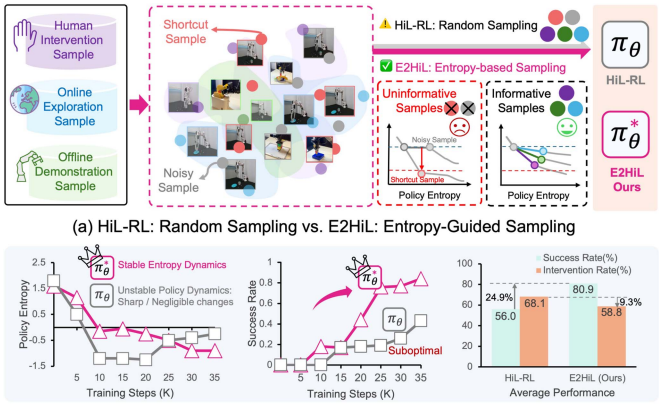
Received 2 December 2025; accepted 22 April 2026. Date of publication 14 May 2026; date of current version 25 May 2026. This article was recommended for publication by Associate Editor G. Pizzuto and Editor W. Pan upon evaluation of the reviewers' comments. This work was supported by the Singapore National Robotics Programme through Research Project DS-RFM M25N4N2009. *(Corresponding author: Ziwei Wang.)*

Haoyuan Deng, Yudong Lin, Yuanjiang Xue, Haoyang Du, Qianzhun Wang, Boyang Zhou, and Ziwei Wang are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: ziwei.wang@ntu.edu.sg).

Zhenyu Wu is with the Beijing University of Posts and Telecommunications, Beijing 100876, China.

The project page can be found at https://e2hil.github.io/.

Digital Object Identifier 10.1109/LRA.2026.3693577



(a) HiL-RL: Random Sampling vs. E2HiL: Entropy-Guided Sampling

(b) Convergence Curve on *Push-Cube* Task and Multi-Task Performance Comparison

Fig. 1. Compared with random uniform sampling, **E2HiL** avoids early entropy collapse through entropy-guided sample selection, achieving a better exploration–performance trade-off and improving success rates with fewer human costs.

spanning multiple embodiments and learning frameworks, achieving a 24.9% higher success rate and 9.3% fewer human interventions compared to state-of-the-art HiL-RL baselines. These results highlight entropy-guided sample selection as a promising and general strategy for efficient real-world HiL-RL, enabling robust skill acquisition with reduced human supervision. Our contributions can be summarized as follows:

1) We propose **E2HiL**, an efficient real-world human-in-the-loop RL framework that improves sample efficiency by actively selecting informative samples, thereby reducing human intervention costs.
2) We characterize sample-induced entropy dynamics through influence functions and apply an entropy-bounded selection rule to remove shortcut and noisy samples, thereby maintaining stable entropy reduction for efficient learning.
3) We evaluate **E2HiL** across 10 real-world tasks and show that its sampling efficiency enables faster convergence under reduced supervision while achieving an effective trade-off between exploration and performance.

II. Related Work

A. Real-World RL for Robotic Manipulation

Real-world reinforcement learning (RL) aims to train manipulation policies directly on physical robots, enabling them to acquire skills that operate reliably under real sensor feedback and physical dynamics. However, it remains highly challenging due to issues of sample efficiency [9], [10], [14], sparse rewards [14], [15], and the high cost of human interventions [2], [16]. To address sample inefficiency, prior works have explored off-policy and model-based RL [17], [18], hybrid methods [9], [19], [20], and offline pretraining with foundation-model priors [21], [22], [23]. Approaches such as reward shaping, inferring rewards from demonstrations or success classifiers [24], and leveraging video-based signals [25] have been proposed to solve the challenge of sparse rewards. To reduce human intervention costs, reset-free learning [26], [27] and autonomous exploration strategies [28] have been introduced. However, these approaches often significantly increase training time and therefore have not achieved widespread adoption in real-world robotic learning. Despite these advances, state-of-the-art methods increasingly adopt HiL-RL, where systems such as HIL-SERL [2] leverage human corrective samples to rapidly acquire precise and dexterous manipulation skills. However, conventional HiL-RL methods treat all intervention samples uniformly, making them unable to distinguish informative guidance from noisy or unhelpful corrections. This leads to inefficient use of human effort and slow policy improvement. In contrast, **E2HiL** addresses this issue by actively selecting the informative samples to guide learning more effectively.

B. Entropy-Regularized Reinforcement Learning

Recent robotic reinforcement learning widely adopts entropy-regularized objectives, most notably through SAC [29], MPO [30] and RLPD [31]. These methods show that entropy bonuses improve exploration and stabilize policy learning in manipulation tasks, but they rely on implicit temperature tuning and do not model entropy at the sample level. Token-level entropy regularization has recently been widely adopted in the training of large language models (LLM) with RL, where regulating entropy dynamics is crucial for balancing exploration and exploitation [32], [33]. Recent studies have proposed methods such as Clip-Higher [13] and staged temperature scheduling [34] to encourage exploration, as well as covariance-based regularizers like Clip-Cov and KL-Cov [12] to mitigate entropy collapse and stabilize learning. In contrast, fine-grained entropy regularization remains largely unexplored in robotic RL, and the fundamental role of entropy in policy learning is still unclear [31]. By analyzing action-level entropy dynamics, we observe patterns similar to those reported in LLM-RL, but with domain-specific differences [35]. To our knowledge, **E2HiL** is the first to introduce sample-level entropy regularization into real-world robotic RL, enabling more sample-efficient learning with faster convergence and reduced human intervention.

III. Methodology

A. Preliminaries and Problem Statement

We formulate real-world robotic manipulation RL as a Markov Decision Process (MDP) $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \rho, \mathcal{P}, r, \gamma\}$, where $\mathcal{S}$ denotes the state space (e.g., visual and proprioceptive observations), $\mathcal{A}$ is the action space (e.g., end-effector motions), and $\rho(s_0)$ is the initial state distribution. The transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ captures the stochastic dynamics of the system, while the reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ specifies task objectives. The discount factor $\gamma \in [0, 1)$ balances immediate and long-term return. We denote by $s_t \in \mathcal{S}$ and $a_t \in \mathcal{A}$ the state and action at time t.

To enhance real-world training efficiency, we employ human-in-the-loop reinforcement learning to leverage both human intervention and autonomous exploration samples for improved sample efficiency [31]:

$$\mathcal{L}_Q(\phi) = \mathbb{E}_{(s_t,a_t,r_t,s_{t+1})\sim\mathcal{D}}\left[\left(Q_\phi(s_t, a_t) - \hat{y}_t\right)^2\right] \quad (1)$$

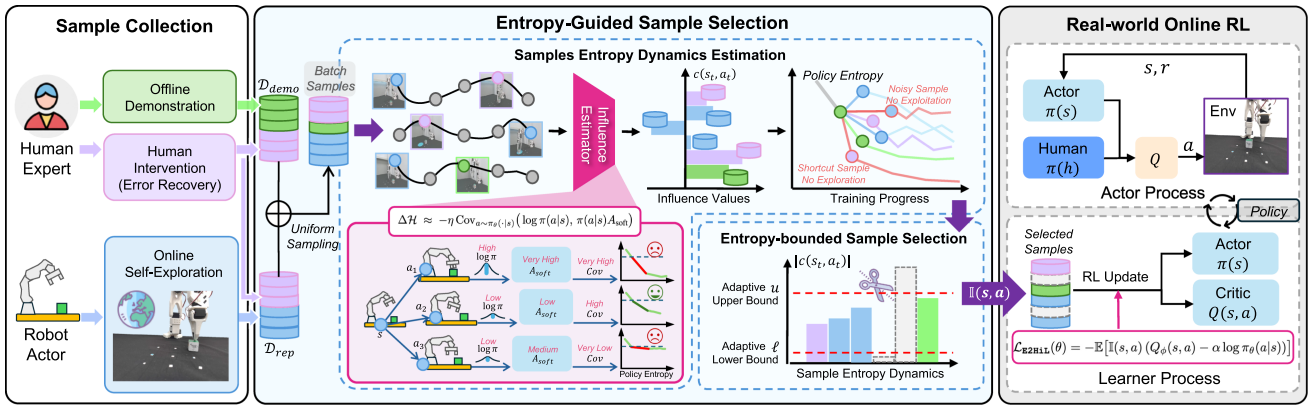


Fig. 2. Entropy-Guided Sample Selection Framework. Our framework consists of two key components: Sample Entropy Dynamics Estimation, where we characterize the entropy dynamics induced by each sample, and Entropy-Bounded Sample Selection, where we prune shortcut and noisy samples by constraining their influence value within dynamic bounds.

where the regression target $\hat{y}_t$ is the bootstrapped Bellman target estimating the expected return under the current policy. A slowly updated target network is used to compute $\hat{y}_t$ to stabilize training. In practice, policy is parameterized by neural networks $\pi_\theta(a_t \mid s_t)$ and the actor parameters θ are updated to maximize the entropy-regularized objective:

$$\mathcal{L}_\pi(\theta) = -\mathbb{E}_{(s_t,a_t)\sim\mathcal{D}}\left[Q_\phi(s_t,a_t) - \alpha \log \pi_\theta(a_t \mid s_t)\right] \quad (2)$$

where α is an adaptive temperature coefficient controlling the balance between exploitation and exploration.

However, conventional HiL-RL methods [2] uniformly sample from replay buffers with high human intervention cost. According to empirical results in [12], stable reduction entropy can improve the exploration and exploitation trade-off with higher sample efficiency, which can reduces the human intervention cost significantly. Therefore, we explicitly formulate our learning objective to incorporate *entropy-guided active sample selection*, wherein training samples contributing to moderate entropy reduction are selected. The objective for HiL-RL is written as follows:

$$\mathcal{L}_{\mathbf{E2HiL}}(\theta) = -\mathbb{E}\left[\mathbb{I}(s_t,a_t)\left(Q_\phi(s_t,a_t) - \alpha \log \pi_\theta(a_t \mid s_t)\right)\right] \quad (3)$$

where $\mathbb{I}(s_t,a_t) \in \{0,1\}$ is an indicator function determined by the entropy-bounded selection criterion. Samples with excessive or negligible entropy contribution are masked out ($\mathbb{I}=0$), removing their gradients from the update. This selective objective preserves only entropy-consistent samples, thereby reducing variance and improving sample efficiency.

B. Overall Pipeline

Our proposed framework aims to stabilize policy entropy dynamics in human-in-the-loop RLPD learning by explicitly regulating the selection and use of replay samples during training. As illustrated in Fig. 2, the entire pipeline follows an iterative loop of sample collection, entropy-guided active sample selection, and online RL policy update. For sample collection, **E2HiL** uses three sources: offline demonstrations before training, robot self-exploration samples during online interaction, and human intervention samples which are corrective takeover for error recovery. All collected transitions (s_t, a_t, r_t, s_{t+1}) are stored in the replay buffer $\mathcal{D}_{rep}$ and demonstration buffer $\mathcal{D}_{demo}$. Before updating the policy, we estimate influence functions of different samples on the policy entropy, which is formulated by influence value $c(s_t,a_t)$: the covariance of action probabilities and soft advantages of policy. Then, we propose entropy-bounded sample selection mechanism that assigns a binary indicator $\mathbb{I}(s_t,a_t)$ to retain samples with moderate influence. Samples with moderate influence on entropy are retained for policy updates, while those causing sharp drops or negligible changes are excluded. This entropy-guided selection stabilizes entropy dynamics and helps maintain a proper balance between exploration and exploitation. By integrating entropy estimation with selective sampling in a closed training loop, the framework improves sample efficiency and enhances convergence stability in real-world human-in-the-loop RL.

C. Sample Entropy Dynamics Estimation

The key challenge in human-in-the-loop RL is improving sample efficiency while avoiding unstable entropy collapse caused by uniform replay. We therefore quantify each sample's effect on policy entropy via a batch-wise influence function derived from entropy dynamics. Specifically, the entropy change of policy π_θ can be approximated as [12]:

$$\Delta\mathcal{H} \approx \mathbb{E}_{s_t\sim d_{\pi_\theta}}\left[-\operatorname{Cov}_{a_t\sim\pi_\theta}\left(\log \pi_\theta(a_t \mid s_t),\, \Delta z_t\right)\right], \quad (4)$$

where Δz_t denotes the logit update for action a_t at state s_t, and $\operatorname{Cov}(\cdot,\cdot)$ is taken under $\pi_\theta(\cdot \mid s_t)$. Intuitively, by relating policy updates to changes in entropy, (4) provides a measurable way to assess each sample's influence on policy entropy.

Logit Update under RLPD: To make (4) applicable in practice, a straightforward approach is to express Δz_t using the actor gradient in RLPD, where policy logits are updated through the standard policy gradient:

$$\Delta z_t \;:=\; z_t^{t+1} - z_t^t \;=\; -\eta\, \nabla_\theta \mathcal{L}_\pi(\theta) \quad (5)$$

where the negative sign arises from minimizing the RLPD actor objective function $\mathcal{L}_\pi$ via gradient descent, and η denotes the learning rate. Let $g(s_t, a'_t) \triangleq Q_\phi(s_t, a'_t) - \alpha \log \pi_\theta(a'_t \mid s_t)$. Here a_t denotes the specific action index associated with the logit parameter θ_{s_t,a_t}, while a'_t denotes actions sampled from the policy $\pi_\theta(\cdot \mid s_t)$ inside the expectation. Using the score-function, the gradient of the actor objective is:

$$\nabla_\theta \mathcal{L}_\pi(\theta) = -\pi_\theta(a_t \mid s_t) \left(\underbrace{g(s_t, a_t) - V_\pi(s_t)}_{A_{\text{soft}}(s_t,a_t)} \right) \tag{6}$$

Here $A_{\text{soft}}(s_t, a_t)$ provides an approximation of the off-policy soft advantage, and $V_\pi(s_t)$ is defined as the entropy-regularized state value:

$$V_\pi(s_t) \;:=\; \mathbb{E}_{a'_t \sim \pi_\theta} \left[Q_\phi(s_t, a'_t) - \alpha \log \pi_\theta(a'_t \mid s_t) \right] \tag{7}$$

Equation (6) naturally appears in RLPD when expressing the policy gradient in terms of the soft advantage, making the actor update depend on the action's value relative to the entropy-regularized state value. Then we obtain the policy logit change for action a_t at state s_t:

$$\Delta z_t = -\eta \, \nabla_\theta \mathcal{L}_\pi(\theta) = \eta \, \pi_\theta(a_t \mid s_t) \, A_{\text{soft}}(s_t, a_t) \tag{8}$$

This decomposition expresses logit updates in terms of policy probabilities and soft advantages, enabling a sample-wise characterization of their influence on entropy.

Policy Entropy Change Estimation: Substituting (8) into (4), the influence function approximation yields:

$$\Delta \mathcal{H} \;\approx\; -\eta \, \text{Cov} \left(\log \pi_\theta(a_t \mid s_t), \; \pi_\theta(a_t \mid s_t) \, A_{\text{soft}}(s_t, a_t) \right). \tag{9}$$

which indicates that entropy dynamics are determined by the covariance between log-probabilities and the soft advantage of policies. This reflects how different samples influence policy entropy: actions with high probability and large soft advantage tend to reduce entropy, while actions with low probability and large soft advantage tend to increase it.

D. Entropy-Guided Sample Selection

To stabilize policy entropy and better utilize costly human interventions, we propose an *entropy-guided sample selection* framework, with two components: *sample entropy dynamics estimation* and *entropy-bounded sample selection*.

Sample Entropy Dynamics Estimation: Given a state–action pair (s_t, a_t), we estimate the covariance term in (9) via Monte Carlo sampling over $\pi_\theta(\cdot \mid s_t)$. Specifically, for each state s_t, we draw K actions $\{a_{t,k}\}_{k=1}^K \sim \pi_\theta(\cdot \mid s_t)$ and compute the covariance. We define the influence value of sample (s_t, a_t) on entropy dynamics as

$$c(s_t, a_t) \;\triangleq\; -\eta \widehat{\text{Cov}}(s_t, a_t). \tag{10}$$

The quantity $c(s_t, a_t)$ is treated as a stop-gradient signal that reflects how sample contributes to the entropy changes.

Entropy-bounded Sample Selection: In practice, we observe that a small portion of samples exhibit extremely large or small covariance magnitudes, far beyond the range of the majority within the same batch. These outlier samples play a dominant role in triggering abnormal entropy fluctuations, which can either drive the policy into premature collapse or stall its progress. Concretely, we categorize some problematic samples: 1) shortcut samples that induce abrupt entropy drops and cause premature policy collapse, and 2) noisy samples that contribute negligibly to entropy change and slow down learning progress (see Section IV-C for empirical verification). To mitigate the impact of samples that cause abrupt entropy changes or contribute negligibly to exploration, we clip the estimated absolute influence value $\{|c(s_t, a_t)|\}$ within a dynamically adjusted interval $[\ell, u]$, where the bounds are computed adaptively from the batchwise percentile range of influence magnitudes. Formally, we define an indicator function that determines whether a sample participates in the policy update:

$$\mathbb{I}(s_t, a_t) = \mathbf{1} \left[\, |c(s_t, a_t)| \in [\ell, u] \, \right]. \tag{11}$$

Samples outside this range are masked out ($\mathbb{I} = 0$), effectively canceling their gradient contributions. Given the per-sample actor loss $l_t \;:=\; Q_\phi(s_t, a_t) \;-\; \alpha \log \pi_\theta(a_t \mid s_t)$, our entropy-aware actor objective is defined as:

$$\mathcal{L}_{\mathbf{E2HiL}}(\theta) \;=\; -\frac{\sum_{(s_t,a_t)\in\mathcal{B}} \mathbb{I}(s_t, a_t) l_t}{\sum_{(s_t,a_t)\in\mathcal{B}} \mathbb{I}(s_t, a_t) \;+\; \varepsilon} \tag{12}$$

where $\mathcal{B}$ denotes a minibatch of size B, and ε is a small constant for numerical stability. This formulation ensures that only entropy-consistent samples contribute to the gradient, leading to smoother entropy evolution and more stable convergence in human-in-the-loop training.

Algorithm 1: RL With Entropy-Guided Sample Selection.

Temperature α, gradient steps G, batch size N
Initialize critics ϕ_i (targets $\phi'_i \leftarrow \phi_i$); initialize actor θ
Initialize replay buffer $\mathcal{D}_{\text{rep}}$ and demo buffer $\mathcal{D}_{\text{demo}}$
Receive initial state s_0
for $t = 0$ **to** T **do**
  if human intervention **then**
    Take action a_t from human operator
    Store (s_t, a_t, r_t, s_{t+1}) in $\mathcal{D}_{\text{rep}}$ and $\mathcal{D}_{\text{demo}}$
  else
    Sample $a_t \sim \pi_\theta(\cdot \mid s_t)$
    Store (s_t, a_t, r_t, s_{t+1}) in $\mathcal{D}_{\text{rep}}$
  for $g = 1$ **to** G **do**
    Sample minibatch b_R of $N/2$ from $\mathcal{D}_{\text{rep}}$
    Sample minibatch b_D of $N/2$ from $\mathcal{D}_{\text{demo}}$
    Set $b \leftarrow b_R \cup b_D$
    Update ϕ minimizing loss from (1):
    Update target networks $\phi'_i \leftarrow \rho\phi'_i + (1-\rho)\phi_i$
  Estimate influence value c_i on b from (10)
  Get indicator function $\mathbb{I}(s_t, a_t)$ from (11)
  Update actor by maximizing objective (12)

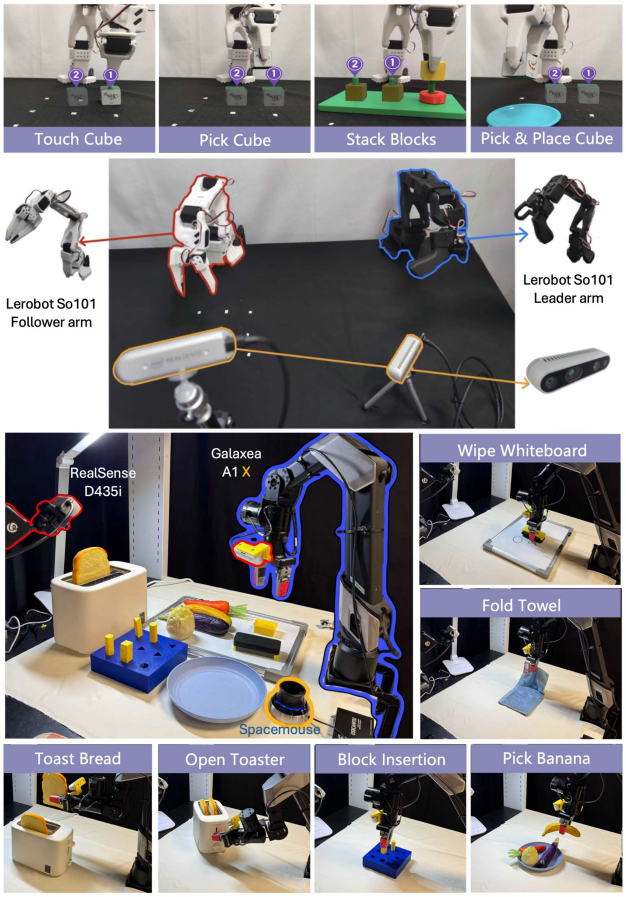


Fig. 3. Real-world experimental setups on Lerobot SO-101 and A1_X arm, comprising 10 manipulation tasks.

Fig. 4. Real-world Manipulation Results on the Lerobot. **E2HiL** significantly outperforms state-of-the-art human-in-the-loop RL approaches by achieving higher success rates with fewer human interventions across four tasks.

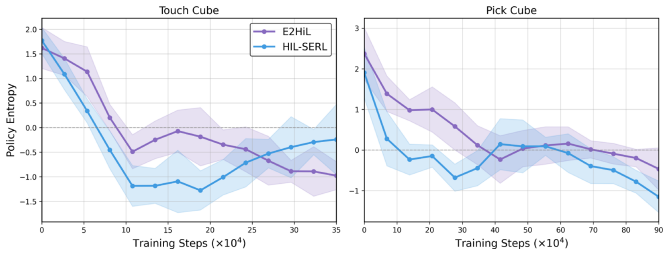


Fig. 5. With entropy-guided sample selection, **E2HiL** exhibits more stable entropy dynamics, avoiding premature entropy collapse and ineffective updates.

IV. Experiments

In this section, we first describe the real-world experimental setups and compare **E2HiL** with state-of-the-art HiL-RL methods. We then validate the accuracy of our entropy-dynamics estimation and the effectiveness of entropy-guided sample selection. Finally, we present ablation studies and hyperparameter analyses to examine the influence of key design choices, including sample pruning behavior, intervention types, and entropy-related parameters, on policy learning in real-world robotic manipulation.

A. Experimental Setup

Baseline: We adopt HIL-SERL [2] and ConRFT [36] as our primary baselines for real-world human-in-the-loop manipulation. HIL-SERL trains policies from scratch with corrective human interventions and entropy regularization from RLPD to stabilize updates. ConRFT initializes policies using Cal-QL [37] and behavior cloning pretraining, followed by RL fine-tuning of vision-language-action policies in a multi-stage paradigm. Both methods are evaluated under identical cross-embodiment settings on multi-object, long-horizon, contact-rich, and non-rigid tasks (Fig. 3).

Evaluation Metrics: We report three main metrics: the final **success rate** and the average **human intervention rate**, which is defined as the ratio between the number of intervention steps and the total robot interaction steps. We also monitor the evolution of the **policy entropy curve** to evaluate the stability of learning and the quality of exploration.

B. Comparisons With the State-of-The-Art

General Comparison on the Lerobot: We compare **E2HiL** with the state-of-the-art method HIL-SERL and report the performances across four position-generalization tasks of increasing difficulty. As shown in Fig. 4, **E2HiL** achieves the best performance in terms of convergence success rate, average human intervention rate, and the stability of the entropy curve. These results demonstrate that entropy-guided sample selection mechanism not only improves task performance and reduces human effort, but also stabilizes the entropy dynamics during training. From Fig. 5, we observe that **E2HiL** achieves a sizable relative improvement of **45%** in success rate across four tasks, while reducing the average human intervention rate by **10.1%**.

General Comparison on the A1_X: We further evaluate **E2HiL** on a broader set of real-world manipulation tasks using the A1_X platform. As shown in Table I, **E2HiL** improves the average success rate by 17.5% while reducing human intervention by 9.4% across all tasks. Table II further shows consistent gains over ConRFT, with a 6.6% increase in

TABLE I
REAL-WORLD MANIPULATION RESULTS ON THE A1_X ROBOT. **E2HiL** SIGNIFICANTLY OUTPERFORMS STATE-OF-THE-ART HUMAN-IN-THE-LOOP RL APPROACHES BY ACHIEVING HIGHER SUCCESS RATES WITH FEWER HUMAN INTERVENTIONS ACROSS SIX TASKS.

Method	Metrics	Multi-object	Long-horizon	Contact-rich			Non-rigid	Average
		Pick Banana	Toast Bread	Open Toaster	Block Insertion	Wipe Whiteboard	Fold Towel	
BC	Success Rate ↑	20.0	15.0	30.0	20.0	30.0	15.0	21.7
HiL-SERL	Success Rate ↑	65.0	55.0	85.0	60.0	70.0	65.0	66.7
	Intervention Rate ↓	58.6	68.1	43.2	54.6	65.3	66.3	59.4
HiL-SERL + **E2HiL**	Success Rate ↑	85.0	75.0	90.0	85.0	90.0	80.0	84.2
	Intervention Rate ↓	48.1	55.2	39.1	46.3	55.6	55.8	50.0

TABLE II
COMPARISON WITH CONRFT [36] ON THE A1_X

Metrics	Task	
	Block Insertion	Pick Banana
Δ Success Rate (%) ↑	+4.5	+8.6
Δ Intervention Rate (%) ↓	−6.7	−8.3

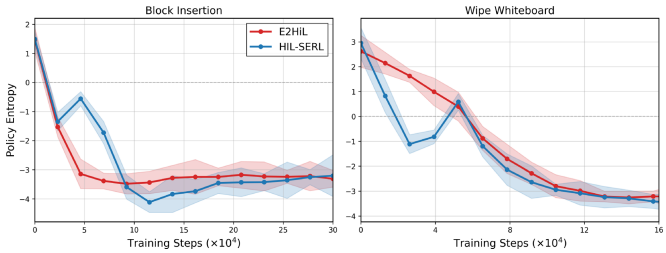


Fig. 6. **E2HiL** exhibits more stable entropy dynamics, avoiding premature entropy collapse.

success rate and a 7.5% reduction in intervention rate. These improvements are consistently observed across different learning frameworks, indicating that **E2HiL** is policy- and embodiment-agnostic, serving as a plug-and-play optimization module without architecture-specific modification.

Case Study Analysis: Fig. 5 shows a case study on the *Touch-Cube* task. During training, the cube is randomly placed at position 1 or 2 to prevent overfitting and require the agent to learn both skills, enabling evaluation of learning efficiency and generalization. Both **E2HiL** and HIL-SERL initially reduce entropy while learning the first position. However, by clipping shortcut samples that would cause abrupt entropy collapse, **E2HiL** maintains a smoother entropy decrease and acquires the first skill at around $10k$ steps. Because entropy remains sufficiently high, it continues exploring, discovers the second position at $20k$ steps, and acquires it by $33k$ steps. In contrast, the baseline converges more slowly and requires higher human intervention. A similar trend is also observed on the A1-X robot, as shown in Fig. 6.

C. Analysis of Samples Entropy Dynamics

Empirical Verification: A natural question is how accurately our method estimates the contribution of samples to policy entropy change. To investigate this, we empirically validate

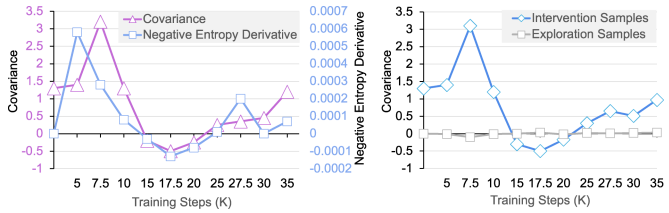


Fig. 7. *Left:* Policy entropy derivative and covariance during HiL-RL training on the *Touch-Cube* task. *Right:* Covariance of human intervention vs. self-exploration samples.

our key result that sample-induced entropy dynamics can be characterized by the covariance between action log-probabilities and soft advantages, as formulated in (9). During HIL-SERL training on the *Touch-Cube* task, we compute the batch-wise average covariance $\mathrm{Cov}(\cdot)$ at each policy update and compare it with the ground-truth entropy derivative $-\Delta\mathcal{H}$. As shown in Fig. 7 *Left*, the two curves exhibit a strong proportional relationship consistent with our theory, particularly during the initial entropy reduction phase (0–10 k steps) and the subsequent entropy increase (10 k–20 k steps). After 20 k steps, when policy entropy stabilizes, covariance values fluctuate around zero, further supporting the validity of our estimation. Deviations observed in the first 5 k steps are likely caused by inaccurate critic Q-value estimates, which bias the computation of A_{soft} in (9). As shown in Fig. 9, although early-stage Q inaccuracies may affect covariance estimation, they have negligible impact on final task success, indicating that **E2HiL** can be applied in a zero-shot manner without additional calibration.

Human Intervention Sample Analysis: Having verified the accuracy of our entropy-dynamics characterization, we next analyze how human intervention and self-exploration samples differ in their influence on policy learning in real-world HiL-RL. During HIL-SERL training on the *Touch-Cube* task, we compute the per-sample covariance c_i and tag samples originating from human interventions. As shown in Fig. 7 *Right*, human intervention samples exhibit significantly larger amplitudes and higher mean covariance values than self-exploration samples, indicating stronger influence on entropy dynamics. This observation is supported by Table III, which reports covariance magnitude distributions over all samples and the two subsets. Human intervention samples dominate the upper tail of the distribution, suggesting a decisive role in driving entropy dynamics and accelerating early policy improvement. However, both sample

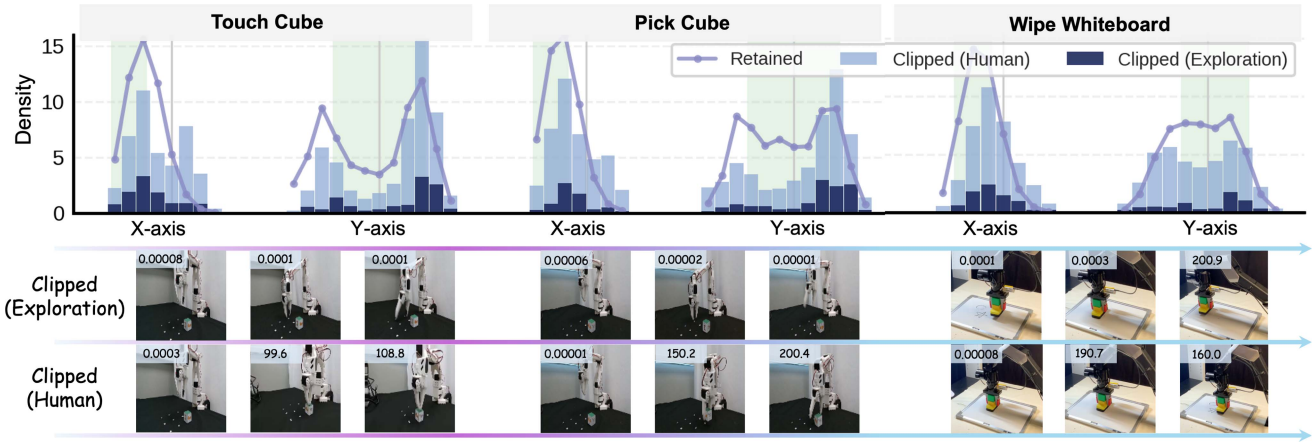


Fig. 8. Retained vs. Clipped Sample Distribution in Three Tasks. Line density of retained samples and histogram of clipped samples shown in task space (gripper position). Most clipped samples come from human interventions, and their covariance visualization highlights out-of-workspace or redundant states.

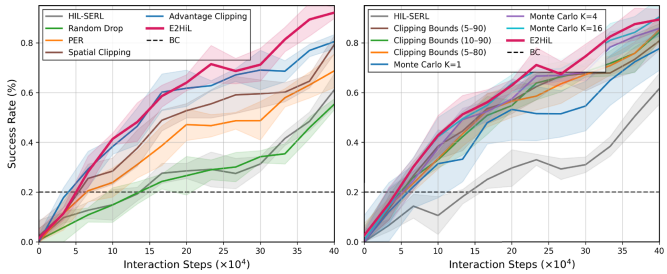


Fig. 9. Ablation and Hyperparameter Analysis. Comparison of sampling strategies and hyperparameter settings demonstrates that **E2HiL** outperforms all baselines.

TABLE III
SAMPLES COVARIANCE DISTRIBUTION. MEAN ABSOLUTE COVARIANCE ACROSS DIFFERENT SAMPLE TYPES.

Subset	All Samples	Intervention Samples	Exploration Samples
Top 2%	269.9	368.1	5.56
Top 5%	127.0	180.2	2.7
Top 10%	67.3	98.6	1.5
Top 20%	34.2	51.1	0.8
Top 50%	13.8	20.6	0.3
Low 20%	0.001	0.002	0.0007
Low 10%	0.0005	0.0007	0.0003
Low 5%	0.0002	0.0003	0.0001
Low 2%	0.00008	0.0001	0.00006
All	6.9	10.3	0.2

types also introduce instability. We identify two problematic cases:

i) shortcut samples, which induce sharp entropy drops and risk premature policy collapse, and *ii) noisy samples* with near-zero covariance that contribute little to entropy change and slow learning. These findings motivate our dynamically adjusted clipping interval $[\ell, u]$, whose bounds are adaptively set per batch as the 5th and 90th percentiles of $\{|c(s_t, a_t)|\}$. We further analyze the sensitivity to clipping ratios through a hyperparameter study, as shown in Fig. 9.

Analysis of Clipped Samples: We analyze the distribution of clipped samples to understand how entropy-guided selection shapes training. Fig. 8 shows accumulated retained (lines) and clipped (histograms) samples over time. Most clipped samples originate from human interventions. Many removed transitions lie outside the effective manipulation region, indicating limited gradient contribution, while others within the workspace correspond to redundant states with minimal entropy impact. Clipping is determined by entropy influence rather than spatial heuristics; consistent with ablation results, simple spatial filtering does not achieve comparable performance. Notably, shortcut and noisy samples exist inside the workspace and are selectively removed. Overall, entropy-bounded clipping mitigates shortcut-induced collapse and suppresses low-informative corrections, leading to stable learning.

D. Ablation Studies and Hyperparameter Analysis

Ablation Study: We conduct mechanism-isolation experiments on the *Insert Block* task to examine whether the performance gains arise from entropy reasoning rather than generic regularization. Specifically, we compare E2HiL with matched-rate Random Drop, Advantage Clipping, Spatial Clipping, and Prioritized Experience Replay (PER). As shown in Fig. 9, none of these variants achieve comparable improvements, indicating that the gains stem from entropy-informed sample selection instead of simple data reduction, gradient magnitude clipping, or heuristic entropy control.

Hyperparameter Analysis: We evaluate sensitivity to clipping bounds (5–90, 10–90, 5–80) and Monte Carlo sample counts $K \in \{1, 4, 8, 16\}$ (Fig. 9). Across reasonable ranges, performance remains stable, demonstrating that E2HiL is robust to entropy threshold and estimator configurations. Using $K = 1$ leads to larger fluctuations, while $K = 8$ (ours) and $K = 16$ achieve nearly identical results, indicating diminishing returns beyond moderate sampling. Although the influence estimator depends on critic accuracy, its effect diminishes as value estimates stabilize during training. We further validate robustness using critic calibration (Cal-QL) and ensemble critics (Table IV), where consistent improvements confirm that E2HiL does not

TABLE IV
COMPARISON WITH Q-ESTIMATION VARIANTS

Metrics	Variant		
	E2HiL	E2HiL-Cal-QL	E2HiL-Ensemble
Success Rate (%) ↑	85	**90**	85
Intervention Rate (%) ↓	**46.3**	48.1	47.8

rely on accurate Q estimates in early phases. Moreover, entropy-guided sample selection stabilizes value updates by suppressing high-variance transitions, enabling stable Q learning; in practice, policy and critic improve in a co-evolving manner rather than depending on a perfectly one at initialization.

V. CONCLUSION

In this letter, we presented **E2HiL**, an efficient real-world human-in-the-loop RL framework with entropy-guided sample selection. By modeling sample-induced entropy dynamics via covariance-based influence functions, **E2HiL** filters shortcut and noisy samples to maintain stable entropy evolution and improve sample efficiency. Extensive real-world experiments across diverse tasks demonstrate that **E2HiL** consistently achieves higher success rates with fewer human interventions than state-of-the-art baselines. Importantly, the method is policy-agnostic and embodiment-agnostic, serving as a plug-and-play optimization module across different learning frameworks and robotic platforms while preserving a favorable exploration–performance trade-off. In future work, we plan to extend **E2HiL** to scalable multi-task real-world RL for VLA models and further improve covariance estimation with more robust value-learning techniques.

REFERENCES

[1] T. Zhang et al., "Empowering embodied manipulation: A bimanual-mobile robot manipulation dataset for household tasks," 2024, *arXiv:2405.18860*.

[2] J. Luo, C. Xu, J. Wu, and S. Levine, "Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning," *Sci. Robot.*, vol. 10, pp. 1–54, 2025.

[3] H. Deng et al., "A survey on reinforcement learning of vision-language-action models for robotic manipulation," *Authorea Preprints*, 2025.

[4] G. Lu et al., "VLA-RL: Towards masterful and general robotic manipulation with scalable reinforcement learning," 2025, *arXiv:2505.18719*.

[5] C.-Y. Hung et al., "NORA-1.5: A vision-language-action model trained using world model-and action-based preference rewards," 2025, *arXiv:2511.14659*.

[6] W. Guo, G. Lu, H. Deng, Z. Wu, Y. Tang, and Z. Wang, "VLA-reasoner: Empowering vision-language-action models with reasoning via online Monte Carlo tree search," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2026, pp. 1–8.

[7] M. Yang et al., "Sim-to-real model-based and model-free deep reinforcement learning for tactile pushing," *IEEE Robot. Autom. Lett.*, vol. 8, no. 9, pp. 5480–5487, Sep. 2023.

[8] L. Chen, Y. Lei, S. Jin, Y. Zhang, and L. Zhang, "RLingua: Improving reinforcement learning sample efficiency in robotic manipulations with large language models," *IEEE Robot. Autom. Lett.*, vol. 9, no. 7, pp. 6075–6082, Jul. 2024.

[9] T. Zhang and H. Mo, "Reinforcement learning for robot research: A comprehensive review and open issues," *Int. J. Adv. Robotic Syst.*, vol. 18, 2021, Art. no. 17298814211007305.

[10] M. Torne et al., "Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation," in *Proc. Robot.: Sci. Syst.*, 2024.

[11] A. Muraleedharan et al., "SPARQ: Selective progress-aware resource querying," in *Proc. CoRL Workshop Resource-Rational Robot Learn.* 2025.

[12] G. Cui et al., "The entropy mechanism of reinforcement learning for reasoning language models," 2025, *arXiv:2505.22617*.

[13] H. Li et al., "SimpleVLA-RL: Scaling VLA training via reinforcement learning," in *Proc. Int. Conf. Learn. Representations*, 2026.

[14] H. Ju, R. Juan, R. Gomez, K. Nakamura, and G. Li, "Transferring policy of deep reinforcement learning from simulation to reality for robotics," *Nature Mach. Intell.*, vol. 4, pp. 1077–1087, 2022.

[15] C. Tang, B. Abbatematteo, J. Hu, R. Chandra, R. Martıń-Martıń, and P. Stone, "Deep reinforcement learning for robotics: A survey of real-world successes," in *Proc. AAAI Conf. Artif. Intell.*, 2025, vol. 39, no. 27, pp. 28694–28698.

[16] J. Spencer et al., "Expert intervention learning: An online framework for robot learning from explicit and implicit human feedback," *Auton. Robots*, vol. 46, no. 1, pp. 99–113, 2022.

[17] I. Clavera, J. Rothfuss, J. Schulman, Y. Fujita, T. Asfour, and P. Abbeel, "Model-based reinforcement learning via meta-policy optimization," in *Proc. Conf. Robot Learn.*, 2018, pp. 617–629.

[18] M. A. Lee et al., "Guided uncertainty-aware policy optimization: Combining learning and model-based strategies for sample-efficient policy learning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 7505–7512.

[19] A. Pinosky, I. Abraham, A. Broad, B. Argall, and T. D. Murphey, "Hybrid control for combining model-based and model-free reinforcement learning," *Int. J. Robot. Res.*, vol. 42, no. 6, pp. 337–355, 2023.

[20] K. Lei et al., "RL-100: Performant robotic manipulation with real-world reinforcement learning," 2025, *arXiv:2510.14830*.

[21] F. Zhong, K. Wu, H. Ci, C. Wang, and H. Chen, "Empowering embodied visual tracking with visual foundation models and offline RL," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 139–155.

[22] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer, "HG-DAgger: Interactive imitation learning with human experts," in *Proc. Int. Conf. Robot. Automat.*, 2019, pp. 8077–8083.

[23] K. Kawaharazuka, T. Matsushima, A. Gambardella, J. Guo, C. Paxton, and A. Zeng, "Real-world robot applications of foundation models: A review," *Adv. Robot.*, vol. 38, no. 18, pp. 1232–1254, 2024.

[24] F. Memarian, W. Goo, R. Lioutikov, S. Niekum, and U. Topcu, "Self-supervised online reward shaping in sparse-reward environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 2369–2375.

[25] Y. Cao et al., "Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 6, pp. 9737–9757, Jun. 2025.

[26] A. Gupta et al., "Reset-free reinforcement learning via multi-task learning: Learning dexterous manipulation behaviors without human intervention," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 6664–6671.

[27] Z. Yang, T. M. Moerland, M. Preuss, A. Plaat, and E. S. Hu, "World models increase autonomy in reinforcement learning," *Trans. Mach. Learn. Res.*, 2024.

[28] R. Dromnelle et al., "Reducing computational cost during robot navigation and human–robot interaction with a human-inspired reinforcement learning architecture," *Int. J. Social Robot.*, vol. 15, no. 8, pp. 1297–1323, 2023.

[29] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Mach. Learn. Res.*, 2018, vol. 80, pp. 1861–1870.

[30] A. Abdolmaleki et al., "Maximum a posteriori policy optimisation," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–20.

[31] P. J. Ball, L. Smith, I. Kostrikov, and S. Levine, "Efficient online reinforcement learning with offline data," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 1577–1594.

[32] D. Cheng et al., "Reasoning with exploration: An entropy perspective," in *Proc. AAAI Conf. Artif. Intell.*, vol. 40, no. 36, 2026, pp. 30377–30385.

[33] D. Tiapkin et al., "Fast rates for maximum entropy exploration," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 34161–34221.

[34] M. Renze, "The effect of sampling temperature on problem solving in large language models," in *Proc. Findings Assoc. Comput. Linguist.*, 2024, pp. 7346–7356.

[35] B. Eysenbach and S. Levine, "Maximum entropy RL (provably) solves some robust RL problems," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–21.

[36] Y. Chen, S. Tian, S. Liu, Y. Zhou, H. Li, and D. Zhao, "ConRFT: A reinforced fine-tuning method for VLA models via consistency policy," in *Proc. Robot.: Sci. Syst.*, 2025, pp. 1–15.

[37] M. Nakamoto et al., "Cal-QL: Calibrated offline RL pre-training for efficient online fine-tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, vol. 36, pp. 62244–62269.