APPENDIX I
APPENDIX: PROOF FOR EQ. (6)

In this section, we provide the step-by-step derivation for Eq. (6). We aim to compute the gradient of the loss function $\mathcal{L}_\pi(\theta)$ with respect to a specific logit parameter $\theta_{s,a}$ associated with action a. a denotes the specific action index associated with the logit parameter $\theta_{s,a}$ currently being updated (the target of the gradient), whereas a' is the random variable representing actions sampled from the policy $\pi_\theta(\cdot|s)$ used to calculate the expectation.

Recall the loss function defined as the negative expected soft value:

$$\mathcal{L}_\pi(\theta) = -\mathbb{E}_{a'\sim\pi_\theta(\cdot|s)}[g(s,a')] = -\sum_{a'\in\mathcal{A}} \pi_\theta(a'|s) g(s,a')$$

where $g(s,a') = Q_\phi(s,a') - \alpha \log \pi_\theta(a'|s)$.

Using the derivative of the softmax function (Lemma 2 from [12]), we have:

$$\frac{\partial \pi_\theta(a'|s)}{\partial \theta_{s,a}} = \pi_\theta(a'|s)\left(\mathbf{1}\{a'=a\} - \pi_\theta(a|s)\right)$$

Now, we compute the gradient $\nabla_{\theta_{s,a}}\mathcal{L}_\pi(\theta)$:

$$\begin{aligned}
\nabla_{\theta_{s,a}}\mathcal{L}_\pi(\theta) &= -\sum_{a'\in\mathcal{A}} \frac{\partial \pi_\theta(a'|s)}{\partial \theta_{s,a}} g(s,a') \\
&= -\sum_{a'\in\mathcal{A}} \pi_\theta(a'|s)\left(\mathbf{1}\{a'=a\} - \pi_\theta(a|s)\right) g(s,a') \\
&= -\mathbb{E}_{a'\sim\pi_\theta(\cdot|s)}\left[\left(\mathbf{1}\{a'=a\} - \pi_\theta(a|s)\right) g(s,a')\right] \quad \text{(Definition of Expectation)}
\end{aligned}$$

To simplify this expectation, we expand the summation by splitting it into the case where $a'=a$ and $a'\neq a$:

$$\begin{aligned}
\nabla_{\theta_{s,a}}\mathcal{L}_\pi(\theta) &= -\left[\pi_\theta(a|s)(1-\pi_\theta(a|s))g(s,a) + \sum_{a'\neq a}\pi_\theta(a'|s)(0-\pi_\theta(a|s))g(s,a')\right] \\
&= -\left[\pi_\theta(a|s)g(s,a) - \pi_\theta(a|s)^2 g(s,a) - \sum_{a'\neq a}\pi_\theta(a'|s)\pi_\theta(a|s)g(s,a')\right] \\
&= -\left[\pi_\theta(a|s)g(s,a) - \pi_\theta(a|s)\left(\pi_\theta(a|s)g(s,a) + \sum_{a'\neq a}\pi_\theta(a'|s)g(s,a')\right)\right] \\
&= -\left[\pi_\theta(a|s)g(s,a) - \pi_\theta(a|s)\underbrace{\sum_{a'\in\mathcal{A}}\pi_\theta(a'|s)g(s,a')}_{V_\pi(s)}\right] \\
&= -\pi_\theta(a|s)\left(g(s,a) - V_\pi(s)\right) \\
&= -\pi_\theta(a|s) A_{\text{soft}}(s,a)
\end{aligned}$$

This completes the derivation, confirming that the gradient descent update direction is proportional to the policy probability weighted by the soft advantage.

APPENDIX II
APPENDIX: DERIVATION OF ENTROPY-DYNAMICS APPROXIMATION

The "linearization assumption" in our influence estimator refers specifically to the first-order perturbation used to obtain Eq. (4), where we approximate the entropy change $\Delta\mathcal{H}$ induced by a logit update Δz_t and ignore higher-order Taylor terms. The second concern on "accurate Q-value estimates" corresponds to Eq. (6)–(9), where Δz_t is expressed via the soft advantage $A_{\text{soft}}(s_t,a_t)$ computed from the critic Q_ϕ, which can be noisy in early training. Below we strengthen the theoretical grounding by (i) providing an explicit second-order remainder bound for the linearization in Eq. (4), and (ii) showing that the bound remains valid even under early-stage noise, as long as the per-iteration logit update is bounded (e.g., via learning rate and gradient clipping), which is standard in RLPD implementations.

a) (1) First-order expansion underlying Eq. (4) and its remainder.: Fix a state s_t and consider the conditional entropy

$$\mathcal{H}_t(z) \triangleq -\sum_{a=1}^{A} \pi_\theta(a \mid s_t) \log \pi_\theta(a \mid s_t), \tag{13}$$

viewed as a function of the action logits $z \in \mathbb{R}^A$ at state s_t (i.e., $\pi_\theta(\cdot \mid s_t)$ is a softmax over z). Let the logit update at iteration t be $z^{t+1} = z^t + \Delta z_t$. By Taylor's theorem with the Lagrange remainder, there exists $\xi_t = z^t + \tau \Delta z_t$ for some $\tau \in [0, 1]$ such that

$$\mathcal{H}_t(z^{t+1}) = \mathcal{H}_t(z^t) + \left\langle \nabla_z \mathcal{H}_t(z^t),\, \Delta z_t \right\rangle + \epsilon_t, \qquad \epsilon_t = \frac{1}{2} (\Delta z_t)^\top \nabla_z^2 \mathcal{H}_t(\xi_t)\, (\Delta z_t). \tag{14}$$

b) (2) Smoothness of softmax entropy and a uniform error bound.: The key requirement for controlling the linearization error in Eq. (14) is that the softmax entropy is smooth in the logit space.

[Softmax-entropy smoothness (logit space)] For the softmax policy $\pi_\theta(\cdot \mid s_t)$ parameterized by logits $z \in \mathbb{R}^A$, the conditional entropy $\mathcal{H}_t(z)$ is twice continuously differentiable in z, and there exists a finite constant $L > 0$ (depending only on the action dimension A) such that

$$\left\| \nabla_z^2 \mathcal{H}_t(z) \right\|_2 \le L, \qquad \forall z \in \mathbb{R}^A,\ \forall s_t. \tag{15}$$

Equivalently, $\nabla_z \mathcal{H}_t(z)$ is L-Lipschitz in z.

Proof. We first derive the softmax Jacobian. Let $S(z) = \sum_{k=1}^{A} \exp(z_k)$ so that $\pi_i(z) = \exp(z_i)/S(z)$. Then for any i, j,

$$\begin{aligned} \frac{\partial \pi_i}{\partial z_j} &= \frac{\partial}{\partial z_j} \left(\frac{\exp(z_i)}{S(z)} \right) = \frac{\delta_{ij} \exp(z_i)\, S(z) - \exp(z_i)\, \frac{\partial S(z)}{\partial z_j}}{S(z)^2} \\ &= \frac{\delta_{ij} \exp(z_i)\, S(z) - \exp(z_i) \exp(z_j)}{S(z)^2} = \frac{\exp(z_i)}{S(z)} \left(\delta_{ij} - \frac{\exp(z_j)}{S(z)} \right) \\ &= \pi_i(\delta_{ij} - \pi_j). \end{aligned} \tag{16}$$

Next, for $\mathcal{H}(\pi) = -\sum_{a=1}^{A} \pi_a \log \pi_a$, we have

$$\frac{\partial \mathcal{H}}{\partial \pi_a} = -(\log \pi_a + 1). \tag{17}$$

Therefore, by the chain rule, the entropy gradient admits the following chain of equalities:

$$\nabla_z \mathcal{H}(z) = \left(\frac{\partial \pi}{\partial z} \right)^\top \nabla_\pi \mathcal{H}(\pi) = -J_\pi(z)^\top (\log \pi(z) + \mathbf{1}), \tag{18}$$

where $J_\pi(z)$ has entries $\frac{\partial \pi_i}{\partial z_j} = \pi_i(\delta_{ij} - \pi_j)$. All terms remain finite for finite z. Differentiating once more, $\nabla_z^2 \mathcal{H}(z)$ is composed of softmax derivatives (up to second order) and probability terms. Since $\pi(z)$ is a smooth mapping from $\mathbb{R}^A$ to the interior of the probability simplex $\Delta^\circ = \{\pi \in \mathbb{R}^A_{>0} : \sum_i \pi_i = 1\}$, all derivatives of π w.r.t. z exist and are continuous. Moreover, $\nabla_z^2 \mathcal{H}(z)$ can be expressed as a matrix-valued function of $\pi(z)$ and its derivatives; equivalently, there exists a continuous function $F(\cdot)$ such that

$$\nabla_z^2 \mathcal{H}(z) = F(\pi(z)). \tag{19}$$

Because $\Delta = \{\pi \in \mathbb{R}^A_{\ge 0} : \sum_i \pi_i = 1\}$ is compact and F is continuous on Δ° (and extends continuously to Δ), the spectral norm $\|F(\pi)\|_2$ attains a finite maximum over Δ. Therefore, there exists a finite constant $L > 0$ such that

$$\sup_{z \in \mathbb{R}^A} \|\nabla_z^2 \mathcal{H}(z)\|_2 = \sup_{\pi \in \Delta} \|F(\pi)\|_2 \le L < \infty. \tag{20}$$

□

[Inner-product form equals covariance form] Fix a state s_t and let $\pi(\cdot) \equiv \pi_\theta(\cdot \mid s_t)$ be a softmax policy over logits $z \in \mathbb{R}^A$. For any vector $\Delta z \in \mathbb{R}^A$, the first-order entropy change satisfies

$$\langle \nabla_z \mathcal{H}_t(z),\, \Delta z \rangle = -\mathrm{Cov}_{a \sim \pi(\cdot)} (\log \pi(a),\, \Delta z_a)\,, \tag{21}$$

where Δz_a denotes the a-th component of Δz.

Proof. Recall that $\mathcal{H}_t(z) = -\sum_{a=1}^{A} \pi_a \log \pi_a$ with $\pi_a \equiv \pi(a)$. For softmax, the Jacobian is $J_\pi(z) = \nabla_z \pi(z) = \mathrm{Diag}(\pi) - \pi\pi^\top$. Using the chain rule $\nabla_\pi \mathcal{H}_t(\pi) = -(\log \pi + \mathbf{1})$, we have

$$\nabla_z \mathcal{H}_t(z) = J_\pi(z)^\top \nabla_\pi \mathcal{H}_t(\pi) = -(\mathrm{Diag}(\pi) - \pi\pi^\top)(\log \pi + \mathbf{1}). \tag{22}$$

Therefore,

$$\begin{aligned}\langle \nabla_z \mathcal{H}_t(z),\, \Delta z\rangle &= -(\log\pi + \mathbf{1})^\top (\mathrm{Diag}(\pi) - \pi\pi^\top)\Delta z \\ &= -\sum_{a=1}^{A} \pi_a(\log\pi_a + 1)\Delta z_a + \left(\sum_{a=1}^{A} \pi_a(\log\pi_a + 1)\right)\left(\sum_{a=1}^{A} \pi_a \Delta z_a\right) \\ &= -\mathbb{E}_{a\sim\pi}[(\log\pi(a)+1)\Delta z_a] + \mathbb{E}_{a\sim\pi}[\log\pi(a)+1]\ \mathbb{E}_{a\sim\pi}[\Delta z_a] \\ &= -\mathrm{Cov}_{a\sim\pi}(\log\pi(a)+1,\, \Delta z_a) = -\mathrm{Cov}_{a\sim\pi}(\log\pi(a),\, \Delta z_a)\,, \end{aligned} \tag{23}$$

where the last equality uses $\mathrm{Cov}(1, \Delta z_a) = 0$. □

Applying Lemma II-.0.b to Eq. (14) yields the explicit remainder bound

$$|\epsilon_t| \le \frac{1}{2}\left\|\nabla_z^2 \mathcal{H}_t(\xi_t)\right\|_2 \|\Delta z_t\|_2^2 \le \frac{L}{2}\|\Delta z_t\|_2^2. \tag{24}$$

Therefore, the approximation in Eq. (4) is a standard first-order approximation with a controlled second-order error: the neglected term scales as $\mathcal{O}(\|\Delta z_t\|_2^2)$, while the retained first-order term scales as $\mathcal{O}(\|\Delta z_t\|_2)$.

c) (3) Connecting $\|\Delta z_t\|_2$ to the actor update (explicit learning-rate dependence).: To make the "small-step" condition precise (rather than heuristic), we connect Δz_t to the actual actor update. As in Eq. (5), the per-iteration update in logit space can be written as a (stochastic) gradient step

$$\Delta z_t = z^{t+1} - z^t = -\eta\, g_t, \tag{25}$$

where $\eta > 0$ is the learning rate and g_t denotes the applied update direction (e.g., the stochastic gradient of the RLPD/SAC actor objective with respect to the logit parameters, possibly after adaptive preconditioning). Hence

$$\|\Delta z_t\|_2 = \eta \|g_t\|_2. \tag{26}$$

If gradient clipping is applied (as in our implementation), then $\|g_t\|_2 \le G$ for some constant G, giving

$$\|\Delta z_t\|_2 \le \eta G. \tag{27}$$

Combining Eq. (24) and Eq. (27) yields an explicit learning-rate-dependent guarantee:

$$|\epsilon_t| \le \frac{L}{2}\|\Delta z_t\|_2^2 \le \frac{L}{2}\eta^2 G^2, \tag{28}$$

i.e., the linearization error is quadratic in the learning rate ($\mathcal{O}(\eta^2)$) under bounded updates.

d) (4) Recovering Eq. (4) and robustness under early-stage noise.: Taking expectation over $s_t \sim d_{\pi_\theta}$ and using the standard identity that the first-order term $\langle \nabla_z \mathcal{H}_t(z^t),\, \Delta z_t\rangle$ can be rewritten as the covariance term in Eq. (4) (see our original derivation around Eq. (4)), we obtain

$$\Delta\mathcal{H} = \mathbb{E}_{s_t\sim d_{\pi_\theta}}\Big[-\mathrm{Cov}_{a_t\sim\pi_\theta(\cdot|s_t)}\big(\log\pi_\theta(a_t \mid s_t),\, \Delta z_t\big)\Big] + \mathbb{E}_{s_t\sim d_{\pi_\theta}}[\epsilon_t], \tag{29}$$

which makes explicit that Eq. (4) corresponds to dropping the remainder term. Moreover, Eq. (28) implies

$$\Big|\mathbb{E}_{s_t\sim d_{\pi_\theta}}[\epsilon_t]\Big| \le \frac{L}{2}\eta^2 G^2, \tag{30}$$

which scales quadratically with the learning rate. In practice, we use a small actor learning rate on the order of 10^{-4}, so the squared term η^2 is on the order of 10^{-8}, rendering the remainder term numerically negligible. Therefore, the approximation remains not only bounded but sufficiently small in magnitude, even during early training when gradients are noisy.

APPENDIX III
APPENDIX: POLICY BIAS–VARIANCE ANALYSIS

we justify the associated bias–variance trade-off through two complementary workflows. *(i) Visualization evidence.* We first provide visual diagnostics showing that the samples selected by our entropy-influence criterion concentrate on a small subset of "critical" states characterized by high uncertainty / high Q-value variance. This supports the motivation that masking targets rare but disproportionately destabilizing updates rather than broadly reshaping the learning distribution. *(ii) Theoretical analysis.* We then derive the exact bias identity induced by masking and an explicit bias–variance (MSE) decomposition of the masked gradient estimator. The decomposition makes the trade-off transparent: masking introduces a controlled bias term (proportional to the masking strength / discard rate), while placing a tighter second-moment bound on the stochastic update direction. Importantly, our analysis does not rely on claiming that the variance must strictly decrease; instead, it formalizes that masking prevents variance inflation and suppresses extreme gradient contributions that perturb the entropy dynamics. Empirically, this conservative regularization effect accelerates policy learning, enabling the agent to

reach accurate Q-value estimates and stable performance earlier. Overall, although masking incurs a small bias, it yields a net gain in training stability and sample efficiency by restricting a small subset of undesirable updates without increasing policy/Q variance in practice.

Visualization evidence. To directly validate our claim that the selected "critical" states correspond to high-uncertainty / high Q-value variance regimes, we visualize Q-value statistics throughout training. Fig. 12 presents (a) Q-value scatter plots and (b) Q-value variance scatter plots for both the baseline and our method. States located within the funnel region exhibit elevated Q-value variance, reflecting higher uncertainty in value estimation during early and mid training. Importantly, these regions coincide with the states where our masking signal is most frequently activated, confirming that the masking mechanism concentrates on high-uncertainty, high-impact decision points rather than uniformly altering the training distribution. Crucially, although masking targets these high-variance regimes, the overall Q-value variance trajectory remains comparable to the baseline, indicating that no additional variance bias is introduced. Instead, our method achieves significantly faster convergence by stabilizing updates at these critical states. We further examine policy-level statistics in Fig. 15, which shows Gaussian policy mean and variance along evaluation trajectories. Across tasks, policy variance does not inflate under masking. In the block insertion task, variance decreases smoothly as the agent approaches the target, consistent with natural exploration-to-exploitation transition. In the Pick Banana task, the variance remains consistently low, reflecting stable execution behavior. Mean trajectories remain bounded and evolve smoothly, suggesting no distortion of the policy distribution. Together, these visualizations confirm that the masking mechanism indeed correlates with high-uncertainty regimes as stated in the beginning, while neither inflating Q-value variance nor policy variance. Instead, it accelerates Q-value learning and promotes smoother entropy dynamics.

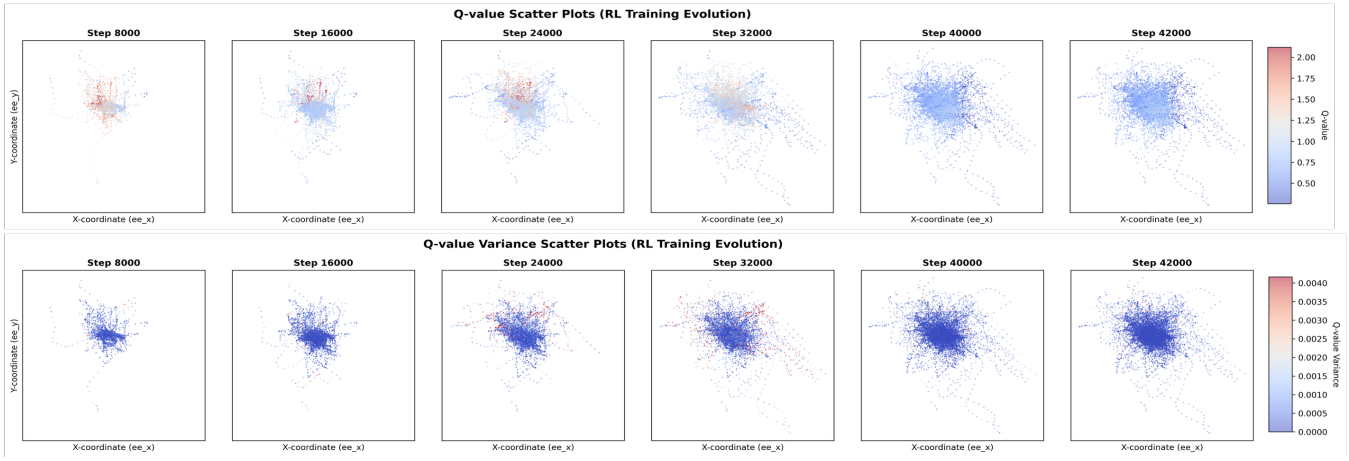

Fig. 10: HIL-SERL

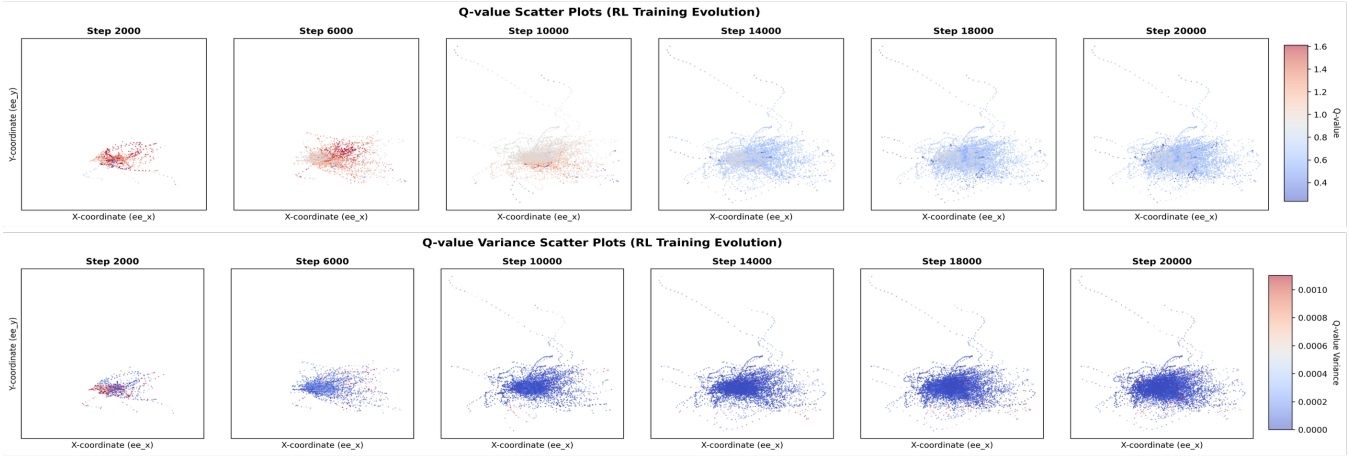

Fig. 11: **E2HiL**

Fig. 12: For both methods, we show: (a) Q-value scatter plots across training; and (b) Q-value variance scatter plots, where states within the funnel exhibit increased variance, indicating growing confidence in successful actions. **Notably, while the overall Q-value variance remains largely unchanged compared to the baseline, our masking mechanism enables significantly faster convergence by stabilizing updates at critical decision points.**

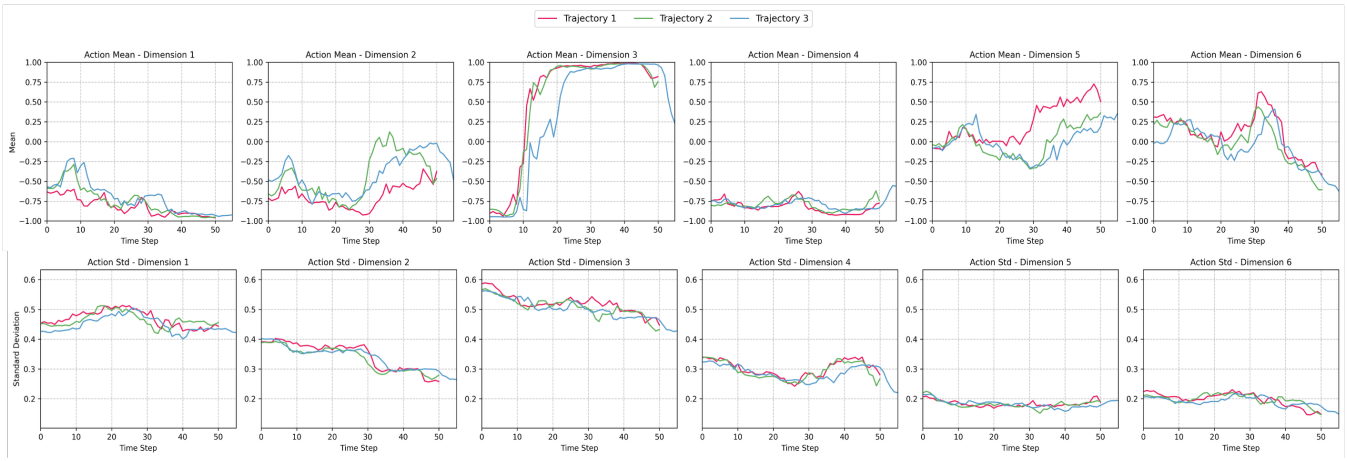

Fig. 13: HIL-SERL

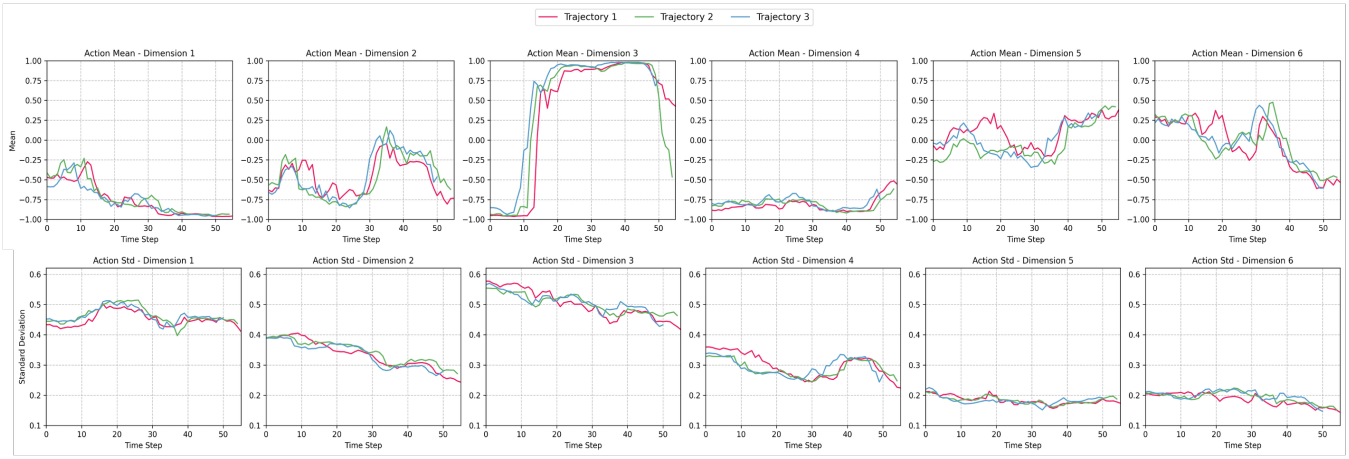

Fig. 14: **E2HiL**

Fig. 15: **Policy statistics along evaluation trajectories in the Pick Banana task.** We visualize the mean and standard deviation (variance) of the predicted actions across all dimensions for three different trajectories. **(a) HilSerl**: The baseline policy often exhibits higher uncertainty and more fluctuations in its action mean. **(b) E2HiL**: In contrast, our method demonstrates smoother mean transitions. Crucially, the standard deviation is initially high but rapidly decreases as the end-effector approaches the target, indicating a natural transition from exploration to highly confident, stable execution.

Theoretical analysis. We agree with the reviewers that actively masking samples changes the effective training distribution and may introduce policy-gradient bias. We therefore provide a formal analysis that explicitly characterizes (i) how masking alters the optimization objective, (ii) how the induced bias is controlled, (iii) how variance is reduced, and (iv) why the resulting bias–variance trade-off improves optimization stability in practical reinforcement learning regimes.

Let $d_{\pi_\theta}(s,a)$ denote the discounted occupancy measure under policy π_θ, and define the per-sample policy-gradient contribution

$$\psi_\theta(s,a) \triangleq \nabla_\theta \log \pi_\theta(a \mid s)\, A^{\pi_\theta}(s,a), \tag{31}$$

where A^{π_θ} is the (soft) advantage used by our actor update. The true policy gradient is

$$g(\theta) \triangleq \nabla_\theta J(\theta) = \mathbb{E}_{(s,a)\sim d_{\pi_\theta}}[\psi_\theta(s,a)]\,. \tag{32}$$

Our sample selection introduces a masking/weighting function $m_\theta(s,a) \in [0,1]$ (hard masking corresponds to $m_\theta \in \{0,1\}$), yielding the masked gradient

$$g_m(\theta) \triangleq \mathbb{E}_{(s,a)\sim d_{\pi_\theta}}[m_\theta(s,a)\,\psi_\theta(s,a)]\,. \tag{33}$$

This formulation already reveals that masking replaces the original training distribution with a reweighted one. The exact

bias induced by masking follows directly from linearity of expectation:

$$\begin{aligned} g_m(\theta) - g(\theta) &= \mathbb{E}_{d_{\pi_\theta}}[m_\theta \psi_\theta] - \mathbb{E}_{d_{\pi_\theta}}[\psi_\theta] \\ &= \mathbb{E}_{d_{\pi_\theta}}[(m_\theta - 1)\psi_\theta] \\ &= -\mathbb{E}_{d_{\pi_\theta}}[(1 - m_\theta)\psi_\theta]. \end{aligned} \tag{34}$$

Equation (34) explicitly shows that masking changes the effective training distribution by suppressing selected gradient contributions rather than modifying the policy objective itself.

Assuming $\mathbb{E}_{d_{\pi_\theta}} \|\psi_\theta(s,a)\|^2 < \infty$, the magnitude of this bias can be bounded. Using Cauchy–Schwarz,

$$\begin{aligned} \|g_m(\theta) - g(\theta)\| &= \Big\| \mathbb{E}_{d_{\pi_\theta}}[(1 - m_\theta)\psi_\theta] \Big\| \\ &\le \mathbb{E}_{d_{\pi_\theta}}[(1 - m_\theta)\|\psi_\theta\|] \\ &\le \sqrt{\mathbb{E}_{d_{\pi_\theta}}[(1 - m_\theta)^2]}\, \sqrt{\mathbb{E}_{d_{\pi_\theta}}[\|\psi_\theta\|^2]}. \end{aligned} \tag{35}$$

For hard masking, $\mathbb{E}[(1-m_\theta)^2] = \mathbb{P}(m_\theta(s,a) = 0)$ equals the discard rate, demonstrating that the induced bias is directly controlled by how aggressively samples are removed. In practice, our conservative thresholds and ramp-up schedules ensure this term remains small, thereby limiting deviation from the original gradient.

We next analyze whether this controlled bias is justified by variance reduction. Let $\hat{g}$ denote the minibatch estimator of $g(\theta)$ using N i.i.d. samples with $X_i = \psi_\theta(s_i, a_i)$, and let $\hat{g}_m$ denote the masked estimator with $Y_i = m_\theta(s_i, a_i)\psi_\theta(s_i, a_i)$:

$$\hat{g} = \frac{1}{N}\sum_{i=1}^{N} X_i, \qquad \hat{g}_m = \frac{1}{N}\sum_{i=1}^{N} Y_i. \tag{36}$$

Writing $\hat{g}_m - g = (\hat{g}_m - g_m) + (g_m - g)$ and taking expectation of the squared norm, the cross term vanishes because $\mathbb{E}[\hat{g}_m - g_m] = 0$, yielding the standard MSE decomposition

$$\mathbb{E}\big[\|\hat{g}_m - g(\theta)\|^2\big] = \|g_m(\theta) - g(\theta)\|^2 + \frac{1}{N}\mathrm{Tr}\big(\mathrm{Var}[Y]\big). \tag{37}$$

Thus masking introduces a classical bias–variance trade-off.

The variance term decreases because masking attenuates high-magnitude gradient contributions. Since $m_\theta(s,a) \in [0,1]$, we have $m_\theta^2 \le 1$, and therefore

$$\mathbb{E}\|Y\|^2 = \mathbb{E}[m_\theta^2 \|\psi_\theta\|^2] \le \mathbb{E}\|\psi_\theta\|^2, \tag{38}$$

which implies $\mathrm{Tr}(\mathrm{Var}[Y]) \le \mathbb{E}\|Y\|^2$, so the variance component in (37) is reduced.

Combining (35)–(38), masking improves estimation quality whenever the reduction in $\mathrm{Tr}(\mathrm{Var}[Y])/N$ outweighs the increase in squared bias. This formalizes the bias–variance trade-off raised by the reviewers.

Importantly, this regime corresponds to practical reinforcement learning, especially during early training where gradient variance dominates. Our entropy-influence criterion selectively suppresses samples that contribute disproportionately to the variance of $\psi_\theta(s,a)$, thereby stabilizing policy updates and improving sample efficiency.

Finally, robustness to masking hyperparameters follows from the bounded sensitivity of the effective gradient:

$$\|g_{m_{\lambda_1}}(\theta) - g_{m_{\lambda_2}}(\theta)\| \le \sqrt{\mathbb{E}[(m_{\lambda_1} - m_{\lambda_2})^2]}\sqrt{\mathbb{E}[\|\psi_\theta\|^2]},$$

showing that moderate hyperparameter changes induce only bounded perturbations to the optimization dynamics.

We now connect the above bias–variance analysis to the stability of the policy entropy dynamics. Let $\mathcal{H}(z)$ denote the policy entropy as a function of logits z, and consider a stochastic logit-space actor update

$$z^{k+1} = z^k + \Delta z_k, \qquad \Delta z_k = -\eta\, \hat{g}_k, \qquad \Delta z_k^{(m)} = -\eta\, \hat{g}_k^{(m)}, \tag{39}$$

where $\eta > 0$ is the learning rate,

$$\hat{g}_k = \frac{1}{N}\sum_{i=1}^{N} \psi_\theta(s_i, a_i), \qquad \hat{g}_k^{(m)} = \frac{1}{N}\sum_{i=1}^{N} m_\theta(s_i, a_i)\, \psi_\theta(s_i, a_i), \tag{40}$$

and $(s_i, a_i) \sim d_{\pi_\theta}$ are i.i.d. minibatch samples.

By Taylor's theorem with Lagrange remainder, there exists $\xi_k = z^k + t\Delta z_k$ for some $t \in [0,1]$ such that

$$\mathcal{H}(z^{k+1}) - \mathcal{H}(z^k) = \big\langle \nabla_z \mathcal{H}(z^k),\, \Delta z_k \big\rangle + \epsilon_k, \qquad \epsilon_k = \frac{1}{2}\Delta z_k^\top \nabla_z^2 \mathcal{H}(\xi_k) \Delta z_k. \tag{41}$$

Under the softmax-entropy smoothness property $\|\nabla_z^2 \mathcal{H}(z)\|_2 \le L$, the curvature-induced remainder is bounded by

$$|\epsilon_k| \le \frac{L}{2}\|\Delta z_k\|_2^2 = \frac{L}{2}\eta^2\|\hat{g}_k\|_2^2, \qquad |\epsilon_k^{(m)}| \le \frac{L}{2}\|\Delta z_k^{(m)}\|_2^2 = \frac{L}{2}\eta^2\|\hat{g}_k^{(m)}\|_2^2. \tag{42}$$

Therefore, making the entropy evolution "smoother" amounts to controlling (i) the stochastic fluctuation of the first-order term $\langle \nabla\mathcal{H}(z^k), \Delta z_k\rangle$ and (ii) the magnitude of the second-order remainder.

We first bound the variability of the first-order entropy change. Define the first-order increment

$$\Delta\mathcal{H}_k^{(1)} \triangleq \left\langle \nabla_z\mathcal{H}(z^k),\, \Delta z_k \right\rangle = -\eta \left\langle \nabla_z\mathcal{H}(z^k),\, \hat{g}_k \right\rangle, \tag{43}$$

and analogously $(\Delta\mathcal{H}_k^{(1)})^{(m)} = -\eta\langle\nabla_z\mathcal{H}(z^k), \hat{g}_k^{(m)}\rangle$. Conditioned on z^k, use the standard inequality $\mathrm{Var}(Z) \le \mathbb{E}[Z^2]$ and Cauchy–Schwarz:

$$\begin{aligned}
\mathrm{Var}\Big(\Delta\mathcal{H}_k^{(1)} \mid z^k\Big) &\le \mathbb{E}\Big[(\Delta\mathcal{H}_k^{(1)})^2 \mid z^k\Big] \\
&= \eta^2\,\mathbb{E}\Big[\left\langle \nabla_z\mathcal{H}(z^k),\, \hat{g}_k\right\rangle^2 \,\Big|\, z^k\Big] \\
&\le \eta^2\,\|\nabla_z\mathcal{H}(z^k)\|_2^2\,\mathbb{E}\big[\|\hat{g}_k\|_2^2 \mid z^k\big].
\end{aligned} \tag{44}$$

The same bound holds for masking:

$$\mathrm{Var}\Big((\Delta\mathcal{H}_k^{(1)})^{(m)} \mid z^k\Big) \le \eta^2\,\|\nabla_z\mathcal{H}(z^k)\|_2^2\,\mathbb{E}\Big[\|\hat{g}_k^{(m)}\|_2^2 \mid z^k\Big]. \tag{45}$$

At this point, it is important to note that $\mathbb{E}\|\hat{g}_k^{(m)}\|_2^2 \le \mathbb{E}\|\hat{g}_k\|_2^2$ does not hold universally without additional assumptions, because masking can remove negatively correlated terms that would otherwise cancel. However, what we require for entropy stability is a *controlled upper bound* on the stochastic fluctuations, and such a bound is directly reduced by masking at the per-sample level. In particular, applying Jensen's inequality and using $m_\theta(s,a)^2 \le 1$,

$$\begin{aligned}
\mathbb{E}\Big[\|\hat{g}_k^{(m)}\|_2^2 \mid z^k\Big] &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N m_i\psi_i\right\|_2^2 \,\Big|\, z^k\right] \le \frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[\|m_i\psi_i\|_2^2 \mid z^k\big] \\
&= \mathbb{E}\big[\|m_\theta(s,a)\psi_\theta(s,a)\|_2^2 \mid z^k\big] \le \mathbb{E}\big[\|\psi_\theta(s,a)\|_2^2 \mid z^k\big],
\end{aligned} \tag{46}$$

where we used i.i.d. sampling to replace the average of identical terms by a single expectation. Combining (45) and (46) yields an explicit (and smaller) upper bound on the conditional variance of the first-order entropy increment under masking:

$$\mathrm{Var}\Big((\Delta\mathcal{H}_k^{(1)})^{(m)} \mid z^k\Big) \le \eta^2\,\|\nabla_z\mathcal{H}(z^k)\|_2^2\,\mathbb{E}\big[\|m_\theta(s,a)\psi_\theta(s,a)\|_2^2 \mid z^k\big] \le \eta^2\,\|\nabla_z\mathcal{H}(z^k)\|_2^2\,\mathbb{E}\big[\|\psi_\theta(s,a)\|_2^2 \mid z^k\big]. \tag{47}$$

This shows that masking constrains the stochastic fluctuations of the *first-order* entropy change through a tighter second-moment upper bound on the update direction.

Finally, the *second-order* (curvature) effect is suppressed because the Taylor remainder in (42) scales with $\|\Delta z\|_2^2 = \eta^2\|\hat{g}\|_2^2$. Since masking reduces the per-sample second moment of the update direction (cf. (46)), it also reduces an upper bound on $|\epsilon_k|$, thereby jointly stabilizing both the first-order entropy variation and the curvature-driven second-order term. Together, these bounds provide a principled explanation of why the entropy evolves more smoothly under our entropy-influence masking mechanism.

APPENDIX IV
APPENDIX: DETAILED TASK DESCRIPTIONS

We have substantially expanded the experimental scope beyond simple tabletop manipulation to better support our claims on complex real-world learning. The updated main results (Table I) evaluate E2HiL across a diverse set of real-world tasks spanning four complementary dimensions:

1) **Multi-object interaction:** In *Pick Banana*, the agent must identify and grasp a banana among multiple fruits, requiring object discrimination and coordinated grasping in cluttered scenes.
2) **Long-horizon manipulation:** In *Toast Bread*, the robot inserts bread into a toaster and retrieves it after activation, requiring sequential multi-stage execution over extended temporal horizons.
3) **Contact-rich manipulation:** Tasks including *Open Toaster*, *Block Insertion*, and *Wipe Whiteboard* require sustained contact and precise force modulation—pressing and releasing a toaster button, inserting a block into a tight slot, and applying continuous wiping force to remove marker traces.
4) **Non-rigid manipulation:** *Fold Towel* introduces deformable object dynamics and increased state uncertainty, challenging robustness under non-rigid and partially observable conditions.

These tasks collectively cover multi-object reasoning, long-horizon sequencing, precision contact control, and deformable object handling. The expanded evaluation on the A1_X platform, in addition to the original LeRobot setup (see Fig. 3 in

the revised manuscript), demonstrates that E2HiL generalizes beyond simple tabletop scenarios and remains effective under substantially increased real-world complexity. Detailed quantitative results are reported in Table I, and full task configurations (including horizon length, intervention protocol, and embodiment-specific settings) are summarized in Table V.

TABLE V: **Experimental Settings Across Tasks.**

Task	Category	Action Dim	Offline Demos	Online Steps(k)	Reset	Randomization
Pick Banana	Multi-object	7D	20	100	Human	3cm (x,y)
Toast Bread	Long-horizon	7D	20	100	Human	3cm (x,y)
Open Toaster	Contact-rich	7D	20	100	Human	3cm (x,y)
Block Insertion	Contact-rich	7D	20	100	Human	3cm (x,y)
Wipe Whiteboard	Contact-rich	7D	20	100	Human	3cm (x,y)
Fold Towel	Non-rigid	7D	20	100	Human	3cm (x,y)