PRIM: Past-Action Regularization for Imitation with Flow Matching

Anonymous authors
Paper under double-blind review

Abstract

Behavior cloning commonly uses Flow Matching to learn distributions over future actions from expert demonstrations. Expert behavior can depend on temporal information absent from the current observation, making past actions a natural input for recovering this information and improving action prediction. However, strong temporal correlations in expert actions can encourage high sensitivity to action history, allowing model-generated errors to be amplified when executed actions enter subsequent policy inputs. We introduce Past-Action Regularization for Imitation with Flow Matching (PRIM), which penalizes the velocity field's Jacobian with respect to action history to reduce the sensitivity of generated actions to history errors. Under stated distributional and regularity assumptions, we show that squared temporal action correlations lower-bound the history sensitivity required by a conditional-mean policy. We also derive the recursive propagation of action errors through environment dynamics and history inputs. Under sufficient conditions, controlling this sensitivity keeps accumulated errors bounded. We evaluate PRIM on LIBERO-90 and find that directly conditioning on past actions can lower success rates, while Jacobian regularization suppresses error accumulation and improves success.

1 Introduction

Behavior cloning enables robots to learn from expert demonstrations by predicting actions from visual observations and robot states. Diffusion and Flow Matching provide expressive frameworks for learning distributions of future action chunks conditioned on these inputs (Chi et al., 2023; Lipman et al., 2023). However, expert decisions may also depend on recent control history. When this temporal information is absent from the policy input, predicting expert actions from the current observation alone imposes a Markov assumption that the demonstrations may not satisfy.

Delayed action prediction provides evidence of this missing temporal information. Prior work reports that predicting the current action from an earlier observation can yield lower error than using the current observation (Lazzati et al., 2026). In our notation, the delayed input contains both the visual observation o_{t-d} and robot state s_{t-d}. This observation suggests that past inputs retain information relevant to the current action that is not fully captured by (o_t, s_t). We design controlled experiments to examine this temporal dependence and find that supplying past actions removes the prediction advantage of delayed observations. These results support a non-Markovian interpretation of expert behavior and motivate conditioning on past actions to improve action prediction.

However, lower action-prediction error need not improve success. Our controlled comparisons reveal this gap, consistent with failures associated with copycat behavior, past-action information leakage, causal confusion, and compounding imitation errors (Wen et al., 2020; Seo et al., 2023; de Haan et al., 2019; Ross et al., 2011). We attribute part of this gap to the repeated reuse of imperfect model outputs. Each action error changes the next state through the environment and enters the next input as action history. These changes affect later predictions, whose errors are again carried forward. History sensitivity controls amplification along the direct writeback path.

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.

Temporal correlations make action history predictive, encouraging policies to rely on it when minimizing prediction error. We formalize this relationship by showing that, under conditional stationarity and regularity assumptions, squared temporal correlations lower-bound the history Jacobian of the conditional-mean policy (Theorem 1). With action variance and the Poincaré constant held fixed, stronger correlations raise this bound. This sensitivity can then amplify errors written back into later inputs.

To use predictive history while limiting error accumulation, we introduce Past-Action Regularization for Imitation with Flow Matching (PRIM). PRIM penalizes the velocity field's Jacobian with respect to action history, controlling generated-action sensitivity through the sampling dynamics. We derive the recursive paths through which action errors affect later states and actions, and establish sufficient conditions under which this regularization keeps accumulated errors bounded during repeated execution.

Our contributions are summarized as follows:

1. **Diagnosing non-Markovian expert demonstrations.** We provide controlled experimental evidence that expert demonstrations exhibit non-Markovian temporal dependence relative to the current observation and robot state. Past actions recover temporal information and reduce action-prediction error, but can also lower success rates.
2. **Regularizing action-history sensitivity with theoretical guarantees.** We relate squared temporal action correlations to a lower bound on history sensitivity under stated assumptions. We introduce PRIM and prove sufficient conditions under which its velocity-field regularization controls generated-action sensitivity and keeps accumulated state and history errors bounded during execution.
3. **Validating prediction benefits and error control.** We conduct matched comparisons on LIBERO-90 and four RoboMimic tasks and find that past actions can reduce prediction error while lowering success. PRIM improves success and surpasses the current-state baseline on LIBERO-90. We further use history replacement and perturbation experiments to confirm continued dependence on past actions, reduced error accumulation, and higher success under shared perturbations.

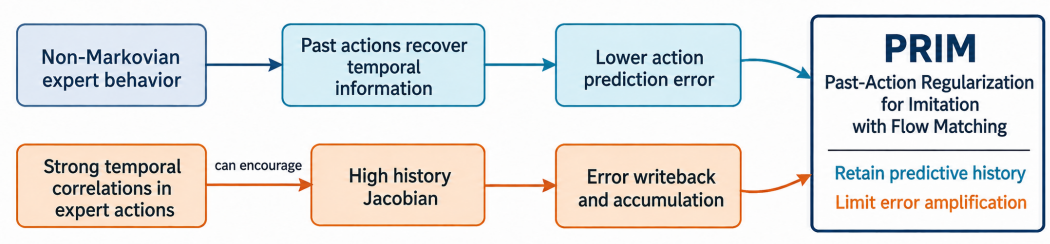


Figure 1: Overview of PRIM.

2 Related Work

Non-Markovian behavior cloning. Observation delay reveals temporal information that may be absent from the current policy input, while action chunking changes both the prediction horizon and the replanning frequency (Zhao et al., 2023; Lazzati et al., 2026; Zhang et al., 2025). Past actions supply recent control context, but they can also induce copycat behavior and causal confusion when a policy learns dependencies that change during deployment (Wen et al., 2020; de Haan et al., 2019; Seo et al., 2023). Prior methods remove past-action information from a learned representation through target-conditioned adversarial training or conditional-independence penalties. PRIM retains past actions as policy inputs and controls the derivative through which their execution errors alter later actions.

Closed-loop error and policy sensitivity. Covariate shift causes individual imitation errors to alter subsequently visited states (Ross et al., 2011). Prior analyses control local

state deviations through policy derivatives or combine policy Jacobians with environment dynamics in an augmented closed-loop system (Pfrommer et al., 2022; Mehta et al., 2025; Block et al., 2023). Related generative-policy methods constrain derivatives with respect to sampled actions, intermediate velocity fields, or external conditions (Abyaneh et al., 2026; Jiang et al., 2025; Dolatpour et al., 2026). PRIM regularizes sensitivity to executed action history and connects it to error accumulation across replanning steps.

3 Preliminaries

We introduce the policy inputs, action execution, and Flow Matching objective used throughout the paper.

3.1 Behavior Cloning with Action History

Behavior cloning learns a policy from expert demonstrations, which record visual observations o_t, robot states s_t, and actions $a_t \in \mathbb{R}^{d_a}$. Let S contain the current observation, robot state, and task condition, let $H = a_{t-K:t-1}$ contain K past actions, and let $A = a_{t:t+N-1}$ denote the future action chunk. A conditional policy $\pi_\theta(A \mid S, H)$ models the expert action distribution given these inputs; setting $K = 0$ removes action history. During execution, the policy predicts a chunk, executes its first $P \leq N$ actions, and replans from the updated input. Consequently, training histories contain expert actions, whereas execution histories contain the model's previously executed outputs.

We write the environment transition as $s_{t+1} = f(s_t, a_t)$, with any additional simulator state implicit. When discussing prediction from observations and states, we omit the fixed task instruction. Let $d \geq 0$ denote the observation delay. Under squared error, the optimal predictor is the conditional mean of the target action given its inputs. The regression experiments below estimate these conditional means on demonstration data.

3.2 Flow Matching for Action Prediction

Flow Matching learns to generate an action chunk by integrating a velocity field from an initial noise sample (Lipman et al., 2023). During training, we independently draw Gaussian noise $\xi \sim \mathcal{N}(0, I)$ and time $\tau \sim \mathcal{U}[0, 1]$ for each demonstration. We construct the interpolation $X_\tau = (1 - \tau)\xi + \tau A$ and train the velocity field to predict $V = A - \xi$ by minimizing

$$\mathcal{L}_{\mathrm{FM}}(v) = \mathbb{E}\left[\frac{1}{Nd_a}\|v(X_\tau, \tau, S, H) - (A - \xi)\|_2^2\right]. \tag{1}$$

We train v_θ by minimizing Eq. (1). At inference, we draw $X_0 \sim \mathcal{N}(0, I)$ and solve the ODE $\mathrm{d}X_\tau/\mathrm{d}\tau = v_\theta(X_\tau, \tau, S, H)$ from $\tau = 0$ to 1 to generate the action chunk X_1. We then execute its first P actions before replanning. Here, τ denotes Flow Matching time and t denotes environment time.

4 Evidence of Non-Markovian Expert Data and Temporal Information Recovery through Action History

We conduct four controlled experiments on LIBERO-90 (Liu et al., 2023) to reproduce the U-shaped delay curve reported in prior work (Lazzati et al., 2026), test whether it reflects non-Markovian expert behavior, and examine whether past actions recover the missing temporal information and improve action prediction.

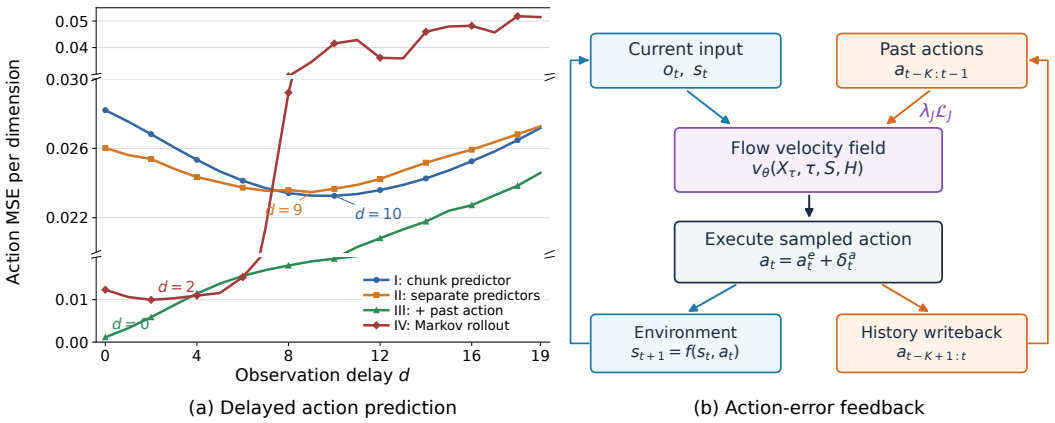
Experiment I. We use an action-chunk predictor conditioned on (o_{t-d}, s_{t-d}) to predict $a_{t-d:t-d+N-1}$, whose conditional mean is $\mathbb{E}[a_{t-d:t-d+N-1} \mid o_{t-d}, s_{t-d}]$. We measure the error at position $d < N$, which is aligned with a_t.

Experiment II. We train a separate predictor for every delay d, with each predictor mapping (o_{t-d}, s_{t-d}) to the same action a_t and estimating $\mathbb{E}[a_t \mid o_{t-d}, s_{t-d}]$.

Experiment III. We additionally supply the action immediately preceding each delayed input and train predictors estimating $\mathbb{E}[a_t \mid o_{t-d}, s_{t-d}, a_{t-d-1}]$.

Experiment IV. We learn and execute a one-step Markov policy $a_t = \pi_{\mathrm{M}}(o_t, s_t)$, then repeat Experiment II on its trajectories to estimate $\mathbb{E}_{\pi_{\mathrm{M}}}[a_t \mid o_{t-d}, s_{t-d}]$. Here the expectation is taken over trajectories generated by π_{M}.

We evaluate each experiment's prediction $\hat{a}_t$ using the mean squared error per action dimension, $\frac{1}{d_a}\mathbb{E}\|\hat{a}_t - a_t\|_2^2$, on held-out trajectories. Figure 2(a) reports the resulting four MSE curves.



(a) Delayed action prediction (b) Action-error feedback

Figure 2: Action history supplies temporal information and carries execution errors. (a) LIBERO-90 delay curves for Experiments I–IV; the vertical scale is expanded between 0.02 and 0.03. (b) PRIM regularizes history sensitivity to limit error propagation through state and history inputs.

We reproduce the U-shaped delay curve in Experiment I (Lazzati et al., 2026), suggesting that past observations and states may contain information more predictive of the current action. However, the curve could also reflect differences in prediction accuracy across positions in an action chunk. To isolate this factor, we train separate predictors in Experiment II and find that the U-shaped curve persists: the prediction MSE of estimators of $\mathbb{E}[a_t \mid o_{t-d}, s_{t-d}]$ does not increase monotonically with delay d.

This finding contrasts with prediction under a Markov assumption. If observations and states form a sufficient Markov state and the expert selects actions from the current input, then $a_t \perp (o_{t-d}, s_{t-d}) \mid (o_{t-d+1}, s_{t-d+1})$ for $d \geq 1$. This conditional independence implies that the conditional-mean predictor using the newer input has no greater optimal MSE than the predictor using the older input. The optimal prediction error must therefore be nondecreasing with delay. The U-shaped curve retained in Experiment II provides direct evidence of non-Markovian expert behavior relative to the current input: the current observation does not fully summarize the information in past experience relevant to action selection.

We therefore add the preceding action in Experiment III to recover this missing information, removing the prediction advantage of delayed inputs. To isolate effects of the environment and the regression procedure, we replace expert behavior with a learned Markov policy in Experiment IV and repeat the regression analysis on its trajectories. The curve becomes nearly nondecreasing with the environment, observation delays, and regression settings unchanged, further supporting non-Markovian behavior in the original LIBERO-90 demonstrations as a source of the U-shaped curve.

However, lower action-prediction error alone does not guarantee higher success. We next examine how errors in executed actions affect later predictions through action history, then evaluate prediction accuracy and success in Section 6.

5 PRIM: Method and Analysis

Section 4 shows that past actions recover temporal information missing from the current observation. During execution, their prediction errors also enter subsequent states and action histories. We derive these two propagation paths and regularize history sensitivity to limit the errors carried into later predictions.

5.1 Action Errors Propagate Through State and History

To illustrate error propagation, we take a policy with one past action and one executed action as an example: $a_t = \pi_\theta(s_t, a_{t-1})$. Let s_t^e, a_{t-1}^e, a_t^e denote the expert state, previous action, and current action. The prediction error under the correct expert input is $\varepsilon_t := \pi_\theta(s_t^e, a_{t-1}^e) - a_t^e$. During execution, write $s_t = s_t^e + \delta_t^s$, $a_{t-1} = a_{t-1}^e + \delta_{t-1}^a$, and $\delta_t^a := a_t - a_t^e$. Define $J_t^s := \delta\pi_\theta/\delta s_t$ and $J_t^a := \delta\pi_\theta/\delta a_{t-1}$ at the expert input. A first-order expansion gives $\delta_t^a = \varepsilon_t + J_t^s\delta_t^s + J_t^a\delta_{t-1}^a + o(\|\delta_t^s\| + \|\delta_{t-1}^a\|)$.

The environment evolves as $s_{t+1} = f(s_t, a_t)$, so $\delta_{t+1}^s = (\delta f/\delta s)\delta_t^s + (\delta f/\delta a)\delta_t^a + o(\|\delta_t^s\| + \|\delta_t^a\|)$. At time $t+1$, the executed a_t becomes a history input. Substituting this state error into the policy expansion at time $t+1$ yields

$$\delta_{t+1}^a = \varepsilon_{t+1} + J_{t+1}^s \frac{\delta f}{\delta s}\delta_t^s + \left(J_{t+1}^s \frac{\delta f}{\delta a} + J_{t+1}^a\right)\delta_t^a + o(\|\delta_t^s\| + \|\delta_t^a\|). \tag{2}$$

Equation (2) shows how action errors accumulate: each error enters the next state and action history, changes the next output, and returns through subsequent policy inputs. The coefficient $J_{t+1}^s(\delta f/\delta a)$ describes the state path, where environment dynamics and policy state sensitivity transmit the error. The coefficient J_{t+1}^a describes direct history writeback. Their sum, $J_{t+1}^s(\delta f/\delta a) + J_{t+1}^a$, is the local gain from δ_t^a to δ_{t+1}^a.

Standard BC and FM primarily reduce the new prediction error ε_{t+1} in Eq. (2), leaving the subsequent propagation terms without a direct constraint. The derivative $\delta f/\delta a$ is determined by environment dynamics. When the state-feedback contribution is relatively weak, the additional history term J_{t+1}^a is a particularly important target for regularization. We therefore introduce past-action sensitivity regularization to retain the prediction benefit of history while limiting the influence of existing action errors on later outputs.

5.2 Temporal Correlations Encourage Policy Sensitivity to Action History

Temporal correlations extending from past to future actions make history predictive. At a fixed current condition, the following result bounds the history sensitivity required to predict the conditional mean of the future actions.

Theorem 1 (Temporal correlation and policy sensitivity to action history)**.** *Fix the current condition S and consider one action coordinate. Suppose the sequence spanning $H = a_{t-K:t-1}$ and $A = a_{t:t+N-1}$ has constant variance $\sigma_a^2 > 0$ and lag-dependent correlations $\mathrm{Corr}(a_r, a_{r+d} \mid S) = \rho_d$. Assume $\mathrm{Var}(g(H) \mid S) \le C_P\mathbb{E}[\|\delta g/\delta H\|_2^2 \mid S]$. For a differentiable conditional-mean policy $\pi^*(S, H) = \mathbb{E}[A \mid S, H]$ in this inequality's domain,*

$$\mathbb{E}\left[\left\|\frac{\delta\pi^*}{\delta H}\right\|_F^2 \middle| S\right] \ge \frac{\sigma_a^2}{C_P}\sum_{d=1}^{N}\rho_d^2. \tag{3}$$

At fixed σ_a^2/C_P, stronger squared temporal correlations raise the sensitivity lower bound. Thus, predicting the conditional mean can require sensitivity to the same history that later carries execution errors in Eq. (2). Appendix C proves the bound and extends it to stochastic policies with the correct conditional mean and to the optimal FM field at $\tau = 0$.

We measure Pearson correlations between a_t and a_{t-d} in LIBERO-90, using pairs within each trajectory, centering actions by task, and averaging over the six continuous action coordinates. The correlations are 0.989, 0.928, 0.840, and 0.740 at lags $d = 1, 3, 5, 7$, respectively, decreasing with lag but remaining high across the history window.

We also test this relation using a Flow policy conditioned on the current observation o_t, state s_t, and three past actions $a_{t-3}, a_{t-2}, a_{t-1}$, trained with the FM objective. We compare squared action correlations with the corresponding squared velocity derivatives at $\tau = 0$ on held-out LIBERO-90 data. After linear adjustment for current inputs and other past actions, the associations are positive in all 90 tasks, with a median Spearman correlation of 0.874. We report the corresponding measurements for four RoboMimic tasks in the appendix.

5.3 The PRIM Action-History Regularizer

To limit the sensitivity J_t^a of generated actions to past actions, we penalize the velocity-field derivative $\delta v/\delta H$. For $K \geq 1$, draw a unit-norm Rademacher direction $u \in \{-1/\sqrt{Kd_a}, +1/\sqrt{Kd_a}\}^{Kd_a}$, with independent equiprobable signs. Define $\mathcal{L}_J(v) := \mathbb{E}\|(\delta v/\delta H)u\|_2^2$. Since $\mathbb{E}[uu^\top] = I/(Kd_a)$, this gives $\mathcal{L}_J(v) = (Kd_a)^{-1}\mathbb{E}\|\delta v/\delta H\|_F^2$. The final objective is

$$\mathcal{L}(v) = \mathcal{L}_{\mathrm{FM}}(v) + \lambda_J \mathcal{L}_J(v), \tag{4}$$

where λ_J controls regularization strength and training uses $v = v_\theta$. The random projection estimates the normalized Frobenius energy without forming the full Jacobian (Hoffman et al., 2019; Bai et al., 2021).

To cover the full Flow-time interval in each update, we divide $[0, 1]$ into m equal intervals and independently sample $\tau_b \sim \mathcal{U}[b/m, (b+1)/m]$ for $b = 0, \ldots, m-1$. For each training example, we share the noise ξ and history direction u across these times. We then estimate the history derivative with paired forward evaluations at perturbation radius $\epsilon > 0$:

$$\widehat{\mathcal{L}}_J = \frac{1}{m}\sum_{b=0}^{m-1} \left\| \frac{v(X_{\tau_b}, \tau_b, S, H + \epsilon u) - v(X_{\tau_b}, \tau_b, S, H)}{\epsilon} \right\|_2^2. \tag{5}$$

Here each X_{τ_b} is constructed on the training bridge. The stratified average estimates the same uniform-time expectation of the finite-difference penalty, while covering every time interval in each update and reducing variation due to time sampling. For small ϵ, the finite difference approximates the directional derivative in $\mathcal{L}_J$. Gradients with respect to the model parameters are obtained by backpropagating through the paired forward evaluations. The FM term uses its usual uniformly sampled training time.

To connect this penalty to the generated actions, we bound the history sensitivity of the sampled endpoint. Consider a continuously differentiable velocity field whose sampling dynamics satisfy $\lambda_{\max}(\mathrm{sym}(\delta v/\delta X_\tau)) \leq \kappa$, where $\mathrm{sym}(M) = (M + M^\top)/2$. Let $c_\kappa = (\int_0^1 e^{2\kappa(1-\tau)}\,\mathrm{d}\tau)^{1/2}$. The initial noise is independent of history and is held fixed when differentiating the output. At a fixed current condition and action history, the generated-action Jacobian satisfies

$$\mathbb{E}_\xi \|J_t^a\|_F^2 \leq c_\kappa^2 \mathbb{E}_\xi \int_0^1 \left\| \frac{\delta v(X_\tau, \tau, S, H)}{\delta H} \right\|_F^2 \mathrm{d}\tau. \tag{6}$$

Here X_τ follows the sampling ODE, and J_t^a contains the rows of $\delta X_1/\delta H$ corresponding to the actions being executed. The history partial derivative of v holds X_τ fixed and is then evaluated along that path. Equation (6) bounds the final action sensitivity by the velocity-field sensitivity during sampling, linking the regularizer to the J_t^a term in Eq. (2).

Penalizing the velocity-field derivative avoids the computation and memory required to differentiate final actions through the multi-step ODE solver. Section 5.4 relates the training penalty to sampling-path sensitivity and bounds the resulting accumulated error.

5.4 Regularization Controls Error Accumulation

Equation (6) bounds the sensitivity of generated actions to history. We now extend this result to repeated execution, where each predicted action enters the history used at the next replanning step. Although each step introduces new prediction errors, sufficiently low history sensitivity keeps their accumulation bounded.

At each replanning step, the policy executes the first P generated actions, $a_{t:t+P-1} = \pi_v^P(s_t, a_{t-K:t-1})$, and writes the final K executed actions into the next history, with $1 \le K \le P \le N$. After r execution segments, the environment time is $t = rP$. Relative to the corresponding expert trajectory, define the state error $\delta^s_{rP} := s_{rP} - s^e_{rP}$ and the history-action error $\delta^a_{rP-K:rP-1} := a_{rP-K:rP-1} - a^e_{rP-K:rP-1}$; these are the input errors at the next replanning step.

Let α bound propagation through the state input, including the policy response J^s_t, and let $\beta > 0$ bound the effect of action changes on the next state and history. Their derivative definitions are given in the appendix. We assume that the normalized history-Jacobian energy along sampling paths is at most $C\mathcal{L}_J(v)$, for a constant $C > 0$, at every relevant local input. Let $\bar{\varepsilon}$ bound the conditional root mean square of each new action-prefix prediction error under correct expert inputs.

Theorem 2 (Informal: sufficient history regularization keeps accumulated errors bounded)**.** *Under the local and sampling assumptions specified in the appendix,* $\mathbb{E}_\xi \|J^a_t\|_F^2 \le c_\kappa^2 C K d_a \mathcal{L}_J(v)$. *The root mean square gain for existing errors is at most* $\rho_v := \alpha + \beta c_\kappa \sqrt{C K d_a \mathcal{L}_J(v)}$. *If* $\alpha < 1$ *and* $\mathcal{L}_J(v) < (1-\alpha)^2/(\beta^2 c_\kappa^2 C K d_a)$, *then* $\rho_v < 1$, *and the long-run mean squared error satisfies*

$$\limsup_{r\to\infty} \mathbb{E}\left[\|\delta^s_{rP}\|_2^2 + \|\delta^a_{rP-K:rP-1}\|_2^2\right] \le \left(\frac{\beta\bar{\varepsilon}}{1-\rho_v}\right)^2. \tag{7}$$

Thus, keeping $\mathcal{L}_J(v)$ below this threshold keeps accumulated state and history errors bounded in mean square under the stated assumptions, even as each replanning step introduces a new prediction error.

6 Experiments

6.1 Policy Training and Evaluation

We train Flow policies on LIBERO-90 using the current visual observation, nine-dimensional robot state, task condition, and K past seven-dimensional actions. Seeds 42, 43, and 44 use batch size 256 and a cosine schedule over 60k training updates. The grid contains $K = 0, \lambda_J = 0$ and $K \in \{1,3,5,7\}, \lambda_J \in \{0,1\}$, giving 27 models. Each task contributes 25 training and 25 held-out demonstrations. We evaluate success after 20k, 40k, and 60k updates with one Flow sample, twenty Euler steps, and $N = 20, P = 10$. Twenty fixed initial states per task and a 400-step limit give 1,800 rollouts per model; further evaluation details appear in Appendix H.2.

We evaluate the history penalty in Eq. (5) on the full predicted action chunk using $m = 4$ time intervals and $\epsilon = 0.1$.

We also adapt PALR and the target-conditioned adversarial (TCA) stage of prior work (Seo et al., 2023; Wen et al., 2020) to the same FM backbone with $K = 5$; their representation losses and training settings are detailed in the appendix. Table 1 reports means and sample standard deviations over three seeds, except $K = 5, \lambda_J = 1$ at 20k (two seeds). Velocity and Jacobian measurements refer to 20k; dashes denote unmeasured losses.

Table 1: LIBERO-90 success rates and losses across methods and training steps.

Policy	K	λ_J	Velocity MSE	Mean $\mathcal{L}_J$	Success@20k	Success@40k	Success@60k
Current state	0	0	.08922	–	76.5±0.75	78.3±0.74	77.9±0.47
Past actions	1	0	.06253	43.1	37.2±8.4	41.6±11.7	44.1±13.0
PRIM	1	1	.08833	1.21×10^{-5}	76.1±3.6	83.3±0.5	84.6±0.8
Past actions	3	0	.05961	34.0	53.1±4.6	67.0±4.1	68.4±4.7
PRIM	3	1	.08761	7.95×10^{-5}	76.5±1.5	83.1±1.3	83.4±1.0
Past actions	5	0	.06028	17.2	56.4±3.3	64.8±1.0	69.2±2.1
PRIM	5	1	.08785	1.64×10^{-4}	76.7±1.7	83.4±0.9	84.6±1.8
Past actions	7	0	.06272	14.3	57.3±6.1	57.2±12.9	60.5±8.9
PRIM	7	1	.08646	3.33×10^{-4}	**76.9±0.4**	**84.1±0.5**	**85.3±1.1**
PALR (adapted)	5	0	–	–	46.4±7.0	50.0±14.2	53.9±13.3
TCA (adapted)	5	0	–	–	48.8±12.4	55.5±12.7	56.2±14.0

Past-action policies achieve lower velocity-prediction MSE than the current-state baseline, yet substantially lower success. With history regularization, PRIM improves success across history lengths and exceeds the current-state baseline with further training. PRIM also exceeds both adapted baselines at every evaluated checkpoint for $K = 5$. We further compare single-sample action MSE for matched $K = 3$, 20k policies and find that FM alone has lower prediction error but lower success than PRIM (see appendix).

To compare with a simpler history regularizer, we evaluate action-history dropout at probability 0.1 in a matched $K = 5$, 20k experiment. Dropout leaves success close to the unregularized policy, while Jacobian regularization improves it (Appendix H.6).

We also study temporal ensembling (Lazzati et al., 2026) and find that independently sampling future actions with the same current observation and action history retains most of its LIBERO-90 success-rate gain without varying observation delay; details appear in Appendix G.

6.2 Three-Seed RoboMimic Results

We next test whether the same action-history regularizer transfers across manipulation benchmarks. We train action-chunk Flow policies from scratch on the Can, Square, ToolHang, and Transport PH datasets in RoboMimic (Mandlekar et al., 2022), using 200 demonstrations per task and three training seeds. The grid contains a current-state policy with $K = 0$, and action-history policies with $K \in \{1, 3, 5\}$ and $\lambda_J \in \{0, 0.01\}$. Each matched history-policy pair shares its architecture, initialization seed, minibatch sequence, Flow noise, and time samples. The regularized model uses a normalized Rademacher action-history direction for each training example. Evaluation uses the 20k EMA checkpoint, one Flow sample, twenty Euler steps, $N = 20, P = 10$, and 250 fixed fresh initial states per task and seed.

Table 2: RoboMimic success at 20k updates, reported as means and sample standard deviations over three seeds with 250 rollouts each. PALR and TCA use $K = 5$ and the FM adaptations described in the appendix, which also reports their 40k results.

	$K=0$	$K=1$		$K=3$		$K=5$		$K=5$, adapted	
Task	$\lambda_J=0$	$\lambda_J=0$	$\lambda_J=.01$	$\lambda_J=0$	$\lambda_J=.01$	$\lambda_J=0$	$\lambda_J=.01$	PALR	TCA
Can	**99.6±0.0**	99.2±0.0	99.3±0.5	99.3±0.5	98.8±0.4	98.7±0.8	98.9±0.2	98.8±0.4	98.4±0.7
Square	**87.9±2.3**	76.9±8.1	84.9±2.3	84.1±2.4	84.9±1.9	85.2±4.2	85.7±3.8	82.1±1.6	81.2±3.8
ToolHang	80.3±4.0	74.4±3.4	79.7±2.7	76.1±5.0	**84.3±2.0**	76.7±2.2	83.2±1.1	72.0±2.0	71.9±3.3
Transport	54.7±2.4	54.3±3.6	63.3±5.1	55.3±1.5	**64.8±3.7**	56.4±1.8	62.1±2.7	53.7±0.5	52.7±1.2

PRIM improves success in most matched task–history comparisons. The gains are largest on ToolHang and Transport, where regularized history policies also exceed the current-state baseline and, at $K = 5$, both adapted baselines. The appendix reports the action correlations and policy sensitivities relevant to Eq. (3).

6.3 Error Accumulation and Recovery

To test history-error propagation in Eq. (2), we perturb each $\lambda_J \in \{0, 1\}$ policy's initial history and compare perturbed and reference branches as they write back predictions under shared expert conditions and sampling noise. Let δ_r^a be their action-prefix difference after r replanning steps and $\mathcal{C}_R = \sum_{r=1}^{R} \|\delta_r^a\|_2^2/(Pd_a)$. PRIM lowers $\mathcal{C}_8$ across all tested seeds and history lengths despite larger seed-averaged initial errors (Figure 3a–b). After an executed-action perturbation, PRIM also lowers cumulative action deviations in most tested conditions (Figure 3c–d).

We compare representation regularization under a common initial history perturbation, replaying expert observations as each branch writes back its actions. PALR and TCA retain subsequent amplification; PRIM's response decays rapidly (Table 3). Representation and velocity sensitivities and RoboMimic diagnostics appear in the appendix.

Table 3: Common history perturbation on LIBERO-90 ($K = 5$, 60k): $\mathbb{E}\|\delta_r^a\|_2^2/(Pd_a)$ averaged over three seeds and 170 held-out windows from 85 tasks; one perturbation of physical-coordinate RMS 0.01.

Method	$r = 1$	$r = 3$	$r = 8$	$\mathbb{E}\mathcal{C}_8$
FM alone	.00310	.01031	.00438	.03842
PALR	.00224	.00856	.000270	.02587
TCA	.00191	.00919	.00467	.04350
PRIM	1.95×10^{-9}	1.55×10^{-12}	3.26×10^{-15}	1.96×10^{-9}

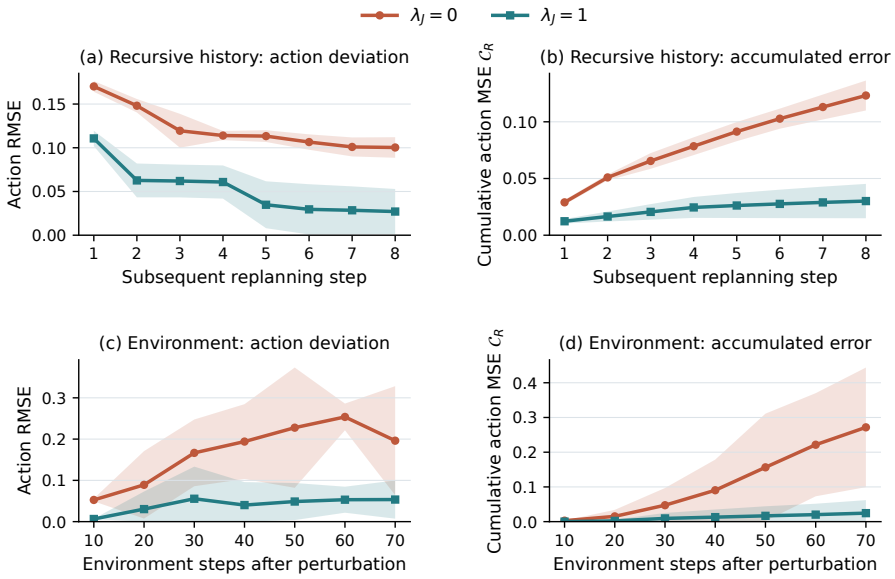


Figure 3: Action deviations: (a–b) history writeback with expert conditions fixed; (c–d) environment execution. Lines and shading show three-seed means and standard deviations.

To test recovery, both policies receive identical perturbations from the same state and history where they succeed unperturbed. PRIM achieves higher success (Table 4).

Table 4: Success after common perturbations: 39 initial conditions with unperturbed success for both policies, two directions each (78 pairs). Gains are percentage points.

Perturbed input or command	Amplitude	$\lambda_J = 0$	PRIM	Gain
History values	0.1	72/78 (92.3%)	77/78 (98.7%)	+6.4
History values	0.3	65/78 (83.3%)	77/78 (98.7%)	+15.4
History values	0.6	59/78 (75.6%)	76/78 (97.4%)	+21.8
Executed action	0.1	70/78 (89.7%)	71/78 (91.0%)	+1.3
Executed action	0.3	52/78 (66.7%)	70/78 (89.7%)	+23.1
Executed action	0.6	49/78 (62.8%)	68/78 (87.2%)	+24.4

7 Conclusion

Past actions recover temporal information absent from the current observation, while their direct writeback lets generated errors influence later predictions. PRIM controls this path through the action-history Jacobian of the Flow velocity field. Under the stated local conditions, its sensitivity bound keeps accumulated state and action-history errors bounded in mean square. PRIM reduces measured error propagation and improves success over FM-only history policies across LIBERO-90 seeds and history lengths; with further training, it also exceeds the current-state baseline.

AI USE STATEMENT

Generative AI tools, including ChatGPT and Codex, assisted with literature discovery, language editing and translation, and drafting and revising portions of this manuscript. They also assisted with refining research hypotheses and experimental designs, implementing methods and evaluation code, analyzing experimental results, and developing and checking mathematical arguments and proofs. The author directed the research and takes responsibility for the final content, including the text, citations, mathematical claims, code, and reported results.

References

Amin Abyaneh, Charlotte Morissette, Mohamad H. Danesh, et al. Contractive Diffusion Policies: Robust Action Diffusion via Contractive Score-Based Sampling with Differential Equations. *ICLR*, 2026.

Shaojie Bai, Vladlen Koltun, and J. Zico Kolter. Stabilizing Equilibrium Models by Jacobian Regularization. *International Conference on Machine Learning*, 2021.

Adam Block, Ali Jadbabaie, Daniel Pfrommer, Max Simchowitz, and Russ Tedrake. Provable Guarantees for Generative Behavior Cloning: Bridging Low-Level Stability and High-Level Behavior. arXiv:2307.14619, 2023.

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

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal Confusion in Imitation Learning. *NeurIPS*, 2019.

Amirreza Dolatpour Fathkouhi, Justin Lee, and Heman Shakeri. Interventional Flow Matching: Prospective Dose-Response Forecasting with Velocity-Field Jacobian Regularization. arXiv:2606.29386, 2026.

Judy Hoffman, Daniel A. Roberts, and Sho Yaida. Robust Learning with Jacobian Regularization. arXiv:1908.02729, 2019.

Sunshine Jiang, Xiaolin Fang, Nicholas Roy, et al. Streaming Flow Policy: Simplifying Diffusion/Flow-Matching Policies by Treating Action Trajectories as Flow Trajectories. *CoRL*, 2025.

Filippo Lazzati, Kyle Stachowicz, William Chen, et al. Why Does Action Chunking Improve Behavioral Cloning Performance in Robotic Control? arXiv:2608.02547, 2026.

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

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. arXiv:2306.03310, 2023.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What Matters in Learning from Offline Human Demonstrations for Robot Manipulation. *Proceedings of the 5th Conference on Robot Learning*, PMLR 164:1678–1690, 2022.

Shaunak A. Mehta, Yusuf Umut Ciftci, Balamurugan Ramachandran, Somil Bansal, and Dylan P. Losey. Stable-BC: Controlling Covariate Shift with Stable Behavior Cloning. *IEEE Robotics and Automation Letters*, 2025.

Daniel Pfrommer, Thomas T. C. K. Zhang, Stephen Tu, and Nikolai Matni. TaSIL: Taylor Series Imitation Learning. *NeurIPS*, 2022.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. *AISTATS*, 2011.

Seokin Seo, HyeongJoo Hwang, Hongseok Yang, and Kee-Eung Kim. Regularized Behavior Cloning for Blocking the Leakage of Past Action Information. *NeurIPS*, 2023.

Chuan Wen, Jierui Lin, Trevor Darrell, Dinesh Jayaraman, and Yang Gao. Fighting Copycat Agents in Behavioral Cloning from Observation Histories. *NeurIPS*, 2020.

Thomas T. Zhang, Daniel Pfrommer, Chaoyi Pan, Nikolai Matni, and Max Simchowitz. Action Chunking and Exploratory Data Collection Yield Exponential Improvements in Behavior Cloning for Continuous Control. arXiv:2507.09061, 2025.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. *Robotics: Science and Systems*, 2023.

Appendix

A Notation and the Flow Sampling Path

Let S contain the current visual observation, robot state, and task condition, let $H = a_{t-K:t-1}$ contain the previous K actions, and let $A = a_{t:t+N-1}$ be the expert action chunk. For $\xi \sim \mathcal{N}(0, I)$ and $\tau \in [0, 1]$, the training bridge is $X_\tau = (1-\tau)\xi + \tau A$, with target velocity $A - \xi$. The learned field is $v(X_\tau, \tau, S, H)$. At inference, the sampled action path solves

$$\frac{\mathrm{d}X_\tau}{\mathrm{d}\tau} = v(X_\tau, \tau, S, H), \qquad X_0 = \xi, \tag{8}$$

and $\pi_v^P(S, H)$ selects the first P actions from X_1.

At a replanning time, write $z_t = (s_t, H_t)$. The map $\mathcal{G}(s_t, H_t, A_t^P)$ composes the following P environment transitions and writes the executed actions into the next history. The complete replanning map is $\mathcal{F}_v(z_t) = \mathcal{G}(s_t, H_t, \pi_v^P(s_t, H_t))$. All local results below apply to a convex set $\mathcal{Z}$ of replanning inputs and the Flow paths that they induce.

B Action Errors Enter the State and History Channels

We first expand the one-step relation stated in the main paper. Consider $a_t = \pi_\theta(s_t, a_{t-1})$ and $s_{t+1} = f(s_t, a_t)$. Let s_t^e, a_{t-1}^e, a_t^e denote the expert quantities, let $\varepsilon_t = \pi_\theta(s_t^e, a_{t-1}^e) - a_t^e$, and write the deployed errors as $s_t = s_t^e + \delta_t^s$ and $a_{t-1} = a_{t-1}^e + \delta_{t-1}^a$. Define $J_t^s = \delta\pi_\theta/\delta s$ and $J_t^a = \delta\pi_\theta/\delta a_{t-1}$ at the expert input. Continuous differentiability gives

$$\delta_t^a = \varepsilon_t + J_t^s \delta_t^s + J_t^a \delta_{t-1}^a + r_t^\pi, \qquad \frac{\|r_t^\pi\|_2}{\|\delta_t^s\|_2 + \|\delta_{t-1}^a\|_2} \to 0. \tag{9}$$

The environment expansion gives

$$\delta_{t+1}^s = \frac{\delta f}{\delta s}\delta_t^s + \frac{\delta f}{\delta a}\delta_t^a + r_t^f, \qquad \frac{\|r_t^f\|_2}{\|\delta_t^s\|_2 + \|\delta_t^a\|_2} \to 0. \tag{10}$$

At time $t+1$, the executed a_t becomes the history input. Substituting Eq. (10) into Eq. (9) at time $t+1$ yields

$$\delta_{t+1}^a = \varepsilon_{t+1} + J_{t+1}^s \frac{\delta f}{\delta s}\delta_t^s + \left(J_{t+1}^s \frac{\delta f}{\delta a} + J_{t+1}^a\right)\delta_t^a + o(\|\delta_t^s\| + \|\delta_t^a\|). \tag{11}$$

The term $J_{t+1}^s(\delta f/\delta a)$ is the state-mediated path. The term J_{t+1}^a is the direct action-history writeback path. Replacing one action with P consecutive actions only changes the dimensions of the composed maps; the same two paths appear in $\mathcal{F}_v$.

C Temporal Correlations Encourage Policy Sensitivity to Action History

Assumptions. Fix a current condition $S = s$ for which conditional second moments exist. For one scalar action coordinate, assume the actions from $t-K$ through $t+N-1$ have a common conditional mean and variance $0 < \sigma_a^2(s) < \infty$, and conditional correlations depending only on lag: $\mathrm{Corr}(a_r, a_{r+d} \mid s) = \rho_d(s)$ whenever both indices lie in this window. This condition extends temporal dependence across the boundary between history and prediction. Let H contain all supplied history coordinates. Its conditional law must satisfy $\mathrm{Var}(g(H) \mid s) \le C_P(s)\mathbb{E}[\|\delta g/\delta H\|_2^2 \mid s]$, with $0 < C_P(s) < \infty$, for the conditional-mean functions used below. Assume these functions have weak derivatives in the corresponding weighted Sobolev space. Infinite derivative energy makes the lower bound immediate; the proof therefore considers finite energy.

Proof of Theorem 1. For $i = 0, \ldots, N-1$, write $\pi_i^*(s, H) = \mathbb{E}[a_{t+i} \mid s, H]$. The previous action a_{t-1} is measurable with respect to H. The tower property therefore gives

$$\mathrm{Cov}(\pi_i^*(s, H), a_{t-1} \mid s) = \mathrm{Cov}(a_{t+i}, a_{t-1} \mid s) = \sigma_a^2(s)\rho_{i+1}(s). \tag{12}$$

Conditional Cauchy–Schwarz bounds the square of this covariance by $\sigma_a^2(s)\,\mathrm{Var}(\pi_i^*(s,H) \mid s)$. Applying the assumed Poincaré inequality then yields

$$\sigma_a^4(s)\rho_{i+1}(s)^2 \le \sigma_a^2(s)C_P(s)\mathbb{E}\left[\left\|\frac{\delta\pi_i^*(s,H)}{\delta H}\right\|_2^2 \middle| s\right]. \tag{13}$$

Divide by $\sigma_a^2(s)C_P(s) > 0$ and sum over i. The sum of squared row gradients is the squared Frobenius norm, proving Eq. (3). For vector actions, apply the argument to each output coordinate using the matching coordinate of a_{t-1}. With coordinate variances $\sigma_{a,c}^2(s)$ and lag correlations $\rho_{d,c}(s)$, the lower bound becomes $C_P(s)^{-1}\sum_{c=1}^{d_a}\sigma_{a,c}^2(s)\sum_{d=1}^{N}\rho_{d,c}(s)^2$. Integrating over S gives the corresponding unconditional bound, retaining the expectation over all these condition-dependent quantities. □

Stochastic policies and Flow Matching. Let $\pi(s,H,\xi)$ be a generated action chunk with noise independent of H given s, and assume its noise-averaged output equals $\pi^*(s,H)$. Assume differentiation under this expectation is justified by an integrable derivative bound. Jensen's inequality gives $\mathbb{E}_\xi\|\delta\pi/\delta H\|_F^2 \ge \|\delta\pi^*/\delta H\|_F^2$ for almost every H, so the theorem also bounds generated-action sensitivity after averaging over history and noise. An exact conditional generative policy satisfies the required mean equality. For the population-optimal FM field, independence of ξ gives $v^*(\xi,0,s,H) = \pi^*(s,H) - \xi$. Its history derivative at $\tau = 0$ therefore obeys the same bound. Combining this result with the endpoint sensitivity bound below gives a bound on the integrated sensitivity: under its assumptions and the conditional-mean equality, $\mathbb{E}[\mathcal{E}_H(v;\omega) \mid s] \ge \sigma_a^2(s)\sum_d \rho_d(s)^2/(C_P(s)c_\kappa^2)$, taking the full output chunk $P = N$.

Future temporal correlations. Conditional stationarity makes $\mathrm{Corr}(a_{t+i}, a_{t+j} \mid s) = \rho_{j-i}(s)$. Consequently, $\sum_{0\le i<j<N}\mathrm{Corr}(a_{t+i},a_{t+j} \mid s)^2 = \sum_{d=1}^{N-1}(N-d)\rho_d(s)^2$. For $N \ge 2$, Eq. (3) also bounds sensitivity below by $\sigma_a^2(s)/[C_P(s)(N-1)]$ times this sum. At a controlled variance-to-Poincaré ratio, stronger squared temporal correlations thus increase the required sensitivity lower bound.

D The Velocity-Field Jacobian Controls Endpoint Sensitivity

For a sampled path $\omega = (X_{0:1}, S, H)$, define

$$\mathcal{E}_H(v;\omega) := \int_0^1 \left\|\frac{\delta v(X_\tau,\tau,S,H)}{\delta H}\right\|_F^2 \mathrm{d}\tau, \qquad c_\kappa := \left(\int_0^1 e^{2\kappa(1-\tau)}\,\mathrm{d}\tau\right)^{1/2}. \tag{14}$$

Lemma 1 (Action-history sensitivity of the sampled endpoint)**.** *Suppose v is continuously differentiable in X and H, the initial noise is independent of H, and every induced path satisfies $\lambda_{\max}(\mathrm{sym}(\delta v/\delta X)) \le \kappa$, where $\mathrm{sym}(M) = (M + M^\top)/2$. Then*

$$\left\|\frac{\delta\pi_v^P(S,H)}{\delta H}\right\|_F \le c_\kappa\sqrt{\mathcal{E}_H(v;\omega)}. \tag{15}$$

Proof. Let $Z_\tau = \delta X_\tau/\delta H$. Since X_0 is independent of H, $Z_0 = 0$. Differentiating Eq. (8) gives

$$\frac{\mathrm{d}Z_\tau}{\mathrm{d}\tau} = \frac{\delta v}{\delta X_\tau}Z_\tau + \frac{\delta v}{\delta H}, \qquad Z_0 = 0. \tag{16}$$

Let $\Phi(r,\tau)$ be the transition matrix of the homogeneous equation. For any vector y, differentiating $\|\Phi(r,\tau)y\|_2^2$ and using the symmetric-part bound gives $\mathrm{d}\|\Phi(r,\tau)y\|_2^2/\mathrm{d}r \le 2\kappa\|\Phi(r,\tau)y\|_2^2$. Gronwall's inequality therefore gives $\|\Phi(1,\tau)\|_2 \le e^{\kappa(1-\tau)}$. Variation of constants and Cauchy–Schwarz yield

$$\|Z_1\|_F \le \int_0^1 e^{\kappa(1-\tau)}\left\|\frac{\delta v}{\delta H}\right\|_F \mathrm{d}\tau \le c_\kappa\sqrt{\mathcal{E}_H(v;\omega)}. \tag{17}$$

Selecting the first P endpoint actions has operator norm at most one, which proves Eq. (15). □

E The Complete Replanning Map and Local Error Accumulation

Write $J_t^s = \delta\pi_v^P/\delta s_t$, $J_t^a = \delta\pi_v^P/\delta H_t$, $G_s = \delta\mathcal{G}/\delta s_t$, $G_H = \delta\mathcal{G}/\delta H_t$, and $G_A = \delta\mathcal{G}/\delta A_t^P$. The chain rule gives

$$\frac{\delta\mathcal{F}_v}{\delta z_t} = \left[G_s + G_A J_t^s,\ G_H + G_A J_t^a\right]. \tag{18}$$

Define $\alpha = \sup_{z_t\in\mathcal{Z}} \|[G_s+G_A J_t^s, G_H]\|_2$ and $\beta = \sup_{z_t\in\mathcal{Z}} \|G_A\|_2$. These constants retain the state-feedback and environment terms while isolating the additional trainable action-history term.

Theorem 3 (Action-history sensitivity bounds replanning error)**.** *Assume the conditions of Lemma 1, $\mathcal{Z}$ is convex, and $\mathcal{E}_H(v;\omega) \le \overline{\mathcal{E}}_H$ on every local inference path. Then*

$$\sup_{z_t\in\mathcal{Z}} \|J_t^a\|_F \le c_\kappa\sqrt{\overline{\mathcal{E}}_H}, \qquad \sup_{z_t\in\mathcal{Z}} \left\|\frac{\delta\mathcal{F}_v(z_t)}{\delta z_t}\right\|_2 \le \alpha + \beta c_\kappa\sqrt{\overline{\mathcal{E}}_H}. \tag{19}$$

If $\alpha+\beta c_\kappa\sqrt{\overline{\mathcal{E}}_H} < 1$, the local replanning map is contractive. If every replanning step injects an error with norm at most $\overline{\eta}$, and the perturbed inputs remain in $\mathcal{Z}$, then

$$\limsup_{r\to\infty} \|\delta z_r\|_2 \le \frac{\overline{\eta}}{1-\alpha-\beta c_\kappa\sqrt{\overline{\mathcal{E}}_H}}. \tag{20}$$

Proof. Lemma 1 gives the first inequality in Eq. (19). Equation (18), the triangle inequality, and submultiplicativity give the second. Let $\rho = \alpha+\beta c_\kappa\sqrt{\overline{\mathcal{E}}_H} < 1$. Convexity of $\mathcal{Z}$ and the fundamental theorem of calculus imply $\|\mathcal{F}_v(z) - \mathcal{F}_v(z')\|_2 \le \rho\|z - z'\|_2$. With a bounded new error at every replanning step, $\|\delta z_{r+1}\|_2 \le \rho\|\delta z_r\|_2 + \overline{\eta}$. Iterating this relation and summing the geometric series proves Eq. (20). □

E.1 Mean Squared Error Under the Training Regularizer

We give the formal statement and proof of the informal mean-square theorem in Section 5.4 of the main paper.

Notation. Consider $1 \le K \le P \le N$, where K is the history length, P is the number of actions executed between consecutive replanning steps, and N is the generated chunk length. After r execution segments, $t = rP$. Fix an expert trajectory obeying the same dynamics, let $H_t = a_{t-K:t-1}$ and $H_t^e = a_{t-K:t-1}^e$, and define $\delta_t^s = s_t - s_t^e$, $\delta_{t-K:t-1}^a = H_t - H_t^e$, and $\delta z_t = z_t - z_t^e = (\delta_t^s, \delta_{t-K:t-1}^a)$. All previous history entries are replaced by generated actions after one execution prefix, so $G_H = 0$. For fixed Flow noise, the state and history derivative columns of the replanning map are $G_s + G_A J_t^s$ and $G_A J_t^a$.

Assumptions on dynamics and sampling. Assume continuously differentiable dynamics and velocity fields, existence of the sampling solution on $[0,1]$, and uniform bounds $\|G_s + G_A J_t^s\|_2 \le \alpha$ and $\|G_A\|_2 \le \beta$ on the convex local input region $\mathcal{Z}$, for all relevant noise draws and action comparisons. Equivalently, the state bound controls $\delta(s_{t+P}, [\pi_v^P]_{P-K:P-1})/\delta s_t$ with history and noise fixed. A sufficient action bound is $\|(\delta s_{t+P}/\delta a_{t:t+P-1})|_{s_t}\|_2^2 + 1 \le \beta^2$, because selecting the final K actions has operator norm one.

Assumption relating the training penalty to sampling paths. At every fixed local input, assume the sampling conditions of Lemma 1 and the comparison

$$\frac{1}{Kd_a}\mathbb{E}_\xi \int_0^1 \left\|\frac{\delta v(X_\tau,\tau,S,H)}{\delta H}\right\|_F^2 \mathrm{d}\tau \le C\mathcal{L}_J(v). \tag{21}$$

Here the left side is conditional on that input and follows sampling paths; $\mathcal{L}_J$ averages over the training bridge. This comparison is assumed uniformly over the input segments used below. The derivative bounds also cover the action segments between the policy and expert prefixes.

Assumptions on new errors and the local region. Define the fresh prefix error by $\varepsilon_{t:t+P-1} = \pi_v^P(s_t^e, H_t^e; \xi) - a_{t:t+P-1}^e$, and assume its conditional root mean square, given the trajectory up to that replanning step, is bounded by $\bar{\varepsilon}$ at each step. Every step samples $\xi \sim \mathcal{N}(0, I)$ independently of the preceding trajectory; input comparisons use the same noise draw. Assume $\mathbb{E}\|\delta z_0\|_2^2 < \infty$. For every replanning step, assume almost surely that the actual and expert inputs, and their connecting segments, lie in $\mathcal{Z}$, and that the action segments joining the generated and expert prefixes lie in the region where the bound on G_A holds. Expectations below include the initial randomness and the successive Flow noise draws.

Theorem 4 (Mean squared error accumulation under history regularization)**.** *Under these conditions,* $\mathbb{E}_\xi\|J_t^a\|_F^2 \le c_\kappa^2 C K d_a \mathcal{L}_J(v)$. *Define* $\rho_v = \alpha + \beta c_\kappa \sqrt{C K d_a \mathcal{L}_J(v)}$. *If* $\alpha < 1$ *and* $\mathcal{L}_J(v) < (1-\alpha)^2/(\beta^2 c_\kappa^2 C K d_a)$, *then*

$$\limsup_{r\to\infty} \mathbb{E}\big[\|\delta_{rP}^s\|_2^2 + \|\delta_{rP-K:rP-1}^a\|_2^2\big] \le \left(\frac{\beta\bar{\varepsilon}}{1-\rho_v}\right)^2. \tag{22}$$

Proof. Lemma 1 and Eq. (21) first give the stated bound on $\mathbb{E}_\xi\|J_t^a\|_F^2$. Applying the chain rule and the triangle inequality to the two derivative columns yields $\|\delta\mathcal{F}_v/\delta z\|_2 \le \alpha + \beta\|J_t^a\|_F$. Minkowski's inequality then gives

$$\left(\mathbb{E}_\xi\left\|\frac{\delta\mathcal{F}_v(z;\xi)}{\delta z}\right\|_2^2\right)^{1/2} \le \alpha + \beta\left(\mathbb{E}_\xi\|J_t^a\|_F^2\right)^{1/2} \le \rho_v. \tag{23}$$

For two fixed local inputs z, z^e, write $z(u) = z^e + u(z - z^e)$, $u \in [0,1]$. The fundamental theorem of calculus expresses their output difference as the integral of $(\delta\mathcal{F}_v(z(u);\xi)/\delta z)(z - z^e)$. Applying Minkowski's inequality to this integral and using the uniform bound above gives

$$\left(\mathbb{E}_\xi\|\mathcal{F}_v(z;\xi) - \mathcal{F}_v(z^e;\xi)\|_2^2\right)^{1/2} \le \rho_v\|z - z^e\|_2. \tag{24}$$

This integral bounds the response to finite input differences.

Let $\eta_t = \mathcal{F}_v(z_t^e;\xi) - z_{t+P}^e$ denote the new error after executing the policy prefix from the correct expert input. The two executions differ only in their action prefixes, so the bound on G_A, integrated along the action segment, gives $\|\eta_t\|_2 \le \beta\|\varepsilon_{t:t+P-1}\|_2$. Thus its conditional root mean square is at most $\beta\bar{\varepsilon}$. The exact error decomposition is

$$\delta z_{t+P} = \mathcal{F}_v(z_t;\xi) - \mathcal{F}_v(z_t^e;\xi) + \eta_t. \tag{25}$$

Conditioning on the preceding trajectory fixes the compared inputs and leaves the fresh noise independent. Apply Eq. (24), then Minkowski's inequality over the preceding trajectory. With $e_r = (\mathbb{E}\|\delta z_{rP}\|_2^2)^{1/2}$, this yields $e_{r+1} \le \rho_v e_r + \beta\bar{\varepsilon}$. Since the sensitivity threshold ensures $\rho_v < 1$, iteration gives

$$e_r \le \rho_v^r e_0 + \beta\bar{\varepsilon}\frac{1-\rho_v^r}{1-\rho_v}. \tag{26}$$

Taking the limit superior and squaring proves Eq. (22), since $\|\delta z_{rP}\|_2^2 = \|\delta_{rP}^s\|_2^2 + \|\delta_{rP-K:rP-1}^a\|_2^2$. □

F The Mean Training Loss and Local Inference Paths

The pathwise result in Theorem 3 uses a bound on every local path. The main paper uses the mean-square result in Theorem 4, with the explicit local comparison in Eq. (21). We also record a distributional bound. Let $m = Kd_a$, and suppose the density ratio of local inference tuples to training-bridge tuples is bounded by C. If $u \in \{-1/\sqrt{m}, +1/\sqrt{m}\}^m$ has independent coordinates, then

$$\mathbb{E}_u\left\|\frac{\delta v}{\delta H}u\right\|_2^2 = \frac{1}{m}\left\|\frac{\delta v}{\delta H}\right\|_F^2. \tag{27}$$

Therefore, for every $q \in (0,1)$, change of measure and Markov's inequality give

$$\Pr\left[\mathcal{E}_H(v;\omega) \le \frac{Cm\mathcal{L}_J}{q}\right] \ge 1-q, \qquad \left\|\frac{\delta \pi_v^P}{\delta H}\right\|_F \le c_\kappa \sqrt{\frac{Cm\mathcal{L}_J}{q}} \tag{28}$$

on those paths. Equation (28) connects the optimized mean loss to typical local inference paths under an explicit coverage condition. The deterministic guarantee in Theorem 3 still uses a bound over every path in the local rollout set.

To verify Eq. (27), note that $\mathbb{E}_u[uu^\top] = I/m$. For $J = \delta v/\delta H$, $\mathbb{E}_u\|Ju\|_2^2 = \mathrm{tr}(J\mathbb{E}[uu^\top]J^\top) = \|J\|_F^2/m$. One JVP is therefore an unbiased estimate of mean Frobenius energy per valid history coordinate (Hoffman et al., 2019; Bai et al., 2021).

G Independent Action Generation as an Extension

PRIM controls propagation across replanning boundaries. This section records an additional experiment on correlated errors within an executed action chunk. A joint Flow policy generates $a_{t:t+P-1}$ from one noisy action chunk under $(s_t, a_{t-K:t-1})$, so one Flow sample couples the future actions.

G.1 Independent Composition

We fix the current condition, replanning boundary, and ten-step execution length. Standard inference executes ten actions from one joint Flow sample. Independent composition draws ten Flow samples and executes position i from sample i. Table 5 reports the same $K = 7, \lambda_J = 1$, 20k policies on the same 1,800 initial states.

Table 5: Standard inference and independent composition for joint 20k models on LIBERO-90.

Seed	Standard	Independent	Gain
42	76.44%	88.94%	+12.50 pp
43	76.94%	89.00%	+12.06 pp
44	77.33%	87.61%	+10.28 pp
Mean	76.91%	88.52%	+11.61 pp

Consider two consecutive action errors $e_0 = \delta_t^a$ and $e_1 = \delta_{t+1}^a$. Linearizing two environment transitions gives $\delta_{t+2}^s \simeq B_0 e_0 + B_1 e_1$, where $B_0 = (\delta f_{t+1}/\delta s)(\delta f_t/\delta a)$ and $B_1 = \delta f_{t+1}/\delta a$. Condition on $(s_t, a_{t-K:t-1})$, let $\mu_i = \mathbb{E}[e_i]$, and let $\Sigma_{ij} = \mathrm{Cov}(e_i, e_j)$. Then

$$\mathbb{E}\|\delta_{t+2}^s\|_2^2 \simeq \|B_0\mu_0 + B_1\mu_1\|_2^2 + \mathrm{tr}(B_0\Sigma_{00}B_0^\top) + \mathrm{tr}(B_1\Sigma_{11}B_1^\top) + 2\,\mathrm{tr}(B_0\Sigma_{01}B_1^\top). \tag{29}$$

The final term is the contribution of cross-action covariance. Independent generation preserves each conditional marginal and removes this cross term by using

$$q(a_{t:t+P-1} \mid s_t, a_{t-K:t-1}) = \prod_{i=0}^{P-1} p_i(a_{t+i} \mid s_t, a_{t-K:t-1}). \tag{30}$$

H Additional Experimental Details

H.1 Stratified Flow Times and Finite Differences

The implemented history penalty uses paired forward evaluations at histories H and $H + \epsilon u$, with the same noisy action, time, current condition, and validity indicators. Model-parameter gradients are computed through both evaluations. For a fixed training example, noise, and direction, let $\ell(\tau)$ denote the squared finite-difference response at time τ. Dividing $[0,1]$ into m equal intervals and drawing one independent uniform time τ_b from each gives

$\mathbb{E}[m^{-1}\sum_b \ell(\tau_b)] = \int_0^1 \ell(\tau)\,\mathrm{d}\tau$. Thus stratification preserves the uniform-time expectation of this finite-difference objective.

Let μ_b and σ_b^2 be the conditional mean and variance of ℓ within interval b, and let $\bar{\mu} = m^{-1}\sum_b \mu_b$. The stratified estimator has variance $m^{-2}\sum_b \sigma_b^2$. The average of m independent uniform draws over the entire interval has variance $m^{-2}\sum_b \sigma_b^2 + m^{-2}\sum_b (\mu_b - \bar{\mu})^2$. Stratification removes the second contribution for this fixed example, noise, and direction. Finite-difference bias remains controlled by the perturbation radius and local smoothness, separately from time-sampling variation. The experiments use $m = 4$ and $\epsilon = 0.1$.

H.2 Data and Evaluation

The LIBERO-90 training and validation partitions each contain 25 demonstrations for every task. The fixed offline set contains 2,048 target windows. Every policy predicts a 20×7 action chunk and executes the first ten actions. Each action contains three translation coordinates, three rotation coordinates, and one gripper coordinate. Missing actions at the start of an episode are zero padded and accompanied by a validity mask; history never crosses an episode boundary.

Success evaluation uses EMA parameters, one Flow sample, twenty Euler steps, and $N = 20, P = 10$. Every task uses twenty newly generated initial states, fixed across compared policies, with a 400-step limit, giving 1,800 rollouts for a complete model evaluation. The recursive-writeback experiment holds the state condition and Flow noise fixed between its two branches, so their difference is caused by the history that is written back after each prediction.

H.3 LIBERO-90 Action Prediction and History Sensitivity

To relate success to prediction accuracy and history sensitivity, we measure both quantities for the 20k policies in Table 6. Action history lowers single-sample prediction error, while models trained with FM alone have a large history finite-difference response. Relative to the FM-only policy at the same history length, setting $\lambda_J = 1$ reduces this measured sensitivity by five to six orders of magnitude and raises success from 37.2% to 76.1% for $K = 1$, and from 53.1% to 76.5% for $K = 3$.

Table 6: Three-seed results for Flow Matching on LIBERO-90 at 20k updates. Action MSE uses one sample, ten Euler steps, and the first ten actions in physical action coordinates. History sensitivity is a Bridge-FD estimate on the first ten velocity outputs, rescaled by the number of valid history coordinates; the perturbation radius is 0.1. Success entries use one sample and twenty Euler steps.

K	λ_J	Action MSE	History sensitivity	Success
0	0	0.0408	–	76.5±0.75%
1	0	0.0275	301.7	37.2%
1	1	0.0354	8.49×10^{-5}	76.1%
3	0	0.0259	713.6	53.1%
3	1	0.0349	1.67×10^{-3}	76.5%

H.4 RoboMimic Action Correlation and History Sensitivity

We measure temporal dependence and fitted-policy sensitivity on four RoboMimic tasks in Table 7. Let $\rho_1 = d_a^{-1}\sum_{d=1}^{d_a} \mathrm{Corr}(a_{t,d}, a_{t-1,d})$, using nonconstant action coordinates. All four datasets have strongly correlated adjacent actions and substantial fitted-policy sensitivity, with different task rankings for the two measurements.

Table 7: RoboMimic action correlation and post-training history sensitivity. Policies use $K = 3$, 20k updates, and the FM objective alone. Jacobian values are estimated on 256 fixed training windows per task with one uniform time in each quarter of $[0, 1]$, using exact input JVPs; entries give three-seed means and sample standard deviations. Actions use the training normalization.

Task	d_a	ρ_1	$\mathbb{E}\|\delta v_\theta/\delta H\|_F^2$
Can	7	.9659	1165.9±136.5
Square	7	.9527	1306.5±7.9
ToolHang	7	.9513	1578.1±181.9
Transport	14	.9720	1042.6±31.4

H.5 PALR and TCA Adaptations

We compare representation regularization with direct control of the velocity field's history sensitivity. PALR and TCA address past-action information in learned representations (Seo et al., 2023; Wen et al., 2020). We apply their auxiliary objectives to the FM policies used in our experiments, retaining the backbone, current observations, five past actions, action normalization, and future action-chunk target. TCA denotes the target-conditioned adversarial stage of the original method; its subsequent information-bottleneck stage is omitted. The auxiliary objectives replace the Jacobian penalty, and their coefficients are fixed at 1 for PALR and 0.1 for TCA across both benchmarks.

Policy representation. Let $\phi_\theta(S, H) \in \mathbb{R}^{d_\phi}$ be the flattened condition representation supplied to the FM velocity network. The LIBERO representation contains the encoded current visual and robot observation, five past actions, and the task condition, giving $d_\phi = 7 \times 512 = 3584$. The RoboMimic representation contains all sixteen projected condition positions, with invalid positions zeroed, giving $d_\phi = 16 \times 512 = 8192$. The auxiliary loss therefore acts on the representation used for action generation. LIBERO uses the current two camera images; RoboMimic uses the current low-dimensional observation. Both policies additionally receive five past actions. The expert current action a_t conditions the auxiliary objective during training only; inference uses (S, H).

PALR. We use the released estimator of the Hilbert–Schmidt conditional independence criterion (HSCIC), which measures dependence between the policy representation and a_{t-1} after conditioning on a_t. With the representation rescaled by $\sqrt{d_\phi}$, the policy objective is

$$\mathcal{L}_{\mathrm{PALR}}(\theta) = \mathcal{L}_{\mathrm{FM}}(\theta) + \widehat{\mathrm{HSCIC}}\Big(\phi_\theta(S, H)/\sqrt{d_\phi},\, a_{t-1} \mid a_t\Big). \tag{31}$$

All three Gaussian kernels use $k(x, y) = \exp(-\|x - y\|_2^2)$. The conditional estimator uses ridge 10^{-5}, added to the conditioning Gram matrix as $n\,10^{-5}I$ for a batch of n examples. Actions use the policy's training normalization. We evaluate the RoboMimic estimator on examples with a valid most recent action, using batches with at least two such examples. Our estimator follows the authors' public implementation, commit bdca9e0.

TCA. Let $g_\psi(\phi_\theta(S, H), a_t)$ predict the previous expert action. Its mean squared error is $\mathcal{L}_{\mathrm{adv}}(\theta, \psi) = d_a^{-1}\mathbb{E}\|g_\psi(\phi_\theta(S, H), a_t) - a_{t-1}\|_2^2$. Each training iteration first makes one adversary update minimizing $\mathcal{L}_{\mathrm{adv}}$ with the policy representation detached. It then freezes ψ and updates the policy by minimizing $\mathcal{L}_{\mathrm{TCA}}(\theta) = \mathcal{L}_{\mathrm{FM}}(\theta) - 0.1\mathcal{L}_{\mathrm{adv}}(\theta, \psi)$. This encourages the policy representation to suppress previous-action predictability beyond that provided by the current expert action. The adversary receives the unscaled representation concatenated with a_t, applies LeakyReLU, a 300-unit linear layer, LeakyReLU, and an action-dimensional linear output. The negative slope is 0.2. Adam uses initial learning rate 4×10^{-4} and moment coefficients $(0.5, 0.999)$; its learning rate follows the policy's relative decay. Representation noise is zero. The auxiliary predictor is discarded at inference. These steps follow the authors' TCA implementation, commit 52dabfd.

Training and evaluation. All baselines use seeds 42, 43, and 44, batch size 256, FP32, and AdamW for the policy, with initial learning rate 10^{-4}, weight decay 10^{-6}, and gradient clipping at norm 1. We train LIBERO policies with 25 demonstrations per task and a 60k cosine schedule ending at 5×10^{-6}, with EMA decay 0.999. We train RoboMimic policies with 200 PH demonstrations per task for 40k updates, with 100 warmup updates, cosine decay to 10^{-5}, and EMA decay 0.995. The RoboMimic backbone has width 512, six Transformer layers, and eight heads; the LIBERO backbone has width 512, three layers, and eight heads. Checkpoint steps count policy updates; TCA additionally performs one adversary update per policy update.

Evaluation uses EMA weights, one Flow sample, twenty Euler steps, and $N = 20, P = 10$. We evaluate the LIBERO 20k, 40k, and 60k checkpoints on the same bank of twenty newly generated initial states per task, preserving the bank's original gripper configuration and the 400-step limit (1,800 rollouts per seed). We evaluate the RoboMimic 20k and 40k checkpoints on the fixed 250-state bank per task used for the main comparisons. The episode limits are 250 for Can, 300 for Square, and 650 for ToolHang and Transport. Executed actions enter subsequent history inputs in both benchmarks. Tables in the main paper report LIBERO results at all three checkpoints and RoboMimic results at 20k; Table 8 supplies the remaining 40k results. Entries use sample standard deviations across training seeds.

Table 8: Adapted PALR and TCA on RoboMimic after 40k policy updates. Each entry reports success (%) over three seeds, with 250 rollouts per seed and $K = 5$.

Task	PALR	TCA
Can	98.9±0.6	98.9±0.5
Square	83.2±2.8	84.0±3.2
ToolHang	73.2±1.7	76.1±5.7
Transport	55.5±2.2	57.7±3.1

H.6 Action-Input Dropout Control

We compare random removal of past-action values with explicit history-sensitivity regularization. All policies in Table 9 use the joint Flow architecture with $K = 5$, 25 training demonstrations per task, batch size 128, seed 20260813, and the same 80k learning-rate schedule. The comparison takes checkpoints at 10k and 20k. Dropout independently sets normalized past-action coordinates to zero with probability 0.1 and divides retained values by 0.9; validity indicators remain unchanged. Dropout is disabled during evaluation. The Jacobian controls use one uniformly sampled Flow time and an automatic-differentiation directional derivative during training.

Evaluation uses EMA weights, one Flow sample, twenty Euler steps, $N = 20, P = 10$, and a 600-step limit. Each task starts from the recorded initial states of demonstrations 25–29, giving 450 rollouts per checkpoint. The data split, model, training budget, and evaluation conditions are matched within this comparison.

Table 9: Action-input dropout and history-sensitivity regularization on LIBERO-90. Each entry reports successes out of 450 rollouts and the corresponding percentage.

Training objective	Success@10k	Success@20k
FM alone	119/450 (26.4%)	219/450 (48.7%)
FM with action dropout, $p = 0.1$	133/450 (29.6%)	216/450 (48.0%)
FM with Jacobian, $\lambda_J = 0.01$	249/450 (55.3%)	324/450 (72.0%)
FM with Jacobian, $\lambda_J = 0.1$	244/450 (54.2%)	333/450 (74.0%)
FM with Jacobian, $\lambda_J = 1$	272/450 (60.4%)	333/450 (74.0%)

At 10k, action dropout improves success by 3.1 percentage points over FM alone. At 20k, the difference is -0.7 points, while the Jacobian controls improve success by 23.3–25.3 points. Thus history-sensitivity regularization produces a substantially larger benefit than action dropout in this matched configuration.

H.7 Repeated History Writeback and Environment Execution

We test the influence of an initial history error on subsequent predictions using 20k policies with $\lambda_J \in \{0, 1\}$, three training seeds, and $K \in \{1, 3, 5, 7\}$. Each model uses the same 172 starting conditions from 86 LIBERO-90 tasks with sufficiently long held-out trajectories. At replanning step r, the reference and perturbed branches receive the same expert condition S_r and Flow noise ξ_r. The reference branch starts from expert history; the perturbed branch starts from the final K actions of a model-generated prefix. Both subsequently write their own generated actions into history.

With $P = 10$, denote the generated prefix by $A_r^q = \pi_\theta^P(S_r, H_r^q; \xi_r)$, where $q \in \{\text{ref}, \text{wb}\}$ identifies the reference or perturbed branch. The next history is $H_{r+1}^q = [A_r^q]_{P-K:P-1}$. Define their action difference and cumulative mean squared deviation as

$$\delta_r^a = A_r^{\text{wb}} - A_r^{\text{ref}}, \qquad \mathcal{C}_R = \sum_{r=1}^{R} \frac{\|\delta_r^a\|_2^2}{Pd_a}. \tag{32}$$

These quantities compare two executions of the same policy. Table 10 gives the history-length breakdown. The initial perturbation is each model's own prediction error; its MSE, averaged over seeds, is larger for the regularized policy at every K. The main figure averages conditions and history lengths within each seed, then reports the mean and one standard deviation across seeds.

Table 10: Eight-step recursive writeback energy across history lengths. Each entry is the mean over three seeds and the same 172 starting conditions from 86 tasks, under matched states and Flow noise.

K	$\mathcal{C}_8$, $\lambda_J = 0$	$\mathcal{C}_8$, $\lambda_J = 1$	Reduction
1	.0971	.0175	82.0%
3	.0878	.0245	72.1%
5	.1541	.0387	74.9%
7	.1537	.0397	74.2%

Across three seeds and $K \in \{1, 3, 5, 7\}$, all twelve matched comparisons have lower eight-step recursive propagation under $\lambda_J = 1$; the mean relative reduction is 74.3%. An additional $K = 3$ comparison evaluates $\lambda_J = 0$ and $\lambda_J = 1$ over 180 paired sequences. The initial history errors are 0.1833 and 0.1842 RMSE. After four replanning steps, the cumulative mean squared propagation energies are 0.0795 and 0.0365, a reduction of 54.1%. Regularization gives lower cumulative propagation on 95.6% of starting points; a one-sided paired Wilcoxon test after task aggregation gives $p = 2.71 \times 10^{-10}$. Extending to eight replanning steps on 86 tasks gives cumulative energies 0.0936 and 0.0261, a reduction of 72.1%.

Environment execution. The main figure also reports simulator interventions with $K = 3$ policies at 20k, the same nine task initializations, and three training seeds. At environment step 29, we add a perturbation of Euclidean norm 0.1 to the six continuous action coordinates, keep the gripper command unchanged, and continue for 70 steps. One perturbed branch uses its own executed history; another receives the unperturbed branch's history at every replanning step. All branches use the same sampling noise. At the first replanning step, the two perturbed branches have the same state, so their action difference isolates the history response. PRIM lowers this response in all 27 task–seed conditions, with seed-averaged reductions of 91.7%–96.4%.

Along the complete perturbed execution, both state and history change. PRIM yields a smaller cumulative action deviation in 25 of the 27 conditions, with a mean reduction of 91.0%. Each point in the main figure summarizes ten subsequent environment steps; lines and shading report the mean and one standard deviation across seeds. We evaluate task completion separately in Appendix H.8.

An additional simulator comparison uses the same intervention time and continuation length. For $\lambda_J = 0$, replacing the recursively generated history with the unperturbed reference

history lowers cumulative action deviation from 0.2517 to 0.0217. For $\lambda_J = 1$, the complete cumulative deviation is 0.0678, 73.1% below the unregularized policy. At the first replanning step after intervention, the contribution associated with history writeback decreases from 0.04946 to 0.00177.

H.8 Task Completion After Shared History and Action Perturbations

We extend the action-deviation experiment to task success and give the two compared policies identical perturbations. The models are the seed42, $K = 3$, 60k EMA checkpoints trained with $\lambda_J = 0$ and $\lambda_J = 1$. The design fixes task indices $0, 3, \ldots, 87$ and initial-state indices 2 and 3 from the existing evaluation bank, giving 60 conditions. Intervention times are balanced across environment steps 20, 50, and 80. A source policy generates the shared trajectory prefix; its regularization setting is also balanced, with ten conditions for each of the six source-policy and intervention-time combinations. We recreate each branch by replaying the saved commands from the same initial state, thereby restoring the same simulator, controller, and gripper state before intervention.

At a branch time t, draw two independent Rademacher directions for each intervention type. A history direction has 18 continuous coordinates across the three past actions; a command direction has six continuous coordinates. Each is normalized to unit Euclidean norm and has zero gripper coordinates. The directions are shared between models and across amplitudes $r \in \{0.1, 0.3, 0.6\}$. A history intervention supplies $H + ru$ at the first replanning step only, in normalized history coordinates, with validity unchanged. Subsequent replanning steps use the actions actually executed by that branch.

For the command intervention, apply a common additive vector to the tenth action after the branch point, at step $t + P - 1$. To respect the command bounds, let $a_d^{(0)}$ and $a_d^{(1)}$ be the two models' clean commands in coordinate d. Set $\ell_d = \max(-1 - a_d^{(0)}, -1 - a_d^{(1)})$, $b_d = \min(1 - a_d^{(0)}, 1 - a_d^{(1)})$, and $\delta_d^a = \min(b_d, \max(\ell_d, ru_d))$. Both models then execute their own command plus this same δ^a, which is also written into history. The gripper coordinate remains zero. The mean realized norms are 0.09994, 0.29922, and 0.59597 for requested amplitudes 0.1, 0.3, and 0.6; their minima are 0.09289, 0.27440, and 0.54799. History amplitudes and command amplitudes are expressed in their respective input coordinates.

Each condition has one unperturbed execution per model and twelve perturbed executions per model: two intervention types, three amplitudes, and two directions. We evaluate all 1,560 executions with an absolute 400-step limit, applying each intervention before episode completion. Without perturbation, the models succeed on 41/60 and 50/60 conditions; 39 conditions are successful for both. The main paper reports success retained on this common subset. Table 11 includes all 60 conditions and their two directions.

Table 11: All starting conditions under common perturbations. Every row has 60 conditions and two directions per condition. Gains and confidence intervals are in percentage points.

Intervention	Amplitude	$\lambda_J = 0$	PRIM	Gain	95% interval
History	0.1	88/120 (73.3%)	99/120 (82.5%)	+9.2	[0.0,18.3]
History	0.3	86/120 (71.7%)	100/120 (83.3%)	+11.7	[1.7,22.5]
History	0.6	72/120 (60.0%)	100/120 (83.3%)	+23.3	[14.2,32.5]
Command	0.1	85/120 (70.8%)	95/120 (79.2%)	+8.3	[−2.5,20.8]
Command	0.3	68/120 (56.7%)	96/120 (80.0%)	+23.3	[13.3,33.3]
Command	0.6	64/120 (53.3%)	93/120 (77.5%)	+24.2	[15.8,31.7]

The two directions share an initial condition. We therefore compute percentile intervals from 2,000 bootstrap replicates that resample entire tasks, retaining all their initial states and directions. The corresponding gain intervals for the 39-condition common-success subset are [1.4,11.8], [7.5,25.6], and [10.3,33.8] points for history perturbations, and [−8.5,12.2], [11.5,35.1], and [14.3,35.5] points for command perturbations. These pointwise intervals are computed separately for each comparison using one training seed and two regularization set-

tings. The task-success measurements support robustness to the tested errors; the recursive experiment separately measures their subsequent action-deviation accumulation.

H.9 Representation Regularization and History Error Propagation

We test whether the PALR and TCA objectives in Appendix H.5 control the history contribution in Eq. (11). We freeze the LIBERO 60k EMA checkpoints of FM, PALR, TCA, and PRIM with $K = 5$ and seeds 42, 43, and 44. We select two distinct held-out demonstrations per task with sufficiently long continuations, giving 170 windows from 85 tasks. Tasks 1, 11, 21, 28, and 70 lack enough eligible windows. Every model uses the same windows, action normalization, Gaussian noise, and perturbation directions. Sampling uses twenty Euler steps, $N = 20$, and $P = 10$.

A common initial perturbation. Both branches start from the expert history. One receives a single Rademacher perturbation with physical-coordinate RMS 0.01 or 0.1 across the 5×7 history coordinates, including the gripper command; validity indicators are unchanged. Each branch uses its own generated history at subsequent replanning steps, while observations and robot states follow the same expert trajectory. Generated commands are clipped to $[-1, 1]$ before history writeback. The branches share a fresh noise draw at each replanning step. Let δ_r^a denote their difference over the first ten actions predicted at replanning step r; the main table reports $\mathbb{E}\|\delta_r^a\|_2^2/(Pd_a)$ and $\mathbb{E}\mathcal{C}_8$, where $\mathcal{C}_8 = \sum_{r=1}^{8} \|\delta_r^a\|_2^2/(Pd_a)$. These differences are measured in physical action coordinates. The expert observations isolate history feedback from environment-state feedback.

At amplitude 0.01, PALR and TCA reach their largest seed-averaged response at the third replanning step, respectively 3.82 and 4.81 times their initial response. Each of their three seeds also exhibits a later response larger than its first. PALR subsequently decays and has lower cumulative deviation than FM, indicating partial suppression; PRIM's response decays rapidly after the first replanning step. These observations establish finite-time amplification under the tested histories. Table 12 provides seed variation.

Table 12: LIBERO history-perturbation diagnostics at 60k. The first two measurements report means and sample standard deviations over three seeds; the last column reports means. Action differences use amplitude 0.01.

Method	Initial action MSE	$\mathbb{E}\mathcal{C}_8$	MSE at step 8, expert/self history
FM alone	.01527±.00118	.03842±.01277	.01771 / .06937
PALR	.01275±.00050	.02587±.00094	.01615 / .06213
TCA	.01404±.00189	.04350±.01928	.01598 / .06509
PRIM	.01814±.00240	$(1.96 \pm 1.29) \times 10^{-9}$	.03045 / .07544

The initial action MSE compares one generated action prefix with the expert prefix. The last column is a separate intervention: one branch receives the correct expert history at every replanning step, while another starts from the same correct history and then writes back its own actions. Their initial predictions agree exactly. At the eighth replanning step, using self-generated history increases expert-action MSE for all four methods, including PRIM. This larger history change can create observation–history combinations outside the expert distribution. Its effect differs from the response to a common small perturbation.

Response before and after the condition representation. Using the condition representation $\phi_\theta(S, H)$ defined in Appendix H.5, we estimate $\mathbb{E}\|(\delta\phi_\theta/\delta H)u\|_2^2$ and $\mathbb{E}\|(\delta v_\theta/\delta H)u\|_2^2$. Here u is a unit Rademacher direction over normalized history coordinates. Both use central finite differences with radius 0.01 on the same 170 windows. Velocity sensitivity covers all twenty predicted actions, with one uniform time sample from each quarter of $[0, 1]$. In the derivative limit, averaging these directional energies over u gives the corresponding squared Frobenius norm divided by Kd_a. We separately measure PALR's HSCIC on 64 common held-out windows per task, estimating it within each task and averaging over all 90 tasks.

Table 13: Condition dependence and history sensitivity of the same LIBERO checkpoints. Entries are means and sample standard deviations over three seeds.

Method	HSCIC	$\mathbb{E}\|(\delta\phi_\theta/\delta H)u\|_2^2$	$\mathbb{E}\|(\delta v_\theta/\delta H)u\|_2^2$
FM alone	$(2.36 \pm .14)\times10^{-3}$	2808.9±647.5	63.97±4.16
PALR	$(4.14 \pm .37)\times10^{-5}$	33.37±9.21	54.52±1.34
TCA	$(9.90 \pm .84)\times10^{-4}$	759.4±126.6	71.28±11.97
PRIM	$(2.11 \pm .17)\times10^{-3}$	.2418±.0186	$(2.14 \pm .43)\times10^{-4}$

PALR lowers HSCIC by 98.25% and representation response energy by 98.81%, while velocity response energy falls by 14.77%. Let F_θ denote the velocity network following the condition encoder, so that $v_\theta = F_\theta(X_\tau, \tau, \phi_\theta(S, H))$. The chain rule gives $\delta v_\theta/\delta H = (\delta F_\theta/\delta\phi_\theta)(\delta\phi_\theta/\delta H)$: stronger response in the later network can compensate for smaller representation changes. Consistently, the mean per-example ratio of velocity to representation response energy is 0.0228 for FM and 1.8038 for PALR. TCA also retains substantial velocity sensitivity despite a smaller representation response. Thus, the adapted representation objectives retain substantial output sensitivity.

We also measure the weighted auxiliary-gradient norm relative to the FM-gradient norm at the final EMA checkpoints, using the same 256 held-out examples and two noise/time assignments with frozen parameters and the saved TCA adversary. The ratios range from 0.000681 to 0.001013 for PALR and from 0.002482 to 0.003607 for TCA. The adversary's previous-action MSE is 0.00388–0.00432, compared with 0.00684 when the current expert action directly approximates the previous action. Previous-action information therefore remains readable from its inputs.

RoboMimic replication. We repeat the history interventions on Can, Square, ToolHang, and Transport using $K = 5$, 20k checkpoints, and three seeds per method. Each task supplies one window from each of forty distinct training demonstrations. FM/PRIM use a 20k learning-rate schedule, whereas the PALR/TCA checkpoints come from their 40k schedule. For each frozen model, the two branches share the same weights, current inputs, and sampling noise while using different action histories. Define ΔMSE_8 as the expert-action MSE with recursive self history minus the MSE with expert history restored at every replanning step. Table 14 reports this difference and the common-perturbation response.

Table 14: RoboMimic history diagnostics. Means and sample standard deviations over three seeds, all multiplied by 10^3; forty demonstration windows per task. Panel B uses a single initial perturbation of physical-coordinate RMS 0.01.

Task	FM alone	PALR	TCA	PRIM
A. Additional expert-action error, $10^3\Delta\text{MSE}_8$				
Can	.569±.125	1.104±.206	1.131±.159	.00222±.00194
Square	.455±.181	.435±.036	.513±.051	.00623±.00682
ToolHang	4.461±2.949	1.186±.650	2.129±.777	.221±.319
Transport	2.864±1.312	1.536±.845	.834±.412	.0649±.0471
B. Accumulated action difference, $10^3\mathbb{E}\mathcal{C}_8$				
Can	1.605±1.505	2.229±1.006	.710±.909	$(9.45 \pm 3.46)\times10^{-5}$
Square	6.759±2.424	3.782±.883	9.517±1.238	$(2.65 \pm 1.83)\times10^{-4}$
ToolHang	6.328±4.365	4.855±2.744	6.005±3.633	.00113±.00060
Transport	2.146±1.796	1.930±1.474	3.394±1.490	$(3.85 \pm .74)\times10^{-4}$

All 24 PALR/TCA task–seed models have positive additional error from recursive history at the eighth replanning step; PRIM has a smaller mean increase on each task. Partial suppression by PALR/TCA varies across tasks and perturbation amplitudes. The diagnostics establish continued history-error propagation and a gap between representation regularization and output sensitivity. However, PALR has lower initial action MSE and smaller accumulated response than FM on LIBERO while achieving lower task success.