DTSA: DUAL-TIMESCALE STABILITY AND ADAPTATION FOR ROBUST OFFLINE REINFORCEMENT LEARNING WITH SPIKING NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Combining offline reinforcement learning with spiking neural networks exposes a distributional instability that a state–action mismatch account does not capture: fixed data support interacts with time-varying neuronal dynamics. We formalize this as a dual-timescale shift: slow offline optimization drifts spike-conditioned action relations away from the supported structure, while deployment perturbations move a frozen policy into neural-dynamic regimes not covered by training. Treating the spike response induced by the decision history as an explicit condition variable, we propose DTSA, whose two mechanisms operate on one shared state–spike condition space: Conditional Structure Preservation (CSP) stabilizes spike-conditioned local action structure during training, and Adaptive Episodic Memory (AEM) corrects actions within an episode through confidence-gated, magnitude-limited fusion with all parameters frozen, functionally inspired by the stability–plasticity principles of biological learning. Both mechanisms are analyzed under a unified dual-timescale error decomposition. Experiments show that DTSA consistently improves the average normalized return across multiple benchmark tasks, by 18.1 points on Adroit over the matched spiking backbone, and exhibits better overall performance retention and distribution-shift robustness under observation noise, sensor dropout, and environment-dynamics shifts.

1 INTRODUCTION

Offline reinforcement learning (Offline RL) learns from previously collected data without further interaction (Prudencio et al., 2024). It is useful when exploration is costly or unsafe but sensitive to distribution shift: optimization may favor weakly supported behaviors whose errors cannot be corrected by new experience, the central problem of unsupported policy improvement (Fujimoto et al., 2019; Kumar et al., 2019; 2020; Fujimoto & Gu, 2021).

For policies built on spiking neural networks (SNNs), offline RL introduces an additional source of distributional mismatch. Temporally evolving membrane states and threshold-based firing make the current response depend on the present input and preceding activity. A small near-threshold membrane change can flip a spike and propagate it to the action (Maass, 1997; Zenke & Ganguli, 2018; Neftci et al., 2019). Offline RL with SNNs therefore exhibits a *dual-timescale shift*. Slow offline updates accumulate unsupported behavioral drift relative to the fixed data support (Kumar et al., 2019), while deployment perturbations rapidly move even a frozen controller into neural-dynamic conditions not covered during training (Morimoto & Doya, 2005; Zhang et al., 2021; Xu et al., 2025a). The two shifts are coupled along one decision chain: the state and history reshape the internal dynamics, and the spikes propagate the change to the action. Both therefore act on the same internal variable, the spike response. Biological learning governs the same tension on two timescales: slow homeostatic processes protect acquired behavior, and faster reversible ones absorb recent context (Turrigiano & Nelson, 2004; McClelland et al., 1995).

To address this problem, we propose DTSA, a dual-timescale stability–adaptation framework for robust offline RL with spiking neural networks. Both timescales act on one shared coordinate, the joint state–spike condition $x = (s, \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ is the time- and layer-averaged spike response induced by the decision history. On the slow timescale, Conditional Structure Preservation (CSP) constrains

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.

locally supported behavioral relations under joint conditioning while retaining the action-learning objective. On the fast timescale, Adaptive Episodic Memory (AEM) keeps parameters frozen and adapts within episodes through confidence-gated, magnitude-limited use of recent experience. Relations stabilized in training can therefore be recovered at deployment. The design mirrors the biological division of labor at the computational level only.

Across multiple benchmarks, DTSA consistently improves the average normalized return over the matched spiking backbone, by 18.1 points on Adroit. It also retains more performance under observation-noise, sensor-dropout, and environment-dynamics shifts.

We summarize our contributions as follows.

- We formalize a dual-timescale distributional shift for offline RL with spiking neural networks, in which fixed data support interacts with time-varying neuronal dynamics along one decision chain. We then show that both timescales act on one shared state–spike condition space.
- We propose DTSA, a dual-timescale framework that couples Conditional Structure Preservation (CSP), a spike-conditioned windowed kernel-divergence regularizer trained on the slow timescale, with Adaptive Episodic Memory (AEM), gated gradient-free transport that keeps every parameter frozen at deployment. Both mechanisms are analyzed under a unified error decomposition.
- Extensive experiments on three D4RL suites show that DTSA improves on its matched SNN backbone across standard benchmarks and under eight perturbations spanning four shift categories. Ablations and a bank-pairing intervention trace the gains to the two timescales.

2 Related Work

2.1 Offline Reinforcement Learning under Distribution Shift

Existing Offline RL methods confront distribution shift through behavior regularization, support constraints, conservative value estimation, or support-aware policy extraction (Kumar et al., 2019; Fujimoto et al., 2019; Kumar et al., 2020; Tarasov et al., 2023; Kostrikov et al., 2022; Wang et al., 2023; Fujimoto & Gu, 2021). These methods characterize mismatch through state–action statistics, policy–behavior divergence, action support, or value uncertainty (Agarwal et al., 2020; An et al., 2021). For temporally stateful controllers, however, the same state can induce different responses through the internal dynamical state, in SNNs via membrane integration and spike generation (Maass, 1997; Eshraghian et al., 2023; Neftci et al., 2019). We make this additional source of shift explicit, and keep deployment adaptation retrieval-based with all parameters frozen.

2.2 Offline Reinforcement Learning with Spiking Neural Networks

SNNs process information through discrete spikes and temporally evolving neuronal states (Maass, 1997; Eshraghian et al., 2023); in RL they range from population-coded actors to fully spiking networks (Tang et al., 2021; Amaya et al., 2024; Xu et al., 2025b; 2026; Huang et al., 2025). Responses depend on internal states, so perturbations alter spike patterns before the action (Zenke & Ganguli, 2018; Neftci et al., 2019). Most prior work targets representation, optimization, or efficiency (Tang et al., 2021; Xu et al., 2026; Huang et al., 2025; Yang et al., 2025; Yang & Chen, 2025; Yang et al., 2026), not the distribution shift of offline RL: the interaction of fixed offline data with time-varying spiking dynamics.

2.3 Homeostatic Plasticity in Biological Neural Systems

Biological neural systems must remain plastic enough to learn while preventing persistent activity changes from destabilizing function: homeostatic plasticity counteracts sustained deviations through changes in synaptic efficacy or intrinsic excitability (Marder & Goaillard, 2006; Turrigiano & Nelson, 2004; Turrigiano, 2008). This stability–plasticity view suggests that stabilization need not operate at the same timescale as learning: slower processes preserve accumulated behavior, while faster ones respond to recent context without rewriting long-term parameters (McClelland et al.,

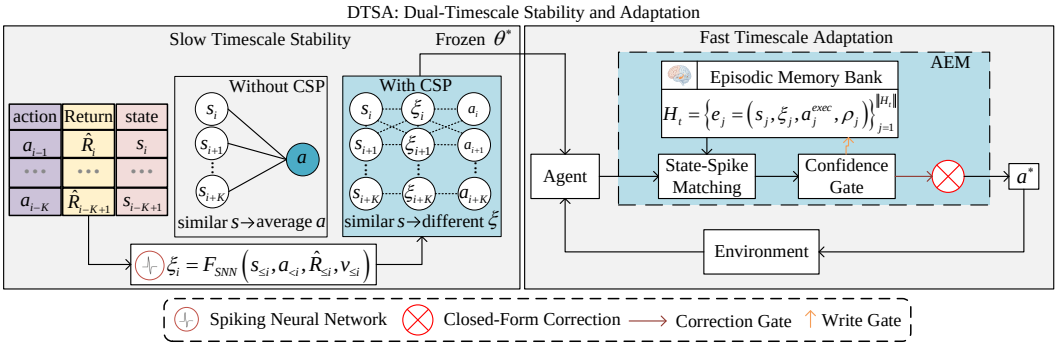


Figure 1: Overview of DTSA. **Slow timescale (left):** the action, return-to-go and state channels drive the spiking network, whose readout Eq. (1) is the spike response $\boldsymbol{\xi}$. Without CSP, nearby states collapse to one averaged action $\mathbf{a}$; with CSP, conditioning on $\boldsymbol{\xi}$ keeps them apart, and the trained policy passes to the right frozen at θ^*. **Fast timescale (right), the dashed AEM region:** the frozen agent queries the episodic bank H_t through state–spike matching, the confidence gate admits a closed-form correction that yields $\mathbf{a}^*$, and the write gate returns post-execution entries to the bank through the environment loop.

1995; Benna & Fusi, 2016; Doya, 2002). How the two rates should be instantiated in an offline spiking controller, and whether the division of labor helps under distribution shift, remains largely unexplored.

3 METHOD

DTSA's overall architecture appears in Figure 1. Both mechanisms instantiate the stability–plasticity principle of §2.3 at the computational level only. Complete proofs are deferred to Appendix H.

3.1 STATE–SPIKE CONDITIONING AND JOINT GEOMETRY

Offline RL setting. Consider a continuous-control MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with action box $\mathcal{A} = \{\mathbf{a} : \|\mathbf{a}\|_\infty \le 1\}$, $|r| \le \bar{r}$, $\bar{H} := \frac{1}{1-\gamma}$, and a fixed offline dataset $\mathcal{D} = \{\tau_n\}_{n=1}^N$, $\tau_n = (\hat{R}_i, s_i, a_i)_{i=1}^{T_n}$, generated by an unknown behavior policy μ ($\hat{R}_i$ the return-to-go token). A conditional sequence policy predicts the action from a length-K context $c_i = (\hat{R}_{i-K+1:i}, s_{i-K+1:i}, a_{i-K+1:i-1})$, $\hat{a}_i = \pi_\theta(c_i) \in \mathcal{A}$, with $J(\pi_\theta) = \mathbb{E}_{\pi_\theta} \sum_t \gamma^t r_t$. Offline, J cannot be evaluated by interaction, so μ is the only reference available.

Spike response as an internal condition. The backbone stacks L spiking layers, each running U internal steps with leaky integrate-and-fire (LIF) dynamics $u_t^l = \beta(1 - z_{t-1}^l)u_{t-1}^l + v_t^l$, $z_t^l = \mathbf{1}[u_t^l \ge v_{\text{th}}]$, $\beta \in (0,1)$, where v_t^l is the normalized pre-spike input, $z_t^l \in \{0,1\}$ the spike output, and the fired membrane is reset multiplicatively (the reset uses the detached previous spike, so gradients propagate within the current step only). The current spike reflects the accumulated input history, in the sense that it is produced by the recurrent dynamics above. We therefore read out a time- and layer-averaged spike response and condition on it jointly with the external state:

$$\boldsymbol{\xi}_i = \frac{1}{L}\sum_{l=1}^{L} \phi_l\Big(\frac{1}{U}\sum_{t=1}^{U} z_t^l\Big) \in \mathbb{R}^{d_\xi}, \qquad x_i = (s_i, \boldsymbol{\xi}_i), \tag{1}$$

with ϕ_l the dimension-alignment map. Throughout, subscripts ξ (dimension d_ξ, bandwidth σ_ξ, covariances $\Sigma_{\xi\xi}$) refer to this spike channel, written bold $\boldsymbol{\xi}$ (plain in Figure 1). $\boldsymbol{\xi}_i = F_{\text{SNN}}(s_{\le i}, a_{<i}, \hat{R}_{\le i}, v_{\le i})$ is an internal variable induced by the history, not an additional observation. Similar states may sit in different neural regimes, so s alone cannot certify shared local behavior. At deployment, $\boldsymbol{\xi}_t$ denotes the same readout at step t, always defined by Eq. (1), never an

instantaneous membrane state. Both timescales share one geometry on $\mathbf{x} = (s, \boldsymbol{\xi})$:

$$\begin{aligned} k^x_{ij} &= k^s_{ij}\, k^\xi_{ij} = \exp\!\big(-\tfrac{1}{2} d^2_x(i,j)\big), \\ d^2_x(i,j) &= \tfrac{\|s_i - s_j\|_2^2}{\sigma_s^2} + \tfrac{\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|_2^2}{\sigma_\xi^2}, \\ k^s_{ij} &= e^{-\|s_i - s_j\|_2^2/2\sigma_s^2}, \qquad k^\xi_{ij} = e^{-\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|_2^2/2\sigma_\xi^2}, \end{aligned} \tag{2}$$

with median-adaptive bandwidths σ_s, σ_ξ. The product kernel requires agreement in both channels. The same d_x drives CSP weighting in §3.2 and AEM retrieval in §3.3.

Throughout, two conventions fix the geometry: the readout Eq. (1) is recomputed from the current θ at every forward pass, with the BC gradient flowing through the same LIF spikes. Within a gradient step the conditioning path is detached (Algorithm 1), so all kernel and weight functionals are evaluated at the realized geometry. The bounds below are therefore conditional on that realized geometry, uniform over $\theta \in \Theta$ at it (Appendix H.1).

The analysis rests on three assumption blocks (*geometry*, *CSP regularity*, and *AEM estimation*), stated in full in Appendix H.1. Joint conditioning refines state-only conditioning: the product kernel Eq. (2) can only shrink local support (Proposition 1, Appendix H.1). A δ_0-sized perturbation, in turn, flips spikes with probability linear in the near-threshold membrane mass (Lemma 4, Appendix H). This is the fast-timescale failure mode that the AEM gates of §3.3 detect.

3.2 Slow timescale: Conditional Structure Preservation

CSP objective. Behavior cloning anchors pointwise prediction, $\mathcal{L}_{\mathrm{BC}}(\theta) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} m_i \|\mathbf{a}_i - \hat{\mathbf{a}}_i\|_2^2$ over valid positions $\mathcal{I}$, but is blind to relations between positions. Those relations are the policy's local action structure, which the CSP regularizer preserves. Within the local set $\mathcal{W}_i$ of Eq. (2), where $M_{ij} \in \{0,1\}$ masks valid causal pairs (one sampled context window, both positions valid) and ω_{ij} is the pair weight below, the weights are

$$w_{ij} = \frac{\omega_{ij}\, M_{ij}\, k^x_{ij}}{\sum_{k\in\mathcal{W}_i} \omega_{ik} M_{ik} k^x_{ik} + \epsilon}, \quad j \in \mathcal{W}_i, \qquad \omega_{ij} = \mathbf{1}[i \neq j] + \omega_0 \mathbf{1}[i = j], \tag{3}$$

and the Gaussian action kernel $\kappa(\mathbf{a}, \mathbf{a}') = \exp(-\|\mathbf{a} - \mathbf{a}'\|_2^2/2\sigma_a^2) = \langle \varphi(\mathbf{a}), \varphi(\mathbf{a}')\rangle_{\mathcal{H}_\mathrm{k}}$, $\|\varphi\|_{\mathcal{H}_\mathrm{k}} = 1$, $\mathcal{H}_\mathrm{k}$ its RKHS. The conditional embeddings $\hat{\mu}_p(x_i) = \sum_j w_{ij}\varphi(\mathbf{a}_j)$ (behavior) and $\hat{\mu}_q(x_i) = \sum_j w_{ij}\varphi(\hat{\mathbf{a}}_j)$ (policy) yield the plug-in potentials $V_p = \sum_j w_{ij}\kappa(\mathbf{a}_i, \mathbf{a}_j)$, $V_q = \sum_j w_{ij}\kappa(\hat{\mathbf{a}}_i, \hat{\mathbf{a}}_j)$, $V_{pq} = \sum_j w_{ij}\kappa(\mathbf{a}_i, \hat{\mathbf{a}}_j)$, and the conditional Cauchy–Schwarz (CS) structural loss

$$\mathcal{L}_{\mathrm{CSP}}(\theta) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \Big[\log V_p(x_i) + \log V_q(x_i) - 2\log V_{pq}(x_i)\Big]. \tag{4}$$

Each bracket estimates the normalized overlap $-2\log \frac{\langle \hat{\mu}_p, \hat{\mu}_q\rangle_{\mathcal{H}_\mathrm{k}}}{\|\hat{\mu}_p\|_{\mathcal{H}_\mathrm{k}}\|\hat{\mu}_q\|_{\mathcal{H}_\mathrm{k}}}$, which is scale-normalized, symmetric, and density-estimation-free, and has relation-preserving (not mean-collapsing) minimizers (Remark 1, Appendix H). The deployed objective $\theta^* = \arg\min_\theta\ \mathcal{L}_{\mathrm{BC}}(\theta) + \lambda_{\mathrm{CSP}}\mathcal{L}_{\mathrm{CSP}}(\theta)$ adds a warmup ramp, a relative-amplitude clamp $\lambda_{\mathrm{CSP}}(t)\mathcal{L}_{\mathrm{CSP}} \le \eta_{\mathrm{cs}}\,\mathrm{sg}(\mathcal{L}_{\mathrm{BC}})$, and a down-weighted self-anchor $\omega_0 < 1$ in Eq. (3) (Algorithm 1).

The loss controls what it targets. Its population form is non-negative, vanishing exactly when the policy's conditional action embedding is proportional to the behavior's. The squared conditional-embedding distortion $\mathbb{E}_x D^2(x)$, $D^2 := \|\mu_p - \mu_q\|^2_{\mathcal{H}_\mathrm{k}}$, is bounded by the loss plus a kernel floor (Appendix H). The loss value penalizes collapse toward averaged actions and is tracked as D_{CS} in §4.

Theorem 1 (CSP generalization and behavior recovery)**.** *Under Assumptions 1 and 2 and the frozen-conditioning convention, with probability $\ge 1-\delta$ over the N_w sampled context blocks ($n \asymp KN_w$ decision positions), the empirical minimizer $\hat{\theta}$ of $\widehat{\mathcal{L}}_{\mathrm{BC}} + \lambda_{\mathrm{CSP}}\widehat{\mathcal{L}}_{\mathrm{CSP}}$ satisfies*

$$\mathcal{L}_{\mathrm{BC}}(\hat{\theta}) + \lambda_{\mathrm{CSP}}\,\mathcal{L}_{\mathrm{CSP}}(\hat{\theta}) \ \le\ \sigma_B^2 + 2\,\varepsilon_{\mathrm{gen}}(n,\delta), \qquad \varepsilon_{\mathrm{gen}} = \widetilde{\mathcal{O}}\big(\sqrt{\bar{p}/n}\big), \tag{5}$$

the uniform generalization inequality behind this bound, with explicit constant C_{gen}, is Eq. (11) (Appendix H). Behavior is recovered, $J(\mu) - J(\pi_{\hat{\theta}}) \le C_R\sqrt{\sigma_B^2 + 2\varepsilon_{\mathrm{gen}}}$, with every target $\varepsilon > C_R\sigma_B$ met at sample cost polynomial in $\bar{p}$ and $1/\varepsilon$ relative to μ (Corollary 1, Appendix H).

3.3 FAST TIMESCALE: ADAPTIVE EPISODIC MEMORY

Episodic memory and joint retrieval. AEM keeps a fixed-capacity bank $H_t = \{e_j = (s_j, \boldsymbol{\xi}_j, \mathbf{a}_j^{\text{exec}}, \rho_j)\}$ with reliability-weighted replacement, cold-start warmup of h_0 steps (retrieval requires at least $h_{\min}$ entries), and write-after-execute causality. The frozen policy proposes $\mathbf{a}_t^0 = \pi_{\theta^*}(c_t)$ with spike response $\boldsymbol{\xi}_t$. Retrieval uses the same d_x Eq. (2) to select the top-K_r neighbors and form weights q_{tj} with adaptive bandwidth τ_t (the median positive neighbor distance), each multiplied by the stored reliability ρ_j (Eq. (12), Appendix H). Weighted moments with the $1 - \sum_j q_{tj}^2$ effective-sample-size correction give the ridge-stabilized covariance $\hat{\Sigma}_{\xi\xi}$ (λ_ξ conditioning near-binary spike directions). The closed-form transport $\hat{B}_t = \hat{\Sigma}_{a\xi}\hat{\Sigma}_{\xi\xi}^{-1}$ maps the neural-state deviation into the local conditional action $\tilde{\mathbf{a}}_t = \hat{\mu}_{a,t} + \alpha_\xi \hat{B}_t(\boldsymbol{\xi}_t - \hat{\mu}_{\xi,t})$ (Eq. (13), gain α_ξ=1 deployed), i.e., the ℓ_2-projection of the retrieved neighborhood onto linear spike–action couplings. The affine model class of Assumption 3(ii) is the estimator itself, so AEM adapts behaviorally rather than in parameters.

Condition (ii) of the AEM block ties departures from the affine form to an observable: the certainty gate uses the residual of this fit, $\Sigma_{a|\xi} = \hat{\Sigma}_{aa} - \hat{\Sigma}_{a\xi}\hat{\Sigma}_{\xi\xi}^{-1}\hat{\Sigma}_{\xi a}$. A curved or multi-valued relation therefore drives $g_{\text{act}} \to 0$, and the method abstains rather than extrapolating. The curvature bias $\rho_B r_\xi^2$ is retained as an explicit term, second order in the neighborhood radius. The behavioral relevance of $\boldsymbol{\xi}$ is analyzed in §4.

Transport, center, and model-bias errors combine into the correction budget ε_e, decaying as $\mathcal{O}(N_{\text{eff}}^{-1/2})$ up to log factors at the effective sample size (Lemma 2, Corollary 2, Appendix H).

Confidence gates. Three statistics decide whether local evidence may act: $g_{\text{sup}} = 1 - e^{-\bar{\rho}_t N_{\text{eff}}/\beta_{\text{sup}}}$ (effective support, $N_{\text{eff}} = 1/\sum_j q_{tj}^2$, $\bar{\rho}_t = \sum_j q_{tj}\rho_j$), $g_{\text{cov}} = e^{-r_t/2}$ with $r_t = d_{\text{nn},t}/\tau_t$ the nearest-memory distance over the median retrieval distance (Euclidean, since Mahalanobis is ill-conditioned on near-binary spike coordinates), and $g_{\text{act}} = \exp\big(-\text{tr}(\Sigma_{a|\xi})/(d_a \bar{v}_t \beta_{\text{act}})\big)$ with $\bar{v}_t$ the retrieved actions' mean per-dimension variance. The gate g_t is their product (Eq. (14)). Below δ_{read} the correction gate is closed and AEM abstains ($\mathbf{a}_t = \mathbf{a}_t^0$). Otherwise, the fusion step Eq. (15) is applied.

Two statements follow (Proposition 2, Appendix H.1). Below δ_{read} the controller reduces exactly to the frozen policy. On the active set, all three gate factors must pass their validation thresholds. The correction is then capped, $\|\mathbf{a}_t - \mathbf{a}_t^0\|_2 \le \rho_{\text{cap}}$, by a certified budget that saturates at $\rho_{\max} R_a$ and shrinks with the estimation part of ε_e as $\mathcal{O}(N_{\text{eff}}^{-1/2})$ (Eq. (9), Appendix H.1). Hence $\|\mathbf{a}_t - \mathbf{a}_{t-1}\|_2 \le \bar{\nu} + 2\rho_{\text{cap}}$ and $\mathbf{a}_t$ is non-anticipative.

Write admission and bounded fusion. The written value $\mathbf{a}_t^{\text{exec}} \equiv \mathbf{a}_t$ ($\mathbf{a}^*$ in Figure 1) may itself have been corrected. The loop is closed by design, and three coupled rules damp it (Algorithm 2). **(W1)** Admission is confidence-gated ($g_t \ge \delta_{\text{write}}$). Corrected steps additionally require the magnitude clip to be inactive, since a saturated clip signals an ill-conditioned local fit. Reliability is recorded at admission: ρ_t=1 for uncorrected steps and ρ_t=g_t<1 for corrected ones. **(W2)** Reliability persists into use, $q_{tj} \propto \rho_j\, e^{-d_x^2(t,j)/2\tau_t^2}$ (Eq. (12)), and at capacity the bank evicts $\arg\min_j \rho_j$. **(W3)** The bank is episode-local. Every admitted action lies within ρ_{cap} of the frozen policy's proposal (Proposition 2(ii)), so the retrieved neighborhood lies within the behavior-supported ρ_{cap}-tube around π_{θ^*}. An erroneous correction must pass all three gate factors repeatedly to enter, and then contributes at reduced weight. The executed action solves the three-anchor proximal problem Eq. (15) (Appendix H), clipped by $\text{SafeClip}_{\rho_{\max}}$ and projected onto the action box.

Corollary 3 (Appendix H) then bounds the fast timescale's value impact: $|\widetilde{J}(\tilde{\pi}) - \widetilde{J}(\pi_{\theta^*})| \le C_R \bar{\rho}$ with $\bar{\rho} \le \min\{\rho_{\text{cap}}, \rho_{\max} R_a \bar{p}_e\}$, a bounded, auditable perturbation of the slow-timescale guarantee.

3.4 UNIFIED DUAL-TIMESCALE ANALYSIS

Both timescales share the occupancy factorization $d^\pi(s, \mathbf{a}) = d_{\mathcal{X}}^\pi(x)\, \rho(\mathbf{a} \mid x)$ along $\mathbf{x} = (s, \boldsymbol{\xi})$: CSP stabilizes the conditional factor on the data condition set. AEM acts in the same geometry

Table 1: D4RL MuJoCo normalized scores (Fu et al., 2020) ("m", "m-r", "m-e": medium, medium-replay, medium-expert). Baselines as reported by Huang et al. (2025). DTSA is from our runs under the same protocol. Bold: best.

MuJoCo Tasks	BC	CQL	DT	FCNet	SpikeGPT	SpikeBERT	DSFormer (PSSA)	DTSA
halfcheetah-m-e	35.8	62.4	86.8±1.3	91.2±0.3	23.6±4.5	24.3±6.0	91.5±0.3	**93.2±0.2**
walker2d-m-e	6.4	98.7	108.1±0.2	108.8±0.1	22.6±4.8	92.5±22.4	108.9±0.1	**110.5±0.7**
hopper-m-e	111.9	111.0	107.6±1.8	110.5±0.5	32.7±5.4	84.1±8.8	110.9±0.2	**112.8±0.4**
halfcheetah-m	36.1	**44.4**	42.6±0.1	42.9±0.4	26.9±0.8	20.0±3.5	42.8±0.3	43.5±0.1
walker2d-m	6.6	79.2	74.0±1.4	75.2±0.5	16.4±10.2	22.9±10.4	75.2±1.4	**82.8±0.3**
hopper-m	29.0	58.0	67.6±1.0	57.8±6.0	25.1±6.4	31.4±4.9	74.1±4.3	**90.2±3.0**
halfcheetah-m-r	38.4	**46.2**	36.6±0.8	39.8±0.8	21.8±2.0	32.2±8.0	38.8±0.7	40.8±0.4
walker2d-m-r	11.3	26.7	66.6±3.0	63.5±7.5	16.7±3.3	21.2±6.4	71.0±3.6	**83.1±0.6**
hopper-m-r	11.8	48.6	82.7±7.0	85.8±1.7	51.5±7.1	30.1±8.6	96.3±1.3	**99.1±0.3**
MuJoCo mean ↑	31.9	63.9	74.7	75.1	26.4	39.9	78.8	**84.0**

when deployment moves $x \mapsto \tilde{x}$, the displacement carried by the L_Φ map and ε_P shifting the occupancy secondarily. Each term below isolates one error source.

Theorem 2 (Dual-timescale error decomposition)**.** *Under Assumptions 1–3, let $\bar{D}_{\mathcal{S}}$ be the diameter of the γL_P-stable recurrent state set and $\bar{r}_x := \mathcal{O}(\varepsilon_o L_\Phi)$ the condition-displacement budget under observation noise ε_o (Lemma 4(b), Appendix H). With probability $\geq 1-\delta$,*

$$\begin{aligned}\widetilde{J}(\mu) - \widetilde{J}(\tilde{\pi}) \leq\ & C_R\sqrt{\sigma_B^2 + 2\varepsilon_{\text{gen}}} + C_R\bar{\varsigma}\Big(\omega_\mu(\bar{r}_x) + \sqrt{\rho_\kappa^2 + \tfrac{2}{\lambda_{\text{CSP}}}\varepsilon_{\text{gen}}}\Big) \\ & + C_R\bar{\rho}(N_{\text{eff}}, \delta) + 2\bar{r}\bar{H}^2 L_P \bar{D}_{\mathcal{S}}\varepsilon_P + C_R\bar{\varsigma}\big(\omega_\mu(\bar{r}_x + \kappa_\rho\bar{\rho}) - \omega_\mu(\bar{r}_x)\big)\ ,\end{aligned} \tag{6}$$

The five terms are (I) slow optimization, (II) relation drift, (III) gate-validated correction, (IV) dynamics floor, and (V) slow–fast coupling. The constants ($\bar{\varsigma}(u) := \varsigma(2u)$, $\kappa_\rho := 1 + L_P\bar{H}$, $\bar{\rho}$ of Corollary 3) are collected in Appendix H.

Mechanism evidence in §4 tracks each term: the matched-backbone ablation quantifies (I), over the backbone, CSP adds 2.0 points on MuJoCo and 6.6 on Adroit, and AEM adds 2.0 and 6.9. D_{CS} and d_{support} probe (II), gate and correction statistics track (III), the frozen-policy floor under shift bounds (IV), and the pairing intervention quantifies (V) directly.

4 Experiments

4.1 Experimental setup

Benchmarks, baselines, and protocol. All experiments use the public D4RL benchmark (Fu et al., 2020) under its standard splits, with mechanism analyses on the Walker2d/Hopper medium and medium-replay tasks (task lists in Appendix C). MuJoCo and Adroit baselines follow Huang et al. (2025); on AntMaze, CQL, DT, and QDT follow Yamagata et al. (2023), while DSFormer (PSSA), the mechanism-free backbone w/o AEM & CSP, and DTSA are our runs under the matched protocol. DSFormer-PSSA is re-run from its official implementation under the same data and evaluation protocol; the mechanism-free backbone w/o AEM & CSP is the matched control that isolates the mechanism gain. The ablations w/o AEM and w/o CSP isolate the two mechanisms, and since AEM does not train, w/o AEM shares DTSA's checkpoint. Every result is mean±std over five random seeds, with three episodes per seed at each deployment level (eight shift settings in Table 7). Critic-based methods are discussed in Appendix D.1, pseudocode and proofs in Appendices B–H, and code and rollout logs will be released.

4.2 Standard offline performance

On MuJoCo (Table 1), DTSA reaches a mean normalized score of 84.0, the best average in the table by 8.9 points over the strongest ANN baseline (FCNet, 75.1) and by more than 44 points over the spiking models that do not share our backbone (SpikeGPT 26.4, SpikeBERT 39.9). It is the best method on 7 of the 9 locomotion tasks, the exceptions being the two halfcheetah splits, where CQL keeps the best mean. The 5.2-point gap over the matched control DSFormer-PSSA (78.8) is attributed rather than asserted: Table 3 splits it into 1.2 points of backbone and protocol difference, since the mechanism-free model already reaches 80.0, and 4.0 points from CSP and AEM.

Table 2: Results on D4RL Adroit tasks. "e", "h", and "c" denote expert, human, and cloned datasets (protocol as in Table 1).

Task	BC	CQL	DT	FCNet	SpikeGPT	SpikeBERT	DSFormer (PSSA)	DTSA
pen-e	85.1	107.0	110.4±20.9	108.0±11.3	30.5±10.3	46.2±19.5	122.0±17.8	**147.3±2.9**
door-e	34.9	101.5	95.5±5.7	102.9±2.9	65.3±16.9	96.4±4.6	105.2±0.1	**106.3±0.1**
hammer-e	125.6	86.7	89.7±24.6	121.1±6.1	51.1±18.7	71.3±16.5	127.2±0.3	**128.6±0.2**
relocate-e	101.3	95.0	15.3±3.6	50.0±6.0	0.7±0.9	0.3±0.5	108.4±2.2	**113.4±0.8**
pen-h	34.4	37.5	-0.2±1.8	57.7±11.1	29.8±11.7	20.0±16.7	75.7±25.1	**115.4±2.2**
door-h	0.5	9.9	0.1±0.0	0.4±0.5	0.1±0.0	0.2±0.0	0.2±0.0	**27.1±3.0**
hammer-h	1.5	4.4	0.3±0.0	1.2±0.0	0.3±0.0	0.3±0.0	0.2±0.0	**14.1±1.7**
relocate-h	0.0	0.2	0.2±0.2	0.0±0.0	0.1±0.0	0.0±0.0	0.0±0.0	**2.9±0.4**
pen-c	56.9	39.2	22.7±17.1	50.4±24.1	17.0±22.0	17.6±29.0	44.8±14.7	**108.2±0.6**
door-c	-0.1	0.4	0.1±0.0	0.2±0.0	0.2±0.0	0.2±0.0	0.0±0.0	**19.9±2.6**
hammer-c	0.8	2.1	0.3±0.0	0.2±0.0	0.3±0.0	0.3±0.5	0.2±0.0	**16.6±6.9**
relocate-c	-0.1	-0.1	**0.3±0.0**	0.2±0.0	0.1±0.0	0.0±0.5	-0.2±0.0	0.1±0.1
Adroit Mean ↑	36.7	40.3	27.8	41.0	16.3	21.1	48.6	**66.7**

Table 3: DTSA ablation on MuJoCo, averaged over five random seeds. The DTSA column differs slightly from Table 1.

MuJoCo Tasks	w/o AEM & CSP	w/o AEM	w/o CSP	DTSA
halfcheetah-m-e	91.5±0.2	92.3±0.1	92.1±0.2	**93.2±0.2**
walker2d-m-e	108.9±0.4	109.4±0.3	109.4±0.2	**110.5±0.7**
hopper-m-e	110.5±0.2	111.5±0.2	111.7±0.3	**112.8±0.4**
halfcheetah-m	43.1±0.2	43.3±0.2	43.3±0.1	**43.5±0.1**
walker2d-m	75.6±0.3	79.6±0.3	79.3±0.6	**82.8±0.3**
hopper-m	80.4±0.3	85.4±0.3	86.2±6.6	**90.2±0.3**
halfcheetah-m-r	38.7±0.0	39.5±0.0	39.6±0.2	**40.8±0.4**
walker2d-m-r	74.9±0.9	78.9±0.9	78.6±0.5	**83.1±0.6**
hopper-m-r	97.1±0.3	98.1±0.3	98.2±0.5	**99.7±0.3**
MuJoCo Mean ↑	80.0	82.0	82.0	**84.0**

Table 4: DTSA Ablation on Adroit (protocol as in left table).

Adroit Task	w/o AEM & CSP	w/o AEM	w/o CSP	DTSA
pen-e	125.8±0.9	136.0±0.9	141.6±3.3	**147.3±2.9**
door-e	105.2±0.2	106.3±0.0	106.2±0.1	**106.9±0.1**
hammer-e	127.2±0.3	127.7±0.5	127.8±0.8	**128.6±0.2**
relocate-e	107.9±1.0	111.2±0.8	111.7±0.9	**113.4±0.8**
pen-h	94.4±14.6	106.5±1.6	105.7±9.8	**115.4±2.2**
door-h	16.9±2.4	20.5±1.9	19.6±4.7	**27.1±3.0**
hammer-h	6.9±0.6	9.7±1.6	9.9±3.6	**14.1±1.7**
relocate-h	1.6±0.2	2.5±0.5	2.3±0.6	**3.0±0.4**
pen-c	86.3±3.2	106.3±1.2	105.6±10.1	**108.2±0.6**
door-c	5.5±2.4	19.6±2.3	18.9±2.0	**20.0±2.6**
hammer-c	3.5±0.4	13.5±2.9	11.4±9.2	**16.6±6.9**
relocate-c	-0.2±0.0	0.0±0.0	0.0±0.1	**0.1±0.1**
Adroit Mean	56.7	63.3	63.6	**66.7**

SpikeGPT and SpikeBERT trail on every task, so the gain is not explained by spiking computation alone.

The location of the advantage is as informative as its size. On Adroit (Table 2), DTSA is best on eleven of the twelve tasks and lifts the mean from the control's 48.6 to 66.7, a split of 8.1 and 10.0 points, with the widest margins on the splits that mix qualitatively different strategies within one dataset: pen-cloned 108.2 against 44.8, pen-human 115.4 against 75.7, and door-human 27.1 against 0.2. MuJoCo repeats the gradient, with hopper-medium at 90.2 against 74.1 and walker2d-medium-replay at 83.1 against 71.0, while on medium-expert every margin remains under two points (93.2, 110.5, and 112.8 against 91.5, 108.9, and 110.9). This pattern is consistent with joint conditioning: the more heterogeneous the supported behavior at a state, the more there is to preserve.

AntMaze (Table 5) is where the sequence baselines fail outright, DT scoring 31.0, 8.2, and 2.3. DTSA improves over its own mechanism-free backbone on all three tasks (37.3 to 49.3, 65.6 to 119.3, and 24.8 to 34.6) and exceeds the officially tuned DSFormer on each. On maze2d-medium it is the best method compared, 119.3 against QDT's 13.3, with AntMaze scores allowed above 100. CQL still leads umaze and large through critic regularization. Read together with Appendix G, the spiking policy also remains cheaper than the floating-point models it replaces, at 298.5 μJ per pass against 410.5 for DT and 1022.0 for FCNet.

Across the two scored suites DTSA is the best method on 18 of the 21 tasks. The three exceptions are informative rather than incidental: the two halfcheetah splits, where critic regularization suffices on a task whose supported behavior is close to unimodal, and relocate-cloned, where every method lands within 0.3 of zero and no ranking there is meaningful.

4.3 Component ablation and training-time mechanism

Tables 3 and 4 decompose the gains. Each timescale is independently effective over the backbone: CSP adds 2.0 points on MuJoCo and 6.6 on Adroit, and AEM adds 2.0 and 6.9. The full model exceeds either single mechanism. The complementarity grows with support dispersion: the

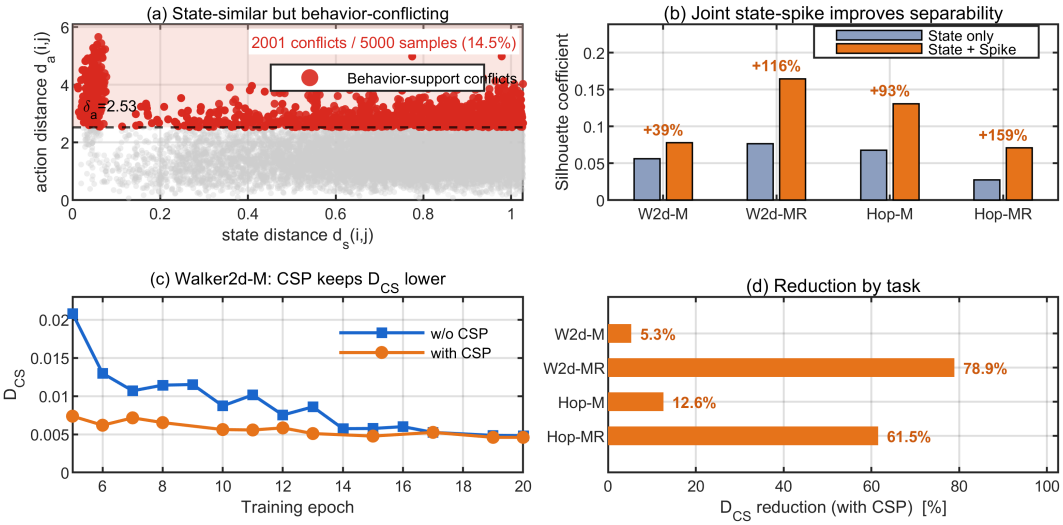


Figure 2: Training-time action path of CSP on the Walker2d and Hopper tasks. **(a)** State-nearest pairs (gray) vs. behavior-support conflicts (red), Walker2d-medium-replay (2001 of 5000 samples). **(b)** Silhouette of proxy behavior modes (state-only vs. state–spike, gains annotated). **(c)** Held-out D_{CS} on Walker2d-medium, both arms plotted to the last shared epoch, over which the CSP curve remains lower. **(d)** Matched-epoch reduction by task (Table 8).

Table 5: Results on AntMaze (maze2d). CQL, DT, and QDT as reported by Yamagata et al. (2023). DSFormer (PSSA), w/o AEM & CSP, and DTSA are our runs under the matched protocol, averaged over five random seeds. Bold: best.

Task	CQL	DT	QDT	DSFormer (PSSA)	w/o AEM & CSP	DTSA
maze2d-umaze	**94.7±23.1**	31.0±21.3	57.3±8.2	32.6±8.2	37.3±7.5	49.3±11.1
maze2d-medium	41.8±13.6	8.2±4.4	13.3±5.6	47.5±11.9	65.6±19.8	**119.3±17.7**
maze2d-large	**49.6±8.4**	2.3±0.9	31.0±19.8	23.1±3.9	24.8±5.7	34.6±10.0

slow timescale supplies a data-supported behavior prior, and the fast timescale adds deployment-conditioned corrections in the same condition space.

Figure 2 traces the CSP action path. **(a) Necessity**: on Walker2d-medium-replay, 2001 state-nearest pairs exceed $\delta_a \approx 2.53$ across 14.5% of samples, so state similarity does not certify compatible support. **(b) Conditioning**: silhouette separability improves with the spike condition on all four mechanism-analysis tasks ($+39\%$ to $+159\%$). **(c,d) Direct action**: at the shared training horizon, training with CSP reduces held-out D_{CS} on every mechanism task, by 5.3% on Walker2d-medium, 12.6% on Hopper-medium, and 61.5% on Hopper-medium-replay, while d_{support} changes by only -0.4% to $+2.5\%$ (Table 8, Appendix D): CSP preserves relations rather than point estimates. The remaining three tasks' curves are shown in Figure 4 (Appendix C).

Same-structure distribution-matching baselines. We test whether the CS form is itself critical by replacing Eq. (4) with three surrogates under identical windows, median bandwidths, detach conditions, and clamp budget η_{cs}: conditional MMD, in-window InfoNCE, and windowed Sinkhorn. The kernel-divergence family improves held-out D_{CS} on AntMaze monotonically in dose (paired t from 2.6 to 3.2). InfoNCE and Sinkhorn remain neutral, and plain conditional MMD is marginally ahead of the CS form. We therefore read Eq. (4) as one parameterization of the windowed kernel-divergence family and place the spike-conditioning advantage where it survives intervention: deployment retrieval (Appendix E.5).

4.4 Deployment Robustness and Behavioral Relevance of Spike Dynamics

With parameters frozen, deployment shifts compare the 2×2 matrix (Figure 3). DTSA attains the highest nominal return in all four panels and remains above the baselines at most severities, with gains concentrated in nominal and low-to-moderate regimes. Its AUC trails the baselines on Walker2d-m-r, and its R_{ID} is below w/o CSP on Hopper-m, reflecting checkpoint starting points. Adaptation is gradient-free and jitter-free (≈ 1.1 MB of state). Spike drift is similar across methods,

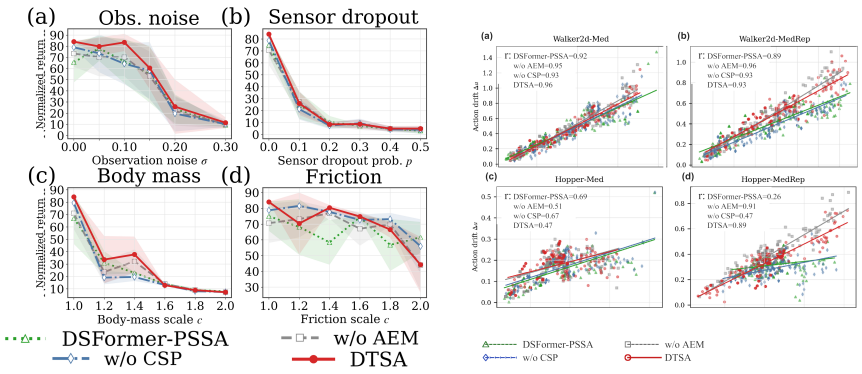


Figure 3: Deployment evaluation. **Left**: normalized return vs. shift severity on Walker2d-medium under (a) observation noise, (b) sensor dropout, (c) body mass, and (d) friction (shading: seed std). **Right**: per-episode $\Delta\boldsymbol{\xi}$ (horizontal) vs. $\Delta\mathbf{a}$ (vertical), $n = 180$/task, colored by method: DSFormer-PSSA (green), w/o AEM (gray), w/o CSP (blue), DTSA (red). Full values are in Table 12 (Appendix D).

but 10 of 16 combinations show per-episode $r \geq 0.88$ between $\Delta\boldsymbol{\xi}$ and action drift (all $p < 0.001$), most pronounced for CSP-trained runs on Hopper-m-r (0.91 and 0.89 against 0.47 and 0.26). This supports $\boldsymbol{\xi}$ as AEM's retrieval coordinate and checks the local-affine assumption, whose violations drive $g_{\text{act}} \to 0$. Tables 9–12 (Appendix D) give the aggregates, adaptation cost, and spike-drift diagnostics.

Bank-pairing intervention. Correlation alone cannot rule out that $\boldsymbol{\xi}$ acts as an extra feature rather than a condition variable. We therefore ran a bank-pairing intervention on antmaze-medium-diverse under a deterministic paired protocol with identical start poses and goals across arms. Gentle spike-conditioned transport adds 5.6 success points over the backbone (n=630, p=0.034), positive in all five seed blocks. Permuting the $(\boldsymbol{\xi}_j, \mathbf{a}_j)$ pairings in the bank (preserving bank size, distance structure, and write volume but breaking only the spike-to-action correspondence) erases the entire gain (-7.1 points, p=0.0018). The shuffled arm is statistically indistinguishable from the backbone, and a transport-off arm is bitwise identical to it. Write-gate statistics supply observable counterparts of ρ_{cap} and ε_e. Gate-threshold ablations stay within noise, so bank composition rather than the confidence threshold is the binding control. Protocol, dose–response, regularizer baselines, and the deterministic evaluation protocol are in Appendices E and F.

Summary of evidence. DTSA improves on its matched backbone across all three suites, with a mechanism share of 4.0 points on MuJoCo and 10.0 on Adroit. CSP acts structurally, since held-out D_{CS} falls while pointwise support distance remains flat. With parameters frozen, the gated bank keeps the highest nominal return under all four shift families, and the pairing intervention provides direct evidence that retrieval exploits the spike–action correspondence.

Limitations. Deployment robustness is evaluated through simulated benchmark shifts, not on a physical robot or in a neuromorphic-chip closed loop. The gradient-free, 1.1 MB footprint makes embedded deployment plausible but not demonstrated. Energy figures are analytical MAC/AC counts, not hardware measurements. Critic-based methods (IQL, TD3+BC, ReBRAC) are compared through published values under their own protocol (Appendix D.1) and are not re-trained under our spiking protocol here.

5 Conclusion

We studied the instability of offline reinforcement learning with spiking neural networks, where fixed data support interacts with time-varying neural dynamics, and formalized it as a dual-timescale shift. DTSA governs both timescales through one shared state–spike condition: CSP preserves behavior-supported action relations during training, and AEM applies gradient-free, gated, bounded corrections within an episode. Over the matched SNN control, DTSA lifts the mean normalized return from 78.8 to 84.0 on MuJoCo and from 48.6 to 66.7 on Adroit, with 4.0 and 10.0 tracing to the two mechanisms. Under observation noise, sensor dropout, and dynamics shifts, DTSA remains above the baselines at most severities.

AI USE STATEMENT

In this work, we used generative AI tools (GLM-5.3, developed by Zhipu AI; Qwen3.8-flash, developed by Alibaba Group; and DeepSeek-V4-Flash, developed by DeepSeek-AI) solely to aid or polish the writing of the manuscript: the tools were used for improving the quality of the language of the text, such as grammar, clarity, coherence, and readability. These tools have not played any role in the development of the ideas presented in the manuscript, in the research methodology, in the proofs, in the interpretation of the results, or in the generation of any data; no synthetic datasets were created. All AI-assisted edits were reviewed line by line and approved by the authors. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only the public D4RL benchmark datasets in simulation. It involves no human subjects and no personal data, releases no new datasets, and its algorithms are evaluated exclusively in MuJoCo, Adroit, and AntMaze simulations. We are not aware of conflicts of interest or of discrimination, privacy, safety, or compliance concerns raised by this submission.

REPRODUCIBILITY STATEMENT

All experiments use the public D4RL benchmark under its standard splits, and every reported result is averaged over five random seeds with the episode, checkpoint-selection, and shift-level protocol described in §4.1. The training and deployment algorithms are given in Appendix B, hyperparameters and shift-setting definitions in Appendix C, per-task results and diagnostics in Appendix D, the bank-pairing intervention protocol in Appendix E, and the deterministic evaluation protocol of the deployment stack in Appendix F. The energy figures in Appendix G are analytical estimates computed by the released estimator, and all theoretical statements carry complete proofs in Appendix H. Code and rollout logs will be released publicly.

REFERENCES

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pp. 104–114. PMLR, 2020.

Camilo Amaya, Evan Eames, Gintautas Palinauskas, Alexander Perzylo, Yulia Sandamirskaya, and Axel von Arnim. Neuromorphic force-control in an industrial task: Validating energy and latency benefits. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 717–724, 2024. doi: 10.1109/IROS58592.2024.10802430.

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified Q-ensemble. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7436–7447, 2021.

Peter L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

Marcus K. Benna and Stefano Fusi. Computational principles of synaptic memory consolidation. *Nature Neuroscience*, 19(12):1697–1706, 2016.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Kenji Doya. Metalearning and neuromodulation. *Neural Networks*, 15(4-6):495–506, 2002.

Jason K. Eshraghian, Max Ward, Emre O. Neftci, Xinxin Wang, Gregor Lenz, Geetika Dwivedi, Mohammed Bennamoun, Doo Seok Jeong, and Wei D. Lu. Training spiking neural networks using lessons from deep learning. *Proceedings of the IEEE*, 111(9):1016–1054, 2023.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20132–20145, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, volume 97, pp. 2052–2062, 2019.

Mark Horowitz. Computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference (ISSCC)*, 2014.

Wei Huang, Qinying Gu, and Nanyang Ye. Decision SpikeFormer: Spike-driven transformer for decision making. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19241–19250, 2025. doi: 10.1109/CVPR52734.2025.01792.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191, 2020.

Wolfgang Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659–1671, 1997.

Eve Marder and Jean-Marc Goaillard. Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience*, 7(7):563–574, 2006.

James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995.

Jun Morimoto and Kenji Doya. Robust reinforcement learning. *Neural Computation*, 17(2):335–359, 2005.

Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

Rafael Figueiredo Prudencio, Marcos R. O. A. Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10237–10257, 2024. doi: 10.1109/TNNLS.2023.3250269.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635, 2011.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Guangzhi Tang, Neelesh Kumar, Raymond Yoo, and Konstantinos P. Michmizos. Deep reinforcement learning with population-coded spiking neural network for continuous control. In *Proceedings of the Conference on Robot Learning*, volume 155, pp. 2016–2029, 2021.

Denis Tarasov, Vladislav Kurenkov, Alexander Nikulin, and Sergey Kolesnikov. Revisiting the minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11592–11620, 2023.

Gina G. Turrigiano. The self-tuning neuron: Synaptic scaling of excitatory synapses. *Cell*, 135(3): 422–435, 2008.

Gina G. Turrigiano and Sacha B. Nelson. Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience*, 5(2):97–107, 2004.

Roman Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *International Conference on Learning Representations*, 2023.

Jiawei Xu, Rui Yang, Shuang Qiu, Feng Luo, Meng Fang, Baoxiang Wang, and Lei Han. Tackling data corruption in offline reinforcement learning via sequence modeling. In *International Conference on Learning Representations*, 2025a.

Zijie Xu, Tong Bu, Zecheng Hao, Jianhao Ding, and Zhaofei Yu. Proxy target: Bridging the gap between discrete spiking neural networks and continuous control. In *Advances in Neural Information Processing Systems*, volume 38, pp. 176646–176672, 2025b.

Zijie Xu, Xinyu Shi, Yiting Dong, Zihan Huang, and Zhaofei Yu. CaRe-BN: Precise moving statistics for stabilizing spiking neural networks in reinforcement learning. In *International Conference on Learning Representations*, 2026.

Taku Yamagata, Ahmed Khalil, and Raul Santos-Rodriguez. Q-learning decision transformer: Leveraging dynamic programming for conditional sequence modelling in offline RL. In *International Conference on Machine Learning*, pp. 38989–39007. PMLR, 2023.

Shuangming Yang and Badong Chen. Effective surrogate gradient learning with high-order information bottleneck for spike-based machine intelligence. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):1734–1748, 2025.

Shuangming Yang, Bernabé Linares-Barranco, Yuzhu Wu, and Badong Chen. Self-supervised high-order information bottleneck learning of spiking neural network for robust event-based optical flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2280–2297, 2025.

Shuangming Yang, Jianshi Tang, Tianhang Zhou, Bernabé Linares-Barranco, Giacomo Indiveri, Xinhu Zheng, Mostafa Rahimi Azghadi, and Badong Chen. Efficient training of neuromorphic electronics. *Nature Electronics*, 9(6):594–610, 2026.

Friedemann Zenke and Surya Ganguli. Superspike: Supervised learning in multilayer spiking neural networks. *Neural Computation*, 30(6):1514–1541, 2018.

Huan Zhang, Hongge Chen, Duane Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. In *International Conference on Learning Representations*, 2021.

A NOTATION

Table 6 collects the symbols used throughout the paper, grouped by the mechanism that introduces them. Constants that appear only inside a single assumption or proof step are defined locally where they occur.

Table 6: Global notation. "Slow" = CSP (training timescale); "fast" = AEM (deployment timescale).

Symbol	Meaning
Problem and data (§3.1)	
$\mathcal{S}, \mathcal{A}, P, r, \gamma$	MDP; action box $\|\mathbf{a}\|_\infty \le 1$; discount
$\bar{r},\ \bar{H}{=}(1{-}\gamma)^{-1}$	reward bound; effective horizon
$\mathcal{D} = \{\tau_n\}_{n=1}^N$	offline dataset; trajectories $\tau_n = (\hat{R}_i, s_i, a_i)$
μ	unknown behavior policy (only reference available)
$\hat{R}_i,\ K,\ c_i$	return-to-go token; context length; context window
$\pi_\theta,\ \hat{a}_i{=}\pi_\theta(c_i),\ J(\pi)$	conditional sequence policy, prediction, objective
Spike channel and joint geometry (§3.1)	
$L,\ U,\ \beta,\ v_t^l,\ u_t^l,\ z_t^l,\ v_{\mathrm{th}}$	layers, internal steps; LIF input, membrane, spike, threshold
ϕ_l	dimension-alignment map per layer
F_{SNN}	spike-response encoder: readout Eq. (1) as a function of the history
$\boldsymbol{\xi}_i \in \mathbb{R}^{d_\xi}$	spike response readout Eq. (1) (time/layer-averaged)
$\mathbf{x} = (s, \boldsymbol{\xi})$	joint condition variable shared by both timescales
$k_{ij}^x,\ d_x,\ \sigma_s,\ \sigma_\xi$	product kernel Eq. (2), joint distance, median bandwidths
Slow timescale: CSP (§3.2)	
$\mathcal{W}_i,\ M_{ij},\ \omega_{ij},\ \omega_0$	local set; validity mask; pair weights (self-anchor ω_0)
w_{ij}	CSP weights Eq. (3)
$\kappa(\mathbf{a}, \mathbf{a}'),\ \varphi(\mathbf{a}),\ \mathcal{H}_{\mathrm{k}}$	Gaussian action kernel, feature map, RKHS
$\hat{\mu}_p(x_i),\ \hat{\mu}_q(x_i)$	conditional action embeddings (behavior / policy)
$V_p,\ V_q,\ V_{pq}$	plug-in potentials
$\mathcal{L}_{\mathrm{BC}},\ \mathcal{L}_{\mathrm{CSP}}$	BC anchor; conditional CS structural loss Eq. (4)
$\lambda_{\mathrm{CSP}},\ \eta_{\mathrm{cs}},\ \epsilon$	CSP weight (warmup+clamp); clamp ratio; denominator floor
$D^2{=}\|\mu_p{-}\mu_q\|_{\mathcal{H}_{\mathrm{k}}}^2,\ D_{\mathrm{CS}}$	conditional-embedding distortion; held-out estimator
$\sigma_B^2,\ \varepsilon_{\mathrm{gen}},\ \bar{p}$	behavior variance; generalization term; min local mass
$N_w,\ n \asymp K N_w,\ C_{\mathrm{gen}},\ C_R$	context blocks; decision positions; bound constants
Fast timescale: AEM (§3.3)	
$\theta^*,\ \mathbf{a}_t^0{=}\pi_{\theta^*}(c_t)$	frozen policy; proposed action
$H_t,\ e_j{=}(s_j, \boldsymbol{\xi}_j, \mathbf{a}_j^{\mathrm{exec}}, \rho_j)$	episodic bank; entries with executed action, reliability
$h_0,\ h_{\min},\ K_r$	warmup steps; min bank size; retrieval neighbors
$q_{tj},\ \tau_t$	retrieval weights; adaptive bandwidth (median distance)
$\hat{\Sigma}_{\xi\xi},\ \lambda_\xi$	ridge-stabilized spike covariance (λ_ξ conditioning)
$\hat{B}_t{=}\hat{\Sigma}_{a\xi}\hat{\Sigma}_{\xi\xi}^{-1},\ \alpha_\xi$	closed-form transport; gain ($=1$ deployed)
$\tilde{\mathbf{a}}_t,\ \hat{\mu}_{a,t}$	locally conditioned action; retrieved action mean
$\Sigma_{a\|\xi}$	residual covariance of the affine fit (gate statistic)
$g_{\mathrm{sup}},\ g_{\mathrm{cov}},\ g_{\mathrm{act}},\ g_t$	support / coverage / actuation gates; product g_t
$N_{\mathrm{eff}}{=}1/\sum_j q_{tj}^2,\ \bar{\rho}_t{=}\sum_j q_{tj}\rho_j$	effective sample size; reliability-weighted mass
$\beta_{\mathrm{sup}},\ \beta_{\mathrm{act}},\ \bar{v}_t,\ r_t,\ d_{\mathrm{nn}}$	gate scales; retrieved action variance; coverage ratio
$\delta_{\mathrm{read}},\ \delta_{\mathrm{write}}$	abstention and write-admission thresholds
$\rho_{\mathrm{cap}},\ \rho_{\max},\ R_a$	certified correction budget; clip fraction; action radius
$\mathbf{a}_t,\ \varepsilon_e,\ \bar{\nu}$	executed action; correction budget; action smoothness
Unified analysis (§3.4, Theorem 2)	
$\widetilde{J}$	regularized (occupancy-based) objective
$d^\pi(s, \mathbf{a}){=}d_{\mathcal{X}}^\pi(x)\rho(\mathbf{a} \mid x)$	occupancy factorization along $\mathbf{x}$
$\varepsilon_o,\ L_\Phi,\ \bar{r}_x$	observation noise; displacement map constant; budget
$\varepsilon_P,\ L_P,\ \bar{D}_{\mathcal{S}}$	transition perturbation; Lipschitz; stable-set diameter
$\omega_\mu(\cdot)$	modulus of continuity of the behavior conditional
$\bar{\varsigma}(u){=}\varsigma(2u),\ \kappa_\rho{=}1{+}L_P\bar{H},\ \rho_\kappa$	coupling constants (collected in Appendix H)
$\bar{\rho} \le \min\{\rho_{\mathrm{cap}}, \rho_{\max} R_a \bar{p}_e\}$	fast-timescale value-perturbation bound

B Training and Deployment Algorithms

This section gives the pseudocode for the two execution paths. Algorithm 1 details the slow-timescale CSP training loop, and Algorithm 2 the fast-timescale gradient-free AEM deployment loop.

Algorithm 1 DTSA–SLOW: offline CSP training

Require: dataset $\mathcal{D}$; window K; local set size W; self weight ω_0; schedule $(t_w, \lambda_{\mathrm{CSP}}^{\max}, T)$; ratio cap η_{cs}
Ensure: frozen policy π_{θ^*} and encoder F_{SNN}
1: **for** minibatch $\{(c_i, \mathbf{a}_i)\}_{i\in\mathcal{I}}$ at step t **do**
2: forward through PSSA/LIF layers with the current θ; collect $\hat{\mathbf{a}}_i = \pi_\theta(c_i)$ and $\boldsymbol{\xi}_i$ via the spike readout, detached on the condition path only (Eq. (1): the BC gradient still flows through the LIF spikes via the surrogate rule)
3: $\sigma_s, \sigma_\xi, \sigma_a \leftarrow$ per-batch median pairwise distances on $\{s_i\}, \{\boldsymbol{\xi}_i\}, \{\mathbf{a}_i\}$ (subsampled to a cap if the batch is large) ▷ adaptive bandwidths
4: $k_{ij}^x \leftarrow$ Eq. (2) on causal local sets $\mathcal{W}_i$ with mask M_{ij}; $w_{ij} \leftarrow$ Eq. (3)
5: $V_p, V_q, V_{pq} \leftarrow$ the plug-in potentials; $\mathcal{L}_{\mathrm{CSP}} \leftarrow$ Eq. (4); $\mathcal{L}_{\mathrm{BC}}$ the masked objective over $\mathcal{I}$
6: $\lambda_{\mathrm{CSP}}(t) \leftarrow 0$ if $t < t_w$, else $\lambda_{\mathrm{CSP}}^{\max} \min\{\frac{t-t_w}{T-t_w}, 1\}$; clamp $\lambda_{\mathrm{CSP}}(t)\mathcal{L}_{\mathrm{CSP}} \le \eta_{\mathrm{cs}}\, \mathrm{sg}(\mathcal{L}_{\mathrm{BC}})$
7: gradient step on $\mathcal{L}_{\mathrm{BC}} + \lambda_{\mathrm{CSP}}(t)\mathcal{L}_{\mathrm{CSP}}$ (global norm clipped)
8: **end for**
9: **return** θ^* frozen

Algorithm 2 DTSA–FAST: gradient-free AEM deployment

Require: frozen π_{θ^*}, encoder F_{SNN}; K_r, capacity C_{bank}, warmup h_0; gates $(\beta_{\mathrm{sup}}, \beta_{\mathrm{act}}, \delta_{\mathrm{read}}, \delta_{\mathrm{write}})$; fusion $(\lambda_\pi, \lambda_e, \lambda_{\mathrm{T}}, \lambda_a, \lambda_\xi, \rho_{\max})$
Ensure: executed action sequence $\{\mathbf{a}_t\}$
1: initialize an episode-local bank $H_\emptyset$ (re-initialized at every episode start); run h_0 steps storing $(s, \boldsymbol{\xi}, \mathbf{a}^0, \rho{=}1)$ ▷ cold-start consolidation
2: **for** each decision step t **do**
3: $(\mathbf{a}_t^0, \boldsymbol{\xi}_t) \leftarrow \pi_{\theta^*}(c_t)$ ▷ single forward pass; no gradients
4: **if** $|H_t| < h_{\min}$ **then**
5: $\mathbf{a}_t \leftarrow \mathbf{a}_t^0$ ▷ insufficient memory: the slow timescale governs
6: **else**
7: $\mathcal{N}_t, \{q_{tj}\} \leftarrow$ Eq. (12) with $\tau_t = \mathrm{median}$ top-K_r distances
8: weighted moments, transport, and conditional action Eq. (13); gates Eq. (14)
9: **if** $g_t < \delta_{\mathrm{read}}$ **then**
10: $\mathbf{a}_t \leftarrow \mathbf{a}_t^0$ ▷ abstain (Proposition 2(i))
11: **else**
12: solve $A_t \mathbf{a}_t^\star = \mathbf{b}_t$ Eq. (15); $\mathbf{a}_t \leftarrow \Pi_{\mathcal{A}}\big(\mathbf{a}_t^0 + \mathrm{SafeClip}_{\rho_{\max}}(\mathbf{a}_t^\star - \mathbf{a}_t^0)\big)$ Eq. (15)
13: **end if**
14: **end if**
15: observe (r_t, s_{t+1}); write $e_t = (s_t, \boldsymbol{\xi}_t, \mathbf{a}_t^{\mathrm{exec}}, \rho_t)$ if $g_t \ge \delta_{\mathrm{write}}$, with $\delta_{\mathrm{write}} \le \delta_{\mathrm{read}}$ (uncorrected steps gated too) and, if corrected, the clip was inactive ($\|\mathbf{a}_t - \mathbf{a}_t^0\| < \rho_{\max}\|\mathbf{a}_t^0\|$); set $\rho_t = 1$ if uncorrected, $\rho_t = g_t$ if corrected; evict $\arg\min \rho_j$ at capacity ▷ post-execution, damped self-reinforcement (W1–W3); no leakage: $e_t \notin H_t$
16: **end for**

C Implementation and Experimental Details

This appendix collects the deployment-shift settings and protocol details referenced from §4.1. Full hyperparameters follow the descriptions in §4.1 and Algorithm 1–2.

Evaluation uses D4RL (Fu et al., 2020): MuJoCo (HalfCheetah, Walker2d, Hopper × medium-expert/medium/medium-replay) and Adroit (Pen, Door, Hammer, Relocate × expert/human/cloned). The mechanism analyses focus on the Walker2d/Hopper medium and medium-replay tasks, whose mixed experience most clearly exposes the ambiguity between state similarity and behavioral conflict. Baseline provenance and the five-seed protocol are given in §4.1 and the Table 5 caption.

Table 7: Deployment-shift settings. Eight perturbations in four groups. The main text reports observation noise, sensor dropout, body-mass scaling, and friction scaling, and uses the remaining four as supplementary checks.

Category	**Shift type**
Observation	sensor bias sensor dropout observation noise
Dynamics	body-mass scaling friction scaling gravity scaling
Initial condition	initial-state perturbation
Composite	noise + mass composite shift

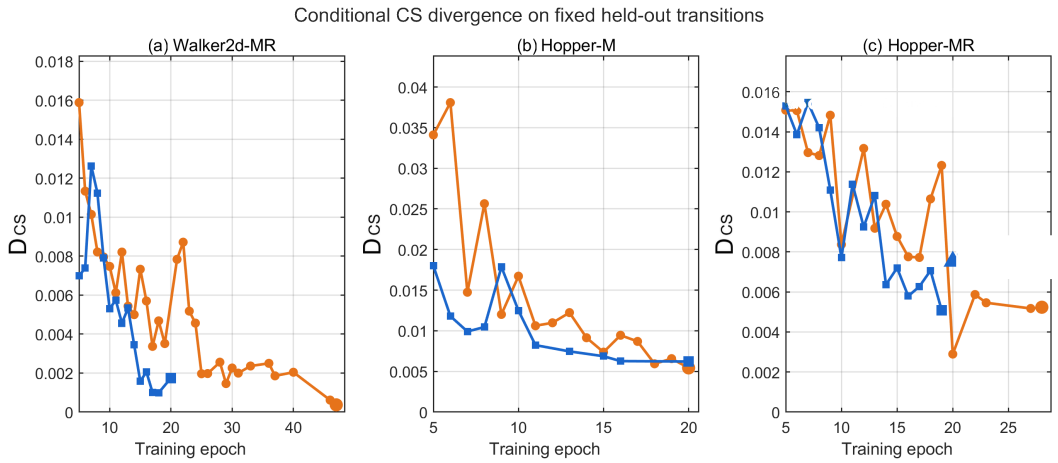


Figure 4: Held-out conditional CS divergence during training on Walker2d-medium-replay, Hopper-medium, and Hopper-medium-replay (blue: w/o CSP, orange: with CSP; see §4.1). The Walker2d-medium curves and the per-task reduction summary are shown in Figure 2.

Ablation configurations and the shared-checkpoint protocol are given in §4.1. Checkpoints are selected by nominal score and then frozen. Reported metrics are mean±std D4RL score, and under deployment R_{ID}, R_{OOD}, retention, ΔR, the return–severity AUC, and action variation and jerk.

The AEM configuration is fixed (capacity 1024, $K_r = 48$, $h_{\min} = 8$, δ_{write}=0.25 $\leq$ δ_{read}, 64-step cold start, $\rho_{\max}$=0.25, action-box projection), and admission follows W1–W3 of §3.3. Fusion weights are selected per task–shift-family under nominal conditions only, and each level runs five random seeds $\times$ three episodes.

D Additional Results

This appendix reports the per-task ablations, robustness aggregates, adaptation-cost accounting, and spike-dynamics diagnostics that support the claims of §4.3 and §4.4. Figure 5 visualizes the conflict samples on Walker2d-medium-replay under the two conditioning schemes.

D.1 Critic-based ANN methods: published averages

The SNN offline-RL literature positions itself against imitation and matched SNN controls: Huang et al. (2025) reports BC, CQL, DT, and spiking baselines and evaluates no critic-based method. We therefore position DTSA against IQL, TD3+BC, and ReBRAC through published numbers rather than re-implementation. These rows are kept out of Table 1, whose matched-backbone comparison carries the paper's mechanism claim. Table 13 lists the Gym-MuJoCo scores of Tarasov et al. (2023) (their Table 2): all three methods re-trained under one protocol (v2 datasets, their evaluator, ten seeds, ANN actor–critic backbones). The corresponding DTSA and DSFormer-PSSA means

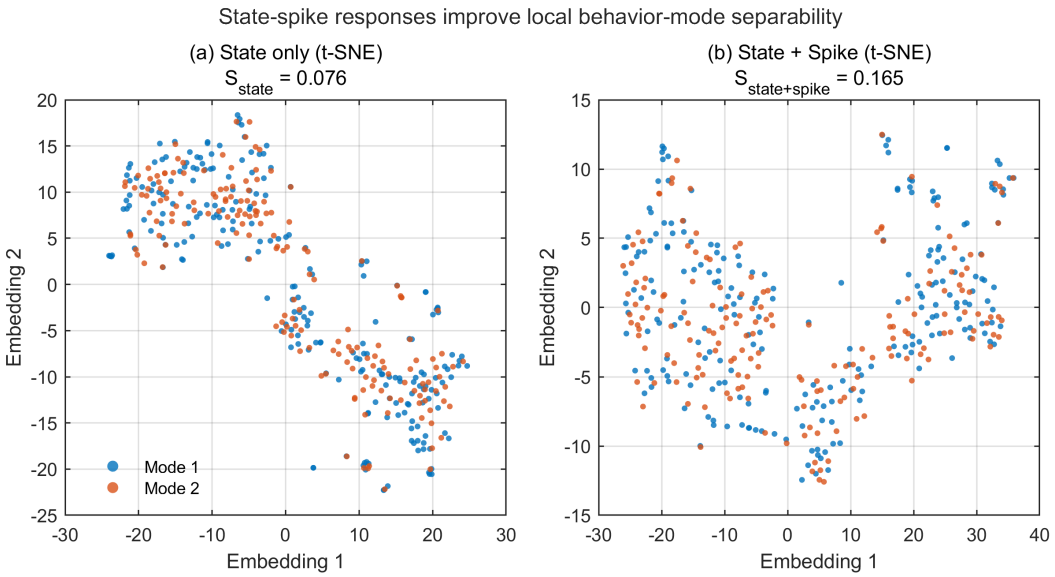


Figure 5: Separation of local behavior modes under state-only vs. state–spike conditioning: t-SNE of the conflict samples on Walker2d-medium-replay, colored by proxy behavior-mode labels. The silhouette gains are annotated in Figure 2(b).

Table 8: Training-time preservation of local behavior support on held-out transitions, with the two arms compared at a shared training horizon. D_{CS} is the quantity directly optimized by CSP; d_{support} is an independent behavioral measure. "w/ CSP" is the w/o AEM configuration, which shares its training run with full DTSA.

Task	Metric	w/o CSP	w/ CSP (w/o AEM)	Rel. imp.
Walker2d-m	$D_{\mathrm{CS}} \downarrow$	0.0049	0.0046	+5.3%
	$d_{\mathrm{support}} \downarrow$	1.7161	1.6924	+1.4%
Walker2d-m-r	$D_{\mathrm{CS}} \downarrow$	0.0017	0.0004	+78.9%
	$d_{\mathrm{support}} \downarrow$	1.8144	1.8099	+0.2%
Hopper-m	$D_{\mathrm{CS}} \downarrow$	0.0062	0.0054	+12.6%
	$d_{\mathrm{support}} \downarrow$	0.8321	0.8352	−0.4%
Hopper-m-r	$D_{\mathrm{CS}} \downarrow$	0.0075	0.0029	+61.5%
	$d_{\mathrm{support}} \downarrow$	0.7231	0.7049	+2.5%

(84.0 and 78.8) come from Table 1 under Huang et al. (2025)'s harness. Cross-protocol values are indicative context, not same-harness comparisons. On raw locomotion the critic-based family reaches a similar band (ReBRAC 88.5 vs. DTSA 84.0), and the deployment axes used in this paper, frozen-parameter robustness under shift (§4.4) and the inference energy of the spiking policy (Appendix G), are not evaluated by those methods.

E Mechanistic Evidence for Spike-Conditioned Retrieval

E.1 Protocol

Evaluation runs under a seeded-reset protocol: before every episode we deterministically reconstruct all random streams that the d4rl AntMaze consumes at reset (each wrapper layer's private generator, reached through both the `env` and `_wrapped_env` links, and the global stream used by goal sampling; see Appendix F). Consequently, two arms evaluated with the same episode seed see identical initial poses and goals, verified end to end: an arm that provably never alters actions (transport disabled) is bitwise identical to the plain backbone over all 70 episodes of the control block. All comparisons below are therefore paired, and we report exact McNemar tests on episode outcomes plus 95% CIs of paired mean differences. Each seed block covers 7 goal-conditioning levels (target return-to-go $\in \{1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.3\} \times R_{\max}$) $\times$ 10–20 episodes. The backbone reference (Appendix E.2) is a single checkpoint of the mechanism-free backbone. The deployment

Table 9: Robustness summary under the four primary deployment shifts. R_{ID} is the nominal (unshifted) return under the deployment-evaluation protocol, whose configuration differs from the standard-protocol scores of Table 3. The two should not be compared directly.

Task	Method	R_{ID} ↑	Mean AUC ↑	Worst-bin ret. ↑
Walker2d-m	DSFormer-PSSA	75.04	33.61	0.0424
	w/o AEM	70.59	35.66	0.0597
	w/o CSP	78.75	35.67	0.0512
	DTSA	**84.06**	**37.88**	0.0588
Walker2d-m-r	DSFormer-PSSA	**88.42**	**25.70**	0.0636
	w/o AEM	78.58	25.49	−0.0012
	w/o CSP	72.17	25.54	**0.0703**
	DTSA	88.36	23.23	0.0016
Hopper-m	DSFormer-PSSA	51.45	18.37	0.1948
	w/o AEM	50.56	18.31	**0.1985**
	w/o CSP	**61.60**	**18.40**	0.1635
	DTSA	57.85	17.78	0.1656
Hopper-m-r	DSFormer-PSSA	85.11	17.76	0.0525
	w/o AEM	86.32	**20.76**	0.0124
	w/o CSP	91.75	17.08	0.0252
	DTSA	**92.06**	19.76	0.0114

Table 10: Deployment-time adaptation cost. "Adapted parameters" is the fraction of policy parameters modified during deployment. "Extra memory" counts within-episode state beyond the frozen policy.

Method	Gradient updates	Adapted parameters	Extra memory
w/o AEM	No	0%	0
DTSA	No	0%	≈ 1.1 MB

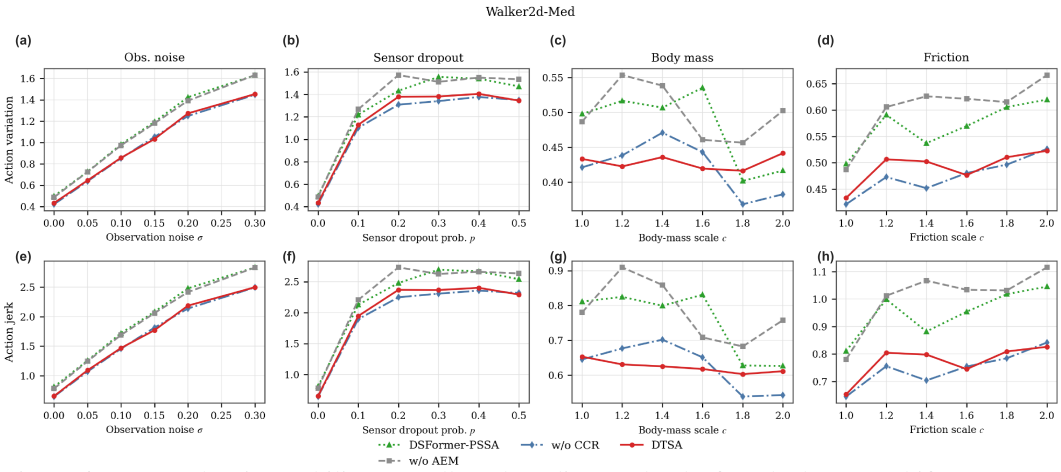


Figure 6: Temporal action stability on Walker2d-medium under the four deployment shifts. Top row: action variation (mean absolute first difference); bottom row: action jerk (mean absolute second difference). Line conventions as in Figure 3.

configuration uses bank capacity 1024, top-k 320, transported-action fusion weight $\lambda_{\mathrm{dsct}} = 0.05$, the confidence gate with eigenvalue flooring, and the safe-write policy of §E.6.

E.2 Checkpoint Selection

All saved epochs of the medium-diverse model were probed with the backbone arm (16 episodes, no deployment stack) before any intervention experiment. The checkpoint used in all deployment experiments was selected solely by this probe score; the probe uses the backbone arm only, so no arm was tuned against it and the selection cannot favor the deployment configuration:

Table 11: Spike-response drift $\Delta\boldsymbol{\xi}$ relative to the nominal reference under deployment shift (mean and maximum over evaluated episodes).

Task	Method	Mean $\Delta\boldsymbol{\xi}$	Max $\Delta\boldsymbol{\xi}$
Walker2d-m	DSFormer-PSSA	0.8569	1.9746
	w/o AEM	0.8903	**1.7417**
	w/o CSP	0.8978	1.7937
	DTSA	**0.8544**	1.7713
Walker2d-m-r	DSFormer-PSSA	0.9077	1.8706
	w/o AEM	0.8361	**1.7865**
	w/o CSP	**0.8296**	1.8222
	DTSA	0.9440	1.8445
Hopper-m	DSFormer-PSSA	0.8415	2.0643
	w/o AEM	**0.8128**	1.6079
	w/o CSP	0.8946	2.0683
	DTSA	0.8292	**1.5999**
Hopper-m-r	DSFormer-PSSA	1.0758	**1.8798**
	w/o AEM	1.0567	2.0552
	w/o CSP	1.0755	1.9029
	DTSA	**1.0343**	2.0421

Table 12: Cross-task Pearson correlation between $\Delta\boldsymbol{\xi}$ and $\Delta\mathbf{a}$ (with $n = 180$ episodes per task covering the four primary shift families, and *** denoting $p < 0.001$).

Task	DSFormer-PSSA	w/o AEM	w/o CSP	DTSA
Walker2d-m	$r = 0.916^{***}$	$r = 0.954^{***}$	$r = 0.931^{***}$	$\boldsymbol{r = 0.964^{***}}$
Walker2d-m-r	$r = 0.892^{***}$	$\boldsymbol{r = 0.962^{***}}$	$r = 0.927^{***}$	$r = 0.931^{***}$
Hopper-m	$\boldsymbol{r = 0.688^{***}}$	$r = 0.507^{***}$	$r = 0.672^{***}$	$r = 0.469^{***}$
Hopper-m-r	$r = 0.263^{***}$	$\boldsymbol{r = 0.910^{***}}$	$r = 0.469^{***}$	$r = 0.888^{***}$

Table 13: Published MuJoCo normalized scores of critic-based methods: single unified re-evaluation of all three under v2 datasets, ten seeds (from Tarasov et al. (2023), Table 2). For context across protocols, Table 1 gives the per-task scores of DSFormer-PSSA (78.8 mean) and DTSA (84.0 mean) under Huang et al. (2025)'s harness. Evaluator versions differ across the two tables, so no ranked comparison is implied.

Task	TD3+BC	IQL	ReBRAC
halfcheetah-m	54.7±0.9	50.0±0.2	65.6±1.0
halfcheetah-m-r	45.0±1.1	42.1±3.6	51.0±0.8
halfcheetah-m-e	89.1±5.6	92.7±2.8	101.1±5.2
hopper-m	60.9±7.6	65.2±4.2	102.0±1.0
hopper-m-r	55.1±31.7	89.6±13.2	98.1±5.3
hopper-m-e	87.8±10.5	85.5±29.7	107.0±6.4
walker2d-m	77.7±2.9	80.7±3.4	82.5±3.6
walker2d-m-r	68.0±19.2	75.4±9.3	77.3±7.9
walker2d-m-e	110.4±0.6	112.1±0.5	111.6±0.3
Mean (9 tasks)	72.1	77.0	88.5

epoch	1	2	3	4	5	6	7
backbone success	0.0	**25.0**	0.0	6.2	12.5	0.0	0.0

The remaining epochs use an observation layout incompatible with the deployment stack and were excluded on that structural basis, not by score.

E.3 Main result: pairing intervention on medium-diverse

Three controls bound alternative explanations. (i) **Specificity:** with transport disabled, retrieval confidence statistics and the full write/reject machinery remain active, yet the arm is bitwise identical to the backbone ($0/70$ discordant on medium-diverse; on umaze the same control and the shuffle-vs-transport-off pair are likewise identical, $0/70$ discordant). The measured effect is therefore created solely by transporting retrieved actions, not by extra computation or memory traffic. (ii) **Dose re-**

Seed block	backbone	λ=.05	ξ-shuffle	gain
1 (n=70/arm)	25.7	37.1	— (λ=.3: 28.6)	+11.4
2 (n=140)	27.1	31.4	26.4	+4.3
3 (n=140)	27.1	32.9	25.0	+5.8
4 (n=140)	27.1	32.1	24.3	+5.0
5 (n=140)	27.1	31.4	23.6	+4.3
Pooled paired tests				
correction vs. backbone	+5.6 pts		147W/112L	$p = 0.034$, CI $[+0.6, +10.5]$
ξ-shuffle vs. correction	−7.1 pts		59W/99L	$p = 0.0018$, CI $[-11.5, -2.8]$
ξ-shuffle vs. backbone	−2.3 pts		107W/120L	$p = 0.43$

Table 14: Antmaze-medium-diverse, deterministic paired evaluation. Shuffling bank (ξ_j, a_j) pairings with a fixed seed preserves bank size, distance structure, and write volume, and breaks only the spike-to-action correspondence; the entire benefit disappears at $p = 0.0018$, and the shuffled arm is again indistinguishable from the backbone ($p = 0.43$).

sponse: $\lambda_{\mathrm{dsct}} = 0.02$ gives $+0.4$ (n=280, p=1.0) and $\lambda_{\mathrm{dsct}} = 0.30$ gives $+2.9$ (n=70, p=0.85), and at λ=0.3 a larger share of applied corrections saturates the 25% magnitude cap (over-cap rate 0.31 vs. 0.26 at λ=0.05, and 0.44 on umaze at λ=0.3). Strong transport therefore conflicts with the policy rather than nudging it, which isolates $\lambda \approx 0.05$ as the regime where the correction magnitude (mean 8.4% of the action norm) remains informative. (iii) **Independence across blocks:** the correction beats the backbone in $5/5$ seed blocks, so the pooled significance is not driven by one block.

E.4 No effect on saturated umaze (negative control)

Arm (antmaze-umaze, block 1, n=70)	Score	vs. backbone (McNemar)
backbone	65.7	—
full (k=320, λ=.3)	64.3	14W/15L, $p = 1.0$
ξ-shuffle	64.3	15W/15L vs. full, $p = 1.0$
transport off	65.7	bitwise ≡ backbone
full, block 2	67.1	16W/14L vs. block 1, p=0.86

Table 15: On umaze every arm collapses into a noise band of ± 3 points around 65%: the task is saturated by BC-quality policies and cannot express the mechanism. We therefore report the mechanism on medium-diverse, where headroom exists.

E.5 Same-structure regularizer baselines (full dose)

The windowed conditional-divergence loss is compared against three same-structure baselines under identical local windows, median bandwidths, and detach conditions, with all four arms given the same budget. The weight λ is set so that the 10%-of-MSE clamp binds throughout training (measured raw ratio $\mathrm{cs}/\mathrm{mse} \approx 1.2 \times 10^{-2}$, binding threshold $\lambda \approx 8.4$), and we use λ=10 plus an intermediate λ=2. Held-out conditional divergence D_{CS} (5% held-out trajectories, original CS estimator, identical evaluation batches across arms ⇒ paired statistics):

Regularizer	ΔD_{CS} vs. none	paired t	best ret. (0–100)
conditional MMD, λ=10	−0.0030	4.9	10
CSP Eq. (4), λ=10	−0.0016	3.2	20
CSP, λ=2	−0.0012	2.6	30
window Sinkhorn, λ=10	−0.0007	1.2	10
ξ-shuffled CSP, λ=10	−0.0006	1.1	10
in-window InfoNCE, λ=10	−0.0004	0.6	40

Table 16: Kernel-divergence regularizers (MMD, CS) improve held-out conditional matching monotonically in dose, with paired t increasing from 2.6 to 3.2 from λ=2 to λ=10, whereas InfoNCE and Sinkhorn are neutral. Returns lie in the evaluation noise band on saturated umaze, so D_{CS} is the discriminating metric. MMD is marginally ahead of the CS form (Δ=0.0013, t=2.57, borderline under series autocorrelation), so the normalization structure specific to CS is not the source of the benefit. Accordingly we describe Eq. (4) as one parameterization of the windowed kernel-divergence family.

E.6 WRITE-GATE OBSERVABLES AND THE SAFE-WRITE POLICY

The AEM write gate is auditable through per-step statistics persisted with every evaluation (gate-open rate, g_t and correction-ratio distributions, rejection counts), which provide the empirical counterparts of the theoretical ρ_{cap} and ε_e quantities. At a representative operating point (medium-diverse, λ=0.05) the gate opens on 0.36 of steps, the mean correction ratio is 0.084, 0.28 of applied corrections saturate the magnitude cap, the median g_t is 0.36, and the run records 123k writes against 36k cap-rejections. Entries rejected by the magnitude cap are replaced in the bank by that step's policy action (safe-write policy): the stored $(s, \boldsymbol{\xi}, \mathbf{a}^{\text{exec}})$ triples remain mutually consistent and cannot bootstrap correction errors, while allowing the bank to reach retrieval size (≈ 1024). An all-reject rule instead would hold ≈ 13 entries per episode, top-k retrieval would degenerate to nearest-neighbor lookup, and no arm would separate from the backbone in paired tests. Gate-threshold ablations (δ_{write}=0, reliability $\rho \equiv 1$, cross-episode persistence) all remain within noise of each other, indicating that bank composition, not the confidence threshold, is the binding control. This is consistent with the theory emphasizing ρ_{cap}.

E.7 LIMITATIONS

The deployment gain is established on antmaze-medium-diverse at the checkpoint selected by the probe of Appendix E.2 and at a correction weight chosen by a prior sweep on umaze; the $5/5$ independent seed blocks and the shuffle control support the protocol. On antmaze-umaze the task ceiling hides the effect. The training-time regularizer improves conditional-matching metrics but, at full dose, not beyond the plain conditional-MMD variant, and returns on saturated benchmarks cannot resolve differences at $n \leq 30$ per cell. Large-diverse and additional seeds are future work.

F DETERMINISTIC EVALUATION PROTOCOL OF THE DEPLOYMENT STACK

Deterministic reset protocol. The d4rl AntMaze stack keeps a private random generator in each wrapper layer, which `env.reset(seed=)` and `unwrapped.seed()` do not propagate through (the d4rl `ProxyEnv` links the inner environment via `_wrapped_env`, not `env`), and maze goal sampling draws from the global random stream. A plain reset therefore does not determine the initial condition across processes. The evaluation protocol reconstructs all random streams deterministically before every episode: each wrapper layer's private generator, reached through both the `env` and `_wrapped_env` links, and the global stream used by goal sampling (Appendix E). Two arms evaluated with the same episode seed thus see identical initial poses and goals, verified end to end: 50-step fixed-action rollout hashes are identical across processes, and the transport-off control is bitwise identical to the backbone over the full 70-episode control blocks in both mazes. All paired statistics in Appendices E and F use this protocol. Because the default d4rl reset path leaves these randomness sources unseeded, the protocol is relevant to any paired evaluation on this stack.

Write-path specification. Bank admission follows W1–W3 of §3.3. The h_0-step cold-start consolidation stores the policy's uncorrected actions with reliability $\rho = 1$, which brings the bank to the retrieval size $h_{\min}$ before the first gating decision; afterwards a step is written when $g_t \geq \delta_{\text{write}}$ ($\delta_{\text{write}} \leq \delta_{\text{read}}$), and corrected steps additionally require the magnitude clip to be inactive. A correction rejected by the magnitude cap is replaced by that step's policy action (the safe-write policy of §E.6), so the stored $(s, \boldsymbol{\xi}, \mathbf{a}^{\text{exec}})$ triples remain mutually consistent while the bank reaches its retrieval capacity. The confidence statistic is computed on the 48-neighbor sample in 256-dimensional spike space with eigenvalue flooring and an effective-rank Mahalanobis distance, which keeps the statistic well posed on near-binary spike coordinates. Per-step gate histograms (g_t, correction ratio, over-cap rates) accompany every evaluation and provide the empirical counterparts of ρ_{cap} and ε_e (§E.6).

Artifacts. Per-arm episode vectors, gate histograms (g_t, correction ratio, over-cap rates, percentiles), seed-block tables, and scripts to regenerate every table will be released with the source code.

Table 17: Operation counts per inference pass, by component and cost class.

Component	Operations	Type
Embedding	$D_{\text{src}} \cdot D \cdot N$	MAC (ANN)
Q, K, V projections	$3UD^2NR_{\text{qkv}}$	SOP (SNN)
PSSA attention $f(Q, K, V)$	$USDN$	SOP (SNN, elementwise)
Attention linear layer	UD^2NR_{attn}	SOP (SNN)
MLP-1 / MLP-2	$4UD^2NR_{\text{mlp}}$ each	SOP (SNN)
Prediction head	$D_{\text{tgt}} \cdot D \cdot N$	MAC (ANN)
Fast adaptation module	closed-form retrieval & statistics	deployment

G Energy Consumption of Inference

This appendix reports the energy DTSA spends on one inference pass. The figures are analytical: they are computed from the operation counts of the deployed graph and from firing rates measured on real evaluation rollouts, not from silicon. Two properties of the implementation fix the accounting. Only the embedding and the prediction head remain in floating point; every matrix inside a PSSA block is driven by spikes, so its cost scales with the spikes that arrive rather than with the dense shape of the layer. And AEM runs after the policy is frozen, so its cost is charged to deployment and never to training, through the four stages episodic memory → joint retrieval → local second-order statistics → evidence gating and action fusion. We therefore report two quantities: the cost of DTSA's own additions, and the total against the floating-point models it replaces. In §G.5 the additions are priced against the DSFormer backbone re-run through our estimator with its official code, and the total is compared against the floating-point baselines.

G.1 Cost model

In our PSSA implementation a spike-gated product selects terms that have already been accumulated rather than forming a floating-point product, so we charge one accumulate (AC) operation for each contributing spike. A floating-point layer is charged one multiply–accumulate (MAC) operation for each dense entry. With N_{ops} the dense operation count of a layer, U its number of internal steps, and R the firing rate of the node driving it,

$$E_{\text{SNN}} = U\,R\,N_{\text{ops}}\,E_{\text{AC}}, \qquad E_{\text{ANN}} = N_{\text{ops}}\,E_{\text{MAC}}, \tag{7}$$

where E_{AC} and E_{MAC} are the per-operation energies. We use the published 45nm CMOS figures $E_{\text{MAC}} = 4.6\,\text{pJ}$ and $E_{\text{AC}} = 0.9\,\text{pJ}$ (Horowitz, 2014). No rate is assumed here: the estimator reads the spike counters of the model being evaluated, so every R in Eq. (7) is measured. Per-pass cost then splits the way the model does,

$$E_{\text{DTSA}} = E_{\text{PSSA}} + E_{\text{fast}}, \tag{8}$$

with E_{PSSA} the backbone (embedding and prediction head as MAC, QKV projections, PSSA attention, attention linear layers and MLPs as AC) and E_{fast} the deployment module of §3.3. Because AEM never enters the training loop, it adds no training-time energy.

G.2 Per-pass operation counts

Each PSSA block exposes six spiking nodes, and the estimator takes one rate per node rather than one rate per block: R_q, R_k, R_v at the three pre-attention LIFs, R_{attn} at the attention output, and R_1, R_2 at the two MLP stages. Two choices in the estimator merit mention. The QKV projection is charged with the average $R_{\text{qkv}} = (R_q + R_k + R_v)/3$, since the three streams share one linear layer. The spike-gated attention operator is charged with the joint rate $\hat{R} = R_q R_k R_v$: a term of $f(Q, K, V)$ contributes only when the corresponding entries of all three spike matrices are active, and the implementation keeps no separate counter for that node, so $\hat{R}$ is the proxy used for it. The two floating-point layers follow the data layout rather than any rate: the embedding reads $D_{\text{src}} = d_s + d_a + 1$ channels (state, action and return-to-go) and the prediction head writes $D_{\text{tgt}} = d_a$, and neither carries an internal time dimension. Table 17 collects the resulting counts at $D = 256$, $U = 4$, context $N = K = 20$, window $S = 8$ (the row sum of the causal local mask), 3 blocks and 2.45M parameters.

Table 18: Measured mean firing rates of our model on the MuJoCo tasks under the fixed evaluation protocol.

Task	Rate	Task	Rate
halfcheetah-medium	0.2778	hopper-medium-expert	0.2616
halfcheetah-medium-expert	0.2786	hopper-medium-replay	0.2585
halfcheetah-medium-replay	0.2704	walker2d-medium	0.2654
hopper-medium	0.2587	walker2d-medium-expert	0.2652
		walker2d-medium-replay	0.2652
Mean: 0.2668			

Table 19: Per-component energy on hopper-medium-replay (μJ): the DTSA pass against the floating-point DT pass it replaces.

Component	DT (ANN)	DTSA
Embedding	0.9	0.35
Self-attention	168.2	10.75
MLP	241.2	37.55
Prediction head	0.2	0.07
Backbone subtotal	410.5	48.72
Episodic memory (write/evict)	–	0.09
Joint retrieval	–	1.37
Local second-order statistics	–	168.58
Gating and action fusion	–	76.28
Fast adaptation subtotal	–	246.31
Total	410.5	295.04

Table 20: Total energy on all nine MuJoCo tasks (μJ, fixed protocol).

Task	PSSA backbone	Fast module	Total
halfcheetah-medium	52.40	249.96	302.36
halfcheetah-medium-expert	52.48	249.96	302.45
halfcheetah-medium-replay	50.57	249.96	300.54
hopper-medium	49.01	245.03	294.04
hopper-medium-expert	49.47	245.02	294.49
hopper-medium-replay	48.72	246.31	295.04
walker2d-medium	50.45	249.93	300.38
walker2d-medium-expert	50.41	249.83	300.24
walker2d-medium-replay	49.76	247.50	297.27
Mean	50.37	248.17	298.53

G.3 Measured firing rates

Table 18 reports the mean measured rates over the nine MuJoCo tasks under the fixed evaluation protocol; averaged across tasks and layers the rate is 0.2668, so roughly one in four presynaptic entries contributes an accumulate operation at any internal step.

G.4 Energy per component and per task

Table 19 decomposes one pass on hopper-medium-replay, where the spike-gated attention and MLP stages cut the backbone to a fraction of the floating-point DT pass, and Table 20 gives all nine tasks. Within DTSA the backbone costs 50.37 μJ on average and the deployment module 248.17 μJ, with the local second-order statistics dominating; writing to the bank and retrieving from it are nearly free.

G.5 Same-protocol comparison and discussion

We report two comparisons of different kinds. The first is measured on the same harness: the DSFormer backbone is put through our estimator with its official implementation, so the two differ only in what DTSA adds. Table 21 gives it per task. The plain PSSA backbone averages 39.05 μJ and the DTSA backbone 50.37 μJ, so spike-conditioned stability is bought with roughly 11 μJ per

Table 21: Same-harness backbone cost: the DSFormer backbone (official implementation) and the DTSA backbone through the identical estimator, constants and protocol ($N{=}K{=}20$, $U{=}4$, $S{=}8$, $D{=}256$, 3 blocks, 10 episodes per target return). Neither column includes a deployment module.

Task	DSFormer backbone	DTSA backbone
halfcheetah-medium	39.01	52.40
halfcheetah-medium-expert	39.96	52.48
halfcheetah-medium-replay	38.51	50.57
hopper-medium	38.21	49.01
hopper-medium-expert	39.43	49.47
hopper-medium-replay	38.15	48.72
walker2d-medium	39.08	50.45
walker2d-medium-expert	40.14	50.41
walker2d-medium-replay	38.95	49.76
Mean	39.05	50.37

Table 22: Total deployment pass against the compared baselines: per-pass energy of DTSA (backbone plus AEM) versus every model whose reported energy is higher, with the MuJoCo mean each obtains. DT, SpikeBERT and FCNet figures as reported by Huang et al. (2025) under their own protocol.

Model	Family	Energy (μJ) $\downarrow$	MuJoCo mean $\uparrow$
DT	ANN	410.5	74.7
SpikeBERT	SNN	806.7	39.9
FCNet	ANN	1022.0	75.1
DTSA	SNN	**298.5**	**84.0**

pass, spent almost entirely in the higher firing rates that CS-conditioned training produces (mean 0.267, Table 18). The second is the deployment-relevant one: Table 22 lists the models whose per-pass energy exceeds DTSA's, with the MuJoCo mean each obtains. With AEM attached, DTSA still costs less than every model in that table and scores higher than all of them.

Within the module's 248.17 μJ, the local second-order statistics take 170.07 μJ and gating with fusion 76.35 μJ, while retrieval takes 1.64 μJ and the writes 0.10 μJ. That is the price of adapting locally without touching a parameter, and it is paid only at deployment. The saving is therefore relative to the floating-point models, not a claim to be the leanest spiking network: DTSA's whole deployment pass, adaptation included, costs less than the DT pass that carries no such module, and most of what it spends goes to the second-order statistics that make adaptation work.

All figures above are analytical per-pass estimates in μJ. Settings: $E_{\mathrm{MAC}} = 4.6$ pJ and $E_{\mathrm{AC}} = 0.9$ pJ at 45nm CMOS (Horowitz, 2014), $D = 256$, $U = 4$, $N = K = 20$ (the context length actually used at deployment), $S = 8$, 3 blocks, with rates averaged over 10 episodes per target return under the fixed protocol.

H Proofs of Main-Text Statements

This appendix collects every statement used in the main text together with its complete proof, grouped by the mechanism it supports. The main text states the definitions, the two mechanisms, and the headline form of the two guarantees; the assumption blocks, the exact constants, the auxiliary propositions, and the kernel-matching remark are stated here.

H.1 Assumptions and deferred statements

Conditioning conventions. The method never assumes $\partial\boldsymbol{\xi}/\partial\theta = 0$: the spike readout Eq. (1) is recomputed from the current θ at every forward pass, and the BC gradient flows through the same LIF spikes via the surrogate rule. The *frozen-conditioning convention* applies one level lower: within a gradient step the conditioning readout is detached (Algorithm 1), so k^x, w_{ij}, and the retrieval weights are measurable functionals of the realized geometry, and bounds are uniform only over $\theta \in \Theta$ at that geometry; the encoder's response to parameter motion is carried by L_Φ (Assumption 1(ii)). The guarantees of Theorem 1, Corollary 1, and Theorem 2 are therefore conditional guarantees:

they hold at any realized geometry, and the training loop of Algorithm 1 re-realizes the geometry at every gradient step; they are not trajectory-wide statements about the full optimization path.

Assumption 1 (Bounded spike dynamics and regular geometry)**.** ***(i)*** *$|v_t^l| \le V$, so the LIF recursion keeps $|u_t^l| \le \bar{u} := V/(1-\beta)$, and membrane pre-threshold masses concentrate, $\mathbb{P}(|u_t^l - v_{\mathrm{th}}| \le \varepsilon) \le L_u\varepsilon$ (the only stochastic condition; the rest hold pathwise);* ***(ii)*** *the encoder $c \mapsto (s, \boldsymbol{\xi})$ is L_Φ-Lipschitz on $\|c\|_\infty \le R_{\mathrm{ctx}}$;* ***(iii)*** *$\|\hat{a}\|_2, \|a\|_2 \le R_a := \sqrt{d_a}$ and $\sigma_s, \sigma_\xi \in [\sigma_{\min}, \sigma_{\max}]$.*

Assumption 2 (CSP regularity: realizability, capacity, and dynamics)**.** ***(i)*** *there exists $\theta_0 \in \Theta$ with $\mathcal{L}_{\mathrm{BC}}(\theta_0) \le \sigma_B^2$ and $\hat{\mu}_q^{\theta_0}(x) \to \hat{\mu}_p(x)$ ($V_q, V_{pq} \to V_p$);* ***(ii)*** *$\|\theta\|_2 \le B_\theta$, the policy is L_π-Lipschitz in θ, pseudo-dimension $\bar{p}$;* ***(iii)*** *r is L_r-Lipschitz, $|r| \le \bar{r}$; P is L_P-Wasserstein-Lipschitz with $\gamma L_P < 1$;* ***(iv)*** *the n decision positions are organized into N_w context blocks sampled independently from the empirical window law ν (each block a contiguous causal window of length $\le K$ as drawn by the training sampler; within-block trajectory dependence is carried explicitly through the concentration argument), with $m(x) := \mathbb{E}[k^x(x, X') \mid x] \ge m_{\min} > 0$;* ***(v)*** *$\mathbb{E}\|\mu_p(x) - \mu_p(x')\|_{\mathcal{H}_k}^2 \le \omega_\mu^2(d_x(x, x'))$ and the comparison modulus $\varsigma(u) := \sup\{\|\mathbf{a} - \mathbf{a}'\|_2 : \|\varphi(\mathbf{a}) - \varphi(\mathbf{a}')\|_{\mathcal{H}_k} \le u\}$ is finite on $\mathcal{A}$.*

Assumption 3 (AEM estimation regularity)**.** *Within a retrieval neighborhood:* ***(i)*** *$(\delta\boldsymbol{\xi}, \delta\mathbf{a})$ are jointly σ_ξ/σ_a-sub-Gaussian with $\lambda_* I \preceq \Sigma_{\xi\xi} \preceq \Lambda_\xi I$;* ***(ii)*** *the behavior conditional mean is affine up to curvature, $\mathbb{E}[\mathbf{a} \mid \boldsymbol{\xi}] = \mu_a + B^\star(\boldsymbol{\xi} - \mu_\xi) + \eta$ with $\mathbb{E}\|\eta\| \le \rho_B \|\boldsymbol{\xi} - \mu_\xi\|^2$, hence $B^\star = \Sigma_{a\xi}\Sigma_{\xi\xi}^{-1}$;* ***(iii)*** *weights are spread, $q_{tj} \le c_q/N_{\mathrm{eff}}$ with $N_{\mathrm{eff}} = 1/\sum_j q_{tj}^2 \in [1, K_r]$, and $q, \mathcal{N}$ are measurable in $(c_{\le t}, H_t)$ only.*

Proposition 1 (Joint conditioning refines state-only conditioning)**.** *Let $p_\mu(\mathbf{a} \mid x)$ and $p_\mu(\mathbf{a} \mid s)$ be the behavior-policy action laws under joint and state-only conditioning.* ***(i)*** *(Consistency) $p_\mu(\mathbf{a} \mid s) = \mathbb{E}_{\boldsymbol{\xi}|s}\big[p_\mu(\mathbf{a} \mid s, \boldsymbol{\xi})\big]$.* ***(ii)*** *(Separation) state-similar sites can carry different joint conditionals.* ***(iii)*** *(Kernel domination) $k_{ij}^x/k_{ij}^s = k_{ij}^\xi$ and, with spike-divergence fraction $\psi := \mathbb{E}\|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2^2/d_\xi$, $\mathbb{E}[k_{ij}^x/k_{ij}^s] \le 1 - \psi\big(1 - e^{-d_\xi/(2\sigma_\xi^2)}\big)$, a bound that decreases monotonically in ψ and stays strictly below 1 for every $\psi > 0$; hence state-only proximity cannot create spurious joint support.*

Proposition 2 (Conservative deployment adaptation)**.** *Under Assumptions 1 and 3:* ***(i)*** *(Fallback) if $g_t < \delta_{\mathrm{read}}$, $\mathbf{a}_t = \mathbf{a}_t^0$, and the deployed controller reduces exactly to the slow-timescale policy.* ***(ii)*** *(Validated, bounded correction) on the active set $\{g_t \ge \delta_{\mathrm{read}}\}$, necessarily $\bar{\rho}_t N_{\mathrm{eff}} \ge \beta_{\sup} \log \frac{1}{1-\delta_{\mathrm{read}}}$, $r_t \le 2\log\frac{1}{\delta_{\mathrm{read}}}$, and $\mathrm{tr}(\Sigma_{a|\xi}) \le d_a \bar{v}_t \beta_{\mathrm{act}} \log\frac{1}{\delta_{\mathrm{read}}}$, and by Lemma 3 the correction obeys the budget*

$$\|\mathbf{a}_t - \mathbf{a}_t^0\|_2 \le \rho_{\mathrm{cap}} := \min\Big\{\rho_{\max} R_a,\ \vartheta^{-1}(1-\vartheta)\big(\varepsilon_e + \bar{\delta} + \bar{\nu}\big)\Big\}, \qquad \vartheta^{-1}(1-\vartheta) = \tfrac{\lambda_{\mathrm{T}} + \lambda_e/\lambda_a}{\lambda_\pi}, \tag{9}$$

with ε_e of Corollary 2 (estimation part $\mathcal{O}(N_{\mathrm{eff}}^{-1/2})$) and $\bar{\delta}, \bar{\nu}$ the local-relation spread and anchor motion.

Remark 1 (What the CS form adds to kernel matching). Embedding matching is a family, not a single loss: replacing the normalized bracket of Eq. (4) by the unnormalized discrepancy $V_p + V_q - 2V_{pq} = \|\hat{\mu}_p - \hat{\mu}_q\|_{\mathcal{H}_k}^2$ recovers a conditional-MMD penalty, which can also descend along solutions where $\hat{\mu}_p, \hat{\mu}_q$ share direction but lose magnitude. The log–ratio form is the squared angle between the conditional embeddings: non-negative, vanishing exactly at proportionality (the identification in §3.2), and insensitive to the local mass $\|\hat{\mu}_p(x)\|_{\mathcal{H}_k}$. It is a pure structure term that can be capped below the BC anchor ($\lambda_{\mathrm{CSP}}\mathcal{L}_{\mathrm{CSP}} \le \eta_{\mathrm{cs}}\,\mathrm{sg}(\mathcal{L}_{\mathrm{BC}})$) and cannot reweight toward low-support regions. Restricted to causal local sets of size W, it costs one Gaussian evaluation per pair, $\mathcal{O}(|\mathcal{I}|\,KW)$.

H.2 CSP CONDITIONING, GENERALIZATION, AND STRUCTURE PRESERVATION

Population form of the CSP loss. With $v_{\min} := e^{-2R_a^2/\sigma_a^2}$, the population loss replaces the plug-ins by the true conditional mean embeddings under the anchor law $\bar{\nu}$; the bracket of Eq. (4) equals $-2\log c(x)$ with $c(x) := V_{pq}/\sqrt{V_p V_q} \in (0, 1]$, so $\mathcal{L}_{\mathrm{CSP}} \ge 0$ vanishing if and only if $\mu_q \propto \mu_p$, realized at θ_0 (Assumption 2(i)).

Lemma 1 (Relational distortion controlled by the loss)**.** *Let $D^2(x) := \|\mu_p(x) - \mu_q(x)\|_{\mathcal{H}_k}^2$ and $c(x) := V_{pq}(x)/\sqrt{V_p(x)V_q(x)} \in (0, 1]$. Then $\mathbb{E}_x D^2(x) \le \rho_\kappa^2 + \mathcal{L}_{\mathrm{CSP}}(\theta)$, with $\rho_\kappa := 1 - \sqrt{v_{\min}}$, and $D^2(x)$ decomposes exactly as $(\sqrt{V_p} - \sqrt{V_q})^2 + 2\sqrt{V_p V_q}\,\big(1 - c(x)\big)$, whose second term is bounded by the pointwise bracket of $\mathcal{L}_{\mathrm{CSP}}$.*

Lemma 2 (Local estimation error at effective sample complexity)**.** *Under Assumptions 1 and 3, on the gate-active set $\{g_t \ge \delta_{\mathrm{read}}\}$, with probability $\ge 1-\delta$ simultaneously over the T_e steps of the episode,*

$$\|\hat{B}_t - B^\star\|_{\mathrm{op}} \le \underbrace{\frac{2\lambda_\xi \|B^\star\|_{\mathrm{op}}}{\lambda_* + \lambda_\xi}}_{\textit{ridge shrinkage}} + \underbrace{C_{\mathrm{est}}\,\sigma_\xi\big(\sigma_a + \|B^\star\|_{\mathrm{op}}\Lambda_\xi^{1/2}\big)\sqrt{\frac{c_q\,(d_\xi + d_a + 1)\,\log(3K_rT_e/\delta)}{N_{\mathrm{eff}}}}}_{\textit{variance}},$$
$$\|\hat{\mu}_{\xi,t} - \mu_\xi\| \le C_2\sigma_\xi\sqrt{\tfrac{c_q\log(6K_rT_e/\delta)}{N_{\mathrm{eff}}}}, \tag{10}$$

with C_{est}, C_2 absolute (sub-Gaussian) constants; λ_* from Assumption 3(i).

Lemma 3 (Fusion is a ϑ-contraction to the frozen anchor)**.** *Let $u' = \tilde{\mathbf{a}}_t - \mathbf{a}_t^0$ and $v' = \mathbf{a}_{t-1} - \mathbf{a}_t^0$. Then $\|\mathbf{a}_t^\star - \mathbf{a}_t^0\|_2 \le (1-\vartheta)\big(\|u'\|_2 + \|v'\|_2\big)$ with $\vartheta = \lambda_\pi/\big(\lambda_\pi + \lambda_{\mathrm{T}} + \lambda_e/\lambda_a\big) \in (0,1)$, and the final $\mathbf{a}_t$ of Eq. (15) satisfies $\|\mathbf{a}_t - \mathbf{a}_t^0\|_2 \le \min\{\rho_{\max}\|\mathbf{a}_t^0\|_2,\ (1-\vartheta)(\|u'\|_2 + \|v'\|_2)\}$.*

Corollary 1 (Offline sample bound and behavior recovery)**.** *Under the conditions of Theorem 1, with probability $\ge 1-\delta$: **(i)** $\mathcal{L}_{\mathrm{BC}}(\hat{\theta}) \le \sigma_B^2 + 2\varepsilon_{\mathrm{gen}}(n,\delta)$ and $\mathcal{L}_{\mathrm{CSP}}(\hat{\theta}) \le 2\varepsilon_{\mathrm{gen}}(n,\delta)/\lambda_{\mathrm{CSP}}$, hence $\mathbb{E}_x D^2(x) \le \rho_\kappa^2 + 2\varepsilon_{\mathrm{gen}}/\lambda_{\mathrm{CSP}}$ by Lemma 1; **(ii)** the frozen policy satisfies the coupling bound $J(\mu) - J(\pi_{\hat{\theta}}) \le C_R\sqrt{\mathcal{L}_{\mathrm{BC}}(\hat{\theta})} \le C_R\sqrt{\sigma_B^2 + 2\varepsilon_{\mathrm{gen}}(n,\delta)}$ with $C_R = 2L_r\bar{H}^2(1 + L_P\bar{H})$; **(iii)** consequently, for every target $\varepsilon > C_R\sigma_B$, whenever $n \ge \left\lceil 4C_\varepsilon^2\,\mathcal{P}_n/\big(\varepsilon^2/C_R^2 - \sigma_B^2\big)^2\right\rceil$ with $\mathcal{P}_n = \bar{p}\log(nB_\theta L_\pi/\delta) + \log(4/\delta)$ and $C_\varepsilon = \max\{C_{\mathrm{gen}}, 8R_a^2\}$ (understood up to the log self-consistency of $\mathcal{P}_n$), we have $J(\mu) - J(\pi_{\hat{\theta}}) \le \varepsilon$.*

Remark 2 (Sample scales n and N_{eff} do not compete). Term (I) decays as $\tilde{\mathcal{O}}(n^{-1/4})$ using all offline positions, while term (III) depends only on $N_{\mathrm{eff}} \le K_r \ll n$ within-episode evidence, so the scales do not compete: on inactive steps (III) and (V) vanish identically, and on the active set the gate's support lower bound keeps every constant uniformly bounded. The shared geometry makes (V) an increment of the same modulus ω_μ at the self-induced displacement $\kappa_\rho\bar{\rho}$, second order in the intervention budget. Setting $g \equiv 0$ collapses Eq. (6) to the "w/o AEM" configuration; $\lambda_{\mathrm{CSP}} \to 0$ forfeits the $\varepsilon_{\mathrm{gen}}/\lambda_{\mathrm{CSP}}$ control ("w/o CSP"); only the full framework bounds all five terms.

Remark 3 (Shared slow–fast condition). $k_{ij}^x = e^{-d_x^2(i,j)/2}$ is the CSP neighborhood metric Eq. (3) and is the AEM retrieval metric Eq. (12): one geometry d_x, two timescales. Relations preserved during training can be recovered by within-episode retrieval, which keeps the coupling term of Theorem 2 a second-order modulus increment.

Lemma 4 (Bounded amplification of threshold spiking)**.** *Let two LIF runs share weights and obey $\|v_t - v_t'\|_2 \le \delta_t$, with $\bar{\varepsilon}_t = \beta^t\|v_0 - v_0'\|_2 + \sum_{k<t}\beta^k\delta_{t-k}$. **(a)** Always $\|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_\infty \le 1$, $\|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2 \le \sqrt{d_\xi}$, and a pre-$t$ spike divergence requires $\min\{|u_t - v_{\mathrm{th}}|, |u_t' - v_{\mathrm{th}}|\} \le \bar{\varepsilon}_t$. **(b)** Under Assumption 1(i) and $\delta_t \le \delta_0$, $\mathbb{E}\|\boldsymbol{\xi}^{(l)} - \boldsymbol{\xi}'^{(l)}\|_2^2 \le d_\xi\min(1, L_u\bar{\varepsilon})$ with $\bar{\varepsilon} = \|v_0 - v_0'\| + \delta_0/(1-\beta)$. For the stacked encoder the aggregate statement is the end-to-end form of Assumption 1(ii), $\|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2 \le L_\Phi\,\mathrm{dist}_{\mathcal{C}}(c,c')$, imposed at the encoder level; the single-layer band argument does not survive cascaded resets and is not extended beyond one layer.*

Proof of Lemma 4. (a) $\boldsymbol{\xi}$ is an average of binary values, so each coordinate shift is at most 1. (b) On the no-prior-divergence event, $z_{t-1} = z_{t-1}'$ and the recursions differ only in the input and leak terms: $u_t - u_t' = \beta(1 - z_{t-1})(u_{t-1} - u_{t-1}') + (v_t - v_t')$. Since $0 \le 1 - z_{t-1} \le 1$, induction gives $\|u_t - u_t'\|_2 \le \beta^t\|u_0 - u_0'\|_2 + \sum_{k=1}^t \beta^{t-k}\delta_k = \bar{\varepsilon}_t$, and a crossing of v_{th} between u_t and u_t' places the threshold within $\bar{\varepsilon}_t$ of both membranes. (b, continued) A flip of cell (l,k,t) requires the band event above; summing over steps, $\mathbb{E}[\text{flip fraction of chain } (l,k)] = (1/U)\sum_t \mathbb{P}(|u_t - v_{\mathrm{th}}| \le \bar{\varepsilon}_t) \le L_u\bar{\varepsilon}$ (Assumption 1(i)); each flip changes coordinate k of the layer average by at most $1/U$ and $|\Delta\xi^{(l),k}| \le 1$, giving $\mathbb{E}\|\boldsymbol{\xi}^{(l)} - \boldsymbol{\xi}'^{(l)}\|_2^2 \le d_\xi\min\{1, L_u\bar{\varepsilon}\}$. Cascades across layers are charged at the encoder level through Assumption 1(ii).

Proof of Proposition 1. (i) is the tower property, since $\boldsymbol{\xi}$ is a function of the same history as c. (ii) Take $\boldsymbol{\xi} \in \{\boldsymbol{\xi}^1, \boldsymbol{\xi}^2\}$ equiprobable given s, with $p_\mu(\mathbf{a} \mid s, \boldsymbol{\xi}^1) = \delta_{\mathbf{a}_1}$ and $p_\mu(\mathbf{a} \mid s, \boldsymbol{\xi}^2) = \delta_{\mathbf{a}_2}$: sites share the mixture $\frac{1}{2}\delta_{\mathbf{a}_1} + \frac{1}{2}\delta_{\mathbf{a}_2}$ marginally, yet the joint conditionals are distinct point masses. (iii)

$t \mapsto e^{-t}$ is convex, and by Lemma 4(a) the support of $T := \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2^2/(2\sigma_\xi^2)$ is $[0, d_\xi/(2\sigma_\xi^2)]$ with fixed mean $\psi\, d_\xi/(2\sigma_\xi^2)$; the extremal expectation of a convex function under a bounded-support, fixed-mean constraint is attained at the endpoints, giving the bound.

Proof of Lemma 1. $D^2 = V_p + V_q - 2V_{pq} = (\sqrt{V_p} - \sqrt{V_q})^2 + 2\sqrt{V_pV_q}\,(1 - c(x))$ with $c(x) = V_{pq}/\sqrt{V_pV_q}$. Since $\sqrt{V_p}, \sqrt{V_q} \in [\sqrt{v_{\min}}, 1]$, $(\sqrt{V_p} - \sqrt{V_q})^2 \le (1 - \sqrt{v_{\min}})^2 = \rho_\kappa^2$; and $1 - t \le -\log t$ gives $2\sqrt{V_pV_q}\,(1 - c(x)) \le \sqrt{V_pV_q}\,\big(-2\log c(x)\big) \le \ell(x)$, as $\sqrt{V_pV_q} \le 1$ and $\ell(x) = -2\log c(x)$ is the bracket of Eq. (4). Integrating completes the proof.

Full constant form of Theorem 1. Under the conditions of the theorem, simultaneously for all $\theta \in \Theta$,

$$\big|\widehat{\mathcal{L}}_{\mathrm{CSP}}(\theta) - \mathcal{L}_{\mathrm{CSP}}(\theta)\big| \le C_{\mathrm{gen}} \sqrt{\frac{\bar{p}\log\big(nB_\theta L_\pi/\delta\big) + \log(4/\delta)}{n}}, \tag{11}$$

with $C_{\mathrm{gen}} = \mathcal{O}\big(\sqrt{K}\, C_W v_{\min}^{-1}(1 + \omega_0^{-1} + L_\pi L_\Phi R_{\mathrm{ctx}})\big)$ and $C_W = W$ for causal windows (N_w sampled context blocks per Assumption 2(iv); block-level McDiarmid coupling and the covering argument of the proof), and the analogous bound for $\widehat{\mathcal{L}}_{\mathrm{BC}}$ with $8R_a^2$ in place of C_{gen} (Hoeffding at the same block resolution). The slack of Eq. (5) is $\varepsilon_{\mathrm{gen}}(n, \delta) := \max\big(C_{\mathrm{gen}}, 8R_a^2\big)\sqrt{\frac{\bar{p}\log(nB_\theta L_\pi/\delta) + \log(4/\delta)}{n}}$, and $C_R = 2L_r\bar{H}^2(1 + L_P\bar{H})$.

Proof of Theorem 1. Fixed θ. Condition on the realized geometry: within a gradient step the conditioning path is detached (Algorithm 1), so the weights w_{ij} are fixed functionals of the sampled conditions, and the only random dependence of each plug-in potential is through which anchors and pair terms the replaced unit contains. Causality caps this: position ℓ lies in at most C_W local sets $\mathcal{W}_i$, and inside each affected row a single point shifts the normalized average by at most $c_V = \mathcal{O}\big((1 + \omega_0^{-1})/n\big)$ (using $m(x) \ge m_{\min}$, Assumption 2(iv), and $0 \le \kappa \le 1$). On $[v_{\min}, 1]$ the bracket ℓ has gradient at most $4/v_{\min}$, hence the influence of one replaced position is $c_\ell \le 4C_w c_V/v_{\min}$. Under the block sampling of Assumption 2(iv) the McDiarmid units are the $N_w = n/K$ independent context blocks: replacing one block removes at most K anchor terms and KC_W pair terms, so $c_{\mathrm{blk}} \le K(1 + C_W)c_\ell$ and $\sum_{\mathrm{blocks}} c_{\mathrm{blk}}^2 = N_w c_{\mathrm{blk}}^2 = \mathcal{O}\big(KC_W^2 v_{\min}^{-2}/n\big)$, whence McDiarmid's bounded-difference inequality (Shalev-Shwartz & Ben-David, 2014) yields $\Pr\{|\widehat{\mathcal{L}}_{\mathrm{CSP}} - \mathcal{L}_{\mathrm{CSP}}| > t\} \le 2e^{-2nt^2 v_{\min}^2/(16C_0KW)^2}$ with C_0 the block-influence constant: the graph coupling costs KW, and within-window trajectory dependence contributes the $\sqrt{K}$ factor of C_{gen}. $\widehat{\mathcal{L}}_{\mathrm{BC}}$ is bounded by $4R_a^2$, giving Hoeffding at the same block resolution. *Uniformizing.* By the Gaussian action kernel, ℓ is $L_\theta = \mathcal{O}(L_\pi L_\Phi)$-Lipschitz in θ (kernel derivatives are bounded by $L_\kappa = \sqrt{2/e}/\sigma_a$, and policy smoothness comes from Assumption 2(ii)). Covering Θ at radius $\epsilon_\theta/(2L_\theta)$ with $|\mathcal{N}| \le (16B_\theta/\epsilon_\theta)^{\bar{p}}$ (Bartlett, 1998), a union bound gives the $\bar{p}\log(nB_\theta L_\pi/\delta)$ term; the net error $L_\theta\epsilon_\theta/(2L_\theta)$ is absorbed into C_{gen}. *Excess.* Both empirical and population objectives are $\varepsilon_{\mathrm{gen}}$-close uniformly, so $\mathcal{L}(\hat{\theta}) \le \widehat{\mathcal{L}}(\hat{\theta}) + \varepsilon_{\mathrm{gen}} \le \widehat{\mathcal{L}}(\theta_0) + \varepsilon_{\mathrm{gen}} \le \mathcal{L}(\theta_0) + 2\varepsilon_{\mathrm{gen}}$; Assumption 2(i) gives $\mathcal{L}(\theta_0) \le \sigma_B^2 + \lambda_{\mathrm{CSP}} \cdot 0$.

Proof of Corollary 1. (i) reads Eq. (5). (ii) Couple $\pi_{\hat{\theta}}$ with μ from the same start: the per-step state divergence obeys $e_{t+1} \le L_P e_t + L_P\|\hat{\mathbf{a}}_t - \mathbf{a}_t\|$ with $\mathbf{a}_t = \mu(c_t)$, so $e_t \le L_P \sum_{k<t} L_P^k\, \mathbb{E}\|\hat{\mathbf{a}} - \mathbf{a}\|$, and $|J(\mu) - J(\pi_{\hat{\theta}})| \le \sum_t \gamma^t L_r(e_t + \mathbb{E}\|\hat{\mathbf{a}}_t - \mathbf{a}_t\|)$; summing the two geometric series gives C_R (reduction coupling, Ross et al. 2011; $\gamma L_P < 1$ by Assumption 2(iii)), and Jensen bounds $\mathbb{E}\|\hat{\mathbf{a}} - \mathbf{a}\| \le \sqrt{\mathcal{L}_{\mathrm{BC}}}$. (iii) Square both sides of (ii) and solve for n.

Proof of Lemma 2. Write $\Sigma_{\xi\xi}^0, \Sigma_{a\xi}^0$ for the unregularized weighted statistics, $\hat{\Sigma}_{\xi\xi}^0 = \Sigma_{\xi\xi}^0 + \Delta$. *Concentration.* $\sum_j q_{tj} = 1$, $q_{tj} \le c_q/N_{\mathrm{eff}}$, $\sum_j q_{tj}^2 = 1/N_{\mathrm{eff}}$: for jointly sub-Gaussian $(\delta\boldsymbol{\xi}, \delta\mathbf{a})$, the weighted second moments deviate at rate $\|\Delta\| \le \varepsilon_\xi$, $\|\hat{\Sigma}_{a\xi}^0 - \Sigma_{a\xi}^0\| \le \varepsilon_{a\xi}$, $\varepsilon_\xi + \varepsilon_{a\xi} \le C\sigma_\xi(\sigma_a + \|B^\star\|\Lambda_\xi^{1/2})\sqrt{c_q(d_\xi + d_a + 1)\log(3K_rT_e/\delta')/N_{\mathrm{eff}}}$ (vector-chaos with incoherent weights; cf. Vershynin, 2012), with the union bound $\delta' = \delta/3K_rT_e$ folded into the log. *Algebra.* The first-order condition of Eq. (13) gives $\hat{B}_t(\Sigma_{\xi\xi}^0 + \Delta + \lambda_\xi I) = \Sigma_{a\xi}^0 + E$, $E := \hat{\Sigma}_{a\xi}^0 - \Sigma_{a\xi}^0$; the normal

equation $B^\star\Sigma^0_{\xi\xi} = \Sigma^0_{a\xi}$ then yields, after substituting $\hat{B}_t = B^\star + (\hat{B}_t - B^\star)$ in the term $\hat{B}_t\Delta$, $\big(\hat{B}_t - B^\star\big)\big(\Sigma^0_{\xi\xi} + \Delta + \lambda_\xi I\big) = E - B^\star\Delta - \lambda_\xi B^\star$. As $\sigma_{\min}(\hat{\Sigma}^0_{\xi\xi} + \lambda_\xi I) \geq \lambda_* + \lambda_\xi - \varepsilon_\xi$, on the good event $\varepsilon_\xi \leq \lambda_*/2$: $\|\hat{B}_t - B^\star\| \leq \frac{2\left(\varepsilon_{a\xi} + \|B^\star\|\varepsilon_\xi\right)}{\lambda_* + \lambda_\xi} + \frac{2\lambda_\xi\|B^\star\|}{\lambda_* + \lambda_\xi}$, which is Eq. (10): the terms match the scalar ridge rate of Caponnetto & De Vito (2007) with $n \to N_{\text{eff}}$; weighted means concentrate by scalar sub-Gaussian averaging (Shalev-Shwartz & Ben-David, 2014).

H.3 Adaptive episodic memory and the unified dual-timescale bound

Corollary 2 (AEM effective-sample bound)**.** *On the gate-active set, let* $\mathbf{a}_t^{\text{loc}} = \mu_a + B^\star(\boldsymbol{\xi}_t - \mu_\xi)$ *be the true local relation and define* $\rho_{\text{cov}} := 2\tau_t \log(1/\delta_{\text{read}})$ *(Euclidean coverage radius, Proposition 2(ii)),* $r_\xi := \rho_{\text{cov}} + C_2\sigma_\xi\sqrt{c_q \log(6K_rT_e/\delta)/N_{\text{eff}}}$*, and*

$$\varepsilon_e := \underbrace{(1 + |\alpha_\xi|)\varepsilon_B\, r_\xi}_{\textit{transport}} + \underbrace{C_2(\sigma_a + \|\hat{B}_t\|_{\text{op}}\sigma_\xi)\sqrt{\tfrac{c_q \log(6K_rT_e/\delta)}{N_{\text{eff}}}}}_{\textit{center}} + \underbrace{|1 - \alpha_\xi|\|B^\star\|\, r_\xi + \rho_B r_\xi^2}_{\textit{model bias}}.$$

Then $\|\tilde{\mathbf{a}}_t - \mathbf{a}_t^{\text{loc}}\|_2 \leq \varepsilon_e$*, whose estimation-driven part decays as* $\mathcal{O}(N_{\text{eff}}^{-1/2})$*; reaching* $\varepsilon_e \leq \varepsilon$ *thus suffices at* $N_{\text{eff}} \geq C\, c_q\, \sigma_\xi^2\big(\sigma_a + \|B^\star\|\Lambda_\xi^{1/2}\big)^2\, d_\Sigma\, \log(K_rT_e/\delta)/\varepsilon^2$ *for a small enough bias budget* $(d_\Sigma = d_\xi + d_a + 1)$*.*

$$\mathcal{N}_t = \text{TopK}_{j \in H_t}\big(-d_x^2(t,j)\big), \qquad q_{tj} = \frac{\rho_j \exp\big(-d_x^2(t,j)/2\tau_t^2\big)}{\sum_{k \in \mathcal{N}_t}\rho_k \exp\big(-d_x^2(t,k)/2\tau_t^2\big)}, \tag{12}$$

$$\hat{\mu}_{\xi,t} = \textstyle\sum_j q_{tj}\boldsymbol{\xi}_j,\ \ \hat{\mu}_{a,t} = \sum_j q_{tj}\mathbf{a}_j,\ \ \hat{\Sigma}_{\xi\xi} = \sum_j q_{tj}\delta\boldsymbol{\xi}_j\delta\boldsymbol{\xi}_j^\top + \lambda_\xi I,\ \ \hat{\Sigma}_{a\xi} = \sum_j q_{tj}\delta\mathbf{a}_j\delta\boldsymbol{\xi}_j^\top,$$

$$\hat{B}_t = \hat{\Sigma}_{a\xi}\hat{\Sigma}_{\xi\xi}^{-1} = \arg\min_B \sum_j q_{tj}\|\delta\mathbf{a}_j - B\,\delta\boldsymbol{\xi}_j\|_2^2 + \lambda_\xi\|B\|_F^2, \qquad \tilde{\mathbf{a}}_t = \hat{\mu}_{a,t} + \alpha_\xi\hat{B}_t\,\delta\boldsymbol{\xi}_t, \tag{13}$$

Proof of Corollary 2. Decompose $\tilde{\mathbf{a}}_t - \mathbf{a}_t^{\text{loc}} = (\hat{\mu}_{a,t} - \mu_a) + \alpha_\xi\big[\hat{B}_t(\boldsymbol{\xi}_t - \hat{\mu}_{\xi,t}) - B^\star(\boldsymbol{\xi}_t - \mu_\xi)\big] + (\alpha_\xi - 1)B^\star(\boldsymbol{\xi}_t - \mu_\xi) + \eta_t$, insert $(\boldsymbol{\xi}_t - \hat{\mu}_{\xi,t}) = (\boldsymbol{\xi}_t - \mu_\xi) + (\mu_\xi - \hat{\mu}_{\xi,t})$, bound $\|\boldsymbol{\xi}_t - \mu_\xi\| \leq \|\boldsymbol{\xi}_t - \boldsymbol{\xi}_{(1)}\| + \|\boldsymbol{\xi}_{(1)} - \mu_\xi\| \leq \rho_{\text{cov}}$ up to the center error, since the gate keeps the nearest memory within ρ_{cov} (Proposition 2(ii)), apply Lemma 2 and the mean bounds, and $\|\eta_t\| \leq \rho_B r_\xi^2$ in expectation (Assumption 3(ii)). Collecting gives ε_e; solve for N_{eff}.

$$\begin{aligned} N_{\text{eff}} &= \tfrac{1}{\sum_j q_{tj}^2}, \quad \bar{\rho}_t = \textstyle\sum_j q_{tj}\rho_j, \quad g_{\text{sup}} = 1 - e^{-\bar{\rho}_t N_{\text{eff}}/\beta_{\text{sup}}}, \\ r_t &= d_{\text{nn},t}/\tau_t, \quad g_{\text{cov}} = e^{-r_t/2}, \\ \Sigma_{a|\xi} &= \hat{\Sigma}_{aa} - \hat{\Sigma}_{a\xi}\hat{\Sigma}_{\xi\xi}^{-1}\hat{\Sigma}_{\xi a}, \\ g_{\text{act}} &= \exp\big(-\text{tr}(\Sigma_{a|\xi})/(d_a\bar{v}_t\beta_{\text{act}})\big), \quad g_t = g_{\text{sup}}\, g_{\text{cov}}\, g_{\text{act}}, \end{aligned} \tag{14}$$

Proof of Proposition 2 (validation part). (i) is the gate–fallback rule Eq. (14). (ii) each factor ≤ 1, so $g_t \geq \delta$ forces each $\geq \delta$; invert the exponentials and the monotone g_{sup}. (iii) is the write-after-execute order of Algorithm 2.

$$\begin{aligned} \mathbf{a}_t^\star = \arg\min_{\mathbf{a}}\ & \tfrac{\lambda_\pi}{2}\|\mathbf{a} - \mathbf{a}_t^0\|_2^2 + \tfrac{\lambda_e g_t}{2}(\mathbf{a} - \tilde{\mathbf{a}}_t)^\top\Lambda_t(\mathbf{a} - \tilde{\mathbf{a}}_t) + \tfrac{\lambda_{\text{T}}}{2}\|\mathbf{a} - \mathbf{a}_{t-1}\|_2^2 \\ \Rightarrow\ & A_t\mathbf{a}_t^\star = \mathbf{b}_t, \qquad \mathbf{a}_t = \Pi_{\mathcal{A}}\big(\mathbf{a}_t^0 + \text{SafeClip}_{\rho_{\max}}(\mathbf{a}_t^\star - \mathbf{a}_t^0)\big), \end{aligned} \tag{15}$$

with $A_t = (\lambda_\pi + \lambda_{\text{T}})I + \lambda_e g_t\Lambda_t \succeq (\lambda_\pi + \lambda_{\text{T}})I$, $\mathbf{b}_t = \lambda_\pi\mathbf{a}_t^0 + \lambda_e g_t\Lambda_t\tilde{\mathbf{a}}_t + \lambda_{\text{T}}\mathbf{a}_{t-1}$; $\Pi_{\mathcal{A}}$ projects to the action box and $\text{SafeClip}_{\rho_{\max}}$ caps $\|\mathbf{a}_t^\star - \mathbf{a}_t^0\|$ at $\rho_{\max}\|\mathbf{a}_t^0\|$ (unit-normalized Λ_t, absorbed into λ_e).

Proof of Lemma 3. $A_t(\mathbf{a}_t^\star - \mathbf{a}_t^0) = \lambda_e g_t \Lambda_t u' + \lambda_{\mathrm{T}} v'$ with $A_t = (\lambda_\pi + \lambda_{\mathrm{T}})I + \lambda_e g_t \Lambda_t$. Diagonalizing Λ_t with eigenvalues $\ell \in [0, 1/\lambda_a]$ ($\|\Lambda_t\|_{\mathrm{op}} \le 1/\lambda_a$, $g_t \le 1$), the induced scalar multipliers are $\alpha(\ell) = \lambda_e g_t \ell/(\lambda_\pi + \lambda_{\mathrm{T}} + \lambda_e g_t \ell)$ and $\beta(\ell) = \lambda_{\mathrm{T}}/(\lambda_\pi + \lambda_{\mathrm{T}} + \lambda_e g_t \ell)$. By monotonicity in $g_t \le 1$, $\alpha(\ell) \le \frac{\lambda_e/\lambda_a}{\lambda_\pi + \lambda_{\mathrm{T}} + \lambda_e/\lambda_a} \le 1 - \vartheta$, while $\beta(\ell) \le \frac{\lambda_{\mathrm{T}}}{\lambda_\pi + \lambda_{\mathrm{T}}}$ and $\Big(1 - \vartheta\Big) - \frac{\lambda_{\mathrm{T}}}{\lambda_\pi + \lambda_{\mathrm{T}}} = \frac{\lambda_\pi\,(\lambda_e/\lambda_a)}{(\lambda_\pi + \lambda_{\mathrm{T}})(\lambda_\pi + \lambda_{\mathrm{T}} + \lambda_e/\lambda_a)} \ge 0$, hence $\beta(\ell) \le 1 - \vartheta$ as well; both suprema are attained at the worst ℓ independently, so $\|\mathbf{a}_t^\star - \mathbf{a}_t^0\| \le (1 - \vartheta)(\|u'\| + \|v'\|)$ without aligning u' and v'. The convex box contains $\mathbf{a}_t^0$ (tanh head), so the firmly-nonexpansive $\Pi_{\mathcal{A}}$ and the ratio cap $\mathrm{SafeClip}_{\rho_{\max}}$ do not enlarge the anchored norm, giving the minimum.

Corollary 3 (Bounded value impact of the fast timescale)**.** *The deployed controller $\tilde{\pi}$ and the frozen policy differ in value by at most the gate-validated correction:* $\big|\widetilde{J}(\tilde{\pi}) - \widetilde{J}(\pi_{\theta^*})\big| \le C_R\,\bar{\rho}, \quad \bar{\rho} := \mathbb{E}\big[\sum_t \gamma^t (1-\gamma)\mathbf{1}\{g_t \ge \delta_{\mathrm{read}}\}\big] \cdot \rho_{\mathrm{cap}} \le \min\{\rho_{\mathrm{cap}},\, \rho_{\max} R_a \bar{p}_e\}$*, for the perturbed environment as well, with the same C_R and $\bar{p}_e$ the intervention rate: adaptation is a bounded, auditable perturbation of the slow-timescale guarantee, reversible at the action layer.*

Proof of Corollary 3. Apply the coupling telescoping of Corollary 1 with per-step action deviation bounded by Proposition 2 ($\le \rho_{\mathrm{cap}}$ when active, 0 otherwise); the same C_R applies since both controllers share contexts and initial conditions.

Proof of Theorem 2. Telescope: $\widetilde{J}(\mu) - \widetilde{J}(\tilde{\pi}) = \big[\widetilde{J}(\mu) - \widetilde{J}(\pi_{\theta^*})\big] + \big[\widetilde{J}(\pi_{\theta^*}) - \widetilde{J}(\tilde{\pi})\big]$. *Second bracket:* Corollary 3 gives $C_R\bar{\rho}$, term (III), with estimation part $\mathcal{O}(N_{\mathrm{eff}}^{-1/2})$ (Corollary 2). *First bracket:* run the coupling of Corollary 1(ii) inside $\widetilde{J}$. For a visited shifted condition $\tilde{x}$, insert $\mu_q(x), \mu_p(x)$ and reproduce through the kernel ($\hat{\mathbf{a}}$ Dirac: embedding $\varphi(\hat{\mathbf{a}})$), then apply the comparison modulus $\bar{\varsigma}$: the action deviation is at most $\bar{\varsigma}$ of the sum of (a) the anchor consistency term $\|\varphi(\hat{\mathbf{a}}(x)) - \mu_q(x)\|_{\mathcal{H}_{\mathrm{k}}} \le L_\kappa\big(\mathcal{L}_{\mathrm{BC}}(\hat{\theta})^{1/2} + L_\pi L_\Phi \,\text{nbhd. radius}\big)$, charged to (I) on the nominal condition set by Corollary 1(i) (reproducing property: μ_q is the kernel mixture of the policy's own Diracs on BC-pinned anchors); (b) $\|\mu_q(x) - \mu_p(x)\|_{\mathcal{H}_{\mathrm{k}}}$, charged to (II) by Lemma 1 as $(\rho_\kappa^2 + 2\varepsilon_{\mathrm{gen}}/\lambda_{\mathrm{CSP}})^{1/2}$; (c) the two continuity terms $\omega_\mu(d_x(x, \tilde{x}))$ (Assumption 2(v)) and $\|\varphi(\hat{\mathbf{a}}(\tilde{x})) - \varphi(\hat{\mathbf{a}}(x))\|_{\mathcal{H}_{\mathrm{k}}} \le L_\kappa L_\pi L_\Phi d_x(x, \tilde{x})$ ($L_\kappa = \sup|\nabla\kappa|^{1/2}$), both $\le \omega_\mu(\bar{r}_x) + L_\kappa L_\pi L_\Phi \bar{r}_x$ and absorbed into (II)'s argument. Shifting the reference return from J to $\widetilde{J}$ for both controllers adds at most $2\bar{r}\sum_t \gamma^t\, t\, \varepsilon_P \bar{D}_{\mathcal{S}} L_P$, i.e. (IV) ($\gamma L_P < 1$ summation). *Coupling.*The slow-timescale charges (a)–(c) hold at visited conditions. Under $\tilde{\pi}$ the correction additionally displaces them, $\tilde{e}_{t+1} \le L_P \tilde{e}_t + L_P \rho_t$, whose discounted solution adds $\kappa_\rho \bar{\rho}$ to ω_μ's argument, so the net extra charge is exactly the increment $\bar{\varsigma}(\omega_\mu(\bar{r}_x + \kappa_\rho\bar{\rho}) - \omega_\mu(\bar{r}_x))$, term (V). The two scales need not be independent. Their interaction is charged explicitly as (V), and the C_R-weighted Gronwall telescoping completes the sum.