BIRD: BELIEF-ENCODER INDEX RESTLESS DIFFUSION FOR PARTIALLY OBSERVABLE RMABS

Anonymous authors
Paper under double-blind review

ABSTRACT

Restless multi-armed bandits (RMABs) provide a principled framework for resource-constrained sequential decision-making, yet classical Whittle-index methods assume fully observable, discrete states and known dynamics. These assumptions break down in partially observable RMABs with continuous latent states and unknown or drifting dynamics, motivating stochastic score models beyond deterministic rankings. We present BIRD (**B**elief-encoder **I**ndex **R**estless **D**iffusion), a scalable score-based algorithm that uses a shared belief encoder and diffusion actor to sample stochastic priority scores, together with a global Top-K projection to enforce the hard budget. We show that the Top-K projection maps score-space policy mirror descent (PMD) to a corresponding update over budget-feasible RMAB actions. Across synthetic and MIMIC-derived environments, BIRD outperforms random, greedy, NeurWIN, PPO, and learned-rollout baselines. Actor ablations isolate the benefit of the shared per-arm diffusion actor over deterministic MLP actor, Gaussian actor, and global joint-arm diffusion variants.

1 INTRODUCTION

Restless multi-armed bandits (RMABs) (Whittle, 1988) model settings in which a decision-maker must repeatedly allocate a fixed budget of interventions across a population of arms whose states evolve over time, regardless of whether they receive an intervention. This framework has been applied to public-health treatment prioritization (Ayer et al., 2019; Mate et al., 2020), wireless scheduling (Nakhleh et al., 2021), recommendation systems (Meshram & Kaza, 2021b), and resource allocation in operations research (Adelman & Mersereau, 2008).

Solving RMABs exactly is PSPACE-hard in general (Papadimitriou & Tsitsiklis, 1999), due to the combinatorial coupling induced by the global budget constraint. As a result, many practical approaches rely on the Whittle-index heuristic (Whittle, 1988), which relaxes the resource constraint via a Lagrangian formulation and decomposes the problem into independent per-arm index computations. Under classical assumptions—discrete Markovian state spaces, full observability, known transition dynamics, and indexability—Whittle-index policies are well studied and have strong empirical and asymptotic theoretical support (Weber & Weiss, 1990; Niño-Mora, 2023).

However, in many applications, the assumptions behind the classical index methods are difficult to satisfy. Arm states may be continuous, only indirectly observed, and governed by unknown or drifting transition dynamics. Partial observability turns the state of each arm into a belief state, i.e. a posterior distribution over latent states conditioned on its observation-action history (Kaelbling et al., 1998).

Prior work has extended Whittle-index ideas to partially observable RMABs, but mostly for structured settings where the belief dynamics are known and indexability or threshold structure can be established (Akbarzadeh & Mahajan, 2022; Meshram & Kaza, 2021a; Ornee & Sun, 2023). These methods provide important structural insight, but they still require model-specific derivations and explicit Whittle-index computation. Partial observability and drifting dynamics also create uncertainty over priority rankings. Similar observation-action histories may correspond to different latent states or arm types. This makes deterministic ranking policies restrictive and motivates stochastic score models that can represent uncertainty over arm priorities.

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.

Our approach. We propose BIRD, which learns stochastic priority scores directly from data by using the expressiveness of diffusion models. Instead of collapsing each arm to a deterministic ranking, BIRD models a distribution over priority scores conditioned on each arm's observation-action history. A shared belief encoder first maps each history to a belief embedding, a shared per-arm diffusion actor samples priority scores, and a global Top-K projection converts these scores into budget-feasible actions. This separates expressive continuous score modeling from discrete hard-budget enforcement, allowing BIRD to handle partially observable and drifting RMABs without explicit belief-state dynamic programming.

Contributions.

- We develop BIRD, a scalable architecture for partially observable RMABs. A shared belief encoder maps observation-action histories to per-arm belief embeddings; a shared per-arm diffusion actor models stochastic priority-score distributions; a Top-K projection enforces the hard budget; and a critic provides Q-values for Q-weighted denoising updates of the diffusion actor.
- We prove (Proposition 5.1) that the Top-K projection pushes the score-space policy mirror descent (PMD) target to a corresponding update over the budget-feasible RMAB action simplex, and use Q-weighted denoising to approximate this target within a shared per-arm actor family.
- We first include a fully observable two-state Markov RMAB sanity check, where exact Whittle indices are available, to verify that BIRD learns behavior close to the classical index benchmark. We then evaluate BIRD across three partially observable RMAB environments of increasing complexity: a synthetic stationary environment, a synthetic drifting environment, and a clinical simulator derived from MIMIC-III. Results show consistent gains over NeurWIN, PPO, greedy, and random baselines, and actor ablations isolate the benefit of the shared per-arm diffusion actor over deterministic MLP actor, Gaussian actor, and global joint-arm diffusion variants, especially under non-stationarity.

2 Related Work

Whittle-index methods. Whittle (1988) introduced the index policy for discrete, fully observable, known-dynamics RMABs. Weber & Weiss (1990) established asymptotic optimality. Extensions to unknown dynamics use optimism (Wang et al., 2023) or Bayesian approaches (Liang et al., 2025). Learning-based methods approximate indices with neural networks (Nakhleh et al., 2021; Xiong & Li, 2023).

Partially observable RMABs. Mate et al. (2020) studied two-state partially observable RMABs where activation reveals state; Akbarzadeh & Mahajan (2022) derived efficient index computation under a restart property. Meshram & Kaza (2025) developed Lagrangian PBVI methods for multi-action partially observable RMABs with *known* dynamics and finite discrete states, noting that belief-MDP indexability is an open problem in general. Our work handles continuous latent states and unknown dynamics without requiring known transition or observation matrices.

Diffusion models for RL. Psenka et al. (2023) proposed Q-score matching, linking the diffusion policy score to action gradients of a learned Q-function. Ma et al. (2025) introduced Reweighted Score Matching (RSM) for online diffusion-policy learning and derived Diffusion Policy Mirror Descent (DPMD). BIRD extends this RSM/DPMD framework to RMABs through a Top-K score-action reparameterization and a shared per-arm critic. Diffusion models have also been used for world-model learning (Li et al., 2025; Rigter et al., 2023) and trajectory planning (Jackson et al., 2024), but not for hard-budget control in partially observable RMABs.

3 Problem Formulation

We consider a partially observable RMAB with N independent arms and a per-period activation budget K. At each time t, arm i has continuous latent state $s_{i,t} \in \mathcal{S}$, receives binary action $a_{i,t} \in \{0, 1\}$, and evolves as $s_{i,t+1} \sim P_i(\cdot \mid s_{i,t}, a_{i,t})$. The controller does not observe $s_{i,t}$ directly; it receives an observation $o_{i,t} \sim O_i(\cdot \mid s_{i,t})$ and a scalar reward $r_i(s_{i,t}, a_{i,t})$. In our setup, the reward is a deterministic function of the latent state. Each individual arm can be viewed as a POMDP, with

the arms coupled exclusively through the hard budget constraint $\sum_{i=1}^{N} a_{i,t} = K$ at every decision step t.

Belief MDP. The controller summarizes the observation and action history of arm i through its belief density $b_{i,t}(s) = \Pr(s_{i,t} = s \mid o_{i,0:t}, a_{i,0:t-1})$. A policy $\pi(\mathbf{a_t} \mid \mathbf{b_t})$ assigns a probability to each feasible activation vector $\mathbf{a_t} = (a_{1,t}, \ldots, a_{N,t})$, given the joint belief $\mathbf{b_t} = (b_{1,t}, \ldots, b_{N,t})$. The expected reward is $R(\mathbf{b_t}, \mathbf{a_t}) = \sum_{i=1}^{N} \mathbb{E}_{s \sim b_{i,t}}[r_i(s, a_{i,t})]$. It then updates each belief using the state transition and observation models. However, when arm dynamics are unknown or drifting, a sufficient belief must also account for latent arm types and time-varying model parameters. This enlarges the belief space and makes exact single-arm planning difficult.

Whittle-index policy. The joint activation budget further complicates planning by coupling action selection across arms. A Whittle-index policy addresses this by assigning each arm a score derived from a price-adjusted single-arm problem, then activating the K arms with the highest scores.

Consider a stationary single-arm belief model. For an activation price $\lambda \in \mathbb{R}$, define the price-adjusted reward (Lagrangian reward) $R_i^{\lambda}(b_i, a_i) = R_i(b_i, a_i) - \lambda a_i$. The corresponding optimal discounted value is

$$V_i^{\lambda}(b_i) = \max_{\pi_i} \mathbb{E}_{\pi_i}\left[\sum_{t=0}^{\infty} \gamma^t \big(R_i(b_{i,t}, a_{i,t}) - \lambda a_{i,t}\big) \,\big|\, b_{i,0} = b_i\right], \qquad 0 < \gamma < 1, \tag{1}$$

where π_i is a single-arm policy. Its action-value is

$$Q_i^{\lambda}(b_i, a) = R_i(b_i, a) - \lambda a + \gamma\, \mathbb{E}[V_i^{\lambda}(b_i') \mid b_i, a].$$

If the single-arm problem is indexable, its Whittle index is

$$W_i(b_i) = \inf\big\{\lambda : Q_i^{\lambda}(b_i, 0) \geq Q_i^{\lambda}(b_i, 1)\big\}. \tag{2}$$

Score-based control. Exact Whittle-index computation is intractable in partially observable RMABs with continuous states and unknown dynamics. We therefore use the index viewpoint as motivation for score-based control: BIRD learns a data-driven priority score for each arm and enforces the original hard budget by a global Top-K projection. These scores are not exact Whittle indices, and we do not assert indexability or convergence to a Whittle-index policy.

4 BIRD: Belief-Encoder Index Restless Diffusion

For each arm i, BIRD separates three quantities: the belief embedding z_i, the continuous priority score w_i, and the binary activation a_i, as shown in Figure 1. A shared BeliefEncoder maps each arm's history to z_i, a shared per-arm diffusion actor samples priority scores $\mathbf{w} \in \mathbb{R}^N$, a deterministic Top-$K$ projection enforces the budget constraint, and the critic evaluates the induced binary actions.

4.1 Architecture

Belief encoder. Since the true belief $b_{i,t}$ is unavailable, BIRD learns a belief embedding from each arm's local interaction history. Let $h_{i,t} = (o_{i,t-L+1:t},\ a_{i,t-L:t-1})$ be an L-step window. A shared encoder f_ϕ maps this history to an embedding:

$$z_{i,t} = f_\phi(h_{i,t}) \in \mathbb{R}^{d_z}. \tag{3}$$

In our implementation, f_ϕ is a Transformer encoder applied independently per arm, with tokens $x_\tau = \mathrm{cat}[o_{i,\tau}, a_{i,\tau}]$. Parameter sharing allows generalization across arms while preserving arm-specific histories.

To encourage z_i to capture latent arm dynamics, we use an auxiliary one-step observation prediction head $g_\phi(z_{i,t}, a_{i,t})$:

$$\mathcal{L}_{\mathrm{aux}} = \mathbb{E}\big[\|g_\phi(z_{i,t}, a_{i,t}) - o_{i,t+1}\|_2^2\big]. \tag{4}$$

This is especially useful when observations are noisy and arm type must be inferred from history.

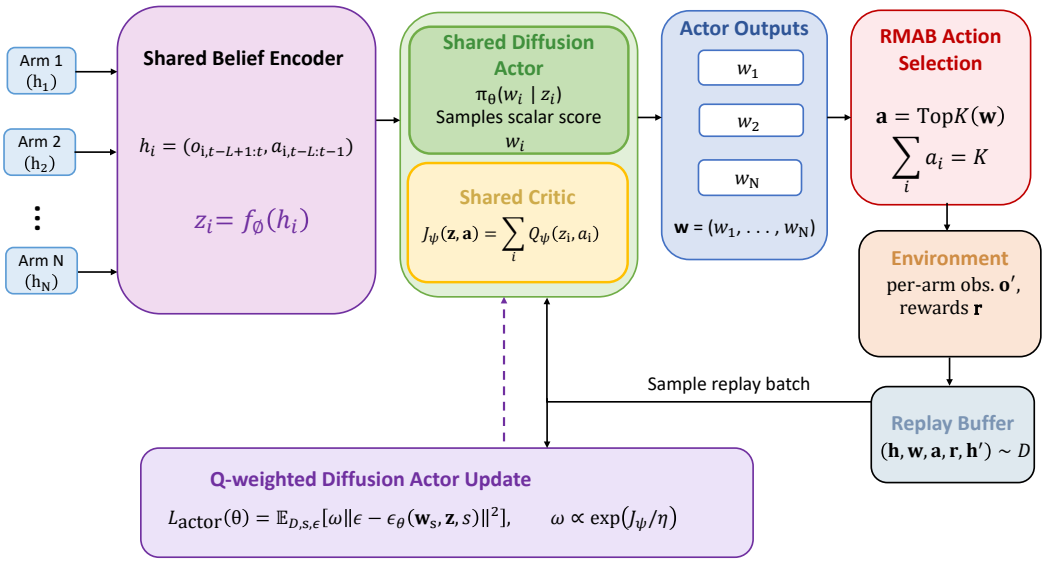


Figure 1: Overview of BIRD. Each arm's local history $h_i = (o_{i,t-L+1:t}, a_{i,t-L:t-1})$ is encoded as $z_i = f_\phi(h_i)$ by a shared belief encoder. A shared diffusion actor samples scalar priority scores w_i, collected into $\mathbf{w}$, and the RMAB action-selection module applies $\mathbf{a} = \mathrm{TopK}(\mathbf{w})$ with $\sum_i a_i = K$. The resulting action is executed in the environment, and replay batches train the shared critic $J_\psi(\mathbf{z}, \mathbf{a}) = \sum_i Q_\psi(z_i, a_i)$, whose values provide Q-weighted denoising updates for the diffusion actor. In the actor loss, s is the diffusion denoising step, ϵ is sampled Gaussian noise, and ω is the critic-derived Q weight. For clarity, the decision time index is shown only in the history window and omitted elsewhere.

Per-arm diffusion score actor. BIRD uses a shared per-arm diffusion actor for scalability and permutation equivariance while modeling stochastic priority-score distributions. Unlike a deterministic actor, diffusion can represent multiple plausible priority scores for the same belief embedding. For each arm i, the actor samples $w_{i,t} \sim \pi_\theta(\cdot \mid z_{i,t})$, and the sampled scores form $\mathbf{w}_t = (w_{1,t}, \ldots, w_{N,t})$. We write $\pi_\theta(\mathbf{w}_t \mid \mathbf{z}_t)$ for the induced distribution over this collected score vector, not a separate joint denoising model.

This stochasticity is motivated by rank uncertainty: in the drifting synthetic simulator, a particle-filter diagnostic shows that posterior latent samples for a fixed observed joint history can move a boundary arm across the Top-K cutoff. On such boundary histories, BIRD's sampled Top-K sets also have higher privileged oracle value than the deterministic MLP actor ablation on average; see Figure 2.

The sampled scores affect the policy only through their induced Top-K actions: at evaluation, $\mathbf{a}_t = \mathrm{TopK}(\mathbf{w}_t)$. During training, stored clean score vectors are reweighted by the critic value of their induced Top-K actions and used as targets in the weighted denoising loss.

Per-arm critic. For each arm i, the twin critic estimates the value of binary activations:

$$Q_\psi(z_{i,t}, a_{i,t}) = \min\{Q_{\psi_1}(z_{i,t}, a_{i,t}), Q_{\psi_2}(z_{i,t}, a_{i,t})\}\,. \tag{5}$$

The global action value is approximated additively as $J_\psi(\mathbf{z}_t, \mathbf{a}_t) = \sum_{i=1}^N Q_\psi(z_{i,t}, a_{i,t})$. This is a structured approximation, not an exact decomposition of the budget-coupled value function. The minimum of two critics reduces overestimation but can be pessimistic; Section 6.3 compares single, twin, and joint critics.

Parameter sharing and budget coupling. The encoder, diffusion actor, and critic are shared across arms. Arms are coupled only through the global TopK projection, which enforces the hard budget constraint.

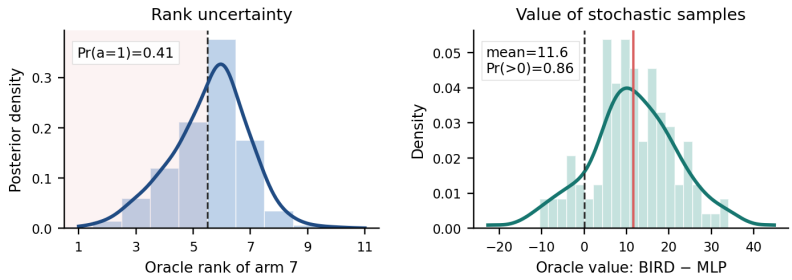


Figure 2: Boundary diagnostic in the drifting synthetic RMAB. Left: posterior latent samples for a fixed observed joint history move a boundary arm across the Top-K cutoff. Right: on boundary histories, BIRD's sampled Top-K actions have higher privileged oracle value than the deterministic MLP actor ablation. The oracle and particle filter are used only for analysis; details are in Appendix D.

4.2 Score-Action Reparameterization

Directly modeling feasible RMAB actions is poorly matched to diffusion modeling, because the feasible action set $\mathcal{A}_K = \{\mathbf{a} \in \{0,1\}^N : \sum_i a_i = K\}$ is discrete and grows as $\binom{N}{K}$. Instead, BIRD applies diffusion in the continuous score space: it samples a continuous score vector $\mathbf{w}_t \sim \pi_\theta(\cdot \mid \mathbf{z}_t)$ and deterministically maps it to a feasible activation vector:

$$\mathbf{a}_t = a(\mathbf{w}_t) = \mathrm{TopK}(\mathbf{w}_t), \qquad a_i(\mathbf{w}_t) = \mathbf{1}\{w_{i,t} \text{ is among the top } K \text{ entries of } \mathbf{w}_t\}.$$

This separates continuous score learning from discrete budget enforcement: the diffusion actor models stochastic priority-score distributions, while the Top-K projection enforces the hard budget exactly.

4.3 Critic Learning

We maintain target actor and critic networks, denoted by $\pi_{\bar{\theta}}$ and $Q_{\bar{\psi}}$, which are updated by exponential moving averages of the online parameters. Given a stored transition $(\mathbf{z}_t, \mathbf{a}_t, \mathbf{r}_t, \mathbf{z}_{t+1})$, the target actor samples $\mathbf{w}'_{t+1} \sim \pi_{\bar{\theta}}(\cdot \mid \mathbf{z}_{t+1})$, inducing $\mathbf{a}'_{t+1} = \mathrm{TopK}(\mathbf{w}'_{t+1})$. The TD target for arm i is $\mathcal{T}_{i,t} = r_{i,t} + \gamma\, Q_{\bar{\psi}}(z_{i,t+1}, a'_{i,t+1})$. The critic loss, jointly with the auxiliary loss, is

$$\mathcal{L}_{\mathrm{critic}}(\psi) = \mathbb{E}\left[\sum_i \left(Q_\psi(z_{i,t}, a_{i,t}) - \mathcal{T}_{i,t}\right)^2\right] + \alpha\, \mathcal{L}_{\mathrm{aux}}. \tag{6}$$

4.4 Algorithm Summary (see Algorithm 1)

BIRD alternates between environment interaction and replay-based modular updates. During interaction, the shared encoder maps each arm's history to a belief embedding, the shared per-arm diffusion actor samples one activation score per arm, and a global Top-K map converts the collected scores into a budget-feasible action. During training, replayed transitions update the critic with target-network TD regression, while stored clean score vectors are reweighted by critic values and used in a Q-weighted denoising loss to update the diffusion actor.

5 Score-Space Policy Mirror Descent for RMABs

Section 4 introduced BIRD's score-action reparameterization: a diffusion actor samples continuous scores $\mathbf{w} \in \mathbb{R}^N$, and $\mathbf{a} = \mathrm{TopK}(\mathbf{w})$ enforces the RMAB budget. We justify this design through policy mirror descent (PMD). PMD updates a policy by maximizing expected Q-value subject to a KL penalty from a reference policy, thereby increasing mass on high-value actions without taking an overly large policy step. The resulting target is an exponentially Q-weighted distribution, which Q-weighted denoising can fit naturally in continuous action spaces. In contrast, RMAB actions are discrete hard-budget binary vectors, making direct diffusion over feasible actions unnatural. We show that the Top-K pushforward of score-space PMD gives the corresponding KL-regularized PMD update over feasible RMAB actions.

Algorithm 1 BIRD Training

Require: arms N, budget K; shared per-arm encoder f_ϕ, actor π_θ, critic Q_ψ, and target networks $\pi_{\bar{\theta}}, Q_{\bar{\psi}}$; replay buffer $\mathcal{D}$; PMD temperature schedule η_t

for each environment step t **do**
  For all arms $i = 1, \ldots, N$: encode $z_{i,t} = f_\phi(h_{i,t})$ and sample $w_{i,t} \sim \pi_\theta(\cdot \mid z_{i,t})$
  Collect $\mathbf{w}_t = (w_{1,t}, \ldots, w_{N,t})$; select $\mathbf{a}_t = \mathrm{TopK}(\mathbf{w}_t)$, with $\sum_i a_{i,t} = K$
  Execute $\mathbf{a}_t$; store $(\mathbf{h}_t, \mathbf{w}_t, \mathbf{a}_t, \mathbf{r}_t, \mathbf{h}_{t+1})$ in $\mathcal{D}$
end for
for each gradient update **do**
  Sample minibatch $(\mathbf{h}, \mathbf{w}_0, \mathbf{a}, \mathbf{r}, \mathbf{h}') \sim \mathcal{D}$, where $\mathbf{w}_0$ is the clean stored score vector
  Re-encode $\mathbf{z} = f_\phi(\mathbf{h})$, $\mathbf{z}' = f_\phi(\mathbf{h}')$
  Update critic by one-step TD regression with target networks
  Set $\mathbf{a}_0 = \mathrm{TopK}(\mathbf{w}_0)$ and compute

$$J_\psi(\mathbf{z}, \mathbf{w}_0) = \sum_i Q_\psi(z_i, a_{0,i})$$

  Compute PMD weights

$$\omega(\mathbf{z}, \mathbf{w}_0) \propto \exp(J_\psi(\mathbf{z}, \mathbf{w}_0)/\eta_t)$$

  Sample diffusion step τ and noise ϵ_i; form $w_{i,\tau} = \sqrt{\bar{\alpha}_\tau} w_{i,0} + \sqrt{1 - \bar{\alpha}_\tau}\epsilon_i$
  Update actor with weighted per-arm denoising:

$$\mathcal{L}_{\mathrm{RSM}} = \omega(\mathbf{z}, \mathbf{w}_0) \sum_i \|\epsilon_i - \epsilon_\theta(w_{i,\tau}, z_i, \tau)\|_2^2$$

  Soft-update target networks $\pi_{\bar{\theta}}, Q_{\bar{\psi}}$; update η_t
end for

5.1 TOP-K PUSHFORWARD OF SCORE-SPACE PMD

Let $\mathcal{A}_K = \{\mathbf{a} \in \{0,1\}^N : \sum_i a_i = K\}$ be the feasible RMAB action set, and let $T(\mathbf{w}) = \mathrm{TopK}(\mathbf{w})$ map a score vector $\mathbf{w} \in \mathbb{R}^N$ to a feasible activation vector.

Following diffusion PMD (Ma et al., 2025), the PMD-improved score policy is

$$\pi^+(\mathbf{w} \mid \mathbf{z}) = \frac{\pi_{\mathrm{old}}(\mathbf{w} \mid \mathbf{z}) \exp(Q^{\pi_{\mathrm{old}}}(\mathbf{z}, T(\mathbf{w}))/\eta)}{Z_{\mathbf{w}}(\mathbf{z})}, \tag{7}$$

where $\eta > 0$ is the PMD temperature and $Z_{\mathbf{w}}(\mathbf{z}) = \int \pi_{\mathrm{old}}(\mathbf{w} \mid \mathbf{z}) \exp(Q^{\pi_{\mathrm{old}}}(\mathbf{z}, T(\mathbf{w}))/\eta)\, d\mathbf{w}$ normalizes the density. The following proposition shows that the Top-K pushforward of this score-space update is exactly the PMD update over budget-feasible RMAB actions.

Proposition 5.1 (Top-K pushforward PMD)**.** *Assume $\pi_{\mathrm{old}}(\cdot \mid \mathbf{z})$ is absolutely continuous, so score ties occur with probability zero. Let*

$$\mu_{\mathrm{old}}(\mathbf{a} \mid \mathbf{z}) = \int_{\{\mathbf{w}: T(\mathbf{w}) = \mathbf{a}\}} \pi_{\mathrm{old}}(\mathbf{w} \mid \mathbf{z})\, d\mathbf{w}$$

be the feasible-action distribution induced by the old score policy, and let μ^+ be the distribution induced by π^+ in (7). Then, for every $\mathbf{a} \in \mathcal{A}_K$,

$$\mu^+(\mathbf{a} \mid \mathbf{z}) = \frac{\mu_{\mathrm{old}}(\mathbf{a} \mid \mathbf{z}) \exp(Q^{\pi_{\mathrm{old}}}(\mathbf{z}, \mathbf{a})/\eta)}{Z_A(\mathbf{z})}, \tag{8}$$

where

$$Z_A(\mathbf{z}) = \sum_{\mathbf{a}' \in \mathcal{A}_K} \mu_{\mathrm{old}}(\mathbf{a}' \mid \mathbf{z}) \exp(Q^{\pi_{\mathrm{old}}}(\mathbf{z}, \mathbf{a}')/\eta)\,.$$

Thus, the Top-K pushforward of score-space PMD is the KL-regularized PMD update over the feasible action simplex $\Delta(\mathcal{A}_K)$.

Proof. See Appendix B. □

Proposition 5.1 justifies the score-action reparameterization. Although BIRD trains the actor over continuous scores, Top-K maps each score sample to a feasible RMAB action, and the resulting ideal target action distribution follows the same PMD update as direct optimization over feasible actions. In practice, $Q^{\pi_{\text{old}}}$ is replaced by the learned critic, and BIRD weights denoising updates using $J_\psi(\mathbf{z}, T(\mathbf{w})) = \sum_{i=1}^{N} Q_\psi(z_i, T_i(\mathbf{w}))$.

5.2 Diffusion Realization by Reweighted Score Matching

The shared per-arm diffusion actor is trained by weighted denoising of stored clean score vectors. For a stored score vector $\mathbf{w}_0 = (w_{1,0}, \ldots, w_{N,0})$, the per-arm forward process samples

$$w_{i,\tau} = \sqrt{\bar{\alpha}_\tau}\, w_{i,0} + \sqrt{1 - \bar{\alpha}_\tau}\, \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, 1),$$

and the shared denoising network $\epsilon_\theta(w_{i,\tau}, z_i, \tau)$ predicts ϵ_i. Under the RSM framework of Ma et al. (2025), the PMD target in (7) can be fit by a weighted denoising loss:

$$\mathcal{L}_{\text{RSM}}(\theta) = \mathbb{E}_{\mathbf{z}, \mathbf{w}_0, \epsilon, \tau}\left[\omega(\mathbf{z}, \mathbf{w}_0) \sum_{i=1}^{N} \|\epsilon_i - \epsilon_\theta(w_{i,\tau}, z_i, \tau)\|_2^2\right], \tag{9}$$

with weight

$$\omega(\mathbf{z}, \mathbf{w}_0) \propto \exp(J_\psi(\mathbf{z}, T(\mathbf{w}_0))/\eta)\,.$$

Thus, denoising is performed per arm with shared parameters, while the sample weight is global because it evaluates the Top-K action induced by the entire stored score vector.

5.3 Why Not a Joint Diffusion Actor?

A natural alternative is to train a joint diffusion actor $\pi_\theta(\mathbf{w} \mid \mathbf{z})$ directly over the joint score space $\mathbb{R}^N$. Although more expressive in principle, such a model scales poorly with the number of arms. It also lacks the natural RMAB inductive bias that the same scoring rule should apply across arms, and risks overfitting to arm ordering or co-activation artifacts under finite replay data. Consequently, the joint diffusion actor underperforms BIRD, as demonstrated in the actor ablation study in Section 6.3.

6 Experiments

6.1 Environments

Fully Observable 2-State Markov RMAB: Each arm is a 2-state Markov chain with known per-arm transition probabilities; true Whittle indices are computable exactly. This tests whether BIRD learns Whittle-style ranking behavior in the simplest fully observable setting.

The main experiments are conducted across three environments of increasing complexity:

Synthetic partially observable RMAB (Stationary): Each arm has a discrete hidden type ($M = 4$) and a 1D continuous latent state observed through noisy measurements. Type-specific transition and treatment parameters are fixed within an episode. This environment tests whether the belief encoder can infer arm type and latent state from history.

Synthetic partially observable RMAB (Drifting): This environment uses the same hidden-type structure, but the transition and treatment-response parameters drift over time. It tests robustness to non-stationarity, where the optimal arm ranking changes dynamically within an episode.

MIMIC-ICU Simulator: Each arm models an ICU patient using parameters derived from the MIMIC-III v1.4 database (Johnson et al., 2016b;a). We construct hourly patient trajectories from adult ICU stays and fit a two-factor latent simulator with hemodynamic and respiratory health states, $[x_{\text{hemo}}, x_{\text{resp}}]$. The policy observes five normalized vitals (HR, SBP, MBP, SpO2, RR), while the latent state, patient type, and transition parameters are hidden. At each step the RMAB policy may intervene on at most K of N simulated patients. Additional cohort construction and parameter-fitting details are given in Appendix C.

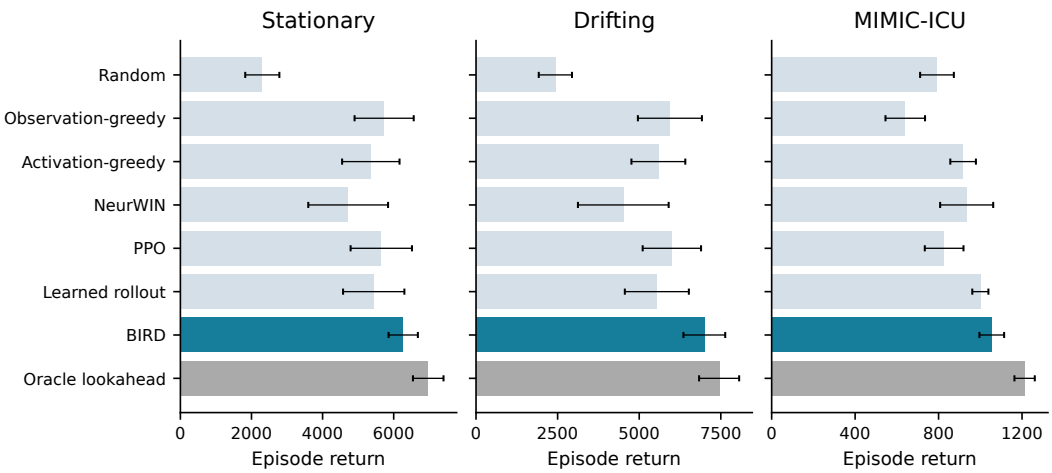


Figure 3: Main comparison at $N = 20, K = 5$. Bars show mean episode return; whiskers show episode standard deviation over 100 evaluations. BIRD is blue; oracle lookahead is a privileged reference.

6.2 Baselines and main comparison

Random selects K arms uniformly. **Observation-greedy** uses current observations. **Activation-greedy** estimates one-step activation advantage using population-average dynamics. **NeurWIN** (Neural Whittle Index Network) (Nakhleh et al., 2021) uses the same belief encoders as BIRD. **PPO** (Proximal Policy Optimization) (Schulman et al., 2017) serves as a standard RL baseline. **Learned rollout** combines the belief encoder with a learned predictive model to select actions through 10-step rollouts, providing a strong planning-based baseline. **Oracle lookahead** accesses latent states and model parameters.

As a sanity check, Appendix Table 12 shows that BIRD is able to learn useful index-like priority scores in the fully observable two-state Markov RMAB.

Figure 3 reports the main comparison results at $N = 20, K = 5$, while Appendix Tables G provide detailed numerical results for $N \in \{20, 40, 100\}$. Across all tested environments and scales, BIRD achieves the highest mean return among the evaluated non-oracle policies. Although some baselines perform competitively in specific environments, none consistently matches BIRD across all tasks.

6.3 Ablation studies

To study the actor design, we compare four variants: (1) **MLP actor**, a deterministic mapping $z_i \mapsto w_i$; (2) **Joint-N**, a single N-dimensional diffusion actor over all arms; (3) **Gaussian actor**, which outputs a parameterized score distribution per arm; and (4) **BIRD**, our proposed per-arm diffusion actor. Figure 4 reports results across the three RMAB environments, with exact values in Appendix Table I. BIRD has the highest mean return in each environment. The weaker Joint-N performance supports our design choice: a joint diffusion actor is more expressive in principle, but is less scalable and lacks the natural permutation-equivariant RMAB inductive bias.

We also studied auxiliary observation prediction, per-arm critic, belief encoder, history window length, observation noise, and dynamics uncertainty, with detailed results in Appendix H). Overall, the ablations support BIRD's shared per-arm diffusion actor, per-arm critic, and auxiliary observation-prediction loss. They also show the value of encoding observation-action histories. BIRD's advantage is most pronounced in the settings with unknown dynamics.

6.4 Scale transfer and computational cost

All training and inference experiments are performed on an NVIDIA GeForce RTX 4070 GPU with 12 GB of memory. Shared per-arm modules allow an $N = 20, K = 5$ checkpoint to be evaluated

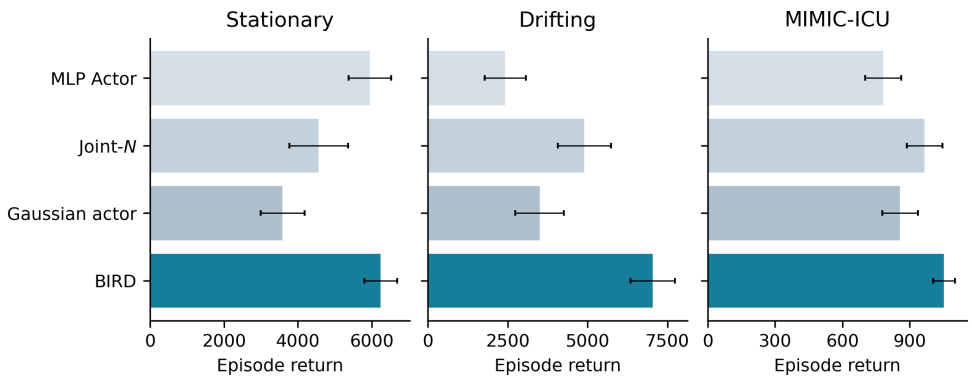


Figure 4: Actor comparison at $N = 20, K = 5$. Bars show mean episode return and whiskers show episode standard deviation over 100 episodes.

at larger N, K without retraining; a joint actor would have an action dimension tied to N and must be retrained for a new scale. Although diffusion sampling incurs substantial inference overhead, a separate timing benchmark shows that latency per full joint decision remains relatively stable across the tested sizes, as the number of arms grows from 20 to 1000. Detailed training times and inference latencies are reported in Appendix F.

7 Conclusion

We presented BIRD, a scalable algorithm for partially observable RMABs with continuous latent states and unknown dynamics. A history encoder and per-arm diffusion actor generate priority scores, which top-K projection converts into budget-feasible actions. Empirically, BIRD outperforms the evaluated non-oracle baselines across the main benchmark environments, with larger gains under non-stationary dynamics. Ablation studies support the contributions of history encoding, the per-arm critic and diffusion actor to BIRD's performance.

Limitations. The PMD interpretation treats the critic and encoder as fixed during the actor update; convergence under jointly learned function approximation and finite replay data remains open. The critic approximates the joint action value by summing per-arm estimates, which may not fully capture capture all interactions induced by the shared activation budget. We establish neither convergence to the Whittle index nor preservation of indexability. Although shared per-arm modules support transfer across problem sizes, diffusion sampling adds inference cost, which may limit its use in applications requiring rapid decisions. Finally, our evaluation is limited to controlled synthetic environments and one data-calibrated simulator. Performance in real-world deployments, particularly under dynamics and observation patterns beyond those tested, remains to be established.

AI Use Statement

In this work, we used generative AI tools for formatting, drafting and editing text, integrating experimental summaries, assisting with literature retrieval and discovery, and providing general coding assistance. Generative AI was not used to generate novel research ideas or prove mathematical theorems. The authors have verified all AI-generated text and code, and take full responsibility for the final content of this work.

Ethics Statement

The MIMIC-ICU application is a simulation with parameters derived from de-identified clinical data. Raw MIMIC-III data require credentialed access and are not redistributed. The reward is a physiological-stability surrogate, not a validated mortality endpoint. Real deployment would require

causal and off-policy evaluation, assessment of subgroup harms and resource-allocation fairness, prospective safety validation, and clinical oversight.

Reproducibility Statement

Algorithm 1 specifies training, Appendix B gives the proof, Appendix C describes simulator construction, and Appendix E reports hyperparameters. Additional experiment protocols and numerical results appear in Appendices H and G. The existing anonymous code link is given in Appendix A.

References

Daniel Adelman and Adam J Mersereau. Relaxations of weakly coupled stochastic dynamic programs. *Operations Research*, 56(3):712–727, 2008.

Nima Akbarzadeh and Aditya Mahajan. Partially observable restless bandits with restarts: Indexability and computation of whittle index. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 4898–4904. IEEE, 2022.

Turgay Ayer, Can Zhang, Anthony Bonifonte, Anne C Spaulding, and Jagpreet Chhatwal. Prioritizing hepatitis c treatment in us prisons. *Operations Research*, 67(3):853–873, 2019.

Matthew Thomas Jackson, Michael Tryfan Matthews, Cong Lu, Benjamin Ellis, Shimon Whiteson, and Jakob Foerster. Policy-guided diffusion. *arXiv preprint arXiv:2404.06356*, 2024.

Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III Clinical Database. *PhysioNet*, September 2016a. doi: 10.13026/C2XW26. URL `https://doi.org/10.13026/C2XW26`. Version 1.4.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016b.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Zongyue Li, Xiao Han, Yusong Li, Niklas Strauss, and Matthias Schubert. Dawm: Diffusion action world models for offline reinforcement learning via action-inferred transitions. *arXiv preprint arXiv:2509.19538*, 2025.

Biyonka Liang, Lily Xu, Aparna Taneja, Milind Tambe, and Lucas Janson. Context in public health for underserved communities: A bayesian approach to online restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28195–28203, 2025.

Haitong Ma, Tianyi Chen, Kai Wang, Na Li, and Bo Dai. Efficient online reinforcement learning for diffusion policy. *arXiv preprint arXiv:2502.00361*, 2025.

Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. Collapsing bandits and their application to public health intervention. *Advances in Neural Information Processing Systems*, 33:15639–15650, 2020.

Rahul Meshram and Kesav Kaza. Indexability and rollout policy for multi-state partially observable restless bandits. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 2342–2347. IEEE, 2021a.

Rahul Meshram and Kesav Kaza. Monte carlo rollout policy for recommendation systems with dynamic user behavior. In *2021 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pp. 86–89. IEEE, 2021b.

Rahul Meshram and Kesav Kaza. Lagrangian relaxation for multi-action partially observable restless bandits: Heuristic policies and indexability. *arXiv preprint arXiv:2509.00415*, 2025.

Khaled Nakhleh, Santosh Ganji, Ping-Chun Hsieh, I Hou, Srinivas Shakkottai, et al. Neurwin: Neural whittle index network for restless bandits via deep rl. *Advances in Neural Information Processing Systems*, 34:828–839, 2021.

José Niño-Mora. Markovian restless bandits and index policies: A review. *Mathematics*, 11(7):1639, 2023.

Tasmeen Zaman Ornee and Yin Sun. Remote estimation of gauss-markov processes over multiple channels: A whittle index policy. *arXiv preprint arXiv:2305.04809*, 2023.

Christos H Papadimitriou and John N Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

Michael Psenka, Alejandro Escontrela, Pieter Abbeel, and Yi Ma. Learning a diffusion model policy from rewards via q-score matching. *arXiv preprint arXiv:2312.11752*, 2023.

Marc Rigter, Jun Yamada, and Ingmar Posner. World models via policy-guided trajectory diffusion. *arXiv preprint arXiv:2312.08533*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Kai Wang, Lily Xu, Aparna Taneja, and Milind Tambe. Optimistic whittle index policy: Online learning for restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10131–10139, 2023.

Richard R Weber and Gideon Weiss. On the gittins index for multiarmed bandits. *The Annals of Applied Probability*, pp. 1024–1033, 1990.

Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.

Guojun Xiong and Jian Li. Finite-time analysis of whittle index based q-learning for restless multi-armed bandits with neural network function approximation. *Advances in Neural Information Processing Systems*, 36:29048–29073, 2023.

A Reproducibility and Code Release

An anonymized code release is available at

`https://anonymous.4open.science/r/bird-rmab-D6ED/.`

The repository includes the synthetic environments, MIMIC-derived simulator, training scripts, evaluation scripts, plotting scripts, and configuration files needed to reproduce the reported results. Raw MIMIC-III data are not redistributed; users must obtain access through PhysioNet credentialed access and follow the MIMIC-III data-use terms.

B Supplementary Details for Section 5

We provide the full proof of Proposition 5.1, which shows that the Top-K mapping from score vectors to feasible RMAB actions transfers the score-space PMD target to the corresponding PMD update over the budget-feasible action simplex.

B.1 Proof of Proposition 5.1

For each feasible action $\mathbf{a} \in \mathcal{A}_K$, define the Top-$K$ preimage region

$$R_{\mathbf{a}} = \{\mathbf{w} \in \mathbb{R}^N : T(\mathbf{w}) = \mathbf{a}\}.$$

Because $T(\mathbf{w}) = \mathbf{a}$ for every $\mathbf{w} \in R_{\mathbf{a}}$,

$$Q_{\text{old}}(\mathbf{z}, T(\mathbf{w})) = Q_{\text{old}}(\mathbf{z}, \mathbf{a}), \qquad \mathbf{w} \in R_{\mathbf{a}}.$$

Therefore,

$$\begin{aligned}
\mu^+(\mathbf{a} \mid \mathbf{z}) &= \int_{R_{\mathbf{a}}} \pi^+(\mathbf{w} \mid \mathbf{z})\, d\mathbf{w} \\
&= \int_{R_{\mathbf{a}}} \frac{\pi_{\text{old}}(\mathbf{w} \mid \mathbf{z}) \exp(Q_{\text{old}}(\mathbf{z}, T(\mathbf{w}))/\eta)}{Z_{\mathbf{w}}(\mathbf{z})} d\mathbf{w} \\
&= \frac{\exp(Q_{\text{old}}(\mathbf{z}, \mathbf{a})/\eta)}{Z_{\mathbf{w}}(\mathbf{z})} \int_{R_{\mathbf{a}}} \pi_{\text{old}}(\mathbf{w} \mid \mathbf{z})\, d\mathbf{w} \\
&= \frac{\mu_{\text{old}}(\mathbf{a} \mid \mathbf{z}) \exp(Q_{\text{old}}(\mathbf{z}, \mathbf{a})/\eta)}{Z_{\mathbf{w}}(\mathbf{z})}.
\end{aligned}$$

Since $\pi_{\text{old}}(\cdot \mid \mathbf{z})$ is absolutely continuous, score ties have probability zero. Thus the regions $\{R_{\mathbf{a}} : \mathbf{a} \in \mathcal{A}_K\}$ cover $\mathbb{R}^N$ up to measure-zero tie boundaries. Hence

$$\begin{aligned}
Z_{\mathbf{w}}(\mathbf{z}) &= \int_{\mathbb{R}^N} \pi_{\text{old}}(\mathbf{w} \mid \mathbf{z}) \exp(Q_{\text{old}}(\mathbf{z}, T(\mathbf{w}))/\eta) d\mathbf{w} \\
&= \sum_{\mathbf{a}' \in \mathcal{A}_K} \int_{R_{\mathbf{a}'}} \pi_{\text{old}}(\mathbf{w} \mid \mathbf{z}) \exp(Q_{\text{old}}(\mathbf{z}, T(\mathbf{w}))/\eta) d\mathbf{w} \\
&= \sum_{\mathbf{a}' \in \mathcal{A}_K} \mu_{\text{old}}(\mathbf{a}' \mid \mathbf{z}) \exp(Q_{\text{old}}(\mathbf{z}, \mathbf{a}')/\eta) = Z_A(\mathbf{z}).
\end{aligned}$$

Substituting $Z_{\mathbf{w}}(\mathbf{z}) = Z_A(\mathbf{z})$ gives

$$\mu^+(\mathbf{a} \mid \mathbf{z}) = \frac{\mu_{\text{old}}(\mathbf{a} \mid \mathbf{z}) \exp(Q_{\text{old}}(\mathbf{z}, \mathbf{a})/\eta)}{Z_A(\mathbf{z})},$$

which is (8).

Finally, (8) is the closed-form solution of

$$\max_{\mu \in \Delta(\mathcal{A}_K)} \left\{ \mathbb{E}_{\mathbf{a} \sim \mu} \left[Q_{\text{old}}(\mathbf{z}, \mathbf{a}) \right] - \eta \, \mathrm{KL}\left(\mu(\cdot \mid \mathbf{z}) \,\|\, \mu_{\text{old}}(\cdot \mid \mathbf{z}) \right) \right\},$$

by the standard Lagrange-multiplier derivation for the Boltzmann PMD update. Thus, the Top-K pushforward of the score-space PMD target is a valid PMD update over budget-feasible actions.

C MIMIC-ICU SIMULATOR CONSTRUCTION

The MIMIC-ICU simulator is built from MIMIC-III v1.4, a de-identified ICU database distributed through PhysioNet (Johnson et al., 2016b;a). We used a BigQuery extraction to form hourly trajectories for adult ICU stays. The exported table contains 20,044 ICU stays and 1,978,699 hourly rows; all exported stays have a recorded length of stay of at least 48 hours. For each hourly bucket, we extracted five vital signs used by the final simulator: heart rate (HR), systolic blood pressure (SBP), mean blood pressure (MBP), oxygen saturation (SpO2), and respiratory rate (RR). Temperature and glucose were also extracted in the preprocessing table, but were sparse in the export and are not used as observations in the final five-dimensional simulator.

The RMAB construction treats each arm as one simulated ICU patient. At time t, arm i has an unobserved two-dimensional latent health state

$$x_{i,t} = \left[x_{i,t}^{\text{hemo}}, x_{i,t}^{\text{resp}}\right],$$

representing hemodynamic and respiratory conditions. Larger latent values correspond to healthier states. The observed vector $y_{i,t} \in \mathbb{R}^5$ is generated by a fitted linear loading model

$$y_{i,t} = Cx_{i,t} + \epsilon_{i,t}, \qquad \epsilon_{i,t} \sim \mathcal{N}(0, \Sigma_y),$$

where the rows of C correspond to HR, SBP, MBP, SpO2, and RR. The loading matrix is fit from the hourly MIMIC trajectories by constructing hemodynamic and respiratory proxy factors from normalized vitals and regressing each vital onto these two factors.

Each patient has a hidden type $\theta_i \in \{1, \ldots, 4\}$, corresponding to four empirical patient subpopulations: septic shock, respiratory failure, hemodynamic instability, and recovering. For each stay, we fit a coupled linear autoregressive model on the two proxy factors,

$$x_{t+1}^{\text{hemo}} = \alpha_{\text{hemo}} x_t^{\text{hemo}} + c_{\text{rh}} x_t^{\text{resp}} + \beta_{\text{hemo}} a_t + d_{\text{hemo}} + \xi_t^{\text{hemo}},$$
$$x_{t+1}^{\text{resp}} = \alpha_{\text{resp}} x_t^{\text{resp}} + c_{\text{hr}} x_t^{\text{hemo}} + \beta_{\text{resp}} a_t + d_{\text{resp}} + \xi_t^{\text{resp}}.$$

Stays with sufficient complete vital history are then clustered into four types using fitted persistence, drift, and observational treatment-response features. The simulator uses the resulting type-level persistence, process-noise, and observation-loading estimates.

During simulation, the type-specific parameters define the mean behavior, but each arm's persistence and treatment-response parameters are allowed to drift over time. Specifically, for arm i, the persistence vector $\alpha_{i,t} = (\alpha_{i,t}^{\text{hemo}}, \alpha_{i,t}^{\text{resp}})$ and treatment-response vector $\beta_{i,t} = (\beta_{i,t}^{\text{hemo}}, \beta_{i,t}^{\text{resp}})$ evolve around their type-specific means according to an Ornstein–Uhlenbeck-style update:

$$\alpha_{i,t+1} = \bar{\alpha}_{\theta_i} + \rho(\alpha_{i,t} - \bar{\alpha}_{\theta_i}) + \sigma_\alpha \xi_{i,t}^{\alpha},$$

$$\beta_{i,t+1} = \bar{\beta}_{\theta_i} + \rho(\beta_{i,t} - \bar{\beta}_{\theta_i}) + \sigma_\beta \xi_{i,t}^{\beta}.$$

The OU drift parameters $\rho = 0.85$, $\sigma_\alpha = 0.03$, and $\sigma_\beta = 0.15$ are calibrated simulator hyperparameters rather than directly estimated from MIMIC. They were chosen to create moderate within-patient non-stationarity: persistence parameters vary slowly around type-specific means, while treatment-response parameters are allowed somewhat larger temporal variation. The resulting parameters are clipped to clinically reasonable simulator ranges.

Because treatment effects are not causally identified from observational ICU data, fitted observational treatment coefficients are used only as descriptive clustering features. The simulator treatment-response magnitudes are calibrated to create clinically meaningful intervention heterogeneity rather than interpreted as causal estimates from MIMIC.

At each simulator step, the action is a binary intervention decision $a_{i,t} \in \{0, 1\}$ for each patient arm, with the hard RMAB budget $\sum_i a_{i,t} \leq K$. The intervention affects both latent dimensions through type-specific treatment-response coefficients. The per-arm reward is the nonnegative health score

$$r_{i,t} = \max\left(0.5 x_{i,t}^{\text{hemo}} + 0.5 x_{i,t}^{\text{resp}}, 0\right),$$

and the episode return sums rewards over arms and time. During learning, the agent observes only past vital vectors and past interventions; the latent state, type, and transition parameters are used only by the simulator and oracle baseline.

As a sanity check, we computed the corresponding observed-vital health proxy on the MIMIC-derived hourly cohort and compared it with 90-day mortality. Survivors had higher mean proxy health than non-survivors (0.282 vs. 0.244); the bootstrap 95% confidence interval for the difference was $[0.031, 0.045]$. The association is directionally consistent but modest (AUC 0.57 using negative proxy health to predict mortality). Therefore, we interpret the reward as a dense physiological-stability surrogate, not as a validated mortality-risk endpoint.

D Boundary Diagnostic Implementation

The diagnostic is run only in the drifting synthetic RMAB, where the simulator transition model and latent variables are available for analysis. These quantities are not used by BIRD during training or evaluation.

Particle-filter rank diagnostic. We first select a fixed L-step observed joint history

$$\mathbf{h}_t = (h_{1,t}, \ldots, h_{N,t}),$$

where each arm history contains the last L noisy observations and past actions. For each arm i, we run an independent bootstrap particle filter conditioned on $h_{i,t}$. Each particle contains a latent state, hidden type, and drifting transition parameters,

$$\left(x_{i,t}^{(s)}, \theta_i^{(s)}, \alpha_{i,t}^{(s)}, \beta_{i,t}^{(s)}\right).$$

Particles are initialized from the simulator prior, propagated using the recorded action sequence, weighted by the likelihood of the observed noisy measurements, and resampled when the effective sample size becomes small. Thus the final particles approximate the posterior over latent simulator variables conditioned on the same observed history.

For each posterior particle, we compute a privileged oracle marginal activation value using the simulator model:

$$\Delta_i^{(s)} = Q_i^{\text{oracle}}\left(x_{i,t}^{(s)}, \theta_i^{(s)}, \alpha_{i,t}^{(s)}, \beta_{i,t}^{(s)}, 1\right) - Q_i^{\text{oracle}}\left(x_{i,t}^{(s)}, \theta_i^{(s)}, \alpha_{i,t}^{(s)}, \beta_{i,t}^{(s)}, 0\right).$$

For each posterior sample s, the oracle marginal values $\{\Delta_i^{(s)}\}_{i=1}^N$ are ranked across arms. We then select a boundary arm whose posterior rank frequently lies near the Top-K cutoff and plot its posterior rank distribution. The activation probability shown in the plot is estimated as

$$\Pr(a_i = 1 \mid \mathbf{h}_t) \approx \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\{i \in \text{TopK}(\Delta^{(s)})\}.$$

This panel visualizes that, for a fixed observed joint history, posterior latent uncertainty can move a boundary arm across the activation cutoff.

Oracle-value boundary comparison. The second diagnostic tests whether stochastic score sampling is useful rather than arbitrary. We collect evaluation histories from the same drifting synthetic simulator and identify boundary histories where the oracle gap between the K-th and $(K+1)$-st largest marginal activation values is small. For each such history, we compute simulator-oracle marginal values Δ_i^{oracle} from the realized latent state and evaluate any Top-K action a using

$$V^{\text{oracle}}(a) = \sum_{i=1}^{N} a_i \Delta_i^{\text{oracle}}.$$

This oracle value is used only as a common evaluator and is not used for training either actor.

For the deterministic MLP actor ablation, we encode the same observed history, compute deterministic arm scores, and form $a^{\text{MLP}} = \text{TopK}(s^{\text{MLP}})$. For BIRD, we encode the same history, sample multiple score vectors from the trained diffusion actor, $\mathbf{w}^{(m)} \sim \pi_\theta(\cdot \mid \mathbf{z}_t)$, and convert each sample to a feasible action $a^{(m)} = \text{TopK}(\mathbf{w}^{(m)})$. We then compare the oracle values

$$V^{\text{oracle}}(a^{(m)}) - V^{\text{oracle}}(a^{\text{MLP}}).$$

The right panel of Figure 2 plots the distribution of the per-history mean value difference across BIRD samples. A positive value means that, on that boundary history, BIRD's stochastic Top-K samples have higher oracle activation value than the deterministic MLP action. This diagnostic supports the interpretation that Q-weighted diffusion does not merely add arbitrary randomness; it can place probability mass on higher-value Top-K activation sets near uncertain activation boundaries.

E HYPERPARAMETERS

Table 1 summarizes the main hyperparameters used for BIRD across the synthetic Stationary/Drifting environments and the MIMIC-ICU Simulator. Unless otherwise stated, the same settings are used for the main comparison and ablation experiments.

Table 1: Hyperparameters used across all experiments.

Hyperparameter	Stationary / Drifting	MIMIC-ICU
Discount γ	0.99	0.99
History length L	40	80
Belief dim d_z	64	64
Encoder layers	3	3
Encoder heads	4	4
Diffusion steps T_{diff}	20	20
Batch size	128	128
Replay size	100K	100K
Actor LR	3e-4	3e-4
Critic LR	3e-4 / 1e-4	1e-4
λ_{init}	2.0	2.0
λ_{target}	1.0	1.0
Action candidates M	4	4
Training epochs	200	200

F TRAINING AND INFERENCE COST

Local training and inference timings use an NVIDIA GeForce RTX 4070 (12GB). For the default N = 20, K = 5 setting, BIRD training takes 131.6 min on synthetic stationary, 128.2 min on synthetic drifting, and 182.8 min on MIMIC-ICU.

Each reported inference time in Table 2 is for one full decision step–scoring all N arms and applying top-K to produce the joint action; for BIRD, this includes diffusion sampling. For transfer across larger numbers of arms, BIRD does not require retraining: Fig. 5 was produced by one model trained at N = 20, K = 5 and evaluated at larger N, K using the same policy set and evaluation protocol. Using this same checkpoint, BIRD inference takes about 98–116 ms per decision at the default scale N = 20, K = 5, and remains about 115–125 ms per decision up to N = 1000, K = 200 across all three environments; these timings are averaged over 100 episodes, i.e., 100 × 100 decision steps.

Table 2: Local transfer latency in milliseconds per joint decision.

N	K	Stationary	Drifting	MIMIC-ICU
20	5	97.571	116.025	116.159
50	12	116.348	115.995	116.604
100	20	118.176	118.214	114.662
1000	200	119.570	119.040	125.088

G Numerical Results

Table 3: Main comparison at $N = 20, K = 5$: mean return $\pm$ episode standard deviation over 100 episodes of length 100. Oracle lookahead uses privileged information. Bold indicates the highest mean return among non-oracle policies.

Policy	Stationary	Drifting	MIMIC-ICU
Random	2304.1 ± 480.0	2431.1 ± 509.1	792.4 ± 80.8
Observation-greedy	5731.4 ± 831.9	5940.2 ± 980.8	640.3 ± 94.8
Activation-greedy	5356.5 ± 807.4	5586.8 ± 825.1	917.6 ± 61.3
NeurWIN	4719.4 ± 1122.1	4512.2 ± 1389.8	934.7 ± 127.2
PPO	5651.3 ± 863.1	6000.3 ± 894.1	826.7 ± 92.7
Learned rollout	5438.2 ± 862.5	5540.2 ± 980.8	1000.2 ± 38.6
BIRD	$\mathbf{6268.1 \pm 411.7}$	$\mathbf{6994.6 \pm 640.5}$	$\mathbf{1055.0 \pm 59.2}$
Oracle lookahead	6971.6 ± 429.7	7448.3 ± 613.5	1212.4 ± 48.8

Table 4: Main comparison at $N = 40, K = 10$.

Policy	Stationary	Drifting	MIMIC-ICU
Random	4489.2 ± 776.1	4718.4 ± 836.0	1576.5 ± 106.1
Observation-greedy	11124.7 ± 1222.2	11608.9 ± 1361.4	1263.1 ± 130.0
Activation-greedy	10540.7 ± 1082.9	11268.5 ± 1007.7	1827.2 ± 79.0
NeurWIN	9342.5 ± 1550.0	9022.6 ± 1988.7	1870.8 ± 172.5
PPO	11155.9 ± 1328.4	12040.5 ± 1283.4	1787.0 ± 185.1
Learned rollout	10768.7 ± 1182.9	11089.3 ± 1513.4	2002.2 ± 51.5
BIRD	$\mathbf{12378.7 \pm 550.6}$	$\mathbf{13827.6 \pm 855.5}$	$\mathbf{2072.8 \pm 70.9}$
Oracle lookahead	14067.2 ± 593.4	15060.9 ± 760.5	2428.5 ± 65.4

Table 5: Main comparison at $N = 100, K = 25$.

Policy	Stationary	Drifting	MIMIC-ICU
Random	11060.9 ± 1213.7	11726.2 ± 1264.9	3894.2 ± 161.4
Observation-greedy	27689.0 ± 1870.6	29204.5 ± 2165.7	3100.3 ± 197.1
Activation-greedy	25881.3 ± 1749.0	27741.8 ± 1776.4	4535.8 ± 127.2
NeurWIN	23875.7 ± 2478.4	23206.4 ± 3025.3	4720.8 ± 245.5
PPO	27880.7 ± 2054.3	30171.5 ± 2001.6	4081.0 ± 207.4
Learned rollout	27192.7 ± 1864.4	27975.5 ± 2193.5	4983.1 ± 72.6
BIRD	$\mathbf{30691.7 \pm 891.0}$	$\mathbf{34244.3 \pm 1358.9}$	$\mathbf{5098.4 \pm 100.1}$
Oracle lookahead	35181.4 ± 963.0	37829.1 ± 1398.6	6071.9 ± 95.7

H Component Ablations

Removing the auxiliary observation prediction loss $\mathcal{L}_{\text{aux}}$ reduces returns in all three environments. The per-arm critic outperforms the joint critic, while a single per-arm critic achieves returns comparable to the twin-critic variant. For the belief encoder, the MLP outperforms the Transformer under stationary dynamics, whereas the Transformer performs better on the drifting and MIMIC-ICU tasks. Longer observation histories are not uniformly beneficial: $L = 80$ improves returns on the stationary and MIMIC-ICU tasks, while $L = 40$ performs best on the drifting task.

We further separate the effects of observation noise and dynamics uncertainty. BIRD's relative advantage is more pronounced under unknown dynamics, while it underperforms some baselines in simpler environments with known dynamics. These findings are consistent with our sanity checks in a two-state Markov environment, where BIRD offers limited benefit over simpler baselines.

Table 6: Auxiliary prediction loss ablation. Mean return ± episode standard deviation over 100 evaluation episodes.

Variant	Stationary	Drifting	MIMIC-ICU
With auxiliary loss	$\mathbf{5908.9 \pm 393.6}$	$\mathbf{6860.4 \pm 656.2}$	$\mathbf{1050.7 \pm 55.9}$
Without auxiliary loss	3502.2 ± 523.8	4635.3 ± 598.7	939.3 ± 74.2

Table 7: Critic ablation; these independently trained checkpoints differ from the main-comparison checkpoints. Mean return ± episode standard deviation; one training seed and 100 evaluation episodes per variant.

Variant	Stationary	Drifting	MIMIC-ICU
Per-arm twin	6181.5 ± 458.9	6886.3 ± 584.9	$\mathbf{987.6 \pm 41.0}$
Per-arm single	$\mathbf{6663.6 \pm 435.0}$	$\mathbf{6952.3 \pm 583.7}$	968.0 ± 42.2
Joint twin	3540.1 ± 762.3	4516.9 ± 1074.5	837.6 ± 109.0

Table 8: History encoder ablation ($L = 40$ synthetic, $L = 80$ MIMIC-ICU). Mean return ± episode standard deviation over 100 evaluation episodes.

Variant	Stationary	Drifting	MIMIC-ICU
Transformer	6250.3 ± 379.8	$\mathbf{7107.4 \pm 566.9}$	$\mathbf{1054.5 \pm 55.4}$
LSTM	5715.0 ± 412.3	6387.8 ± 547.2	966.8 ± 38.4
MLP	$\mathbf{6632.0 \pm 406.7}$	7059.1 ± 705.4	954.0 ± 44.4

Table 9: History-window sensitivity. Mean return ± episode standard deviation over 100 episodes; No $L = 20$ MIMIC result is reported.

L	Stationary	Drifting	MIMIC-ICU
20	6283.3 ± 491.0	6748.0 ± 588.3	—
40	5908.9 ± 393.6	$\mathbf{6886.3 \pm 584.9}$	999.1 ± 50.0
80	$\mathbf{6434.2 \pm 637.1}$	6758.9 ± 599.8	$\mathbf{1052.1 \pm 57.3}$

Table 10: Factor stress tests. Known: noisy observations with fixed homogeneous dynamics. Unknown: noiseless observations with hidden heterogeneous dynamics (noiseless vitals in MIMIC-ICU). $L = 40$, $N = 20$, $K = 5$; 100 evaluation episodes

Setting / policy	Stationary	Drifting	MIMIC-ICU
Known / Observation-greedy	$\mathbf{5186.1 \pm 71.5}$	$\mathbf{5186.1 \pm 71.5}$	678.1 ± 21.2
Known / Activation-greedy	4495.1 ± 204.6	4495.1 ± 204.6	809.1 ± 18.2
Known / Learned rollout	5181.6 ± 72.4	5179.2 ± 76.2	847.8 ± 16.9
Known / BIRD	3735.0 ± 150.2	2249.2 ± 170.0	$\mathbf{859.0 \pm 18.3}$
Unknown / Observation-greedy	5306.0 ± 827.6	5517.5 ± 915.5	609.1 ± 97.8
Unknown / Activation-greedy	5118.8 ± 761.6	5434.5 ± 834.2	951.8 ± 62.5
Unknown / Learned rollout	5360.5 ± 793.2	5605.2 ± 917.4	1005.9 ± 36.4
Unknown / BIRD	$\mathbf{6259.0 \pm 365.6}$	$\mathbf{6392.0 \pm 599.6}$	$\mathbf{1046.3 \pm 46.1}$

Table 12: Fully Observable 2-State Markov RMAB ($N = 50$, $K = 10$, 100 episodes with different seeds). Mean episode return $\pm$ std and % of the random-to-Whittle gap captured. Observation-greedy prioritizes state-0 arms. Activation-greedy ranks arms by their one-step activation advantage using the true per-arm transition probabilities.

Policy	Mean Return	% of Whittle gap
Random	1518.5 ± 38.1	0.0%
Observation-greedy	1515.7 ± 40.2	-1.2%
Activation-greedy	1740.3 ± 40.0	$+98.1\%$
PPO	1548.2 ± 31.2	$+13.2\%$
NeurWIN	$\mathbf{1744.7 \pm 39.2}$	$\mathbf{+100.1\%}$
BIRD (ours)	1721.5 ± 40.0	$+89.8\%$
True Whittle (oracle)	1744.5 ± 40.7	100.0%

I Actor Comparison

Table 11: Actor comparison at $N = 20, K = 5$: mean return $\pm$ episode standard deviation over 100 episodes. Bold indicates the highest mean return in each environment.

Environment	MLP Actor	Joint-N	Gaussian actor	BIRD
Stationary	5959.2 ± 577.4	4567.8 ± 800.0	3589.9 ± 596.7	$\mathbf{6247.4 \pm 445.4}$
Drifting	2425.3 ± 651.8	4902.9 ± 830.8	3502.1 ± 762.3	$\mathbf{7042.5 \pm 693.9}$
MIMIC-ICU	782.6 ± 81.0	966.4 ± 80.0	856.9 ± 79.7	$\mathbf{1052.9 \pm 48.7}$

J Sanity Check

Table 12 presents results on the Fully Observable 2-State Markov RMAB. BIRD outperforms the random baseline, observation-greedy, and PPO, but has a slightly lower return than NeurWIN, activation-greedy, and the true Whittle index policy . In this setting, policies that exploit the known transition model are highly competitive. Nevertheless, these results confirm that BIRD is able to learn useful index-like priority scores.