EXPLORING WHAT MATTERS: VALUE-AWARE EXPLORATION FOR DREAMER

Anonymous authors
Paper under double-blind review

ABSTRACT

Dreamer improves data efficiency by learning a world model from real experience and training its actor and critic on imagined trajectories. These imagined updates depend on the experience collected, making efficient exploration essential under a limited interaction budget. Predictive disagreement can waste interaction on dynamics with little effect on return. Immature value estimates can also overlook useful experience, limiting task-relevant model learning and policy improvement in imagination. We propose Value-aware Exploration (VEX), which directs exploration in DreamerV3 through value-priced predictive uncertainty. VEX evaluates alternative transition predictions through a shared task readout of reward, continuation, and future value, measuring their disagreement in terms of task return. We derive a Bayesian allocation priority under imperfectly observed task sensitivity, recovering uncertainty-only exploration in the uninformative limit. VEX's exploration score directly weights imagined actor updates to guide experience collection. Across three continuous-control tasks, both VEX operating points improve aggregate sample efficiency and final return over DreamerV3 and same-backbone exploration baselines. Ablations demonstrate the contributions of its key components.

1 INTRODUCTION

Learning effective control from limited environment interaction is a central goal of reinforcement learning. Dreamer addresses this challenge by learning a world model and using it to train an actor and critic in imagination (Hafner et al., 2019; 2020; 2025). The learned world model predicts dynamics, rewards, and continuation, supplying imagined trajectories for repeated policy updates without further environment interaction. The quality of these updates depends on the real experience used to train the model.

DreamerV3's stochastic actor and entropy regularization encourage action diversity, but do not explicitly assess uncertainty in the learned dynamics (Hafner et al., 2025). Plan2Explore seeks unfamiliar transitions by maximizing future disagreement among auxiliary predictors in imagination (Sekar et al., 2020). Yet predictive disagreement alone can direct a limited interaction budget toward dynamics that have little effect on task return. Consider two actions in a balance task whose predicted next states are similarly dispersed. For one action, predictions differ mainly in a state component that scarcely affects reward or future return. For the other, they disagree over whether the agent remains upright or loses balance. An uncertainty-only score can assign similar priorities to both actions, although predicted returns vary little in the first case and substantially in the second. Exploration should therefore consider how candidate differences affect task evaluation, alongside the amount of predictive uncertainty. An immature critic may not yet reveal these task effects.

We introduce *Value-aware Exploration* (VEX), which integrates predictive uncertainty and task sensitivity through a common value-priced allocation principle. For a state–action context $x = (s, a)$, learned task sensitivity provides imperfect evidence of how transition errors affect value. Our Bayesian observation-allocation model yields the priority

$$\text{priority}_\rho(x) = \underbrace{U_x}_{\text{uncertainty}} \underbrace{\sqrt{(1-\rho)L^2 + \rho^2 \widetilde{P}_x^2}}_{\text{posterior value price}}. \tag{1}$$

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.

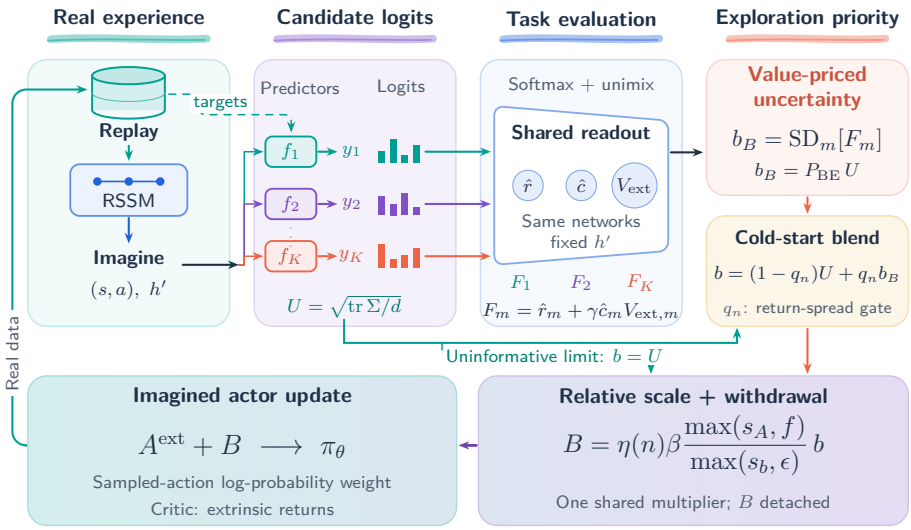

Figure 1: **Value-aware Exploration (VEX) in DreamerV3.** Predictors map each imagined state–action pair to K candidate logits $y_m \in \mathbb{R}^d$ with sample covariance Σ. Using shared successor memory h', the task readout F_m combines predicted reward, continuation, and extrinsic value. Its dispersion gives the BE score $b_B = P_{\mathrm{BE}}U$: uncertainty U weighted by value price P_{BE}. The dashed uncertainty bypass uses $b = U$ at $\rho = 0$. Shared scaling converts score b into correction B for normalized extrinsic advantage A^{ext}, with withdrawal controlled by $\eta(n)$. Here s_A, s_b track advantage and score RMS, and β sets correction strength.

Here U_x measures predictive uncertainty, $\widetilde{P}_x$ is the estimated local value price, L sets a common prior sensitivity scale, and ρ describes sensitivity reliability. This priority minimizes expected squared local value error in the allocation model (Section 4.1). Reliable sensitivities recover local task-readout dispersion. In the uninformative limit $\rho = 0$, the price is constant and allocation depends only on uncertainty. Price heterogeneity governs both the gains from informative pricing and the risks of trusting unreliable sensitivities.

To apply this principle in DreamerV3, VEX defines a shared *task readout* F from the reward, continuation, and extrinsic-value heads. Its standard deviation across candidate transitions gives the Bellman-ensemble (BE) score b_B. This score retains nonlinear readout effects and recovers the reliable-limit priority for affine readouts. The score directly corrects the normalized actor advantage, increasing the update weight of imagined actions whose successor transitions have higher priority. With short imagined rollouts, reward-level bonuses depend on learned value bootstraps to convey their longer-horizon contribution. Advantage-level injection sets the immediate correction and preserves the critic's extrinsic objective. Relative scaling controls the correction's strength, and gradual withdrawal shifts late training toward task-policy refinement (Figure 1). We evaluate two operating points of this architecture: finite Bellman pricing, VEX (priced), and the uninformative limit, VEX ($\rho = 0$), which uses the same correction with a constant transition price.

Our contributions are:

- We introduce a value-priced exploration framework for DreamerV3 that evaluates candidate transitions through a shared task readout and applies their priorities directly to imagined actor updates.
- We derive a Bayesian allocation rule that accounts for imperfect task sensitivity, explains the benefit and risk of selective pricing, and recovers uncertainty-only exploration as its uninformative limit.
- Across three control tasks, both VEX operating points achieve higher IQM scores for sample efficiency and final return than DreamerV3 and same-backbone exploration baselines. Ablations identify the contributions of its key components.

2 Related Work

Uncertainty-driven exploration. Classical exploration bonuses use counts, prediction errors, or information gain to seek informative experience (Bellemare et al., 2016; Pathak et al., 2017; Burda et al., 2018; Houthooft et al., 2016). MAX plans information-seeking actions with dynamics ensembles (Shyam et al., 2019). Plan2Explore maximizes imagined disagreement through a separate exploration actor–critic (Sekar et al., 2020). SOMBRL combines extrinsic reward with epistemic uncertainty and adapts their weight in its Dreamer implementation (Sukhija et al., 2025). Uncertainty signals identify unfamiliar transitions, but uncertainty magnitude alone does not establish which transitions matter for control. VEX evaluates candidate successors through a task readout that combines predicted reward, continuation, and value, pricing their disagreement by its effect on return.

Task-directed exploration and optimism. Value estimates guide exploration through optimistic bonuses and posterior sampling (Strehl & Littman, 2008; Osband et al., 2016). Model-based uncertainty Bellman equations quantify uncertainty in cumulative value (Luis et al., 2023). DreamerV3-XP uses disagreement in reward predictions, with decaying or return-dependent reward mixing (Bierling et al., 2025). H-UCRL and Optimistic World Models place optimism in planning and dynamics learning, respectively (Curi et al., 2020; Mete et al., 2026). The usefulness of task-directed priorities also depends on the reliability of the learned value response. Our Bayesian allocation analysis characterizes this dependence through imperfect task sensitivity, recovering uncertainty-only exploration in the uninformative limit. The BE score evaluates the task readout on finite candidates, retaining nonlinear effects that a local gradient omits.

Value-aware model learning and replay. VAML, VaGraM, and calibrated value-aware learning emphasize the value impact of errors when fitting dynamics (Farahmand et al., 2017; Voelcker et al., 2022; 2025). Curious Replay instead prioritizes stored experience using replay statistics and model losses (Kauvar et al., 2023). These interventions improve the fitting or reuse of collected observations. VEX applies value-aware priorities to the acquisition of new experience: its score weights imagined actor updates, influencing which real transitions enter replay while retaining Dreamer's model-fitting objective and replay sampler.

3 Method: Value-aware Exploration

Value-aware Exploration (VEX) prices uncertainty by how alternative next-state predictions change the agent's task readout, and uses the resulting score to weight imagined actor updates. Replay-trained predictors supply the candidates, shared task heads evaluate them, and relative scaling controls the correction before its gradual withdrawal. Figure 1 shows the learning loop. Algorithm 1 specifies the update sequence.

3.1 Uncertainty and Its Value Price

DreamerV3's recurrent state-space model (RSSM) represents a state as $s_t = (h_t, z_t)$, with deterministic memory h_t and categorical latent z_t (Hafner et al., 2025). For each imagined (s_t, a_t), K auxiliary predictors produce next-latent logits $y_{m,t+1} = f_m(s_t, a_t^{\text{eff}}) \in \mathbb{R}^d$, using the RSSM's componentwise effective action $a_t^{\text{eff}} = a_t / \max(|a_t|, 1)$ in both training and scoring. The candidates share the deterministic successor h_{t+1}. The RSSM prior generates the actual imagined successor. Scoring these alternatives at each step requires no separate candidate rollouts.

To evaluate their task consequences, VEX maps each prediction to a probability feature $p(y)$ using groupwise softmax with Dreamer's uniform mixture. The resulting candidate $\tilde{s}_{m,t+1} = (h_{t+1}, p(y_{m,t+1}))$ is evaluated by the shared reward, continuation, and current extrinsic-value heads:

$$F_{m,t} = \hat{r}(\tilde{s}_{m,t+1}) + \gamma \hat{c}(\tilde{s}_{m,t+1}) V_{\text{ext}}(\tilde{s}_{m,t+1}). \tag{2}$$

Here $\gamma \in (0, 1)$ is the discount, and $F_x(y)$ denotes the task readout for fixed $x = (s_t, a_t)$ and h_{t+1}. It is evaluated on soft features, so its output generally differs from an expectation over categorical states. With bars denoting candidate means, the readout and coordinate dispersions are

$$b_{B,t} = \sqrt{\frac{\sum_{m=1}^{K} (F_{m,t} - \bar{F}_t)^2}{K - 1}}, \qquad U_t = \sqrt{\frac{\sum_{m=1}^{K} \|y_{m,t+1} - \bar{y}_{t+1}\|_2^2}{d(K - 1)}}. \tag{3}$$

The factorization $b_B = P_{\mathrm{BE}} U$ defines a finite value price $P_{\mathrm{BE}} = b_B/U$ for $U > 0$: task-readout sensitivity per unit of predictive uncertainty. VEX computes b_B directly from candidate evaluations, retaining nonlinear readout effects without a separate price network.

A return-spread cold-start gate introduces the BE score as the task heads begin to learn:

$$b_t = (1 - q_n)U_t + q_n b_{B,t}, \qquad q_n = \mathrm{clip}(\Delta_R / L_R, 0, 1). \tag{4}$$

Here n counts environment frames, Δ_R is Dreamer's running percentile spread of imagined extrinsic returns, and L_R is its return-scale floor. For $U_t > 0$, the score is $b_t = P_t U_t$ with effective price $P_t = (1 - q_n) + q_n P_{\mathrm{BE},t}$. Fixing $P_t = 1$ gives $b_t = U_t$, realizing the uninformative limit within the same update up to a common scale. The gate q_n schedules finite pricing but does not estimate reliability ρ.

3.2 Learning and Scaling the Exploration Signal

For each real replay transition (s_t, a_t, o_{t+1}), the observation-conditioned RSSM posterior supplies next-state target logits. Centering them within each categorical group gives y_{t+1}^{data} and removes the common offset that leaves softmax unchanged. Independently initialized predictors fit these targets on resampled bootstrap subsets $\mathcal{B}_m$:

$$\mathcal{L}_m = \mathbb{E}_{\mathcal{B}_m}\left[\frac{1}{d}\left\| f_m(\mathrm{sg}(s_t), \mathrm{sg}(a_t^{\mathrm{eff}})) - \mathrm{sg}(y_{t+1}^{\mathrm{data}})\right\|_2^2\right], \tag{5}$$

where sg denotes stop-gradient. A separate optimizer updates the auxiliary networks. The encoder and RSSM retain Dreamer's model objective. Appendix A specifies the masking and hyperparameters.

VEX uses the normalized extrinsic advantage $A_t^{\mathrm{ext}} = (G_t^\lambda - V_{\mathrm{ext}}(s_t))/S_R$, where $S_R = \max(\Delta_R, L_R)$ and G_t^λ is Dreamer's imagined extrinsic λ-return (Equation (16)).

To control correction strength independently of transition preferences, we update exponential moving averages s_b and s_A of the batch RMS values of b and A^{ext}, then form

$$B_t = c_n b_t, \qquad c_n = \eta(n)\beta \frac{\max(s_A, f)}{\max(s_b, \epsilon)}. \tag{6}$$

Here β controls correction strength, $f > 0$ provides a cold-start advantage-scale floor, and $\epsilon > 0$ stabilizes normalization. Shared scaling preserves relative scores, with target RMS ratio $\beta\eta(n)$ to the extrinsic advantage under stationary scales and inactive floors.

The withdrawal factor $\eta(n)$ stays at one initially, then decreases linearly to zero late in training to favor extrinsic-policy refinement. Applying it outside the RMS trackers prevents normalization from cancelling withdrawal.

3.3 Policy Learning with an Extrinsic Task Reference

From replay posterior states, the actor and RSSM imagine H transitions and compute extrinsic returns G_t^λ. Each correction B_t weights the action whose successor candidates were evaluated. Writing $\pi_t = \pi_\theta(\cdot \mid s_t)$, the corrected advantage and actor loss are

$$\widetilde{A}_t = A_t^{\mathrm{ext}} + B_t, \qquad \mathcal{L}_\pi = -\mathbb{E}\Big[w_t\big(\log \pi_t(a_t)\,\mathrm{sg}(\widetilde{A}_t) + \alpha_{\mathrm{ent}}\mathcal{H}(\pi_t)\big)\Big], \tag{7}$$

where w_t is the discounted continuation weight and α_{ent} the entropy coefficient. Imagined states, sampled actions, weights, and B_t are detached. The actor learns through log probabilities and entropy. Transitions collected by the actor in the real environment replenish replay for world-model and predictor training.

Equal-RMS corrections need not produce the same update. In a two-action categorical illustration with equal probabilities, $(2, 1)$ and $(1, 2)$ favor opposite actions (Appendix B.5). Critic targets remain extrinsic, and world-model objectives are unchanged. At $\eta = 0$, the actor returns to Dreamer's extrinsic objective.

Reward-level bonuses reach the actor through finite λ-returns, with longer-horizon effects depending on learned critic bootstraps. Advantage-level injection supplies the correction directly and retains an extrinsic critic (Appendix A.4).

Algorithm 1 VEX: value-aware exploration in DreamerV3

Require: World model ϕ, actor θ, extrinsic critic ψ, predictors $f_{1:K}$, replay $\mathcal{D}$, pricing setting

Initialize parameters, the slow critic, return statistics, and RMS trackers.
while the interaction budget is not exhausted **do**
  Collect real transitions using π_θ and append them to $\mathcal{D}$.
  for each scheduled learner update **do**
    Refresh the slow critic when scheduled and sample replay sequences from $\mathcal{D}$.
    Infer posterior states and form the standard model losses.
    Train $f_{1:K}$ on detached next-posterior targets with resampled bootstrap masks.
    Imagine H transitions from detached posterior states using the RSSM and actor.
    Compute extrinsic G^λ; update return statistics and obtain A^{ext}.
    Query $y_{1:K,t+1}$ at each (s_t, a_t^{eff}); compute U_t and set $b_t \leftarrow U_t$.
    if the priced operating point is selected **then**
      Evaluate $F_{1:K,t}$ with shared h_{t+1} and compute $b_{B,t}$.
      Obtain q_n from return statistics; set $b_t \leftarrow (1-q_n)U_t + q_n b_{B,t}$.
    end if
    Update RMS trackers from b, A^{ext}; compute $\eta(n)$ and B_t by Equation (6).
    Form the corrected actor loss and extrinsic critic losses.
    Apply the shared backbone optimizer to model, actor, and critic losses.
  end for
end while

4 Theoretical Analysis

We analyze VEX to explain both which uncertain transitions should receive priority and how those priorities influence policy learning. A Bayesian allocation model with imperfect task sensitivity identifies the gains and risks of selective pricing and its uncertainty-only limit. We connect the optimal allocation to the BE score, then characterize the actor-gradient contribution of advantage-level injection and the role of relative scaling. Full proofs are in Appendix B.

4.1 Allocation under Uncertain Sensitivity

Simulation-lemma arguments relate a fixed policy's value error under a learned model to accumulated errors in one-step Bellman evaluations (Kearns & Singh, 2002). To first order, a successor-mean error δ_x changes the task readout by $g_x^\top \delta_x$. We therefore study how fresh observations should be allocated to reduce aggregate local evaluation error when task sensitivity is estimated from learned reward, continuation, and value heads. The allocation model idealizes this choice through directly queryable transition contexts. Section 4.2 then characterizes how VEX turns exploration scores into actor-update preferences within imagination.

Consider finitely many contexts $x = (s, a)$ with unknown successor-vector means μ_x and equal observation cost. Frozen candidates $y_{1:K}$, with mean $\bar{y}$, define

$$\Sigma_x = \frac{1}{K-1}\sum_m (y_m - \bar{y})(y_m - \bar{y})^\top, \qquad U_x^2 = \frac{\operatorname{tr}\Sigma_x}{d} > 0. \tag{8}$$

Conditional on this frozen pilot, fresh observations are mutually independent with covariance $\kappa\Sigma_x$ for a common $\kappa > 0$. Allocate N observations using the continuous relaxation $n_x = N\nu_x$, where $\nu_x \geq 0$ and $\sum_x \nu_x = 1$. This stratified estimation model (Carpentier & Munos, 2011) minimizes the expected square of the aggregate local evaluation error $\sum_x g_x^\top(\widehat{\mu}_x - \mu_x)$.

Theorem 1 (Bayes-optimal value-priced allocation)**.** *Conditional on the frozen covariances, let $g_x \sim \mathcal{N}(0, L^2 I/d)$ and $\varepsilon_x \sim \mathcal{N}(0, \sigma_e^2 I)$ be mutually independent and independent of fresh observation noise, with common prior sensitivity scale $L > 0$ and $\widetilde{g}_x = g_x + \varepsilon_x$. Define the reliability $\rho = L^2/(L^2 + d\sigma_e^2)$ and estimated local value price $\widetilde{P}_x = \sqrt{\widetilde{g}_x^\top \Sigma_x \widetilde{g}_x}/U_x$. Conditional on $\widetilde{g}$, the mean squared local value error is*

$$\mathcal{R}_\rho(\nu) = \frac{\kappa}{N}\sum_x \frac{v_x}{\nu_x}, \qquad v_x = \rho^2 \widetilde{g}_x^\top \Sigma_x \widetilde{g}_x + (1-\rho)L^2 U_x^2. \tag{9}$$

If $\sum_x v_x > 0$, its minimizing allocation satisfies

$$\nu_x^\star = \frac{\sqrt{v_x}}{\sum_z \sqrt{v_z}} \;\propto\; U_x \underbrace{\sqrt{(1-\rho)L^2 + \rho^2 \widetilde{P}_x^2}}_{\text{posterior value price}}, \qquad \mathcal{R}_\rho^\star = \frac{\kappa}{N}\left(\sum_x \sqrt{v_x}\right)^2. \tag{10}$$

Zero-risk contexts receive zero mass. The endpoint $\rho = 0$ denotes the uninformative limit.

Proof sketch. Gaussian conditioning gives $g_x \mid \widetilde{g}_x \sim \mathcal{N}(\rho\widetilde{g}_x, (1-\rho)L^2 I/d)$. Its quadratic expectation is $\mathbb{E}[g_x^\top \Sigma_x g_x \mid \widetilde{g}] = v_x$. Independent, centered sample-mean errors give Equation (9). Cauchy–Schwarz yields $(\sum_x \sqrt{v_x})^2 \le \sum_x v_x/\nu_x$, with equality at $\nu_x \propto \sqrt{v_x}$.

Remark (context-specific reliability). The allocation rule also holds when sensitivity-estimation noise varies across contexts, yielding context-specific reliability ρ_x (Appendix B.1). At an uninformative context, the posterior price reduces to the common prior scale L. Informative sensitivity estimates support a transition-dependent price. Thus price reliability can vary across contexts even when their predictive uncertainty is similar.

Corollary 2 (Gain and risk of plug-in pricing)**.** *Let $w_x = U_x/\sum_z U_z$, $\nu^U = w$, and $\nu_x^P \propto U_x \widetilde{P}_x$. For positive prices, writing $\mathrm{CV}_w(P) = \sqrt{\mathrm{Var}_w(P)}/\mathbb{E}_w[P]$,*

$$\frac{\mathcal{R}_1(\nu^U)}{\mathcal{R}_1^\star} = 1 + \mathrm{CV}_w(\widetilde{P})^2, \qquad \frac{\mathcal{R}_0(\nu^P)}{\mathcal{R}_0^\star} = \mathbb{E}_w[\widetilde{P}]\mathbb{E}_w[1/\widetilde{P}] \ge 1. \tag{11}$$

Both equal one exactly when prices are constant on the support. With at least two positive-weight contexts, the second ratio diverges if one price vanishes while another is bounded away from zero, holding the weights fixed.

Proof sketch. Substituting the two allocations into Equation (9) gives $\mathbb{E}_w[\widetilde{P}^2]/\mathbb{E}_w[\widetilde{P}]^2$ and $\mathbb{E}_w[\widetilde{P}]\mathbb{E}_w[1/\widetilde{P}]$. The first is $1 + \mathrm{CV}_w^2$. Cauchy–Schwarz bounds the second, whose cross-term $w_i w_j \widetilde{P}_j/\widetilde{P}_i$ proves divergence.

Remark (connection to VEX). For an affine task readout $F_x(y) = c + \widetilde{g}_x^\top y$, centering the readout evaluations gives $b_B = \sqrt{\widetilde{g}_x^\top \Sigma_x \widetilde{g}_x}$, exactly the reliable-limit priority, retaining sensitivity magnitude and directional uncertainty. The finite identity $b_B = P_{\mathrm{BE}} U_x$ remains exact for nonlinear readouts, whereas the projection is local for smooth F_x with $\widetilde{g}_x = \nabla F_x(\bar{y})$. At the uninformative limit, the common price L gives the priority LU_x. These relations connect VEX's priced operating point and uninformative limit (Appendices B.3–B.4).

4.2 Separating Scores from Correction Strength

VEX uses the gated exploration score b_t from Equation (4), or $b_t = U_t$ in the uninformative limit. The actor correction $B_t = c_n b_t$ weights each sampled action by its transition score, and c_n controls correction strength.

Proposition 3 (Actor correction as local score ascent)**.** *For a realized batch of M imagined transitions, let $\psi_i = \nabla_\theta \log \pi_\theta(a_i \mid s_i)$. With the stop-gradients in Equation (7), the extra raw ascent gradient is*

$$\widehat{g}_B = \frac{c_n}{M}\sum_i w_i b_i \psi_i, \qquad c_n = \eta(n)\beta \frac{\max(s_A, f)}{\max(s_b, \epsilon)}. \tag{12}$$

Fix the states, weights, coefficient, and score map $\bar{b}$, with $\bar{b}(s_i, a_i) = b_i$. Under common policy support, finite expectations, and valid differentiation under the action integral, fresh actions from π_θ give an unbiased estimate of $\nabla J_{\mathrm{fr}}(\theta)$, where

$$J_{\mathrm{fr}}(\vartheta) = \frac{c_n}{M}\sum_i w_i \mathbb{E}_{a\sim\pi_\vartheta(\cdot|s_i)}[\bar{b}(s_i, a)]. \tag{13}$$

Conditional on identical batches, gates, advantage scales, and withdrawal schedules, multiplying the entire score history and its RMS-tracker initialization by $a > 0$ leaves $c_n b_i$ unchanged whenever the denominator floor is inactive before and after scaling.

Proof sketch. Differentiating the detached correction term gives Equation (12). With the states, weights, coefficient, and score map frozen before fresh sampling, the score-function identity gives $\nabla J_{\mathrm{fr}}(\theta)$. Positive homogeneity of RMS and its consistently scaled EMA gives $s'_b = a s_b$, hence $c'_n = c_n/a$ when the floors are inactive.

Remark. The realized gradient identity permits batch-updated scales; the frozen expectation holds them fixed before fresh sampling. Transition scores b_i set action preferences, and c_n controls strength and withdrawal. Constant factors such as L cancel under the stated normalization. Appendix B.5 details the sampling dependence.

5 Experiments

We evaluate VEX as an exploration framework for DreamerV3 on continuous-control tasks. Our experiments test whether VEX improves sample efficiency and final performance, and how its exploration score, injection path, and withdrawal schedule contribute to these gains. We compare VEX with DreamerV3 and representative same-backbone exploration baselines, then use ablations to examine the design shared by its two operating points. Full experimental protocols and additional results appear in Appendices C–D.

5.1 Experimental Setup

Our evaluation uses three proprioceptive DeepMind Control Suite tasks (Tassa et al., 2018): Acrobot swing-up, Cheetah run, and Walker run, covering swing-up and locomotion. Figure 2(a–c) shows representative renderings of these environments. All tasks retain their native rewards. Each method trains for $510{,}000$ environment frames, corresponding to $255{,}000$ agent decisions with action repeat 2. We measure final performance by averaging the last five evaluation checkpoints and sample efficiency by a uniform-grid estimate of the normalized area under the learning curve (AUC). Cross-task summaries use normalized scores and the interquartile mean (IQM), with 95% task-stratified percentile bootstrap intervals (Agarwal et al., 2021).

We compare VEX with DreamerV3 (Hafner et al., 2025) and three exploration baselines: Optimistic World Models (OWM; Mete et al., 2026), disagreement reward, and random network distillation reward (RND; Burda et al., 2018). OWM modifies world-model learning through an optimistic transition objective. Disagreement reward uses the same ensemble uncertainty as VEX ($\rho = 0$), following the reward-exploration approach of Dreamer-Optimistic (Sukhija et al., 2025). RND supplies prediction-error novelty through the same reward interface. Both reward baselines add exploration rewards in imagination and train the critic on the resulting returns. These comparisons cover optimism in model learning and intrinsic rewards in policy learning.

All methods use the same 12M DreamerV3 backbone and shared training configuration in R2-Dreamer (Morihira et al., 2026), retaining the Dreamer representation objective. We evaluate VEX (priced) and VEX ($\rho = 0$) with exploration hyperparameters shared across tasks. Experiments run on NVIDIA RTX 4090 GPUs. Appendix A.3 gives the architecture and hyperparameters; Appendix C specifies evaluation settings and baseline adaptations.

5.2 Main Results

Table 1 reports final returns and cross-task IQM scores. Both VEX operating points exceed all compared baselines in mean final return on each task. VEX (priced) achieves the highest return on Acrobot swing-up (240), while VEX ($\rho = 0$) leads on Cheetah run (719) and Walker run (661). Their final-return IQMs are 0.54 and 0.58, respectively, compared with 0.45 for DreamerV3 and 0.50 for OWM, the strongest baseline on this aggregate metric. The final-return intervals of both VEX operating points lie above DreamerV3's interval under task-stratified resampling. The task-wise comparison favors VEX (priced) on Acrobot and VEX ($\rho = 0$) on the two locomotion tasks.

Figure 2(d–f) shows how these gains develop during training. On Walker, both VEX operating points separate from the baselines during the middle of training; on Cheetah, their advantage becomes clearer later in the budget. The improvements extend beyond the final evaluations: the AUC IQMs reach 0.31 for VEX (priced) and 0.34 for VEX ($\rho = 0$), compared with 0.27 for DreamerV3 and

Table 1: **Control performance at 510,000 environment frames.** Task columns report mean ± standard deviation of final return across seeds, with each seed's return averaged over its last five evaluations. Aggregate columns report IQM final return and AUC after division by 1000, with 95% task-stratified percentile bootstrap intervals in brackets. Bold marks the highest mean or IQM in each column. Both VEX operating points use hyperparameters fixed across tasks.

Method	Acrobot swing-up	Cheetah run	Walker run	IQM final	IQM AUC
DreamerV3	208 ± 24	547 ± 82	533 ± 104	.45 [.38, .52]	.27 [.25, .29]
OWM	183 ± 33	587 ± 68	617 ± 81	.50 [.45, .55]	.29 [.25, .32]
Disagreement reward	212 ± 42	575 ± 277	575 ± 135	.44 [.31, .58]	.28 [.23, .34]
RND reward	199 ± 36	620 ± 89	520 ± 84	.46 [.41, .52]	.29 [.26, .32]
VEX (priced)	**240 ± 8**	655 ± 54	629 ± 44	.54 [.53, .57]	.31 [.28, .35]
VEX ($\rho = 0$)	227 ± 41	**719 ± 38**	**661 ± 16**	**.58** [.57, .60]	**.34** [.32, .35]

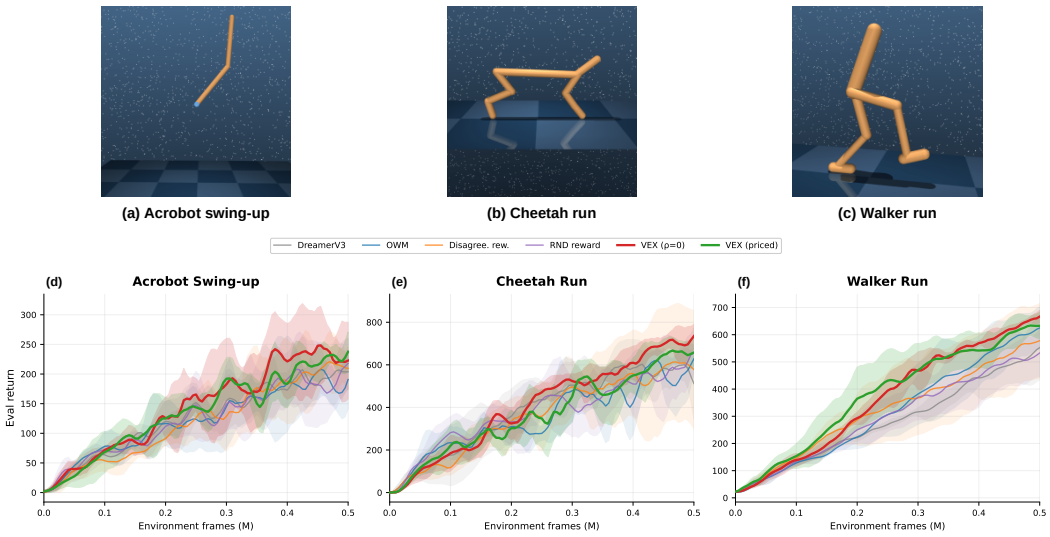

Figure 2: **Environments and learning curves.** (a–c) Representative views of Acrobot swing-up, Cheetah run, and Walker run. (d–f) Corresponding learning curves: lines show smoothed mean evaluation return, and shading indicates one standard deviation across seeds.

0.28–0.29 for the exploration baselines. Thus VEX improves aggregate performance over the training budget and at its end. Per-task AUC values are reported in Appendix D.

5.3 Ablation Study

Table 2 examines how the exploration schedule, injection path, and uncertainty signal affect final return through cumulative ablations. Starting from VEX ($\rho = 0$), we first remove withdrawal, then move the exploration signal from the actor advantage to the imagined reward, and finally replace ensemble disagreement with RND novelty. Each comparison retains the preceding changes. VEX ($\rho = 0$) provides a common uncertainty signal for the schedule and injection-path comparisons. We separately remove withdrawal from VEX (priced) to evaluate its contribution under Bellman pricing.

Withdrawal provides the largest final-return gains in the cumulative ablation. Compared with no withdrawal, it adds 71 points for VEX ($\rho = 0$) and 37 points for VEX (priced). Thus both operating points benefit from reducing exploration late in training, with a larger effect for VEX ($\rho = 0$). This pattern supports a gradual shift from collecting exploratory experience to refining the task policy, rather than maintaining the same correction strength throughout training.

Table 2: **Cumulative component analysis.** The first three rows follow VEX ($\rho = 0$) $\rightarrow$ no withdrawal $\rightarrow$ disagreement reward $\rightarrow$ RND reward; each step retains the preceding changes. Entries are mean differences in final return, first configuration minus second, pooled across tasks.

Comparison	Change	Δ final
VEX ($\rho = 0$) $-$ no withdrawal	Remove schedule	$+71$
VEX ($\rho = 0$) no withdrawal $-$ disagreement reward	Actor to reward	$+11$
Disagreement reward $-$ RND reward	Change signal	$+8$
VEX (priced) $-$ no withdrawal	Remove schedule	$+37$

With the ensemble signal held fixed in the injection-path comparison, advantage-level injection adds 11 points in final return over the reward interface. This smaller but positive contribution supports delivering the correction directly to the actor instead of through a finite imagined return and its learned value bootstrap. Within the reward interface, disagreement exceeds RND by 8 points in final return. The signal comparison therefore shows a modest final-performance benefit; the larger gains come from how the correction is delivered and, especially, when it is withdrawn.

6 Discussion

Limitations and future directions. Our evaluation focuses on proprioceptive control. Visual observations and longer-horizon tasks are natural extensions. Online estimation of sensitivity reliability could make pricing adaptive. Adaptive withdrawal could align exploration duration with learning progress. Discrete successor expectations offer a complementary extension of the soft task readout. Extending the allocation analysis to jointly evolving policies and world models would further connect adaptive pricing and withdrawal to long-term data collection.

Conclusion. We introduced VEX, a value-priced exploration framework for DreamerV3 that links experience acquisition to the task consequences of uncertain dynamics. Its central contribution is to make both the value of resolving uncertainty and the reliability of that value estimate part of exploration design. The Bayesian allocation analysis places selective pricing and uncertainty-only exploration within a common principle, explaining when task information can improve the use of a limited interaction budget and when broader coverage is preferable.

VEX translates this principle into exploration scores and direct corrections to imagined actor updates. Both operating points improve cross-task IQM scores over DreamerV3 and same-backbone exploration baselines on three continuous-control tasks. Ablations highlight the contribution of withdrawal, with additional gains from advantage-level injection. More broadly, the framework provides a basis for studying how learned task knowledge should guide data collection without prematurely excluding unfamiliar experience. By connecting the information an agent seeks to the decisions that information can improve, this work advances exploration as a means of making world-model learning more efficient and opens a path toward more capable agents under costly environment interaction.

AI USE STATEMENT

In this work, we used generative AI tools solely for writing assistance and language polishing, including improving the grammar, wording, clarity, and readability of the manuscript. We did not use these tools to develop research ideas, design the methodology or experiments, formulate or prove theoretical claims, implement methods, generate or process data, or analyze and interpret results. All AI-assisted text was reviewed and revised as needed by the authors to ensure that it preserved the intended technical meaning and accurately reflected the methods, results, and conclusions of this work. The authors take full responsibility for the final content of the paper, including all AI-assisted text.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A specify the scores, optimization objectives, training sequence, and hyperparameters. Appendix B gives full proofs, and Appendix A.4 derives the reward-to-advantage calculation. Appendix C specifies task definitions, training budgets, evaluation metrics, baseline configurations, and ablation protocols. Appendix D provides supplementary results.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

Lukas Bierling, Davide Pasero, Jan-Henrik Bertrand, and Kiki Van Gerwen. DreamerV3-XP: Optimizing exploration through uncertainty estimation. *arXiv preprint arXiv:2510.21418*, 2025.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Alexandra Carpentier and Rémi Munos. Finite time analysis of stratified sampling for Monte Carlo. *Advances in Neural Information Processing Systems*, 24, 2011.

Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. *Advances in Neural Information Processing Systems*, 33:14156–14170, 2020.

Amir-massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Artificial Intelligence and Statistics*, pp. 1486–1494. PMLR, 2017.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

Isaac Kauvar, Chris Doyle, Linqi Zhou, and Nick Haber. Curious replay for model-based adaptation. *arXiv preprint arXiv:2306.15934*, 2023.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Carlos E Luis, Alessandro G Bottero, Julia Vinogradska, Felix Berkenkamp, and Jan Peters. Model-based uncertainty in value functions. In *International Conference on Artificial Intelligence and Statistics*, pp. 8029–8052. PMLR, 2023.

Akshay Mete, Shahid Aamir Sheikh, Tzu-Hsiang Lin, Dileep Kalathil, and P. R. Kumar. Optimistic world models: Efficient exploration in model-based deep reinforcement learning. *arXiv preprint arXiv:2602.10044*, 2026.

Naoki Morihira, Amal Nahar, Kartik Bharadwaj, Yasuhiro Kato, Akinobu Hayashi, and Tatsuya Harada. R2-Dreamer: Redundancy-reduced world models without decoders or augmentation. *arXiv preprint arXiv:2603.18202*, 2026.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. *Advances in neural information processing systems*, 29, 2016.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International conference on machine learning*, pp. 8583–8592. PMLR, 2020.

Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In *International conference on machine learning*, pp. 5779–5788. PMLR, 2019.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Bhavya Sukhija, Lenart Treven, Carmelo Sferrazza, Florian Dörfler, Pieter Abbeel, and Andreas Krause. SOMBRL: Scalable and optimistic model-based RL. *arXiv preprint arXiv:2511.20066*, 2025.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. DeepMind Control Suite. *arXiv preprint arXiv:1801.00690*, 2018.

Claas Voelcker, Victor Liao, Animesh Garg, and Amir-massoud Farahmand. Value gradient weighted model-based reinforcement learning. *arXiv preprint arXiv:2204.01464*, 2022.

Claas Voelcker, Anastasiia Pedan, Arash Ahmadian, Romina Abachi, Igor Gilitschenski, and Amir-massoud Farahmand. Calibrated value-aware model learning with probabilistic environment models. *arXiv preprint arXiv:2505.22772*, 2025.

A Notation and Algorithm Details

This appendix supplements Section 3 and Algorithm 1. We collect the notation, specify how replay supplies the predictor targets, and summarize the architecture and return calculations used in training.

A.1 Notation

Table 3 groups the symbols by their role in the method and the supporting analyses. The index t denotes a transition in replay or imagination, as specified by context. Candidate indices run from 1 to K. The counter n measures environment frames, i indexes imagined batch transitions, and x indexes the transitions in the observation-allocation model.

Table 3: Notation for the method and analysis.

Symbol	Meaning
World model and candidate predictions	
$o_t, a_t, s_t = (h_t, z_t)$	Observation, action, and RSSM state with deterministic memory h_t and categorical latent z_t.
a_t^{eff}	Componentwise effective action $a_t / \max(\lvert a_t \rvert, 1)$, shared by the RSSM and auxiliary predictors.
$x = (s, a), h'$	Scored state–action pair and its shared deterministic successor.
$\phi, \theta, \psi;\ \mathcal{D}$	World-model, actor, and extrinsic-critic parameters; replay buffer.
$S, C, d = SC;\ K$	Number of categorical groups, classes per group, logit dimension, and candidate predictors.
$f_m, y_{m,t+1}$	Predictor m and its next-logit output $f_m(s_t, a_t^{\text{eff}})$.
$\ell^{\text{post}}, y^{\text{data}}$	Posterior logits and their group-centered regression target.
$\mathcal{B}_m, M_{m,t}, v_t$	Bootstrap subset, Bernoulli inclusion mask, and valid-transition mask.
$p(y), \epsilon_{\text{mix}}$	Soft categorical feature and its uniform-mixture coefficient.
$\tilde{s}_{m,t+1}$	Candidate successor $(h_{t+1}, p(y_{m,t+1}))$.
$\hat{r}, \hat{c}, V_{\text{ext}}$	Predicted reward, continuation probability, and current extrinsic value.
$F_x(y), F_{m,t}, \bar{F}_t$	Task readout, candidate evaluations, and their mean at step t.
$\bar{y}, \Sigma_x$	Candidate logit mean and sample covariance with denominator $K-1$.
b_B, U	BE score (standard deviation of candidate task readouts) and RMS coordinate uncertainty.
P_{BE}	Finite task-readout sensitivity per unit of predictive uncertainty.
$D_x, \Pi_{\widetilde{g}_x}$	Matrix of pairwise logit differences and projector $\widehat{g}_x \widehat{g}_x^\top$ onto the learned sensitivity direction.
Policy learning and score scaling	
π_θ, π_t, H	Actor, distribution $\pi_\theta(\cdot \mid s_t)$, and imagination horizon.
$\gamma, \lambda, G_t^\lambda$	Discount, return-mixing parameter, and extrinsic imagined return.
$\iota, \delta r_{t+1}, \delta A_t$	Constant intrinsic reward increment, per-step reward bonus, and resulting normalized-advantage increment.
$A_t^{\text{ext}}, \widetilde{A}_t$	Normalized extrinsic advantage and corrected advantage.
q_n, Δ_R, S_R, L_R	Cold-start gate for the priced setting, running return percentile spread, return scale, and its floor.
P_t, b_t, B_t, c_n	Effective price, exploration score $b_t = P_t U_t$, actor correction $B_t = c_n b_t$, and common scaling coefficient.
$s_b, s_A;\ \beta, f, \epsilon$	Score/advantage EMA RMS trackers; correction strength, advantage-scale floor, and denominator floor.
$n, \eta(n)$	Environment-frame counter and exploration withdrawal schedule.
$w_t, \alpha_{\text{ent}}, \mathcal{H}$	Discounted continuation weight, entropy coefficient, and policy entropy.
$\mathcal{L}_m, \mathcal{L}_\pi, \text{sg}$	Predictor loss, actor loss, and stop-gradient operator.
RMS, SD	Root mean square and sample standard deviation (denominator $K-1$ for candidates).
$M, \psi_i, \widehat{g}_B$	Imagined batch size, policy score $\nabla_\theta \log \pi_\theta(a_i \mid s_i)$, and additional ascent-gradient estimate.

Table 3 continued

Symbol	Meaning
$\vartheta, \bar{b}, J_{\mathrm{fr}}, \widehat{J}_{\mathrm{fr}}$	Trial actor parameters, frozen score map, its expected local objective, and importance-weighted empirical surrogate.
$\mathcal{I}_i$	Imagined history before drawing a_i, including s_i and w_i.
Finite-candidate and observation-allocation analyses	
$\mu_x, \widehat{\mu}_x, \kappa\Sigma_x$	True successor mean, its fresh-observation estimate, and single-observation noise covariance.
δ_x	Successor-mean error entering the local task-readout expansion.
N, n_x, ν_x	Fresh-observation budget, stratum count, and allocation fraction n_x/N.
κ	Common factor relating fresh-observation noise covariance to candidate covariance.
$g_x, \widetilde{g}_x, \varepsilon_x, \sigma_e^2$	True local readout sensitivity, noisy estimate, Gaussian estimation noise, and its isotropic variance.
L, ρ	Prior RMS sensitivity and Bayesian reliability $L^2/(L^2 + d\sigma_e^2)$.
$\sigma_{e,x}^2, \rho_x$	Context-specific sensitivity-noise variance and reliability $L^2/(L^2 + d\sigma_{e,x}^2)$.
$\mathrm{priority}_\rho(x)$	Unnormalized Bayesian allocation priority in Equation (1).
$\widetilde{P}_x$	Estimated local value price $\sqrt{\widetilde{g}_x^\top \Sigma_x \widetilde{g}_x}/U_x$.
$v_x, \nu^\star, \mathcal{R}_\rho^\star$	Posterior expected task variance, optimal allocation, and minimum Bayes risk.
ν^U, ν^P	Allocations proportional to uncertainty and plug-in task dispersion.
$w_x, \mathrm{CV}_w, \mathcal{R}_\rho$	Normalized uncertainty weight, weighted coefficient of variation, and posterior local-error risk.
$\widehat{g}_x$	Unit direction $\widetilde{g}_x/\|\widetilde{g}_x\|_2$ for nonzero learned local sensitivity.
H_F, L_F	Local gradient-Lipschitz and readout-Lipschitz bounds.
δ_m	Centered candidate logit $y_m - \bar{y}$.
$\widehat{F}_m, F_m^\star, e_m;\ R_m$	Estimated and reference readouts, their difference, and the local Taylor remainder.

A.2 Predictor Targets and Replay Objective

Section 3.2 trains the auxiliary predictors on real transitions so that their disagreement can be queried in imagination. We spell out the target and bootstrap weighting here. Softmax is invariant to a common logit offset within a categorical group. Centering therefore provides a consistent regression target. For posterior logits $\ell_{t+1}^{\mathrm{post}}$ and group j with C classes,

$$y_{t+1,j,c}^{\mathrm{data}} = \ell_{t+1,j,c}^{\mathrm{post}} - \frac{1}{C}\sum_{c'=1}^{C} \ell_{t+1,j,c'}^{\mathrm{post}}. \tag{14}$$

The action paired with s_t is the one producing the observed successor at $t+1$. Before concatenation with the state feature, it is transformed to $a_t^{\mathrm{eff}} = a_t/\max(|a_t|, 1)$ componentwise, matching the RSSM transition input. The same transformation is used when querying the predictors in imagination. The actor log probability still uses the original sampled action. Let v_t indicate that both states belong to the same episode, and sample $M_{m,t} \sim \mathrm{Bernoulli}(0.8)$ independently for each predictor and replay update. The masked form of Equation (5) is

$$\mathcal{L}_m = \frac{\sum_t v_t M_{m,t} \, \|f_m(\mathrm{sg}(s_t), \mathrm{sg}(a_t^{\mathrm{eff}})) - \mathrm{sg}(y_{t+1}^{\mathrm{data}})\|_2^2/d}{\max(1, \sum_t v_t M_{m,t})}. \tag{15}$$

Here t ranges over the replay batch. A separate optimizer minimizes $K^{-1}\sum_m \mathcal{L}_m$ using detached inputs and targets. Predictor outputs remain in the ambient d-dimensional space, with no additional centering projection.

A.3 Architecture and Training Details

Table 4 summarizes the architecture and hyperparameters shared across tasks. Auxiliary predictors use RMS normalization and SiLU activations, followed by a linear logit output. Candidate features follow Dreamer's groupwise uniform mixture, $p(y)_{j,c} = (1 - \epsilon_{\text{mix}})\,\text{softmax}(y_j)_c + \epsilon_{\text{mix}}/C$, and share the successor memory and task heads in Equation (2).

For the H imagined transitions, Dreamer's extrinsic return uses successor-indexed rewards and continuation:

$$G_t^\lambda = \hat{r}(s_{t+1}) + \gamma \hat{c}(s_{t+1})\big[(1-\lambda)V_{\text{ext}}(s_{t+1}) + \lambda G_{t+1}^\lambda\big], \qquad G_H^\lambda = V_{\text{ext}}(s_H). \tag{16}$$

The actor weights are $w_t = \prod_{j=0}^{t} \gamma \hat{c}(s_j)$, and the critic retains Dreamer's extrinsic imagined and replay targets. The gate uses the running 95th–5th percentile return spread. RMS estimates are updated before each actor correction from the unwithdrawn score and extrinsic advantage. Withdrawal is applied after scaling.

Table 4: Backbone and exploration hyperparameters.

Parameter	Value	Parameter	Value
Deterministic dimension	2048	Stochastic groups S	32
Categories per group C	16	Predictor count K	10
Predictor hidden width	256	Predictor hidden layers	2
Bootstrap inclusion probability	0.8	Predictor learning rate	10^{-4}
Imagination horizon H	15	Discount γ	$332/333$
Return parameter λ	0.95	Entropy coefficient α_{ent}	3×10^{-4}
Correction strength β	0.3	Advantage floor f	0.03
Scale EMA update fraction	0.01	Return-scale floor L_R	1
Uniform mixture ϵ_{mix}	0.01	Backbone learning rate	4×10^{-5}

A.4 Reward-to-Advantage Transfer

To clarify the injection-path comparison in Section 3.3, consider a constant imagined-reward increment ι with continuation equal to one. Holding critic predictions, the terminal bootstrap, and the return scale fixed, unrolling Equation (16) gives the direct advantage increment

$$\delta A_t = \frac{\iota}{S_R} \sum_{j=0}^{H-t-1} (\gamma\lambda)^j = (1-\gamma)B \sum_{j=0}^{H-t-1} (\gamma\lambda)^j. \tag{17}$$

The second equality uses the reward baseline's conversion $\iota = (1-\gamma)S_R B$ for a constant correction B. With $H = 15$, $\gamma = 332/333$, and $\lambda = 0.95$, the coefficient multiplying B is approximately 0.032 at the first imagined position and 0.019 when averaged uniformly over all 15 positions. This direct advantage increment depends on the finite return horizon. As the critic learns, bootstraps also transmit the intrinsic reward. Advantage-level injection applies B directly, setting the current actor correction independently of this propagation and retaining an extrinsic critic.

B Proofs and Supplementary Analysis

B.1 Bayesian Allocation

The theorem assumes the stated isotropic sensitivity prior conditional on the frozen pilot candidates and their covariances. Fresh observations are independent of this pilot, of the sensitivity observations, and across strata, conditional on the means and sensitivities. The fresh-observation covariance is $\kappa\Sigma_x$ with a common calibration factor κ. For fixed integer counts $n_x \geq 1$, the sample means have covariance $\kappa\Sigma_x/n_x$. The optimization uses the continuous relaxation and excludes pilot cost. Zero-risk strata can be omitted, with $0/0$ interpreted as zero in the risk.

Proof of Theorem 1. Gaussian conditioning gives

$$g_x \mid \widetilde{g}_x \sim \mathcal{N}\left(\rho\widetilde{g}_x, (1-\rho)\frac{L^2}{d}I\right),$$
$$\mathbb{E}[g_x^\top \Sigma_x g_x \mid \widetilde{g}] = \rho^2 \widetilde{g}_x^\top \Sigma_x \widetilde{g}_x + (1-\rho)\frac{L^2}{d}\operatorname{tr}\Sigma_x = v_x. \tag{18}$$

Independent, centered sample-mean errors imply Equation (9). Cauchy–Schwarz gives $(\sum_x \sqrt{v_x})^2 \le (\sum_x v_x/\nu_x)(\sum_x \nu_x)$, with equality exactly at $\nu_x \propto \sqrt{v_x}$ on positive-risk strata. Factoring U_x gives Equation (10). The endpoints follow by the zero-noise and uninformative limits. □

Context-specific noise. For fixed variances $\sigma_{e,x}^2$ and independent errors $\varepsilon_x \sim \mathcal{N}(0, \sigma_{e,x}^2 I)$, Gaussian conditioning applies separately at each context:

$$g_x \mid \widetilde{g}_x \sim \mathcal{N}\left(\rho_x\widetilde{g}_x, (1-\rho_x)\frac{L^2}{d}I\right), \qquad \rho_x = \frac{L^2}{L^2 + d\sigma_{e,x}^2}. \tag{19}$$

Consequently $v_x = \rho_x^2 \widetilde{g}_x^\top \Sigma_x \widetilde{g}_x + (1-\rho_x)L^2U_x^2$. The same conditional risk $\kappa N^{-1}\sum_x v_x/\nu_x$ and Cauchy–Schwarz argument give

$$\nu_x^\star \;\propto\; U_x\sqrt{(1-\rho_x)L^2 + \rho_x^2\widetilde{P}_x^2}. \tag{20}$$

Each context's uninformative limit recovers the priority LU_x. Its reliable limit recovers the plug-in dispersion $\sqrt{\widetilde{g}_x^\top \Sigma_x \widetilde{g}_x}$. All other observation and calibration assumptions remain those of Theorem 1.

B.2 The Two Endpoint Risk Ratios

Proof of Corollary 2. Write $S = \sum_x U_x$ and $p_x = \widetilde{P}_x$. At $\rho = 1$,

$$\mathcal{R}_1(\nu^U) = \frac{\kappa S^2}{N}\mathbb{E}_w[p^2], \qquad \mathcal{R}_1^\star = \frac{\kappa S^2}{N}\mathbb{E}_w[p]^2.$$

At $\rho = 0$, substituting $\nu_x^P = w_x p_x/\mathbb{E}_w[p]$ gives

$$\mathcal{R}_0(\nu^P) = \frac{\kappa L^2 S^2}{N}\mathbb{E}_w[p]\mathbb{E}_w[1/p], \qquad \mathcal{R}_0^\star = \frac{\kappa L^2 S^2}{N}.$$

The first ratio is $1 + \mathrm{Var}_w(p)/\mathbb{E}_w[p]^2$. The second is at least one by Cauchy–Schwarz. Equality in either holds exactly for constant positive prices. If $w_i, w_j > 0$, $p_i \to 0$ and $p_j \ge c > 0$, the second ratio is at least $w_i w_j c/p_i \to \infty$. □

The first identity also extends to some zero prices provided $\mathbb{E}_w[p] > 0$. In the uninformative limit, assigning zero mass to a positive-U_x stratum gives infinite risk.

B.3 Exact Finite Pricing

For $\bar{y} = K^{-1}\sum_m y_m$ and $\delta_m = y_m - \bar{y}$, define

$$\Sigma_x = \frac{1}{K-1}\sum_m \delta_m\delta_m^\top, \qquad U_x^2 = \frac{\operatorname{tr}\Sigma_x}{d}. \tag{21}$$

The pairwise variance identity, applied to the task-readout evaluations and logits, gives

$$b_B(x)^2 = \frac{\sum_{i<j}(F_x(y_i) - F_x(y_j))^2}{K(K-1)}, \qquad U_x^2 = \frac{\sum_{i<j}\|y_i - y_j\|_2^2}{dK(K-1)}. \tag{22}$$

For $U_x > 0$, taking their ratio gives

$$P_{\mathrm{BE}}(x) = \left[d\,\frac{\sum_{i<j}(F_x(y_i) - F_x(y_j))^2}{\sum_{i<j}\|y_i - y_j\|_2^2}\right]^{1/2}, \qquad b_B(x) = P_{\mathrm{BE}}(x)U_x. \tag{23}$$

If all logits coincide, both dispersions vanish and we set $P_{\mathrm{BE}} = 0$. For distinct pairs, P_{BE}^2/d averages the squared secants $[(F_x(y_i) - F_x(y_j))/\|y_i - y_j\|_2]^2$ with weights proportional to $\|y_i - y_j\|_2^2$.

For affine $F_x(y) = c + \widetilde{g}_x^\top y$ with $U_x > 0$ and $\widetilde{g}_x \neq 0$, write $\widehat{g}_x = \widetilde{g}_x/\|\widetilde{g}_x\|_2$. Centering the readout gives

$$b_B(x) = \sqrt{\widetilde{g}_x^\top \Sigma_x \widetilde{g}_x}, \qquad P_{\mathrm{BE}}(x) = \|\widetilde{g}_x\|_2 \sqrt{\frac{d\, \widehat{g}_x^\top \Sigma_x \widehat{g}_x}{\operatorname{tr} \Sigma_x}}. \tag{24}$$

Equivalently, let D_x stack the pairwise differences as rows and $\Pi_{\widetilde{g}_x} = \widehat{g}_x \widehat{g}_x^\top$. Then

$$P_{\mathrm{BE}}(x) = \sqrt{d}\, \frac{\|D_x \widetilde{g}_x\|_2}{\|D_x\|_F} = \|\widetilde{g}_x\|_2 \sqrt{d}\, \frac{\|D_x \Pi_{\widetilde{g}_x}\|_F}{\|D_x\|_F}. \tag{25}$$

B.4 Local Dispersion and Readout Stability

Let $\widetilde{g}_x = \nabla F_x(\bar{y})$ and assume ∇F_x is H_F-Lipschitz on a convex neighborhood containing the candidates. Then

$$\left| b_B(x) - \sqrt{\widetilde{g}_x^\top \Sigma_x \widetilde{g}_x} \right| \leq \frac{H_F}{2} \left(\frac{\sum_m \|\delta_m\|_2^4}{K-1} \right)^{1/2}. \tag{26}$$

Taylor's theorem gives $F_x(y_m) = F_x(\bar{y}) + \widetilde{g}_x^\top \delta_m + R_m$, with $|R_m| \leq H_F \|\delta_m\|_2^2/2$. Applying the reverse triangle inequality to the centered evaluation vectors bounds the left side by $\sqrt{\sum_m (R_m - \bar{R})^2/(K-1)} \leq \sqrt{\sum_m R_m^2/(K-1)}$, which proves the bound. A zero local gradient can therefore coexist with a nonzero BE score.

More generally, let $\widehat{F}_m = F_m^\star + e_m$ be two readouts of the same candidates. With sample standard deviations using denominator $K-1$, the same centered-vector argument gives

$$|\operatorname{SD}(\widehat{F}) - \operatorname{SD}(F^\star)| \leq \operatorname{SD}(e). \tag{27}$$

Common offsets cancel. Only candidate-dependent readout error changes the BE score. Finally, an L_F-Lipschitz readout satisfies $b_B \leq L_F \sqrt{d}\, U_x$ by Equation (22). This is an envelope: among affine readouts with $\|g\|_2 \leq L_F$, the exact largest dispersion is $L_F \sqrt{\lambda_{\max}(\Sigma_x)}$.

B.5 Local Score Ascent and the Actor Gradient

For the current actor θ and a realized batch, freeze the states, weights, coefficient, and score map at their current values. The local importance-weighted surrogate is

$$\widehat{J}_{\mathrm{fr}}(\vartheta) = \frac{c_n}{M} \sum_i w_i b_i \frac{\pi_\vartheta(a_i \mid s_i)}{\pi_\theta(a_i \mid s_i)}. \tag{28}$$

Proof of Proposition 3. With the imagined states, actions, weights, and correction detached, the additional term in the ascent objective is $M^{-1} \sum_i w_i \log \pi_\theta(a_i \mid s_i) B_i$. Differentiating and substituting $B_i = c_n b_i$ gives $\widehat{g}_B = (c_n/M) \sum_i w_i b_i \psi_i$. The derivative of the likelihood ratio in Equation (28) at $\vartheta = \theta$ is ψ_i, proving the realized-batch identity in Equation (12).

For the expected-score identity, fix the contexts s_i, weights w_i, coefficient c_n, and score map $\bar{b}$, and draw fresh actions $a_i' \sim \pi_\theta(\cdot \mid s_i)$. Assume common policy support and differentiation under the action integral, with finite relevant expectations. The corresponding surrogate $\widehat{J}_{\mathrm{fr}}'$ replaces each a_i, b_i by $a_i', \bar{b}(s_i, a_i')$. Importance weighting gives $\mathbb{E}[\widehat{J}_{\mathrm{fr}}'(\vartheta)] = J_{\mathrm{fr}}(\vartheta)$ and hence

$$\begin{aligned} \mathbb{E}\Big[\nabla_\vartheta \widehat{J}_{\mathrm{fr}}'(\vartheta)\Big|_{\vartheta=\theta} \Big] &= \nabla J_{\mathrm{fr}}(\theta) \\ &= \frac{c_n}{M} \sum_i w_i \mathbb{E}_{a \sim \pi_\theta(\cdot \mid s_i)} [\bar{b}(s_i, a) \nabla_\theta \log \pi_\theta(a \mid s_i)]. \end{aligned} \tag{29}$$

Finally, batch RMS is positively homogeneous, and its EMA is positively homogeneous when its initialization and entire input history are scaled together. Thus $b_i' = a b_i$ gives $s_b' = a s_b$ and, with both denominator floors inactive, $c_n' = c_n/a$ and $c_n' b_i' = c_n b_i$. □

The realized-batch identity holds for the actual batch-updated gate and normalization. In the frozen-objective interpretation, these quantities are fixed before the fresh actions are drawn. Detaching a statistic computed from the original actions leaves its sampling dependence intact, so the actual correction has the following general covariance form. Let $\mathcal{I}_i$ be the history before drawing a_i, including s_i and the continuation weight w_i. Under the same regularity assumptions, the score-function identity (Sutton et al., 1999) gives $\mathbb{E}[\psi_i \mid \mathcal{I}_i] = 0$, hence

$$\widehat{g}_B = \frac{1}{M}\sum_i w_i B_i \psi_i, \qquad \mathbb{E}[\widehat{g}_B] = \frac{1}{M}\sum_i \mathbb{E}[w_i \operatorname{Cov}(B_i, \psi_i \mid \mathcal{I}_i)]\,. \tag{30}$$

The full B_i, including batch-updated gates and RMS statistics, stays inside the covariance because it can depend on the sampled actions and later imagined states. This identity concerns the gradient supplied to the optimizer. Its state and preconditioning determine the parameter step.

For one state with two equiprobable actions, $w = 1$, and fixed detached corrections B_1, B_2, the expected additional ascent gradient for the first logit is $(B_1 - B_2)/4$. Corrections $(2, 1)$ and $(1, 2)$ have equal RMS but produce gradients $+1/4$ and $-1/4$, respectively.

C Experimental Protocol

C.1 Tasks and Evaluation

The three DeepMind Control Suite tasks in Table 5 use proprioceptive observations and native rewards. Budgets count training environment frames, including action repeats, and exclude evaluation. We report three training runs per method and task, keeping VEX hyperparameters fixed across tasks.

Table 5: Reporting tasks. Each method uses three training seeds and action repeat 2.

Task	Domain	Suite task	Frame budget
Acrobot swing-up	`acrobot`	`swingup`	510,000
Cheetah run	`cheetah`	`run`	510,000
Walker run	`walker`	`run`	510,000

Evaluation is scheduled at initialization and every 10,000 environment frames. Each evaluation averages ten episodes, with actions set to the mode of the actor's distribution.

C.2 Configurations and Baselines

All methods share R2-Dreamer's 12M DreamerV3 backbone, Dreamer representation objective, replay settings, batch size 16, sequence length 64, and learner-to-data ratio. Table 4 gives the shared architecture and exploration parameters. With the scores and gate of Section 3.1, VEX (priced) uses $b = (1 - q_n)U + q_n b_B$, and VEX ($\rho = 0$) uses $b = U$, the uninformative $\rho = 0$ limit. Their withdrawal multiplier remains one through 340,000 frames and decreases linearly to zero at 510,000, after relative scaling. No-withdrawal variants keep it at one.

DreamerV3 sets the exploration coefficient to zero. OWM adds an optimistic transition-prior loss with $\alpha = 10^{-4}$ and $\eta_{\mathrm{OWM}} = 3 \times 10^{-6}$ (Mete et al., 2026): sampled imagined-successor log probabilities are weighted by detached normalized extrinsic advantages, with prior entropy regularization and fixed recurrent states. Its actor receives no VEX correction.

Disagreement reward uses the same uncertainty signal as VEX ($\rho = 0$) and relative scaling with $\beta = 0.3$. The detached normalized correction B_t becomes imagined reward $\delta r_{t+1} = (1 - \gamma) S_R B_t$, converting a nominal steady-state return increment to reward units. The critic learns recomputed shaped λ-returns. World-model objectives are unchanged. The RMS reference s_A is computed from advantages normalized by the return scale S_R, using the current critic and returns before the current bonus is added. This ordering prevents the current bonus from inflating its own scaling reference. This same-backbone adaptation of Dreamer-Optimistic's reward exploration (Sukhija et al., 2025) uses one task actor–critic. The comparison changes conversion, return propagation, and critic learning (Appendix A.4).

RND replaces disagreement with detached mean squared prediction error against a fixed random target on next-state latent features (Burda et al., 2018), retaining the same reward path and scaling. Both networks have two 256-unit hidden layers and 128 outputs. The predictor learns from real replay at rate 10^{-4}.

C.3 Metrics and Statistical Comparisons

Final return averages the last five evaluations of a completed run. AUC averages returns linearly interpolated at 20,000-frame intervals from 0 to 500,000 frames. Per-task summaries and component comparisons retain native score units.

For cross-task aggregation, final returns and AUC values are divided by 1000. IQM is the 25%-trimmed mean of the nine normalized task–seed scores. Its 95% percentile intervals use 10,000 bootstrap replicates, resampling three runs with replacement independently within each fixed task. Component comparisons in Table 2 report mean differences in final return between configurations, pooled across tasks. The withdrawal interaction compares the two final-return benefits.

C.4 Ablation Design

The ladder starts from VEX ($\rho = 0$) to share the uncertainty signal with the reward baselines. The cumulative sequence VEX ($\rho = 0$) $\rightarrow$ no withdrawal $\rightarrow$ disagreement reward $\rightarrow$ RND reward tests the schedule, injection path, and signal in turn. The no-withdrawal variant of VEX (priced) supplies the corresponding schedule contrast. With M denoting final return, the interaction between the two withdrawal contrasts is

$$\begin{aligned}\mathcal{D}_M = &[M(\text{VEX } (\rho = 0)) - M(\text{VEX } (\rho = 0), \text{ no withdrawal})] \\ &- [M(\text{VEX (priced)}) - M(\text{VEX (priced), no withdrawal})].\end{aligned} \tag{31}$$

Positive values indicate a larger withdrawal benefit for VEX ($\rho = 0$). We report the interaction alongside both schedule contrasts.

D Extended Results

This appendix supplements the main results with per-task learning efficiency and withdrawal interactions.

D.1 Per-Task Learning Efficiency

Table 6 complements the aggregate AUC in Table 1 by showing performance over the training budget on each task.

Table 6: **Per-task learning efficiency.** Time-normalized AUC is reported as mean $\pm$ standard deviation across three runs in native score units. Bold marks means within one point of the column maximum. The evaluation grid is specified in Appendix C.3.

Method	Acrobot	Cheetah	Walker
DreamerV3	127 ± 16	379 ± 71	288 ± 27
OWM	123 ± 33	366 ± 24	322 ± 61
Disagreement reward	119 ± 29	371 ± 95	327 ± 87
RND reward	126 ± 29	387 ± 8	302 ± 57
VEX (priced)	**139 ± 20**	372 ± 44	**382 ± 64**
VEX ($\rho = 0$)	**140 ± 32**	**426 ± 25**	372 ± 18

Both VEX operating points improve AUC over all baselines on Acrobot and Walker. On Acrobot, their means are close (139 and 140), compared with 127 for DreamerV3. VEX (priced) achieves the highest AUC on Walker (382), followed by VEX ($\rho = 0$) (372), exceeding the strongest baseline (327). On Cheetah, VEX ($\rho = 0$) leads with 426, while VEX (priced) reaches 372, within the baseline

range of 366–387. The two evaluation metrics capture different aspects of learning: VEX (priced) leads in Walker AUC, although VEX ($\rho = 0$) achieves the higher final return. The AUC breakdown therefore complements the final-return comparison by distinguishing sustained learning performance from the policy quality reached at the end of training.

D.2 Withdrawal Interaction

The two schedule contrasts in Table 2 quantify the final-return benefit of withdrawal for each operating point. Their difference, Equation (31), measures how this benefit changes with pricing.

Table 7: Difference between the withdrawal benefits for VEX ($\rho = 0$) and VEX (priced), measured in final return.

Comparison	Δ final
$\mathcal{D}_M$ from Equation (31)	$+34$

The interaction is $+34$ points in final return, equal to the difference between the 71- and 37-point schedule benefits in Table 2. Both operating points benefit from withdrawal, with a larger gain for VEX ($\rho = 0$) as training approaches its budget.