WHAT SHOULD I TRY NEXT? DECISION-RELEVANT COUNTERFACTUAL INFORMATION GAIN FOR ACTIVE WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

World models enable agents to evaluate actions through imagined futures, yet these predictions are useful only when the underlying environment dynamics are sufficiently identified. Existing exploration methods typically seek novel states, large prediction errors, local model disagreement, or generic information gain. Such objectives may spend costly interactions resolving uncertainty that has little effect on subsequent decisions. We introduce Active Epistemic World Modeling (AEWM), a framework that asks which real interaction should be performed now to make relevant counterfactual futures more reliable. AEWM distinguishes irreducible outcome noise from reducible uncertainty about latent mechanisms and targets decision-relevant counterfactual information gain. Its practical Counterfactual Disagreement Reduction (CDR) score estimates query-set variance reduction by sampling hypothetical outcomes and reweighting ensemble members without inner-loop retraining. In a matched five-seed study on 60 held-out zero-delay hidden-physics worlds, CDR lowers one-step prediction RMSE at 32 interactions by 19.7% relative to local disagreement. It modestly improves budget-curve area relative to random selection, but matches random at the endpoint and uses roughly $3.2\times$ more scoring time. Historical model ablations expose a delayed-action training-target mismatch, and an all-candidate audit finds weak fidelity of the cheap CDR rankings. The completed study establishes a one-step acquisition result; planner-generated queries and downstream control remain to be tested.

1 INTRODUCTION

What should an agent try next when its world model is uncertain? A model-based agent ultimately needs trustworthy predictions for the actions it might choose, rather than precise knowledge of every dynamical parameter. Exploring an uncertain action can be wasteful if its observation does not clarify the *other* futures that matter to the decision. This is the distinction between local uncertainty and targeted experimental design (Lindley, 1956; Chaloner & Verdinelli, 1995; Bickford Smith et al., 2023; Hübotter et al., 2024).

We organize this question as *Active Epistemic World Modeling* (AEWM). The framework separates stochastic variation within a mechanism from uncertainty about which persistent mechanism governs the world; specifies counterfactual queries that the agent wants to predict; and values an experiment by the expected reduction of uncertainty over those queries. In the ideal version, the queries can be induced by a planner and the agent can alternate between informative experiments and control. Our implemented instantiation instead uses uniformly drawn one-step reward queries and a fixed predictor. That distinction makes the tested claim precise without discarding the broader motivation from uncertainty-aware and latent world-model planning (Chua et al., 2018; Hafner et al., 2019; 2020; Janner et al., 2019).

The practical bottleneck is estimating how every possible observation would change predictions. We study *Counterfactual Disagreement Reduction* (CDR): sample outcomes under an ensemble, likelihood-reweight its members, and measure the expected fall in disagreement across the query set. This is an inexpensive surrogate for a full posterior/context update, not a new information-theoretic identity. It makes AEWM executable with many candidate actions, while creating a testable

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.

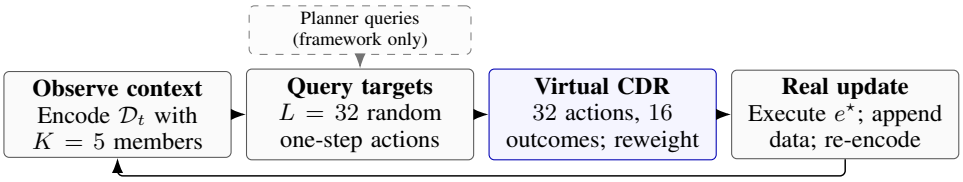

Figure 1: AEWM acquisition loop in the completed experiment. CDR scores candidate interactions by virtual ensemble reweighting at fixed one-step queries; after a real interaction, member contexts are re-encoded. The dashed planner-query input is a framework extension, not part of the reported test.

approximation error because an actual observation also changes the context-conditioned prediction of each member. Figure 1 separates the tested acquisition loop from the planner-conditioned extension.

Our matched experiment uses five independently trained structured-state predictors and 60 held-out zero-delay hidden-physics worlds. It isolates acquisition by sharing a checkpoint and canonical evaluation state within each seed. At 32 interactions, CDR has 19.7% lower one-step RMSE than local disagreement, with the same sign across all five seeds and 50/60 world averages. It does not beat random selection at that endpoint; local disagreement is stronger at the earliest budgets. The budget curve, old 100-world diagnostic sweep, fixed-state architecture ablations, and score-fidelity audit together reveal when the method helps and where the current implementation fails.

Contributions. We provide (i) a world-model formulation of decision-relevant counterfactual acquisition that explicitly distinguishes epistemic from aleatoric uncertainty, while positioning it relative to predictive and transductive active learning; (ii) a batched, fixed-prediction ensemble approximation and an audit of its candidate-ranking fidelity; and (iii) a matched interaction-budget evaluation together with historical model-component diagnostics. The validated outcome is better late-budget *one-step prediction than local disagreement*; planner-conditioned acquisition and multi-step control remain untested.

2 Related Work

Prediction- and decision-targeted design. Classical Bayesian design and active learning select measurements for information gain or expected variance reduction (Lindley, 1956; Chaloner & Verdinelli, 1995; MacKay, 1992; Cohn et al., 1996). Information-directed sampling relates information acquisition to decision regret, and later work studies which learning target an agent should pursue (Russo & Van Roy, 2014; Arumugam & Van Roy, 2021). Expected predictive information gain (EPIG) targets a future input distribution rather than all model parameters (Bickford Smith et al., 2023); transductive active learning targets a designated prediction set (Hübotter et al., 2024). Targeted active learning and goal-oriented design directly value downstream decisions or quantities of interest (Filstroff et al., 2024; Zhong et al., 2026). AEWM shares this established design principle. Its intended setting is a dynamically reachable experiment that identifies hidden mechanisms for world-model queries; the tested CDR score is a variance surrogate, not an implementation of EPIG, optimal-decision information gain, or exact Bayesian design.

Active dynamics identification. Latent task-context inference appears in PEARL (Rakelly et al., 2019). Most directly related, Achterhold & Stueckler (2021) choose action sequences informative about the latent context of a context-conditional dynamics model. Gaussian-process designs also account for trajectory constraints while learning dynamics (Buisson-Fenet et al., 2020). VIME targets information about dynamics parameters (Houthooft et al., 2016); local disagreement and imagined novelty drive other model-based exploration methods (Pathak et al., 2019; Shyam et al., 2019; Sekar et al., 2020). CEE-US, CASCADE, OPAX, and WAKER address structured exploration, limited deployments, active identification, or robust world-model curricula (Sancaktar et al., 2022; Xu et al., 2022; Sukhija et al., 2023; Rigter et al., 2024). CDR differs operationally by scoring the expected reduction of disagreement at a separate query set. That distinction is not evidence of superiority: no faithful matched comparison to these complete algorithms has been run.

What matters for control. PETS propagates uncertainty through probabilistic ensemble rollouts (Chua et al., 2018); PlaNet and Dreamer learn latent models for planning or imagination-based behavior (Hafner et al., 2019; 2020). MBPO studies when model rollouts remain trustworthy, and selective Dyna planning asks where an imperfect model should be used (Janner et al., 2019; Abbas et al., 2020). Goal-aware prediction allocates model capacity to task-relevant futures (Nair et al., 2020). H-UCRL and LOVE use epistemic optimism or value-aware ensemble uncertainty to seek useful interactions, while informational rewards target task-relevant hidden state in POMDPs (Curi et al., 2020; Seyde et al., 2022; Xie et al., 2024). These methods connect prediction quality to decisions more directly than our completed random-query, one-step test.

Uncertainty and calibration. Aleatoric–epistemic decomposition has been used for active learning and risk-sensitive decisions (Kendall & Gal, 2017; Depeweg et al., 2018). Deep ensembles are practical uncertainty proxies, not automatically posterior samples (Lakshminarayanan et al., 2017); regression calibration itself requires explicit assessment (Kuleshov et al., 2018). We therefore distinguish an ideal physical-mechanism posterior from learned context embeddings and outer ensemble weights, and audit CDR rankings rather than treating them as calibrated posterior utilities.

3 Problem Formulation

We consider a family of partially observed controlled environments indexed by a latent mechanism $m \in \mathcal{M}$. The mechanism can encode physical parameters, transition rules, object-specific interaction laws, or other persistent/slowly varying factors. A world is sampled from a prior $p(m)$ and generates observations and rewards through

$$s_{t+1} \sim p_m(s_{t+1} \mid s_t, a_t), \qquad o_t \sim p(o_t \mid s_t), \qquad r_t \sim p(r_t \mid s_t, a_t, m). \tag{1}$$

The agent receives history $h_t = (o_{\leq t}, a_{<t}, r_{<t})$ and dataset $\mathcal{D}_t$ but not m. Context-conditioned dynamics models provide one way to infer such latent factors from recent transitions (Rakelly et al., 2019; Achterhold & Stueckler, 2021). An ideal Bayesian formulation maintains $p(m \mid \mathcal{D}_t)$ and a sufficient state $z_t = f(h_t)$. The learned ensemble below is not a calibrated approximation to that posterior. For delayed actions, a sufficient state must include pending controls; the reported implementation omits them, as discussed in Section 10.

A *counterfactual query* $q = (z_t, a^q_{t:t+H-1}, g_q)$ asks for a statistic $Y_q = g_q(\tau_q)$ of a future trajectory τ_q. In general, Y_q might be cumulative reward or a terminal event. In the completed experiment it is only the predicted one-step reward at a uniformly sampled action. A planner-generated query set would make relevance depend on current decisions, but that extension is not evaluated here.

A *real experiment* $e \in \mathcal{E}_t$ is an action or short action sequence executed in the environment, yielding outcome Y_e and incurring cost $C(e) \geq 0$. Costs can represent physical trials, elapsed time, energy, risk, or consumed in-game resources. The central problem is to choose at most B such experiments before or interleaved with control.

We target *epistemic* rather than total predictive uncertainty, following the established distinction used in active and risk-sensitive learning (Kendall & Gal, 2017; Depeweg et al., 2018). For a query variable Y_q, the law of total variance gives

$$\mathrm{Var}(Y_q \mid \mathcal{D}_t) = \underbrace{\mathbb{E}_{m\sim p(m|\mathcal{D}_t)}[\mathrm{Var}(Y_q \mid m, \mathcal{D}_t)]}_{\text{aleatoric}} + \underbrace{\mathrm{Var}_{m\sim p(m|\mathcal{D}_t)}[\mathbb{E}[Y_q \mid m, \mathcal{D}_t]]}_{\text{epistemic}}. \tag{2}$$

The first term contains stochasticity that additional identification cannot necessarily remove; the second captures disagreement induced by uncertainty over the world mechanism. We therefore define

$$U_{\mathrm{epi}}(q; \mathcal{D}_t) = \mathrm{Var}_{m\sim p(m|\mathcal{D}_t)}\left[\mu_q(m)\right], \qquad \mu_q(m) = \mathbb{E}[Y_q \mid m, \mathcal{D}_t]. \tag{3}$$

The objective is to use a limited interaction budget to improve predictions for the specified query set, not necessarily to identify every coordinate of m. In a fully decision-conditioned AEWM agent, a planner would supply candidate futures and their query weights; a control-versus-experiment rule could then compare the value of acting with the value of obtaining another observation. The completed study isolates the acquisition part: queries are random, the gate is disabled, and no planner-induced distribution is evaluated.

4 Active Epistemic World Modeling

Predictive model. The implemented model has five independently initialized members trained on the same bootstrap transitions. Each encodes up to 16 observed transitions with a two-layer Transformer into a 16-dimensional diagonal-Gaussian mechanism embedding. A conditional MLP predicts the mean and standard deviation of the one-step state change and reward from state, action, and embedding. A 20-noise-level denoising head is used only for an auxiliary training loss; acquisition and evaluation use the Gaussian head, not sampled diffusion rollouts. The general H-step model

$$p_\theta(z_{t+1:t+H}, r_{t:t+H-1} \mid z_t, a_{t:t+H-1}, m) \tag{4}$$

is *not* tested for $H > 1$: here $H = 1$, $z_t = s_t \in \mathbb{R}^4$, and the pending-action queue is omitted.

Candidate-action queries. At each identification step we draw $L = 32$ one-step actions uniformly from $[-1, 1]^2$; these are not planner trajectories. The primary selector uses uniform query weights $w_t(q) = 1/|\mathcal{Q}_t|$. A reward-gap-weighted variant is a secondary ablation (Appendix A.1). For a query set fixed before observing the next experiment, define

$$U_{\mathcal{Q}_t}(\mathcal{D}_t) = \sum_{q \in \mathcal{Q}_t} w_t(q) U_{\mathrm{epi}}(q; \mathcal{D}_t). \tag{5}$$

Decision-relevant information gain. Let $G_{\mathcal{Q}_t}(M) = (\mu_q(M))_{q \in \mathcal{Q}_t}$ collect the mechanism-dependent predictions needed for the query set. The ideal AEWM objective is the information an experiment e yields about this functional, rather than about the entire mechanism:

$$\text{DR-CIG}(e) = \frac{I(G_{\mathcal{Q}_t}(M); Y_e \mid \mathcal{D}_t, e)}{(C(e) + \varepsilon)^\gamma}. \tag{6}$$

We do not estimate this mutual information. For a fixed prediction map, queries, and weights, a variance-reduction design instead uses $U_{\mathcal{Q}_t}(\mathcal{D}_t) - \mathbb{E}_{Y_e}[U_{\mathcal{Q}_t}(\mathcal{D}_t \cup \{e, Y_e\})]$; Section 5 shows its standard Bayesian identity. CDR approximates *this variance surrogate* by reweighting ensemble members. Information gain and variance reduction need not rank experiments alike, and neither objective has been shown to optimize decision loss in the completed study.

Training and swap ablation. Each member is trained with one-step Gaussian negative log likelihood and denoising loss, plus a mechanism-embedding KL term, variance-calibration penalty, and context-half consistency. The full variant adds a simulator-generated mechanism-swap target,

$$\mathcal{L}_{\mathrm{swap}} = \mathbb{E}_{i,j,a}\left[-\log p_\theta(s'_{i \leftarrow j} \mid s_i, a, \bar{m}(\mathcal{D}_j))\right]. \tag{7}$$

The current target applies a immediately under mechanism m_j even if that mechanism delays actions. This mismatch makes the swap ablation diagnostic, not evidence that the loss helps. No observation reconstruction, termination, or long-horizon reward loss is trained. Architecture and coefficients are in Appendix A.

5 Ideal Bayesian Interpretation

The following identity explains the variance surrogate used by CDR, not the information-gain objective in Eq. 6. It is a standard property of coherent Bayesian conditioning, not a new exploration guarantee. Fix the candidate queries, their weights, the conditional prediction map μ_q, and model parameters before seeing an experiment outcome. For a latent mechanism M with posterior $p(M \mid \mathcal{D})$ and $X_q = \mu_q(M)$, the law of total variance gives

$$\mathrm{Var}(X_q \mid \mathcal{D}) = \mathbb{E}_{Y_e \mid \mathcal{D}, e}[\mathrm{Var}(X_q \mid \mathcal{D}, Y_e, e)] + \mathrm{Var}_{Y_e \mid \mathcal{D}, e}[\mathbb{E}[X_q \mid \mathcal{D}, Y_e, e]]. \tag{8}$$

Thus the expected posterior variance reduction for q is

$$\Delta_e(q) = \mathrm{Var}_{Y_e \mid \mathcal{D}, e}[\mathbb{E}[X_q \mid \mathcal{D}, Y_e, e]] \geq 0. \tag{9}$$

An observation is valuable when its possible outcomes move the posterior mean of a query prediction. The same design perspective can be written as $I(G_{\mathcal{Q}}(M); Y_e \mid \mathcal{D}, e)$ for a query-relevant functional

$G_{\mathcal{Q}}(M)$ (Lindley, 1956; Chaloner & Verdinelli, 1995). CDR uses ensemble variance instead of estimating this information directly.

Neither identity applies literally to the learned procedure: independently initialized ensemble members are not posterior draws from a specified prior, their Gaussian likelihoods may be misspecified, and a real observation changes context-encoded predictions as well as outer weights. Expected reduction also does not imply reduction for every realized observation or lower held-out error. Appendix A.6 gives the assumptions and a simple decision-error bound for context; Appendix A.3–A.4 audits the virtual-to-context-update gap.

6 A Practical AEWM Instantiation: Counterfactual Disagreement Reduction

The learned system has K independently initialized model members f_k and outer likelihood weights ω_k, with $\sum_k \omega_k = 1$. Each member also infers an internal diagonal-Gaussian mechanism embedding from the observed context. These three objects differ: an ideal posterior over physical mechanisms, a member's context embedding, and the outer weight on that member. Because members share training data and are not draws from a specified mechanism prior, ω_k is a heuristic confidence distribution, not a calibrated mechanism posterior (Lakshminarayanan et al., 2017). For a fixed context and query q, member k predicts $\mu_k(q)$, and the weighted reward disagreement is

$$D_{\mathcal{Q}}(\mathcal{D}) = \sum_{q \in \mathcal{Q}} w(q) \sum_{k=1}^{K} \omega_k (\mu_k(q) - \bar{\mu}(q))^2, \quad \bar{\mu}(q) = \sum_k \omega_k \mu_k(q). \tag{10}$$

For each candidate experiment e, the members produce Gaussian outcome likelihoods $p_k(y \mid e, \mathcal{D})$. We sample S hypothetical outcomes from their weighted mixture and compute likelihood-reweighted weights

$$\widetilde{\omega}_k(y, e) = \frac{\omega_k p_k(y \mid e, \mathcal{D})}{\sum_j \omega_j p_j(y \mid e, \mathcal{D})}. \tag{11}$$

This has the algebraic form of Bayes' rule but only the operational meaning of reweighting the learned members. Let $\widetilde{D}_{\mathcal{Q}}(y, e)$ use these weights while keeping all $\mu_k(q)$ fixed. The implemented score is

$$\mathrm{CDR}(e) = \frac{\max\left\{0,\ D_{\mathcal{Q}}(\mathcal{D}) - S^{-1} \sum_{s=1}^{S} \widetilde{D}_{\mathcal{Q}}(y^{(s)}, e)\right\}}{(C(e) + \varepsilon)^{\gamma}}. \tag{12}$$

The maximum clips negative finite-Monte-Carlo estimates; it is not part of the ideal objective. A real transition is appended to context and re-encoded by every member, so both ω_k *and* $\mu_k(q)$ subsequently change. Neural parameters remain frozen. Equation 12 therefore estimates only a cheap virtual update; its candidate-ranking fidelity is audited in Appendix A.4.

The matched comparison disables the optional reward-margin gate and uses uniform query weights. Its weighted-query variant and the historical gated variant are described in Appendix A.1 and Appendix A.5, respectively. Query and candidate predictions are batched, avoiding $N \times S$ neural forward passes. The historical runner needlessly scored CDR for random and local selectors; the matched run omits that work. The reported score-region timing excludes query generation and environment stepping.

7 Experiments

We test whether CDR acquires interactions that improve held-out one-step counterfactual prediction. The completed benchmark is structured-state hidden physics; pixel observations, combinatorial rules, and valid multi-step planning remain outside the present empirical scope.

Environment and training. Each world fixes mass, friction, damping, actuator gain, and an action delay of zero, one, or two steps. The agent sees position and velocity but not the mechanism. A shared bootstrap dataset contains 1,279,874 transitions from 10,000 training worlds. The primary

Algorithm 1 Matched-study identification with uniform-query CDR

Require: trained members $\{f_k\}_{k=1}^{K}$, initial context $\mathcal{D}$, budget B
for $b = 1, \ldots, B$ **do**
  infer each member's mechanism embedding from $\mathcal{D}$
  draw 32 one-step query actions $\mathcal{Q}$; set $w(q) = 1/|\mathcal{Q}|$
  compute fixed-context predictions and $D_{\mathcal{Q}}(\mathcal{D})$
  draw 32 candidate experiment actions $\mathcal{E}$
  for all $e \in \mathcal{E}$ **do**
    sample 16 hypothetical outcomes from the ensemble mixture
    reweight members by Eq. 11 for each outcome
    score e with Eq. 12, holding query predictions fixed
  execute $e^\star = \arg\max_{e \in \mathcal{E}} \mathrm{CDR}(e)$; observe $y^\star$
  update outer weights from the observed Gaussian likelihood
  append $(e^\star, y^\star)$ to $\mathcal{D}$; re-encode context at the next step
return context and outer weights

comparison reuses the five existing *no-swap* ensemble checkpoints; within each training seed, every selector uses exactly the same checkpoint. We selected all 60 zero-delay worlds in the previously pinned 100-world ID-test bank before examining selector outcomes. Restricting this analysis to $d = 0$ makes the current one-step fork sensitive to the chosen action; the delayed-world evaluator does not preserve its pending-action queue.

Matched acquisition comparison. For each of five seeds and each zero-delay world, we compare random selection, local ensemble disagreement, global CDR with uniform query weights, and decision-weighted CDR. All use the same 32 candidate-action seed rule; the uncertainty gate is disabled. Identification budgets are $B \in \{0, 1, 2, 4, 8, 16, 32\}$. Held-out evaluation uses the same canonical reset state, 64 unseen actions per world/budget, and 16 common-random-number simulator repeats per action. The neural checkpoint is frozen during identification. This isolates selector behavior on the same predictor and evaluation states, although collected trajectories and contexts necessarily differ. The full four-selector results and provenance are in Appendix A.1.

Metrics and uncertainty. The prediction metric is one-step root-mean-square error (RMSE) of predicted state change and reward, computed across four state-change coordinates and reward in their native simulator units, then averaged over worlds within each seed. We emphasize the global-CDR versus local-disagreement endpoint at $B = 32$ because it is the main observed effect; this focus follows the initial sweep and is *not* a preregistered primary contrast. We also report normalized trapezoidal area under the RMSE–budget curve on $[0, 32]$ (AUC; lower is better). Means and sample standard deviations treat the five independently trained checkpoints as replicates. Comparisons use seed-paired differences and two-sided 95% Student-t intervals with four degrees of freedom; worlds are averaged within each seed. The intervals are descriptive, exploratory, and unadjusted for multiple comparisons. A crossed seed/world bootstrap and world-level effects provide sensitivity analyses in Appendix A.2; five trained checkpoints remain the limiting independent replication unit. Logged acquisition time covers candidate prediction and scoring, but not query generation, environment stepping, or evaluation. Random and local selectors skip unused CDR scoring in this sweep.

Historical diagnostics. The earlier 100-world acquisition sweep appears in Table 2, and fixed-state model/gate ablations in Table 3. The former used different checkpoints and trajectory-dependent evaluation states across methods. The latter includes delayed worlds and exposes a mechanism-swap target mismatch. Neither provides valid delayed-world planning regret. Faithful Plan2Explore- and OPAX-style baselines have not been run.

8 RESULTS

The matched zero-delay experiment gives a more interpretable acquisition comparison than the earlier trajectory-state sweep. All methods start from the same prediction at $B = 0$, share a checkpoint

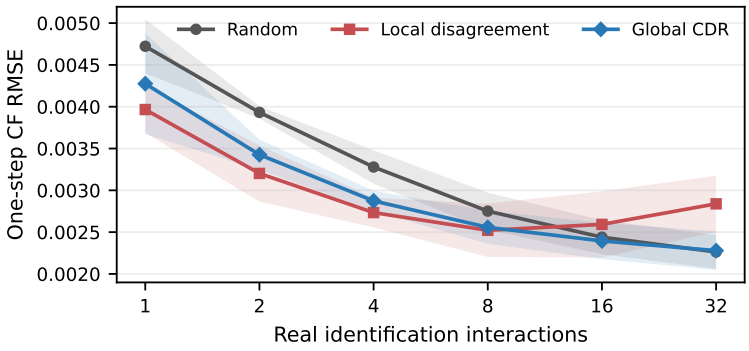

Figure 2: One-step held-out RMSE by real-interaction budget for the matched zero-delay comparison. Bands are two-sided 95% t intervals across five seed-level world means. $B = 0$ is omitted from the plot because all selectors share the same checkpoint and prediction. The horizontal positions show successive tested budgets, not linear budget spacing.

within each seed, and are evaluated at the same canonical state in each world. Table 1 reports five-seed means $\pm$ sample standard deviations; the complete four-selector curve is in Appendix A.1.

Table 1: Matched one-step prediction on 60 held-out zero-delay worlds (five training seeds; lower is better). AUC is normalized over budgets 0–32. Timing is the per-seed median candidate-prediction/scoring time, averaged across seeds; it excludes query generation and environment interaction.

Selector	RMSE at $B = 32 \downarrow$	RMSE AUC $\downarrow$	Score ms $\downarrow$
Random	0.002263 ± 0.000164	0.002776 ± 0.000129	3.58 ± 0.06
Local disagreement	0.002838 ± 0.000270	0.002825 ± 0.000229	3.63 ± 0.05
Global CDR	0.002279 ± 0.000185	0.002651 ± 0.000153	11.65 ± 0.05

At $B = 32$, global CDR has 19.7% lower RMSE than local disagreement (0.002279 versus 0.002838). The seed-paired global-minus-local difference is -0.000560, with a 95% interval of $[-0.000838, -0.000282]$. The budget curve reveals the tradeoff hidden by that endpoint: local disagreement has lower mean error at $B = 1, 2, 4$, while global CDR improves relative to it at larger budgets. Their AUC difference is -0.000174 with interval $[-0.000360, 0.000012]$, which includes zero. The endpoint effect has the same sign for all five training seeds and for 50/60 worlds after averaging over seeds. A crossed seed/world bootstrap gives a descriptive interval $[-0.000821, -0.000315]$; it does not create more independent checkpoints or correct for the post-hoc endpoint focus (Appendix A.2).

Against random selection, global CDR has lower normalized AUC (0.002651 versus 0.002776); the paired difference is -0.000125 with interval $[-0.000248, -0.000002]$. At $B = 32$ there is no reliable difference: the global-minus-random difference is $+0.000016$ with interval $[-0.000103, 0.000135]$. The marginal AUC interval is unadjusted across multiple comparisons and should be treated as exploratory. CDR also takes roughly $3.2\times$ the candidate-scoring time of random or local selection in the measured region (Table 1). Thus the evidence supports a late-budget advantage over local disagreement, not a universal advantage over random acquisition or a proven interaction–compute win.

Decision-weighted CDR does not improve reliably over uniform/global CDR: at $B = 32$ their paired RMSE difference is -0.000011 with interval $[-0.000026, 0.000004]$. Its full matched curve is in Appendix A.1.

Table 2 restores the full AEWM–baseline comparison. In this historical protocol, AEWM's one-step RMSE is 0.0207 ± 0.0097, versus 0.0140 ± 0.0002 for random and 0.0161 ± 0.0017 for global CDR. The apparent ordering is descriptive: selectors had separately trained checkpoints, visited different

Table 2: AEWM and baselines in the historical 100-world ID-test suite at $B = 32$ (five-seed means $\pm$ sample SD; lower is better). Counterfactual (CF) metrics use each run's recorded trajectory states; ECE pools budgets 0–32. Checkpoints differ between methods. Passive and MAX-like are selection-branch aliases of random and local disagreement, respectively, not independent algorithms.

Method	CF NLL	CF RMSE	ID NRMSE	ECE
Random	-2.329 ± 0.008	0.0140 ± 0.0002	0.1967 ± 0.0017	0.0210 ± 0.0002
Passive	-2.332 ± 0.008	0.0140 ± 0.0003	0.1988 ± 0.0040	0.0209 ± 0.0003
Local disagreement	-2.258 ± 0.039	0.0243 ± 0.0056	0.2487 ± 0.0113	0.0211 ± 0.0002
MAX-like	-2.258 ± 0.039	0.0243 ± 0.0056	0.2487 ± 0.0113	0.0211 ± 0.0002
VIME-like	-2.315 ± 0.012	0.0172 ± 0.0031	0.2104 ± 0.0045	0.0210 ± 0.0002
Global CDR	-2.326 ± 0.014	0.0161 ± 0.0017	0.2090 ± 0.0070	0.0210 ± 0.0002
AEWM (weighted CDR)	-2.272 ± 0.116	0.0207 ± 0.0097	0.2313 ± 0.0483	0.0209 ± 0.0002

Table 3: Historical fixed-state AEWM component test at $B = 32$ on 100 ID worlds (five seeds; lower is better). Unlike Table 1, these worlds include delayed actions. Selector/gate rows reuse the full checkpoint; no-swap and deterministic models are retrained. The large no-swap gain exposes a mismatched delayed-action swap target, not a benefit of the full architecture. ECE is a predictive calibration measure.

Variant	One-step CF RMSE↓	ECE↓
Full AEWM	0.0147 ± 0.0001	0.0221 ± 0.0002
No decision weights	0.0147 ± 0.0001	0.0221 ± 0.0002
Local uncertainty only	0.0172 ± 0.0005	0.0228 ± 0.0003
No mechanism swap	0.0029 ± 0.0002	0.0253 ± 0.0001
No uncertainty gate	0.0147 ± 0.0001	0.0221 ± 0.0002
Deterministic transition	0.0158 ± 0.0002	0.0707 ± 0.0001

states, and were evaluated at their own trajectory states. It therefore cannot replace the matched acquisition comparison in Table 1.

Table 3 shows which proposed AEWM components are doing work in the older system. Removing decision weights or the optional gate leaves the one-step endpoint nearly unchanged; local-only acquisition is worse on that historical protocol. The deterministic transition has worse calibration (0.0707 versus 0.0221 ECE), while removing mechanism-swap supervision sharply lowers RMSE (0.0029 versus 0.0147). The swap target incorrectly applies an action immediately in delayed worlds. The old one-step OOD planning-regret output is therefore withheld: its counterfactual fork also resets the pending-action queue. These diagnostics inform model design but do not demonstrate a control advantage.

Acquisition-score diagnostics. Supplementary audits of realized selected actions and model-internal counterfactual updates find limited fidelity of the fixed-prediction CDR score to context-reencoded candidate rankings (Appendix A.3– A.4). These diagnostics do not measure which candidate would best improve held-out prediction.

9 Discussion

AEWM asks which interaction makes a *set of future predictions* more dependable. The matched budget curve gives this idea a concrete, if limited, empirical test: global CDR is worse than local disagreement very early, then overtakes it at the 32-interaction endpoint. The $B = 32$ difference is consistent across the five checkpoints and most held-out worlds. But random acquisition reaches essentially the same endpoint with less measured scoring time; the modest AUC advantage over random is exploratory. Thus the result supports target-set acquisition as a useful direction, not a general interaction-efficiency improvement.

The component study separates attractive architecture from effective architecture. The stochastic variant has better ECE than the deterministic variant, although this comparison also removes the auxiliary denoiser; the optional gate and reward-gap weighting add little in this test. Removing swap supervision yields a large gain because the training target ignores delayed-action queues. This is an

actionable model-design result: a mechanism-aware loss must use a state sufficient for the mechanism being transferred. It is not a reason to claim that mechanism swapping is intrinsically harmful.

The acquisition audit isolates the other bottleneck. The virtual CDR update reweights members but freezes their query predictions; real data alter both weights and context embeddings. All-candidate contextual rankings agree weakly with the cheap virtual rankings, and rescoring with more Monte Carlo outcomes changes many top choices. A robust next instantiation of AEWM should test candidate utility against *actual held-out prediction gains*, account for context changes, and only then couple the acquisition rule to a planner. Our current evidence concerns one-step state prediction, not improved control.

10 Limitations

The evaluation has one simulator, 60 zero-delay worlds, five trained checkpoints, random one-step queries, native-unit aggregate RMSE, and no measured control return. The main $B = 32$ contrast was emphasized after examining the sweep; intervals are exploratory and unadjusted. A matched per-component error breakdown is not reported here, and an independent second-environment replication is unavailable. No faithful matched OPAX-, VIME-, or Plan2Explore-style baseline was run, so superiority to established active identification cannot be inferred.

CDR depends on ensemble diversity and likelihood quality, neither of which is guaranteed by independently initialized members sharing data. Outer weights are not a physical-mechanism posterior. The supplementary all-candidate audit finds weak agreement with a context-reencoded hypothetical score, but does not measure real-outcome ranking or improvement in held-out error. The preliminary no-swap variance–error correlations are weak. The decision-weighting heuristic can ignore underestimated actions. Score-region timing is only a partial compute measure.

The one-step learned state and evaluator omit pending action queues; the mechanism-swap target applies actions immediately even for delayed mechanisms. These defects invalidate delayed-world planning-regret evidence. Pixel observations, combinatorial mechanisms, queue-preserving multi-step planning, and decision-conditioned query generation are outside the completed study.

11 Conclusion

AEWM reframes active world modeling around a practical question: which real interaction will make the counterfactual predictions needed for subsequent decisions more trustworthy? Its CDR instantiation turns a query-set variance surrogate into a batched ensemble acquisition score. In a matched hidden-physics test, CDR improves late-budget one-step prediction over local disagreement, while matching random acquisition at the endpoint. Historical ablations and score audits expose two clear engineering targets: queue-consistent mechanism supervision and an acquisition update that tracks context-dependent predictions. These results establish a measured starting point for decision-relevant active world models; validating the full planner-conditioned loop is the next step.

Reproducibility Statement

Anonymous code and stage-by-stage instructions are available at `https://anonymous.4open.science/r/AEWM-2FEF/`. The accompanying `reproduce.md` covers environment setup, deterministic data generation, training, evaluation, and aggregation for the reported experiments.

AI Use Statement

Generative AI tools assisted with conceptual framing, experiment and code development, literature identification, interpretation of results, and drafting and editing this manuscript. Reported numerical results were checked against the available experiment artifacts. The authors take responsibility for the final content, code, and submission materials.

REFERENCES

Zaheer Abbas, Samuel Sokota, Erin Talvitie, and Martha White. Selective dyna-style planning under limited model capacity. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1–10, 2020. URL https://proceedings.mlr.press/v119/abbas20a.html.

Jan Achterhold and Joerg Stueckler. Explore the context: Optimal data collection for context-conditional dynamics models. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3529–3537, 2021. URL https://proceedings.mlr.press/v130/achterhold21a.html.

Dilip Arumugam and Benjamin Van Roy. The value of information when deciding what to learn. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/517da335fd0ec2f4a25ea139d5494163-Abstract.html.

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented bayesian active learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7331–7348, 2023. URL https://proceedings.mlr.press/v206/bickfordsmith23a.html.

Mona Buisson-Fenet, Friedrich Solowjow, and Sebastian Trimpe. Actively learning gaussian process dynamics. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 5–15, 2020. URL https://proceedings.mlr.press/v120/buisson-fenet20a.html.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/3de568f8597b94bda53149c7d7f5958c-Abstract.html.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996. doi: 10.1613/jair.295. URL https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume4/cohn96a-html/statmodels.html.

Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/a36b598abb934e4528412e5a2127b931-Abstract.html.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1184–1193, 2018. URL https://proceedings.mlr.press/v80/depeweg18a.html.

Louis Filstroff, Iiris Sundin, Petrus Mikkola, Aleksei Tiulpin, Juuso Kylmäoja, and Samuel Kaski. Targeted active learning for bayesian decision-making. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=KcPjuiMqmm.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual_2020/poster_S1lOTC4tDS.html.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/abd815286ba1007abfbb8415b83ae2cf-Abstract.html.

Jonas Hübotter, Bhavya Sukhija, Lenart Treven, Yarden As, and Andreas Krause. Transductive active learning: Theory and applications. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e17fe6fe9990fffb637b42c98c005515-Abstract-Conference.html.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5faf461eff3099671ad63c6f3f094f7f-Abstract.html.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804, 2018. URL https://proceedings.mlr.press/v80/kuleshov18a.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Dennis V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992. doi: 10.1162/neco.1992.4.4.590. URL https://authors.library.caltech.edu/records/efefp-2j353.

Suraj Nair, Silvio Savarese, and Chelsea Finn. Goal-aware prediction: Learning to model what matters. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7207–7219, 2020. URL https://proceedings.mlr.press/v119/nair20a.html.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019. URL https://proceedings.mlr.press/v97/pathak19a.html.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5331–5340, 2019. URL https://proceedings.mlr.press/v97/rakelly19a.html.

Marc Rigter, Minqi Jiang, and Ingmar Posner. Reward-free curricula for training robust world models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0a2b3e9107efed3d361b29f300a903ff-Abstract-Conference.html.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/90720a2fcc41f9332e6a1558da327089-Abstract.html.

Cansu Sancaktar, Sebastian Blaes, and Georg Martius. Curious exploration via structured world models yields zero-shot object manipulation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/98ecdc722006c2959babbdbdeb22eb75-Abstract.html.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 2020. URL https://proceedings.mlr.press/v119/sekar20a.html.

Tim Seyde, Wilko Schwarting, Sertac Karaman, and Daniela Rus. Learning to plan optimistically: Uncertainty-guided deep exploration via latent model ensembles. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1156–1167, 2022. URL https://proceedings.mlr.press/v164/seyde22b.html.

Pranav Shyam, Wojciech Jaskowski, and Faustino Gomez. Model-based active exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019. URL https://proceedings.mlr.press/v97/shyam19a.html.

Bhavya Sukhija, Lenart Treven, Cansu Sancaktar, Sebastian Blaes, Stelian Coros, and Andreas Krause. Optimistic active exploration of dynamical systems. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/77b5aaf2826c95c98e5eb4ab830073de-Abstract-Conference.html.

Annie Xie, Logan Mondal Bhamidipaty, Evan Zheran Liu, Joey Hong, Sergey Levine, and Chelsea Finn. Learning to explore in POMDPs with informational rewards. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54414–54429, 2024. URL https://proceedings.mlr.press/v235/xie24a.html.

Yingchen Xu, Jack Parker-Holder, Aldo Pacchiano, Philip Ball, Oleh Rybkin, Stephen Roberts, Tim Rocktaschel, and Edward Grefenstette. Learning general world models in a handful of reward-free deployments. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ab6a2c6ee757afe43882121281f6065c-Abstract-Conference.html.

Shijie Zhong, Wanggang Shen, Tommie Catanach, and Xun Huan. Goal-oriented bayesian optimal experimental design for nonlinear models using markov chain monte carlo. *SIAM/ASA Journal on Uncertainty Quantification*, 14(1):19–47, 2026. doi: 10.1137/24M1649344. URL https://doi.org/10.1137/24M1649344.

A Appendix: Implementation and Audit

A.1 Complete matched zero-delay selector comparison

Table 4 adds the decision-weighted variant to the matched comparison in the main text. Table 5 gives all tested budgets, and Table 4 gives score-region timings. The accompanying code artifact includes the aggregation script and `reproduce.md`. The aggregation verifies 60 common world IDs, seven budgets, five shared checkpoint hashes across selectors, canonical evaluation states, and matching evaluation counts. Negative paired differences indicate lower error for the first selector.

Decision-weighted minus global CDR has a paired RMSE difference of -0.000011 at $B = 32$ (95% interval $[-0.000026, 0.000004]$) and -0.000004 in AUC (interval $[-0.000011, 0.000003]$). There is no reliable incremental benefit from the value-gap weights in this experiment.

Table 4: All four selectors on the same 60 zero-delay worlds and no-swap checkpoint within each seed. Values are means ± sample standard deviations over five seeds. AUC is normalized over budgets 0–32.

Selector	RMSE at $B = 32$	RMSE AUC	Score ms
Random	0.002263 ± 0.000164	0.002776 ± 0.000129	3.58 ± 0.06
Local disagreement	0.002838 ± 0.000270	0.002825 ± 0.000229	3.63 ± 0.05
Global CDR	0.002279 ± 0.000185	0.002651 ± 0.000153	11.65 ± 0.05
Decision-weighted CDR	0.002267 ± 0.000184	0.002647 ± 0.000154	11.66 ± 0.07

Table 5: Complete matched one-step RMSE budget curve (five-seed means). The same 60 worlds and checkpoint within each seed are used across selectors.

B	Random	Local	Global CDR	Weighted CDR
0	0.008969	0.008969	0.008969	0.008969
1	0.004723	0.003966	0.004275	0.004274
2	0.003932	0.003202	0.003426	0.003444
4	0.003279	0.002734	0.002874	0.002874
8	0.002751	0.002522	0.002557	0.002560
16	0.002440	0.002593	0.002395	0.002387
32	0.002263	0.002838	0.002279	0.002267

The secondary value weighting uses the ensemble's predicted mean reward $\hat{V}(q)$ and $\hat{V}^\star = \max_{q \in \mathcal{Q}} \hat{V}(q)$:

$$w(q) = \frac{\exp[-\beta(\hat{V}^\star - \hat{V}(q))]}{\sum_{q' \in \mathcal{Q}} \exp[-\beta(\hat{V}^\star - \hat{V}(q'))]}, \qquad \beta = 5. \tag{13}$$

This heuristic can suppress a truly good action that the current predictor undervalues; the small weighted-versus-uniform difference above offers no evidence that this form improves identification.

A.2 Seed, world, and calibration sensitivity

We reconstructed the seed-by-world matched endpoint and AUC differences from the raw evaluator records. Table 6 reports the five seed-level differences for the main endpoint and the fraction of worlds with lower seed-mean error. The 20,000-draw crossed bootstrap resamples training seeds and worlds separately with replacement, preserving the matched selector comparison within each selected seed/world cell. Its percentile intervals reflect heterogeneity across both axes but remain exploratory with only five independent training runs. None of the analyses was preregistered or corrected for the multiple examined contrasts and budgets.

Table 6: Global CDR minus comparator at $B = 32$; negative is better for CDR. Seed entries average the same 60 worlds. World fraction counts negative differences after averaging over seeds.

Comparator	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4	World fraction	Crossed 95% CI
Local	-0.000487	-0.000209	-0.000665	-0.000651	-0.000788	50/60	$[-0.000821, -0.000315]$
Random	$+0.000087$	$+0.000054$	-0.000151	$+0.000027$	$+0.000062$	32/60	$[-0.000109, +0.000137]$

At $B = 32$ we also compared world-level ensemble reward variance with one-step RMSE using Spearman correlation within each seed. The seed-mean correlations (sample SD) are 0.094 ± 0.138 for random, 0.198 ± 0.262 for local, 0.178 ± 0.101 for global CDR, and 0.167 ± 0.133 for weighted CDR. These weak associations are not evidence of calibrated epistemic uncertainty or of disagreement as a reliable error predictor. The accompanying code artifact regenerates the per-seed values and evaluator ECE.

A.3 Selected-action virtual-to-real update audit

For each of five global-CDR runs, we reconstructed the first two executed interactions on all 60 matched worlds. At each pre-action state we regenerated the logged 32 one-step reward queries,

verified their pre-action disagreement against the acquisition trace, and read the selected candidate's logged expected post-update disagreement. We then measured disagreement on the *same pre-action state and queries* after the observed transition, first changing only outer weights and then re-encoding the enlarged context with the new weights. All neural parameters stayed frozen. This compares the structural approximation in Eq. 12; it does not evaluate a new post-action state, held-out prediction error, or unexecuted candidates.

Table 7: Spearman correlation, across 60 worlds within each seed, between the selected action's predicted reduction and realized reduction after context re-encoding. The fall at step 2 shows that first-step agreement does not establish stable score fidelity.

	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4	Mean $\pm$ SD
Step 1	0.761	0.725	0.662	0.293	0.765	0.641 ± 0.199
Step 2	0.185	0.334	0.116	0.202	0.371	0.242 ± 0.107

Across the 600 selected actions, 19.5% have negative realized reduction after context re-encoding. A pooled two-step correlation would be confounded by the much larger disagreement and reduction at step 1, so we report within-step values. The weight-only and re-encoded realized reductions also differ, as expected when appending context changes each member's prediction. The accompanying code artifact regenerates this audit from the matched runs. The next subsection compares all candidate rankings under a hypothetical context update; real-outcome held-out error remains unmeasured.

A.4 All-candidate score fidelity and Monte Carlo stability

On the same five global-CDR checkpoints, we selected a deterministic 12-world subset of the 60 matched zero-delay worlds and reconstructed the first two identification steps (120 world–seed–step snapshots). For each snapshot, the logged candidate seed reproduces all 32 actions, the original selected top-scoring action, and its expected virtual disagreement. Checkpoint, run-manifest, and acquisition-trace hashes agree with the source runs. For each candidate we reuse its 16 sampled hypothetical outcomes twice: the original score changes only outer member weights, whereas the contextual score also appends the hypothetical transition and re-encodes every member's mechanism context. Both scores use the same pre-action state, query actions, and uniform query weights. This comparison isolates a model-internal approximation gap; neither score is a real-outcome or held-out-error oracle.

Table 8: Candidate-score audit on 12 worlds per seed, five seeds, and 32 candidates per snapshot. Means $\pm$ sample SD are across seed-level means. Top-5 overlap is the fraction of five candidates shared. The $S = 256$ weight-only Monte Carlo score is a higher-sample reference, not ground truth; the $S = 16$ and $S = 64$ rows average four independent rescoring draws per snapshot.

Metric	Step 1	Step 2	Both steps
Virtual vs. contextual Spearman ρ	0.379 ± 0.348	0.157 ± 0.048	0.268 ± 0.164
Virtual vs. contextual top-1 agreement	0.267 ± 0.208	0.083 ± 0.059	0.175 ± 0.080
Virtual vs. contextual top-5 overlap	0.390 ± 0.178	0.233 ± 0.039	0.312 ± 0.084
$S = 16$ vs. $S = 256$ top-1 agreement	0.188 ± 0.077	0.083 ± 0.033	0.135 ± 0.047
$S = 64$ vs. $S = 256$ top-1 agreement	0.388 ± 0.123	0.138 ± 0.032	0.263 ± 0.069

The contextual top choice agrees with the original virtual score in 16/60 step-1 and 5/60 step-2 snapshots (21/120 overall). Agreement is low despite using identical hypothetical outcomes, so resampling noise alone cannot explain the virtual-to-context discrepancy. Separately, increasing Monte Carlo outcomes from 16 to 64 improves but does not stabilize the virtual top choice relative to $S = 256$. Before clipping, the estimated reduction is negative for 21.8% of candidates under the cheap score and 20.6% under the contextual score, averaging over snapshots; clipping did not change the top-1 choice in the independent $S = 16/64$ rescoring trials. These are finite-sample diagnostics rather than estimates of true expected reduction. The accompanying code artifact regenerates the per-seed records and summary. The mean contextual score gap between its top choice and the virtual top choice is 1.92×10^{-7} in disagreement units, but this gap is not calibrated to one-step prediction

error. A real-outcome, held-out-error comparison is still needed to determine whether these rank changes affect acquisition utility.

A.5 Historical trajectory sweep and model ablations

Table 2 in the main text preserves the earlier acquisition sweep, including the AEWM and baseline rows. Its selectors used separately trained checkpoints and evaluated at states reached by their own trajectories, so its cross-method and cross-budget comparisons are confounded by evaluation state. Historical `passive` and `MAX-like` runs use the same selection branches as random and local disagreement, respectively, and are not independent baselines.

Table 3 in the main text preserves fixed-canonical-state ablations on the original 100 ID worlds, including delayed worlds. Selector/gate variants reuse the full checkpoint; no-swap and deterministic variants are retrained. Removing swap reduces RMSE from 0.0147 to 0.0029 (paired difference -0.011786, 95% interval $[-0.011919, -0.011653]$), concentrated in delayed worlds. The implemented swap target applies the current action immediately regardless of the sampled mechanism's delay; hence this is evidence of a training-target mismatch, not a general rejection of swap supervision. The gate selects a direct candidate action on only 2.1% of full-variant steps; uniform query weights and gate removal have nearly no effect on the one-step endpoint.

The previous one-step OOD planning-regret output is withheld. All 100 selected extrapolative worlds have delay one or two, while the evaluator resets the pending-action queue on every fork. The evaluated action therefore does not change the next physical state; only immediate action cost changes. A queue-preserving multi-step evaluation is required before claiming any planning benefit. As a separate diagnostic on the historical full variant, the seed-mean Spearman correlation between world-budget-level ensemble variance and one-step RMSE is -0.046 ± 0.095 (sample SD), which does not support calibrated epistemic disagreement.

A.6 What the variance identity assumes

Let $X_q = \mu_q(M)$ for a *fixed* query and a fixed conditional prediction map. Conditional on $\mathcal{D}$ and the chosen experiment e, the law of total variance yields

$$\mathrm{Var}(X_q \mid \mathcal{D}, e) = \mathbb{E}_Y[\mathrm{Var}(X_q \mid \mathcal{D}, e, Y)] + \mathrm{Var}_Y\big(\mathbb{E}[X_q \mid \mathcal{D}, e, Y]\big).$$

Choosing e does not itself alter the current belief, so $\mathrm{Var}(X_q \mid \mathcal{D}, e) = \mathrm{Var}(X_q \mid \mathcal{D})$ under non-anticipatory Bayesian design. This proves Eq. 9 for a coherent posterior update. It does not prove that a learned model's prediction error will improve after data collection, that a finite particle approximation is unbiased, or that a particular realized observation reduces variance. The implemented CDR update holds the neural predictor and query weights fixed; after a real observation it updates particle weights and re-encodes the context, not neural parameters.

The corresponding entropy-form target is $I(G_{\mathcal{Q}}(M); Y_e \mid \mathcal{D}, e)$, with $G_{\mathcal{Q}}(M) = (\mu_q(M))_{q \in \mathcal{Q}}$. This distinguishes query-relevant mechanisms rather than all mechanism coordinates. It is a conceptual alternative, not an estimated metric in our experiments.

For completeness, suppose a finite action set has true values $V(a)$ and estimates $\hat{V}(a)$, with calibrated radii satisfying $|V(a) - \hat{V}(a)| \le b\sigma(a)$ simultaneously for every action. If $\hat{a} = \arg\max_a \hat{V}(a)$ and $a^\star = \arg\max_a V(a)$, then

$$V(a^\star) - V(\hat{a}) \le b[\sigma(a^\star) + \sigma(\hat{a})] \le 2b \max_a \sigma(a). \tag{14}$$

This elementary inequality only motivates attention to decision-relevant uncertainty. It does not derive Eq. 13, establish that our ensemble radii are calibrated, or imply a planning-regret improvement.

A.7 Pairwise decision uncertainty is untested

For two actions, uncertainty of their *difference* includes cross-action covariance:

$$\mathrm{Var}[V_M(a_i) - V_M(a_j)] = \mathrm{Var}[V_M(a_i)] + \mathrm{Var}[V_M(a_j)] - 2\,\mathrm{Cov}[V_M(a_i), V_M(a_j)].$$

Equation 13 ignores this covariance and can down-weight an underestimated optimal action. Pairwise-margin or expected-decision-loss acquisition is a future comparator; no result for it is claimed here.

A.8 SIMULATOR, SPLITS, AND EVALUATION

The simulator state is (x, y, v_x, v_y) and each action is a two-dimensional force clipped to $[-1, 1]^2$. It uses time step 0.08, additive velocity noise with standard deviation 0.01, and horizon 64. The reward is negative distance to $(1, 1)$ minus $0.01\|a\|_2^2$. The fixed mechanism comprises mass scale, friction scale, damping scale, actuator gain, and action delay $d \in \{0, 1, 2\}$. The four continuous training/ID-test ranges are $[0.65, 1.45]$, $[0.45, 1.55]$, $[0.55, 1.45]$, and $[0.65, 1.40]$, respectively; the extrapolative ranges are disjoint intervals specified in `aewm/envs/continuous_hidden.py`. Training and ID-test delay probabilities are $(0.55, 0.30, 0.15)$; extrapolative worlds use $(0, 0.35, 0.65)$.

The split manifest has 10,000 training, 1,000 validation, 2,000 ID-test, 1,000 compositional-OOD, and 1,000 extrapolative-OOD worlds. The historical evaluation uses a pinned 100-world subset of each of the ID-test and extrapolative-OOD splits. Bootstrap collection takes two random-action episodes per training world and produced 1,279,874 transitions. The train script further reserves 10% of the bootstrap worlds for checkpoint validation. Mechanism vectors are stored only for simulator pairing/evaluator metadata, not supplied as model observations. Each identification transition contributes an 11-dimensional feature $(s, a, s' - s, r)$. The matched zero-delay study uses 60 preselected ID worlds, 64 held-out actions per world/budget, and 16 simulator-noise repeats per action. The historical study used 128 actions and 32 repeats. Table 2 evaluates at recorded trajectory states; Tables 3 and 1 use canonical reset states. Their different world/state/evaluation protocols make numeric values non-interchangeable.

A.9 MODEL AND TRAINING CONFIGURATION

Each of five particles contains a structured-history token MLP, a two-layer four-head Transformer context encoder, a 16-dimensional diagonal-Gaussian mechanism posterior, and a conditional MLP mean/standard-deviation head for five outputs $(s' - s, r)$. Hidden width is 256; the full ensemble has 7,399,915 trainable parameters (1,479,983 per particle). The stochastic variant also has an auxiliary 20-step denoiser, but Gaussian heads are used for acquisition and evaluation. Ablation and matched-checkpoint training uses 100 epochs, batch size 12,288, AdamW learning rate 3×10^{-4} and weight decay 10^{-5}, with bfloat16 autocast on the reported CUDA runs. The historical sweep used batch size 4,096 for all random seeds and passive seeds 0–2, whereas its other final run configurations used 12,288; some historical runs resumed after a batch-size change. This is another reason not to interpret the historical rows as a controlled acquisition comparison. The loss weights are 1.0 for swap in the full variant, 0.1 for variance calibration, 0.1 for context-half consistency, and 0.001 for posterior KL. The no-swap variant sets the first weight to zero; the deterministic variant replaces the standard-deviation head with fixed standard deviation 0.1 and removes the denoiser. All reported checkpoints are selected by validation loss. The training loss omits the Gaussian normalization constant, while the evaluation NLL includes it; either can be negative for sufficiently concentrated continuous densities. Only values computed with the same metric convention are compared.

A.10 ACQUISITION CONFIGURATION AND COMPUTE CAVEAT

The tested CDR selector uses $K = 5$ ensemble members, $L = 32$ random one-step reward queries, $N = 32$ uniformly sampled candidate actions, $S = 16$ hypothetical outcomes per candidate, unit interaction cost, $\gamma = 1$, and $\varepsilon = 10^{-3}$. The primary global CDR uses uniform query weights; the secondary decision-weighted variant uses $\beta = 5$. Query and candidate predictions are batched; virtual updates are NumPy operations over member likelihoods. The optional gate uses a top-two reward margin greater than 1.96 times the sum of member-weighted standard deviations; it is not empirically calibrated as a confidence level and was disabled in the matched study. The configured 16-candidate prefilter and 8-step CEM planner are not called by the identification runner. Unlike the historical runner, the matched study skips unused CDR scoring for random and local selectors. Its logged time includes candidate prediction/scoring but excludes query generation, environment stepping, and evaluation; it is therefore a partial compute comparison, not total wall-clock interaction cost.

A.11 STATISTICAL UNIT AND UNRESOLVED TESTS

Tables report means and sample standard deviations over five training seeds. Paired differences average over worlds within each seed and use Student-t intervals with four degrees of freedom; seed–world observations are not treated as independent training replicates. The emphasis on the global-versus-local $B = 32$ contrast was chosen after inspecting the initial results, and the crossed bootstrap is a sensitivity analysis rather than a remedy for that selection. The historical budget-curve artifact uses pooled world-level intervals and contains one-step action regret, so it is excluded from the present paper. The matched acquisition comparison now addresses one-step prediction on zero-delay worlds. The central missing experiments are a corrected swap target and state representation carrying the action queue, queue-preserving multi-step planning, faithful external active-identification baselines, and a second environment. None of these is imputed from current one-step results.