WHICH VALUE HEAD MISSED? OUTCOME-RESOLVED PROPER-SCORE EXCESS FOR WORLD-MODEL CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

A world-model forecast can be uncertain without a consequential outcome, while a confident reward, cost, or termination forecast can resolve with unexpectedly large proper loss. The environment response therefore exposes a learning signal unavailable at forecast time: which value head realized loss above that head's self-expected score. We define outcome-resolved value score (ORVS) as the positive, nominally standardized proper-loss residual and introduce Outcome-Resolved Score Control (ORSC), which propagates immutable collection-time labels through a terminal-masked continuation critic into guarded actor updates. Under a common late-update path, ORVS raises shifted return over uncertainty by $43.9\,[30.5, 57.3]$ points with similar nominal return; a separate ranking analysis finds that its top 10% contains 0.61 ± 0.04 of subsequent harmful continuations, versus 0.49 ± 0.03 for predicted cost and 0.42 ± 0.04 for uncertainty. The control effect recurs with a TD-MPC-style agent ($36.6\,[23.4, 49.8]$ shifted-return points) and unseen visual distractions (0.744 ± 0.022 retention), while SafetyPoint and SafetyCar cost rates reach 2.43 and 3.07 per 1,000 steps.

1 INTRODUCTION

An ensemble can disagree strongly about a transition whose reward remains intact, while a low predicted-cost transition can end in collision or premature termination. Both cases are ambiguous before the environment responds; afterward they differ in which value head was inaccurate and by how much its realized scoring loss exceeded the forecast's own expected difficulty. This temporal asymmetry motivates a concrete question for world-model control: can resolved value-head scores provide more useful supervision for later policy updates than forecast-time ambiguity alone?

We operationalize the resolved signal with outcome-resolved value score (ORVS). For the reward, cost, and termination heads used in value accounting, ORVS subtracts each predictive distribution's self-expected proper loss, normalizes by a nominal head-wise scale, and rectifies the residual. The result identifies an unusually difficult realized outcome for the issuing forecast. Its sign is a score comparison rather than a utility label: unexpectedly high reward and unexpectedly safe binary outcomes can also produce positive residuals. This distinction makes future control relevance an empirical question, tested below with simulator continuations rather than assumed from the scoring rule.

Outcome-Resolved Score Control (ORSC) turns ORVS into an actor-level learning interface. A snapshot emits forecasts before observing the transition outcome; that outcome is scored once and stored with the snapshot identifier; a terminal-masked critic learns the continuation of the immutable label; and imagined actor updates trade return against predicted score exposure. Return and visitation guards are designed to preserve task utility and state coverage. This chronology prevents the outcome from recalibrating the forecast against which it is scored and keeps outcome-time supervision separate from evaluation feedback.

The experiments ask whether the implemented ORVS substitution improves control, whether its ranking enriches for later simulator-defined harm, and whether the effect recurs across backbones and observations. Cross-family baselines provide operating context, while uncertainty, predicted cost, unsigned surprise, and a zero-signal wrapper use the same continuation-critic and guard path. ORVS im-

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.

proves shifted return over uncertainty by $43.9\,[30.5, 57.3]$ points and retention by $0.059\,[0.039, 0.079]$ in the Dreamer-style comparison; the shifted-return difference is $36.6\,[23.4, 49.8]$ points with TD-MPC. At 10% coverage, it captures 0.61 ± 0.04 of harmful continuations from a common state pool, versus 0.42 ± 0.04 for uncertainty.

Contributions. The technical contribution is a snapshot-bound, affine-reward-scale-normalized value-head score and a guarded control path that carries its realized residual into differentiable imagined updates. The empirical contribution is a controlled selector-substitution study with paired effects, independent simulator-continuation labels, head and weight analyses, a second latent-control backbone, and unseen visual distractions. Together they show that post-response value-head scoring is a productive policy-learning interface in the evaluated world-model regimes.

2 Forecast- and Outcome-Time Value Scores

We consider an episodic controlled process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, r, c, d, \rho_0, \gamma)$, where d_t distinguishes true termination from time-limit truncation, and a world model with predictive distributions $\hat{P}_\theta^{(k)}(\cdot \mid s_t, a_t)$ for three value heads,

$$\mathcal{Y} = \{r_t, c_t, \mathbb{1}[\mathrm{term}_t]\}. \tag{1}$$

Safety Gymnasium supplies its native binary simulator cost. For DMControl and Distracting Control, c_t is a predeclared binary adverse-outcome cost: reward below the task's 10th nominal-training percentile or non-time-limit termination. The percentile is fixed before model fitting and is not recomputed under shift. Let $\mathcal{F}_t^-$ denote the information available after the model emits these distributions but before the environment reveals y_t, and let

$$\mathcal{F}_t^+ = \sigma(\mathcal{F}_t^-, y_t) \tag{2}$$

denote the resolved information set. Uncertainty and predicted-cost selectors are $\mathcal{F}_t^-$-measurable. A realized loss residual is $\mathcal{F}_t^+$-measurable and, whenever its positive excess has nonzero conditional variance, cannot be recovered from $\mathcal{F}_t^-$. Equation (2) therefore separates a distribution over possible outcomes from knowledge of which value-head score was realized.

For head k, let $\ell_k(\hat{P}, y)$ be a lower-is-better strictly proper loss and define its self-expected loss

$$\bar{\ell}_t^{(k)} = \mathbb{E}_{Y \sim \hat{P}_\theta^{(k)}(\cdot \mid s_t, a_t)}[\ell_k(\hat{P}_\theta^{(k)}, Y)]. \tag{3}$$

To place heterogeneous losses on a dimensionless scale, the nominal calibration split fixes

$$s_k = \left(\mathbb{E}_{(s,a,y) \sim \mathcal{D}_{\mathrm{cal}}} \left[\ell_k(\hat{P}, y) - \bar{\ell}^{(k)} \right]^2 \right)^{1/2} + 10^{-8}, \qquad Z_t^{(k)} = \frac{\ell_k(\hat{P}_t^{(k)}, y_t^{(k)}) - \bar{\ell}_t^{(k)}}{s_k}. \tag{4}$$

The scale is computed separately for each task and active head, then frozen before selector-budget selection and all shifted evaluation.

Definition 1 (Outcome-resolved value score)**.** *After observing $y_t^{(k)}$, the head-wise score residual and aggregate selector are*

$$R_t^{(k)} = \left[Z_t^{(k)} \right]_+, \qquad R_t = \sum_{k=1}^{K} w_k R_t^{(k)}. \tag{5}$$

For the categorical reward head with ordered support $b_1 < \cdots < b_M$ and CDF values F_m, the implementation uses

$$\ell_r(F, y) = \sum_{m=1}^{M-1} (b_{m+1} - b_m) \left(F_m - \mathbb{1}[y \le b_m] \right)^2, \quad \bar{\ell}_r = \sum_{m=1}^{M-1} (b_{m+1} - b_m) F_m (1 - F_m). \tag{6}$$

Binary cost and termination use Bernoulli negative log loss, whose self-expected loss is binary entropy, with probabilities clipped as specified in Appendix A. Equation (5) is one-sided: a head activates when its realized proper loss exceeds its own self-expected loss. Positive refers to the centered score residual, not utility direction; high reward, absence of a predicted cost, or survival under high

predicted termination probability can also activate a head. ORVS is thus an outcome-resolved score-excess signal whose relation to later control outcomes is measured empirically. Uniform $w_k = 1/3$ is fixed across suites; Section 7 tests reward-, cost-, and termination-heavy alternatives.

A model-implied forecast-time analogue averages the same rectified residual over predicted outcomes:

$$\widetilde{R}_t^- := \sum_{k=1}^{K} w_k \mathbb{E}_{Y \sim \hat{P}_t^{(k)}} \left[\frac{\ell_k(\hat{P}_t^{(k)}, Y) - \bar{\ell}_t^{(k)}}{s_k} \right]_+, \tag{7}$$

which averages over outcomes rather than identifying the one that occurred. It equals $\mathbb{E}[R_t \mid \mathcal{F}_t^-]$ when each predictive marginal $\hat{P}_t^{(k)}$ equals the corresponding conditional outcome law; otherwise it remains a model-implied pre-outcome quantity. For any positive affine reward recoding $y' = ay + b$, CRPS, its self-expected value, and the RMS term in s_r scale by a when the support is transformed with the reward. Thus $Z_t^{(r)}$ is affine-scale invariant up to the 10^{-8} numerical stabilizer; binary-head losses are already dimensionless.

3 Outcome-Resolved Score Control

ORSC learns a score-continuation critic $C_\psi(s, a)$ with target

$$y_t^C = R_t + \gamma(1 - d_t) C_{\bar{\psi}}(s_{t+1}, a_{t+1}), \qquad a_{t+1} \sim \pi_\phi(\cdot \mid s_{t+1}). \tag{8}$$

Here d_t is the environment terminal flag, timeout truncations bootstrap, and the target network receives Polyak updates. The critic minimizes

$$\mathcal{L}_C(\psi) = \frac{1}{2} \mathbb{E}_{\mathcal{D}} \left[\left(C_\psi(s_t, a_t) - \mathrm{sg}(y_t^C) \right)^2 \right], \tag{9}$$

where sg stops gradients and every replayed R_t remains attached to the snapshot that produced its forecast.

The ORSC actor maximizes return while limiting predicted continuation score exposure and preserving the return of a frozen matched reference:

$$\max_\phi J(\pi_\phi) \quad \text{s.t.} \quad G_R(\pi_\phi) \le \epsilon_R, \qquad J(\pi_\phi) \ge J(\pi_{\mathrm{ref}}) - \delta, \tag{10}$$

where C^π denotes the discounted continuation of collection-time labels and

$$G_R(\pi; \mathcal{D}) = \mathbb{E}_{s_0 \sim \mathcal{D}, \pi} \left[\sum_{h=0}^{H-1} \alpha_h C^\pi(s_h, a_h) \right], \qquad \alpha_h = \frac{(1-\gamma)\gamma^h}{1 - \gamma^H}. \tag{11}$$

The labels depend on their stored forecast snapshots, so Eq. (11) is a replay-rooted score-exposure surrogate. It yields the saddle objective

$$\max_\phi \min_{\lambda, \mu \ge 0} \mathcal{L} = J(\pi_\phi) - \lambda(G_R - \epsilon_R) + \mu(J(\pi_\phi) - J(\pi_{\mathrm{ref}}) + \delta). \tag{12}$$

Each actor step starts B latent rollouts from replay states and imagines $H = 15$ steps. $\widehat{J}$ is the batch mean of normalized Dreamer λ-returns with trace parameter 0.95, and $\widehat{G}_R = B^{-1} \sum_b \sum_h \alpha_h C_\psi(s_{b,h}, a_{b,h})$ estimates Eq. (11) on the same imagined states and actions. The frozen 50-episode reference mean is mapped through the same random-to-reference normalization as $\widehat{J}$. The actor differentiates $(1 + \mu)\widehat{J} - \lambda \widehat{G}_R$ plus the backbone's unchanged entropy term through reparameterized actions and latent dynamics, while world-model, return-critic, and score-critic parameters are held fixed for this step. Projected dual descent gives

$$\lambda \leftarrow [\lambda + \eta_\lambda(\widehat{G}_R - \epsilon_R)]_+, \quad \mu \leftarrow [\mu + \eta_\mu(\widehat{J}_{\mathrm{ref}} - \delta - \widehat{J})]_+. \tag{13}$$

Returns are normalized by each task's nominal random-to-reference span before applying δ.

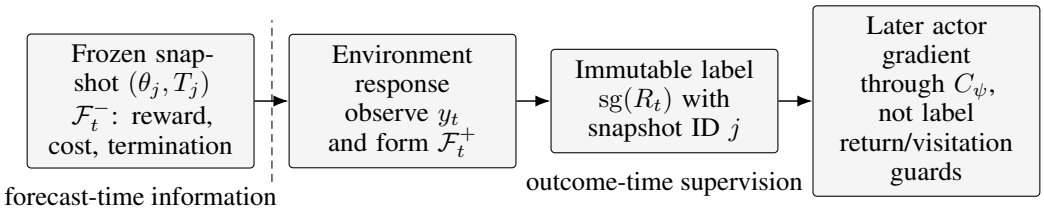

Figure 1: ORSC chronology and gradient boundary. A frozen snapshot issues value-head forecasts; the response creates a stop-gradient, snapshot-bound score; and later actor gradients pass through its learned continuation while model and label remain fixed.

Chronology. Before collection, a seed-fixed partition assigns each focal-policy interaction to training replay or a chronological calibration stream; assignments are shared across selector rows and included in the 2M-step focal budget. Snapshot (θ_j, T_j) emits every predictive distribution before the environment reveals y_t. Replay-assigned transitions are scored immediately and stored with immutable head-wise ORVS and snapshot identifier j; calibration-assigned transitions are scored by the same snapshot but never enter replay. Every 5,000 interactions, temperatures are refit on the latest 20,000 calibration transitions, and T_{j+1} applies prospectively. The world model and continuation critic then update from replay, followed by actor and projected dual updates. A separately trained 2M-step reference policy supplies the frozen nominal return target. If recent-window occupancy entropy falls more than τ below that reference, the positive update to λ is suppressed while μ continues to enforce the return margin; otherwise Eq. (13) applies unchanged. Evaluation freezes every learned parameter and supplies no training feedback.

Matched selector contract. Every row in the selector swaps uses the same active heads, collection protocol, calibration split, frozen scales, continuation-critic architecture and target in Eq. (8), return and visitation guards, actor estimator, 2M-step focal-policy budget, and update timing. The alternatives replace only R_t:

$$q_{\mathrm{ORVS}} = \sum_k w_k [Z_t^{(k)}]_+, \qquad q_{\mathrm{US}} = \sum_k w_k |Z_t^{(k)}|,$$
$$q_{\mathrm{U}} = \sum_k w_k \ \mathrm{JS}\left(\{\hat{P}_{t,m}^{(k)}\}_{m=1}^5\right), \qquad q_{\mathrm{C}} = \hat{P}_t^{(c)}(Y^{(c)} = 1). \tag{14}$$

q_{US} is outcome-resolved but symmetric; q_{U} is Jensen–Shannon disagreement among five prediction-ensemble members on the same value heads; and q_{C} is the ensemble-mean forecast probability of the binary cost event. ORVS and unsigned surprise use the corresponding ensemble-mean predictive distributions. Every selector row trains the same five-member heads, so the alternative definitions change access to shared outputs rather than predictive capacity. Each scalar is multiplied by one nominal-validation constant so its root-mean-square magnitude matches q_{ORVS} before training. This normalization establishes a common nominal amplitude unit; the estimand remains complete selector substitution under the shared $\epsilon_R = 0.12$, including each selector's induced activation and dual dynamics. "No extra selector" keeps the wrapper and guards but sets $q = 0$ and $\lambda = 0$, whereas "DreamerV3" is the native backbone without the wrapper.

Figure 1 exposes the two boundaries used by ORSC: outcome resolution creates the label, and only its learned continuation participates in a later actor gradient.

4 Relation to Forecast-Time Risk and Model Error

Calibration methods improve the forecast available in $\mathcal{F}_t^-$, ensembles quantify disagreement, and pessimistic or conservative objectives penalize unsupported model rollouts (Gneiting & Raftery, 2007; Dawid, 1982; Kuleshov et al., 2018; Malik et al., 2019; Chua et al., 2018; Buckman et al., 2018; Yu et al., 2020; Kidambi et al., 2020; Kumar et al., 2020; Frauenknecht et al., 2024). ORVS instead supplies the standardized loss of the outcome that occurred. Equation (7) is its model-implied pre-outcome average; ORVS selects one realized positive residual after the response.

Safe RL and safe world models optimize predicted costs, feasibility, or tail risk during planning and policy improvement (Tamar et al., 2015; Chow et al., 2018; Achiam et al., 2017; Huang et al., 2024; As et al., 2024; Huang et al., 2025). ORVS instead supplies a post-response value-head score to a later update. The predicted-cost row in Eq. (14) provides a forecast-time cost contrast inside that update path, while the cross-family methods contextualize task-level utility rather than serving as component-matched controls.

Dreamer and TD-MPC learn predictive latent dynamics without exposing a snapshot-bound realized head score as an actor variable (Hafner et al., 2020; 2021; 2025; Hansen et al., 2022). Value-aware model learning aligns model fit with downstream value, while prioritized replay and curiosity use TD error or surprise for sampling and exploration (Farahmand et al., 2017; Schaul et al., 2016; Pathak et al., 2017). ORSC changes the actor objective through the continuation of a rectified, value-head residual. Unsigned surprise tests rectification within the same centered heads, and the zero-signal row tests the wrapper path suggested by model-misspecification analyses (Pan et al., 2022; Eysenbach et al., 2022).

5 Evaluation Protocol

Questions. The evaluation tests three hypotheses: ORVS substitution improves shifted return and safety cost with similar nominal utility (H1); high ORVS ranks states with harmful subsequent simulator continuations (H2); and the selector effect recurs under a second latent-control backbone and held-out visual distractions (H3). Same-stage selector swaps provide the primary controlled comparison; algorithm-family baselines provide task-level operating points.

State-control suite. Walker-walk, Hopper-hop, and Cheetah-run from DMControl use gravity $0.7g$ and $1.3g$, a $2\times$ friction coefficient, and normalized-input Gaussian noise with $\sigma = 0.1$ (Tassa et al., 2018). Shifted return averages gravity directions within family, then gravity, friction, and noise within task. The binary cost-head target is the predeclared low-reward-or-termination event in Section 2; reported DMControl outcomes remain environment return. SafetyPointGoal-v0 and SafetyCarGoal-v0 use ten 1,000-step episodes per seed (Ji et al., 2023; Ray et al., 2019); cost per 1,000 steps is mean episodic native simulator cost.

Visual transport. Walker-walk, Cheetah-run, and Finger-spin use pixel observations from Distracting Control Suite under held-out medium color, background-video, and camera distractions (Stone et al., 2021). Clean validation selects hyperparameters; distraction frames are never used for tuning. Each task-distraction-seed cell contains 20 episodes, yielding 600 clean and 1,800 shifted episodes per selector.

Training, selection, and uncertainty. Each focal policy receives 2M environment interactions and uses ten independent seeds. A selector-wrapper run also uses a separately trained 2M-step reference policy; that cost is included in total accelerator-hours and the same reference is reused across selector comparisons for a task and seed. Calibration transitions are routed from focal collection into a chronological stream excluded from training replay, so they do not add interactions. ORSC and selector swaps train on nominal dynamics; contextual domain randomization alone observes the declared perturbation families during training. Dreamer-style comparisons share a latent backbone, and TD-MPC transport uses its corresponding latent-control backbone. SAC, PPO, PETS, contextual domain randomization, and CVaR-RL retain task-standard architectures (Haarnoja et al., 2018; Schulman et al., 2017; Chua et al., 2018; Benjamins et al., 2021; Tamar et al., 2015). Hyperparameter validation and test episodes are disjoint, and all settings are frozen before evaluation. Experiments use $8\times$A100-80GB GPUs with CUDA 12.4. Pooled statistics average episodes within condition, families within task, and tasks within seed before forming two-sided 95% t-intervals over matched seed aggregates; retention is the equally weighted task-wise ratio of shifted to nominal return.

Table 1: Same-stage swaps under one Dreamer-style backbone and Eq. (14). Retention and rates are hierarchical 10-seed aggregates; only the selector definition changes.

Selector	DM nominal	DM shifted	Retention	Point cost/1k	Car cost/1k
DreamerV3	674.7 ± 11.4	513.3 ± 15.8	0.747 ± 0.024	3.98 ± 0.49	4.81 ± 0.56
ORVS (ORSC)	670.3 ± 11.8	603.0 ± 13.6	0.890 ± 0.018	2.43 ± 0.32	3.07 ± 0.39
Uncertainty-only	668.9 ± 12.8	559.1 ± 14.7	0.831 ± 0.021	2.84 ± 0.37	3.62 ± 0.44
Predicted-cost-only	666.1 ± 13.4	547.8 ± 16.1	0.814 ± 0.023	2.58 ± 0.31	3.41 ± 0.40
Unsigned surprise	669.7 ± 12.7	566.4 ± 15.0	0.839 ± 0.022	2.97 ± 0.38	3.73 ± 0.45
No extra selector	672.1 ± 12.2	528.4 ± 17.0	0.771 ± 0.026	3.63 ± 0.45	4.29 ± 0.50

Table 2: Paired ORVS effects relative to each matched selector. The upper panel reports return effects; the lower panel reports safety effects and wins. Brackets are paired 95% intervals; d is the mean paired seed difference divided by its sample standard deviation. Wins count positive paired effects across nine task–shift-family cells and two safety tasks.

Comparator	Shifted return Δ	Retention Δ	d
Uncertainty	$43.9\,[30.5, 57.3]$	$0.059\,[0.039, 0.079]$	1.44
Predicted cost	$55.2\,[39.7, 70.7]$	$0.076\,[0.052, 0.100]$	1.52
Unsigned surprise	$36.6\,[22.9, 50.3]$	$0.051\,[0.031, 0.071]$	1.24
No selector	$74.6\,[57.8, 91.4]$	$0.119\,[0.091, 0.147]$	1.67

Comparator	Point reduction	Car reduction	Wins
Uncertainty	$0.41\,[0.16, 0.66]$	$0.55\,[0.24, 0.86]$	8/9; 2/2
Predicted cost	$0.15\,[0.03, 0.27]$	$0.34\,[0.10, 0.58]$	9/9; 2/2
Unsigned surprise	$0.54\,[0.27, 0.81]$	$0.66\,[0.34, 0.98]$	8/9; 2/2
No selector	$1.20\,[0.85, 1.55]$	$1.22\,[0.83, 1.61]$	9/9; 2/2

6 Results

6.1 Utility under Dynamics Shifts and Safety Constraints

The absolute task-level results appear in Appendix B. ORSC leads all three shifted DMControl tasks with similar nominal utility: its shifted mean is 603, versus 522 for contextual domain randomization and 513 for DreamerV3, and mean retention rises from 0.776 and 0.747 to 0.890. The family decomposition in Appendix I shows positive margins in all nine task–shift cells, including a smallest paired advantage of 17 $[4, 30]$ on Hopper-hop gravity.

On SafetyPointGoal-v0 and SafetyCarGoal-v0, ORSC attains cost rates of 2.43 and 3.07 per 1,000 steps at returns of 34.8 and 29.0. The selector-swap effects below provide the controlled safety comparison with matched seeds and a common late-update path.

6.2 Matched Selector Attribution

Table 1 holds the late-update path fixed and substitutes only the selector, estimating the effect of the implemented ORVS substitution.

Table 2 then uses matched seeds to quantify ORVS effects at the suite and task-family levels.

Within the common late-update path, ORVS improves retention and both safety rates against every tested alternative; all reported paired intervals exclude zero. Unsigned surprise shares the centered value heads but replaces positive-residual rectification by an absolute residual, so its lower return and higher safety cost identify rectification as a consequential part of the implemented selector contrast. The head-wise results in Appendix H show that full ORVS has the strongest joint operating point.

Table 3: Simulator-continuation ranking at 10% coverage and $H = 15$. Return deficit and event incidence are computed within selected states; capture is the fraction of all harmful events selected from the common pool.

Selector	Normalized return deficit	Event incidence	Event capture	AUPRC
ORVS	0.42 ± 0.02	0.47 ± 0.03	0.61 ± 0.04	0.566 ± 0.028
Uncertainty-only	0.30 ± 0.02	0.32 ± 0.03	0.42 ± 0.04	0.371 ± 0.026
Predicted-cost-only	0.34 ± 0.02	0.38 ± 0.03	0.49 ± 0.03	0.438 ± 0.025
Unsigned surprise	0.32 ± 0.02	0.35 ± 0.03	0.46 ± 0.04	0.402 ± 0.027
Uniform	0.17 ± 0.02	0.08 ± 0.01	0.10 ± 0.01	0.078 ± 0.009

6.3 Simulator-Continuation Ranking

The judge measures association with future harm on a common timeline. A frozen Dreamer-style evaluation checkpoint supplies one shifted transition pool, its forecast snapshots, and one policy π_{eval}; every selector is recomputed from the same head outputs and realized y_t. Candidates are nonterminal transitions with simulator-cloneable s_{t+1}, so the triggering outcome used to form ORVS lies outside the judge horizon. From each clone, π_{eval} and its matched reference π_{ref} receive common simulator randomness. With $G_H(\pi) = \sum_{h=0}^{H-1} \gamma^h r_{t+1+h}$, their normalized return deficit is

$$D_H(s_{t+1}) = \frac{[G_H(\pi_{\text{ref}}) - G_H(\pi_{\text{eval}})]_+}{J_{\text{ref}} - J_{\text{random}} + 10^{-8}}. \tag{15}$$

A harmful event is the union of D_H exceeding the task-specific 90th percentile fixed on nominal validation continuations, any native simulator-cost event in the π_{eval} continuation, or true termination before step H; time-limit truncation is not a terminal event. Simulator outcomes alone define the label. All selectors rank the same 45,000 states, disjoint from calibration and tuning. A preassigned state order resolves ties; aggregation averages within task and then across tasks within seed before pairing seeds and continuation randomness. At 10% coverage, each selector returns 4,500 states from a pool containing 3,450 positive events.

ORVS captures 12 percentage points more harmful events than predicted cost and 19 percentage points more than uncertainty at identical coverage. Across the 5/10/20% coverage and $H \in \{5/15/30\}$ grid in Appendix C, ORVS leads predicted cost in eight of nine operating points; predicted cost leads by one point at $H = 5$ and 5% coverage. This held-out association is strongest once selection or continuation extends beyond the most local event.

6.4 Backbone and Observation-Regime Transport

Appendix D reports the complete transport matrix. On the TD-MPC-style backbone, ORVS improves shifted return over uncertainty by $36.6\,[23.4, 49.8]$, retention by $0.045\,[0.026, 0.064]$, and the Point/Car cost rates by $0.40\,[0.17, 0.63]$ and $0.50\,[0.22, 0.78]$. In pixel control, the pooled paired retention advantage over unsigned surprise is $0.032\,[0.011, 0.053]$. A localized Finger-spin video reversal—0.681 ± 0.031 for unsigned surprise versus 0.668 ± 0.030 for ORSC—coexists with the positive pooled effect. The displayed ORVS contrasts therefore recur across a second backbone and unseen visual shifts.

7 Analysis and Scope

Full ORVS has higher retention, lower cost rates, and greater within-trajectory event capture than each individual head in Appendix H. Among nonuniform weights, reward-heavy settings favor shifted return and cost-heavy settings favor safety rates; uniform weights provide the strongest joint point. Across 400 checkpoints per run, ORSC maintains 0.97 ± 0.02 reference-relative coverage and a -4.4 ± 3.1 return gap; its $17.6 \pm 4.1\%$ focal-run wall-time overhead accompanies an 89.7-point shifted-return gain, with separate reference training included in total accelerator-hours.

The evidence supports complete ORVS selector substitution in world-model control regimes with trustworthy reward, cost, and termination feedback. Nominal RMS alignment supplies a common

amplitude unit while retaining selector-specific activation and dual dynamics; positive affine reward recodings leave ORVS unchanged up to the numerical stabilizer, whereas nonlinear recodings can change it. The continuation judge quantifies finite-horizon simulator harm, and cross-family methods supply contextual operating points, so the controlled claims rest on the same-stage late-update comparisons rather than a formal safety guarantee. Within this scope, the head and weight analyses explain the joint operating point, and the backbone and visual tests establish transport beyond one representation.

8 Conclusion

An environment response turns a predictive distribution into a value-head score that can supervise later control. ORVS extracts the positive standardized score residual, and ORSC propagates its immutable label through a terminal-masked critic into guarded actor updates. Under a common late-update path, ORVS adds $43.9\,[30.5, 57.3]$ shifted-return points over uncertainty and lowers both matched safety-cost rates; the return advantage repeats with TD-MPC, while held-out visual control shows a positive pooled retention effect over unsigned surprise. Its 0.61 ± 0.04 harmful-event capture at 10% coverage further shows that realized value-head scores identify policy-relevant regions of subsequent simulator behavior.

AI Use Statement

Generative AI assisted with editing and reorganizing the manuscript's prose for readability, formatting, and presentation. The authors reviewed the revisions and take responsibility for the scientific content and final text.

Reproducibility Statement

Sections 2–5 specify the selector equations, critic and actor estimators, matched alternatives, update chronology, evaluation hierarchy, aggregation order, and uncertainty convention. Appendix A records optimization settings; the remaining appendices give task-level transport, operating grids, and mechanism analyses.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Yarden As, Bhavya Sukhija, and Andreas Krause. Safe exploration using bayesian world models and log-barrier optimization. *arXiv preprint arXiv:2405.05890*, 2024. doi: 10.48550/arXiv.2405.05890. URL https://arxiv.org/abs/2405.05890.

Carolin Benjamins, Theresa Eimer, Frederik Schubert, André Biedenkapp, Bodo Rosenhahn, Frank Hutter, and Marius Lindauer. Carl: A benchmark for contextual and adaptive reinforcement learning. *arXiv preprint arXiv:2110.02102*, 2021. doi: 10.48550/arXiv.2110.02102. URL https://arxiv.org/abs/2110.02102. Workshop on Ecological Theory of Reinforcement Learning, NeurIPS 2021.

Jacob Buckman, Danijar Hafner, George Tucker, Eugene Brevdo, and Honglak Lee. Sample-efficient reinforcement learning with stochastic ensemble value expansion. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8224–8234. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/f02208a057804ee16ac72ff4d3cec53b-Abstract.html.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18 (167):1–51, 2018. URL https://jmlr.org/papers/v18/15-636.html.

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, pp. 4754–4765. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/3de568f8597b94bda53149c7d7f5958c-Abstract.html.

A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982. doi: 10.1080/01621459.1982.10477856. URL https://doi.org/10.1080/01621459.1982.10477856.

Benjamin Eysenbach, Alexander Khazatsky, Sergey Levine, and Ruslan Salakhutdinov. Mismatched no more: Joint model-policy optimization for model-based RL. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23230–23243. Curran Associates, Inc., 2022. doi: 10.52202/068431-1688. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/935151cc6cb5d8b6816133b75233775a-Abstract-Conference.html.

Amir-Massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1486–1494. PMLR, 2017. URL https://proceedings.mlr.press/v54/farahmand17a.html.

Bernd Frauenknecht, Artur Eisele, Devdutt Subhasish, Friedrich Solowjow, and Sebastian Trimpe. Trust the model where it trusts itself – model-based actor-critic with uncertainty-aware rollout adaption. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13973–14005. PMLR, 2024. URL https://proceedings.mlr.press/v235/frauenknecht24a.html.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.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://openreview.net/forum?id=S1lOTC4tDS.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0oabwyZbOu.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Nicklas A Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8387–8406. PMLR, 2022. URL https://proceedings.mlr.press/v162/hansen22a.html.

Dongchi Huang, Jiaqi Wang, Yang Li, Chunhe Xia, Tianle Zhang, and Kaige Zhang. PIGDreamer: Privileged information guided world models for safe partially observable reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25859–25875. PMLR, 2025. URL https://proceedings.mlr.press/v267/huang25ai.html.

Weidong Huang, Jiaming Ji, Chunhe Xia, Borong Zhang, and Yaodong Yang. Safedreamer: Safe reinforcement learning with world models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tsE5HLYtYg.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety-gymnasium: A unified safe reinforcement learning benchmark. In *Advances in Neural Information Processing Systems*, volume 36, pp. 18964–18993. Curran Associates, Inc., 2023. doi: 10.52202/075280-0831. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3c557a3d6a48cc99444f85e924c66753-Abstract-Datasets_and_Benchmarks.html.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21810–21823. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f7efa4f864ae9b88d43527f4b14f750f-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. PMLR, 2018. URL https://proceedings.mlr.press/v80/kuleshov18a.html.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html.

Ali Malik, Volodymyr Kuleshov, Jiaming Song, Danny Nemer, Harlan Seymour, and Stefano Ermon. Calibrated model-based deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4314–4323. PMLR, 2019. URL https://proceedings.mlr.press/v97/malik19a.html.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=JYtwGwIL7ye.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2778–2787. PMLR, 2017. URL https://proceedings.mlr.press/v70/pathak17a.html.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. OpenAI technical report, 2019. URL https://cdn.openai.com/safexp-short.pdf.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.05952.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. doi: 10.48550/arXiv.1707.06347. URL https://arxiv.org/abs/1707.06347.

Austin Stone, Oscar Ramirez, Kurt Konolige, and Rico Jonschkowski. The distracting control suite – a challenging benchmark for reinforcement learning from pixels. *arXiv preprint arXiv:2101.02722*, 2021. doi: 10.48550/arXiv.2101.02722. URL https://arxiv.org/abs/2101.02722.

Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the CVaR via sampling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1):2993–2999, 2015. doi: 10.1609/aaai.v29i1.9561. URL https://ojs.aaai.org/index.php/AAAI/article/view/9561.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018. doi: 10.48550/arXiv.1801.00690. URL https://arxiv.org/abs/1801.00690.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14129–14142. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/a322852ce0df73e204b7e67cbbef0d0a-Abstract.html.

A Implementation and Optimization Details

This appendix collects the numerical choices shared by ORSC and every selector swap. Equation (14) defines the intervention.

Table 4: Fixed optimization settings. Hyperparameters are selected on nominal validation and shared by selector swaps.

Setting	Value	Setting	Value
Replay capacity	1,000,000 transitions	Calibration window	20,000 transitions
World-model optimizer	Adam, 3×10^{-4}	Actor optimizer	Adam, 8×10^{-5}
Critic optimizer	Adam, 10^{-4}	Dual optimizer	projected SGD, 5×10^{-3}
Model / actor batch	256 / 256	Discount γ	0.99
Continuation horizon	15	Target update	100 gradient steps
Polyak coefficient	0.01	ORVS budget ϵ_R	0.12
Return margin δ	0.03 normalized return	Entropy tolerance τ	0.10 nats
Head weights	$(1/3, 1/3, 1/3)$	Bernoulli clipping	$[10^{-6}, 1 - 10^{-6}]$
Return trace parameter	0.95	Head scaling	Frozen nominal RMS
Dreamer-style RSSM	1024 deterministic, 32×32 categorical	TD-MPC latent width	512
Actor / critic MLP	4×512 ELU	Reward/cost/term. heads	2×512 ELU
Updates per environment step	1	Global gradient clip	100
Reference estimate	50 nominal episodes	Recalibration cadence	5,000 environment steps

The focal and reference runs use identical latent size, imagination horizon, and update-to-data ratio within each backbone. The reference is trained once per task and seed; its final 50-episode nominal mean supplies $\widehat{J}_{\mathrm{ref}}$. The seed-fixed replay/calibration partition is established before collection. Temperature fitting minimizes held-out CRPS or Bernoulli negative log loss without actor gradients, and a calibration window is eligible only after all its transitions have received scores from the preceding snapshot.

B Absolute Task-Level Operating Points

The main text emphasizes controlled selector substitutions. The following tables retain the reported algorithm-family operating points for each state-control task.

Table 5: DMControl return, mean $\pm$ 95% interval over 10 seeds. Shifted scores equally average gravity, friction, and observation-noise families.

	Walker-walk		Hopper-hop		Cheetah-run	
Method	Nominal	Shifted	Nominal	Shifted	Nominal	Shifted
SAC	962 ± 12	684 ± 38	245 ± 18	152 ± 22	811 ± 15	597 ± 31
PPO	934 ± 16	658 ± 42	228 ± 21	139 ± 26	783 ± 19	564 ± 37
PETS	921 ± 20	712 ± 35	237 ± 19	168 ± 20	798 ± 17	631 ± 29
Context-Rand.	948 ± 14	738 ± 30	241 ± 17	179 ± 19	805 ± 16	649 ± 27
CVaR-RL	897 ± 22	701 ± 36	219 ± 23	163 ± 24	768 ± 21	612 ± 33
DreamerV3	958 ± 13	724 ± 32	248 ± 16	174 ± 21	818 ± 14	642 ± 28
ORSC	955 ± 14	849 ± 24	243 ± 17	208 ± 16	813 ± 15	752 ± 22

ORSC's nominal returns remain near the strongest task-specific operating points while its shifted return is highest on every task. Safety Gymnasium supplies a complementary native-cost view over the same ten-seed evaluation hierarchy.

Table 6: Safety Gymnasium outcomes under ten 1,000-step episodes per seed. Cost is cumulative over 10,000 evaluation steps; Cost/1k is mean episodic cost.

	SafetyPointGoal-v0			SafetyCarGoal-v0		
Method	Return	Cost	Cost/1k	Return	Cost	Cost/1k
SAC	34.2 ± 1.8	42.1 ± 5.3	4.21 ± 0.53	28.7 ± 2.1	51.3 ± 6.1	5.13 ± 0.61
PPO	31.8 ± 2.0	38.7 ± 4.8	3.87 ± 0.48	26.1 ± 2.4	47.2 ± 5.7	4.72 ± 0.57
PETS	32.5 ± 1.9	35.4 ± 4.5	3.54 ± 0.45	27.3 ± 2.2	43.8 ± 5.2	4.38 ± 0.52
Context-Rand.	33.1 ± 1.7	31.2 ± 4.0	3.12 ± 0.40	27.8 ± 2.0	39.5 ± 4.8	3.95 ± 0.48
DreamerV3	35.1 ± 1.6	39.8 ± 4.9	3.98 ± 0.49	29.2 ± 1.9	48.1 ± 5.6	4.81 ± 0.56
ORSC	34.8 ± 1.5	24.3 ± 3.2	2.43 ± 0.32	29.0 ± 1.8	30.7 ± 3.9	3.07 ± 0.39

The controlled reductions relative to uncertainty, predicted cost, unsigned surprise, and the zero-signal wrapper are reported in Table 2.

C Simulator-Continuation Coverage and Horizon Grid

The primary judge setting in Section 6 is one point in a predeclared coverage–horizon grid. The same score, next-state clone, reference continuation, event union, and aggregation order are reused in every cell below.

Table 7: Harmful-event capture under the simulator-continuation judge. Rows vary fixed selection coverage; columns vary continuation horizon.

Selector	Coverage	$H = 5$	$H = 15$	$H = 30$
ORVS	5%	0.30 ± 0.03	0.36 ± 0.03	0.35 ± 0.03
Uncertainty	5%	0.23 ± 0.03	0.25 ± 0.03	0.24 ± 0.03
Predicted cost	5%	0.31 ± 0.03	0.29 ± 0.03	0.27 ± 0.03
Unsigned surprise	5%	0.24 ± 0.03	0.27 ± 0.03	0.26 ± 0.03
Uniform	5%	0.05 ± 0.01	0.05 ± 0.01	0.05 ± 0.01
ORVS	10%	0.53 ± 0.04	0.61 ± 0.04	0.59 ± 0.04
Uncertainty	10%	0.39 ± 0.04	0.42 ± 0.04	0.40 ± 0.04
Predicted cost	10%	0.50 ± 0.04	0.49 ± 0.03	0.45 ± 0.04
Unsigned surprise	10%	0.43 ± 0.04	0.46 ± 0.04	0.44 ± 0.04
Uniform	10%	0.10 ± 0.01	0.10 ± 0.01	0.10 ± 0.01
ORVS	20%	0.72 ± 0.04	0.79 ± 0.03	0.77 ± 0.04
Uncertainty	20%	0.61 ± 0.04	0.64 ± 0.04	0.61 ± 0.04
Predicted cost	20%	0.70 ± 0.04	0.68 ± 0.04	0.64 ± 0.04
Unsigned surprise	20%	0.66 ± 0.04	0.69 ± 0.04	0.67 ± 0.04
Uniform	20%	0.20 ± 0.02	0.20 ± 0.02	0.20 ± 0.02

Across coverage and horizon, ORVS leads the forecast-time selectors most clearly once the judge includes more than the most local event. The next table resolves the primary operating point by task to verify that pooled capture is not driven by one environment.

Table 8: Task-level harmful-event capture at the primary 10% coverage and $H = 15$ simulator-continuation setting.

Selector	Walker-walk	Hopper-hop	Cheetah-run	SafetyPoint	SafetyCar
ORVS	0.58 ± 0.05	0.55 ± 0.06	0.63 ± 0.05	0.66 ± 0.04	0.62 ± 0.05
Uncertainty	0.39 ± 0.05	0.37 ± 0.05	0.43 ± 0.05	0.47 ± 0.05	0.44 ± 0.05
Predicted cost	0.46 ± 0.05	0.42 ± 0.05	0.48 ± 0.05	0.55 ± 0.04	0.54 ± 0.05
Unsigned surprise	0.44 ± 0.05	0.40 ± 0.05	0.47 ± 0.05	0.50 ± 0.05	0.49 ± 0.05
Uniform	0.10 ± 0.02	0.10 ± 0.02	0.10 ± 0.02	0.10 ± 0.02	0.10 ± 0.02

D Backbone and Observation-Regime Transport

The complete matrix separates TD-MPC-style state control from Dreamer-style pixel control. Each panel retains the training protocol of its native backbone, while selector rows share the late-update path within that panel.

Table 9: Transport results. Top: TD-MPC-style backbone on five state-control tasks. Bottom: three pixel-control tasks under held-out medium distractions.

TD-MPC-style backbone					
Selector	DM nominal	DM shifted	Retention	Point cost/1k	Car cost/1k
Backbone baseline	676.1 ± 12.6	525.4 ± 16.8	0.776 ± 0.024	3.71 ± 0.41	4.35 ± 0.49
ORVS (ORSC)	671.8 ± 12.9	589.6 ± 14.2	0.869 ± 0.019	2.47 ± 0.33	3.18 ± 0.38
Uncertainty-only	669.5 ± 13.1	553.0 ± 15.1	0.824 ± 0.021	2.87 ± 0.36	3.68 ± 0.42
Predicted-cost-only	667.4 ± 13.3	545.1 ± 15.7	0.812 ± 0.022	2.63 ± 0.32	3.46 ± 0.40
No extra selector	674.2 ± 12.5	533.8 ± 16.4	0.787 ± 0.025	3.55 ± 0.43	4.18 ± 0.47
Distracting Control Suite					
Selector	Clean	Color	Video	Camera	Retention
Backbone baseline	698 ± 18	458 ± 25	391 ± 27	432 ± 24	0.612 ± 0.028
ORVS (ORSC)	691 ± 19	548 ± 22	481 ± 25	514 ± 23	0.744 ± 0.022
Uncertainty-only	687 ± 20	510 ± 24	447 ± 26	474 ± 24	0.694 ± 0.025
Unsigned surprise	684 ± 21	519 ± 23	455 ± 25	487 ± 23	0.712 ± 0.024
No extra selector	693 ± 19	471 ± 25	406 ± 27	441 ± 25	0.633 ± 0.027

The ORVS–uncertainty return and safety effects remain positive on TD-MPC, while the visual panel shows a positive pooled ORVS–unsigned-surprise retention effect. The next sections resolve these aggregate effects by task.

E Pixel-Control Task Effects

On Walker-walk, the pixel retention effect over unsigned surprise is $0.063\,[0.030, 0.096]$. The corresponding effects are $0.030\,[0.004, 0.056]$ on Cheetah-run and $0.003\,[-0.019, 0.025]$ on Finger-spin. The final interval reflects the Finger-video reversal reported in Section 6, while the pooled matched-seed effect remains positive.

F Task-Level Backbone Transport

Task-level TD-MPC results resolve the pooled transport effect in Appendix D by locomotion environment.

Table 10: TD-MPC-style shifted return by task. Each row shares the 2M-step focal-policy, 10-seed protocol of Appendix D.

Selector	Walker-walk	Hopper-hop	Cheetah-run
Backbone baseline	714 ± 31	171 ± 20	691 ± 29
ORVS (ORSC)	821 ± 26	201 ± 17	747 ± 23
Uncertainty-only	760 ± 29	188 ± 18	711 ± 25
Predicted-cost-only	744 ± 30	181 ± 19	710 ± 26
No extra selector	726 ± 31	176 ± 20	699 ± 28

G Selector Mechanism, Weighting, and Efficiency

Three analyses test whether ORSC's gains coincide with healthy occupancy, complementary value heads, and acceptable training cost. The visitation table reports all guard diagnostics at the checkpoint level.

Table 11: Checkpoint visitation diagnostics. Entropy is the plug-in entropy over 512 latent-state centroids in a 50,000-transition window; 400 checkpoints are recorded per run.

Selector	Entropy gap	Coverage ratio	Guard trigger fraction	Return gap to reference
ORVS	-0.06 ± 0.03	0.97 ± 0.02	0.08 ± 0.03	-4.4 ± 3.1
Uncertainty	-0.17 ± 0.04	0.91 ± 0.03	0.19 ± 0.04	-5.8 ± 3.5
Predicted cost	-0.11 ± 0.03	0.94 ± 0.03	0.14 ± 0.04	-8.6 ± 3.8
No selector	-0.22 ± 0.05	0.88 ± 0.04	0.23 ± 0.05	-2.6 ± 3.3

ORVS stays closest to reference occupancy among the learned selectors and triggers the visitation guard least often. The following fixed-weight sweep asks whether that joint point depends on one dominant head.

Table 12: Fixed ORVS head-weight sensitivity without shifted-test retuning. The upper panel reports return outcomes; the lower panel reports safety and within-trajectory capture.

Weight setting	$(w_r, w_c, w_{\text{term}})$	DM shifted	Retention
Uniform	$(1/3, 1/3, 1/3)$	603.0 ± 13.6	0.890 ± 0.018
Reward-heavy	$(0.6, 0.2, 0.2)$	596.2 ± 14.1	0.879 ± 0.020
Cost-heavy	$(0.2, 0.6, 0.2)$	585.1 ± 14.8	0.866 ± 0.021
Termination-heavy	$(0.2, 0.2, 0.6)$	591.4 ± 14.4	0.873 ± 0.020

Weight setting	Point cost/1k	Car cost/1k	Event capture
Uniform	2.43 ± 0.32	3.07 ± 0.39	0.57 ± 0.04
Reward-heavy	2.81 ± 0.35	3.48 ± 0.42	0.53 ± 0.04
Cost-heavy	2.35 ± 0.30	2.94 ± 0.37	0.56 ± 0.04
Termination-heavy	2.52 ± 0.33	3.10 ± 0.39	0.55 ± 0.04

Uniform weighting delivers the strongest combined return, safety, and capture point, while specialized settings expose the return–cost tradeoff. Efficiency accounting separates focal-policy interactions from reference-policy training.

Relative to DreamerV3, ORSC adds 89.7 shifted-return points and lowers Point/Car cost rates by 1.55/1.74. Including the separate reference run, the 64.0 additional accelerator-hours correspond to 1.40 shifted-return points per added accelerator-hour in this implementation.

Table 13: Mean efficiency per task-seed run over 50 runs per method on $8\times$A100-80GB. Wall time and per-GPU peak memory describe the focal run; total accelerator-hours include reference training for selector methods.

Method	Focal steps	Ref. steps	Focal wall h	Peak GB	Total accel. h
DreamerV3	2M	–	6.8 ± 0.5	42.6 ± 1.9	54.4 ± 4.0
ORSC	2M	2M	8.0 ± 0.6	49.8 ± 2.1	118.4 ± 6.3
Uncertainty-only	2M	2M	7.5 ± 0.6	47.2 ± 2.0	114.4 ± 6.1
Predicted-cost-only	2M	2M	7.6 ± 0.6	46.9 ± 2.0	115.2 ± 6.2

H Within-Trajectory and Head-Wise Diagnostics

The within-trajectory ranking supplies a second outcome-based view beside the cloned-state judge, and the lower panel separates the contribution of each ORVS head under the same update path.

Table 14: Top: held-out within-trajectory ranking at fixed top-10% coverage and horizon 15. Bottom: head-wise ORVS decomposition.

Signal	Normalized return deficit	Bad-outcome incidence	Bad-event capture
ORVS	0.39 ± 0.02	0.44 ± 0.03	0.57 ± 0.04
Uncertainty-only	0.29 ± 0.02	0.31 ± 0.03	0.41 ± 0.04
Predicted-cost-only	0.33 ± 0.02	0.37 ± 0.03	0.48 ± 0.03

Signal	DM shifted	Retention	Point cost/1k	Car cost/1k	Event capture
Reward-only	581.2 ± 14.5	0.857 ± 0.020	3.06 ± 0.38	3.79 ± 0.44	0.46 ± 0.04
Cost-only	561.7 ± 15.2	0.839 ± 0.022	2.69 ± 0.34	3.31 ± 0.39	0.50 ± 0.04
Termination-only	573.4 ± 14.8	0.851 ± 0.021	2.82 ± 0.35	3.42 ± 0.40	0.52 ± 0.04
Full ORVS	603.0 ± 13.6	0.890 ± 0.018	2.43 ± 0.32	3.07 ± 0.39	0.57 ± 0.04

I Family-Resolved Dynamics Shifts

The family decomposition tests whether the task-level shifted averages conceal a reversal under gravity, friction, or observation noise.

Table 15: Perturbation-family return over 10 seeds and 20 episodes per seed. The final column is the paired ORSC difference from the stronger comparator.

Method	Task	Gravity	Friction	Observation noise	ORSC paired margin
Context-Rand.	Walker	721 ± 34	704 ± 32	789 ± 28	–
DreamerV3	Walker	689 ± 35	713 ± 31	770 ± 29	–
ORSC	Walker	803 ± 30	836 ± 26	908 ± 23	$82/123/119$
Context-Rand.	Hopper	172 ± 21	187 ± 18	178 ± 19	–
DreamerV3	Hopper	161 ± 22	181 ± 20	180 ± 18	–
ORSC	Hopper	189 ± 19	223 ± 16	212 ± 15	$17/36/32$
Context-Rand.	Cheetah	618 ± 30	667 ± 26	662 ± 25	–
DreamerV3	Cheetah	607 ± 31	655 ± 27	664 ± 25	–
ORSC	Cheetah	716 ± 26	779 ± 21	761 ± 22	$98/112/97$

The gravity column averages $0.7g$ and $1.3g$ before paired task-family aggregation. The corresponding gravity/friction/noise intervals for the ORSC margins are $[50, 114]/[94, 152]/[95, 143]$, $[4, 30]/[22, 50]/[18, 46]$, and $[70, 126]/[88, 136]/[72, 122]$. The nine cells are correlated views of the same ten seeds and are not treated as independent tests.