BOUNDARYWORLD: DECISION-REGIME-SUFFICIENT WORLD MODELS FOR RELIABLE LONG-HORIZON PLANNING

Anonymous Authors

ABSTRACT

World models are often optimized for state or observation fidelity, yet a planner does not need every detail of a predicted future. For a fixed downstream decision problem, the consequential distinctions are those that can change action rankings or feasibility. We formalize this requirement with *Decision-Regime Trajectory Sufficiency* (DRTS), which measures whether a learned model preserves the action-conditioned joint distribution of future decision regimes induced by that problem. A regime-abstraction residual measures the utility loss induced by representing a trajectory through its regime sequence and yields a continuous-plan regret bound that separates abstraction error from predictive error. BOUNDARYWORLD adds a rollout-conditioned joint regime-trajectory objective to a standard latent world model while retaining the same CEM search mechanics. An analytic diagnostic illustrates the distinction; on Safety-Gymnasium and MetaDrive, BoundaryWorld reduces $H{=}20$ planning regret by 19.3% and 21.7%, respectively, relative to the strongest controlled alternative in each domain.

Keywords: world models; model-based reinforcement learning; decision sufficiency; state abstraction; long-horizon planning.

1 INTRODUCTION

World models predict the consequences of candidate actions and are now standard in latent control, robotics, and autonomous driving (Hansen et al., 2024; Hafner et al., 2025; Chen et al., 2025). Their quality is often assessed through state, latent, or reconstruction fidelity, whereas a planner ultimately uses predicted futures to compare actions (Zhang et al., 2026a; Hansen et al., 2022).

Figure 1 shows the resulting mismatch in a stylized braking example. Moving a future pedestrian from 0.97 to 1.03 m barely changes the scene but can flip the maneuver; a less accurate background prediction may still preserve the correct action. Such boundary sensitivity is especially consequential in safety-constrained environments such as MetaDrive and Safety-Gymnasium (Li et al., 2023; Ji et al., 2023).

Several approaches reduce unnecessary prediction. Decoder-free latent MPC removes pixel reconstruction (Hansen et al., 2022; 2024); recent world models emphasize event structure, task-critical objects, sparse imagination, or compact task features (Peng et al., 2026; Zhang et al., 2026b; Chun et al., 2026; Fu et al., 2026); MIST-WM learns task-specific factors for decision making (Feng et al., 2026). These approaches emphasize selective content retention; our focus is the plan-conditioned distribution of decision changes along an entire candidate horizon.

A complementary line defines model quality through downstream decisions. Value-aware learning and value equivalence preserve selected value computations (Farahmand et al., 2017; Voelcker et al., 2025; Grimm et al., 2020); state abstraction, homomorphism, and bisimulation preserve compatible reward-transition structure (Ferns et al., 2004; Li et al., 2006; Ravindran and Barto, 2003); BS-MPC applies a bisimulation objective to latent control (Shimizu and Tomizuka, 2025). Structural certification identifies transitions needed by long-horizon goals (Lu et al., 2026). These criteria address values, abstract transitions, or selected predictive factors, but do not explicitly model the *joint action-conditioned sequence of future decision structures* over a candidate-plan horizon.

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.

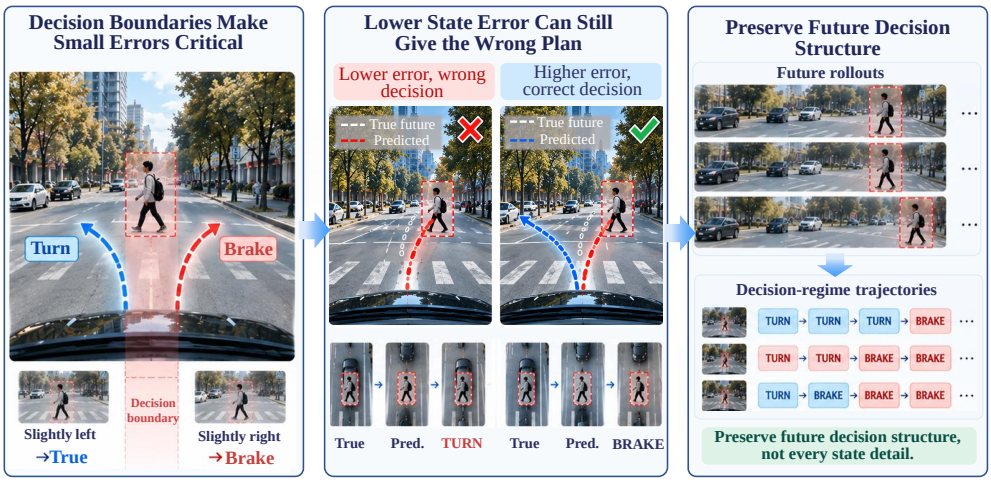

Figure 1: Motivation: small state errors can cross decision boundaries, while preserving future decision regimes supports reliable planning.

A fixed decision problem D induces *decision regimes* over the planner's information state, for example maneuver branches or active-constraint patterns. In driving, a regime can combine the maneuver branch (keep, brake, or lane change) with whether a safety constraint is active, turning a candidate plan into a sequence such as *keep* $\rightarrow$ *keep* $\rightarrow$ *brake*. Candidate action sequences therefore induce distributions over future regime trajectories. *Decision-Regime Trajectory Sufficiency* (DRTS) measures whether a learned model preserves those distributions.

A regime sequence can still incur utility approximation error under the chosen readout. The residual $\eta_D(\rho)$ measures this loss; combined with DRTS error, it bounds planning regret.

BOUNDARYWORLD learns this predictive object from counterfactual latent rollouts using planner-induced regime labels, while deployment retains the same CEM proposal and refinement mechanics across controlled methods. Our contributions are:

(1) **We introduce DRTS**, which preserves the joint action-conditioned distribution of future decision regimes rather than full future states or only selected value computations. We characterize the bounded regime-measurable functionals controlled by DRTS and relate it to full-dynamics fidelity, compatible abstraction/bisimulation, and restricted value equivalence.

(2) **We propose BoundaryWorld**, which couples latent dynamics to a rollout-conditioned joint regime-sequence model. The regime-trajectory likelihood backpropagates through predicted rollouts into the encoder and dynamics, rather than acting only as a post-hoc classifier, while retaining the same CEM search mechanics.

(3) **We connect abstraction to planning** through an explicit regime residual and a continuous-plan regret bound. Controlled experiments compare regime-trajectory fidelity with state and aggregate-value fidelity under matched model and planning budgets.

2 RELATED WORK

World Models and Planning. Latent world models range from stochastic predictive control (Hafner et al., 2019; Chua et al., 2018) to decoder-free MPC (Hansen et al., 2022; 2024) and scalable imagination-based control (Hafner et al., 2025). Recent methods add event structure, task-critical objects, sparse imagination, closed-loop evaluation, and 4D driving simulation or representation (Peng et al., 2026; Zhang et al., 2026b; Chun et al., 2026; Zhang et al., 2026a; Zhao et al., 2025).

Value Equivalence and Abstraction. VAML/CVAML optimize models through downstream values (Farahmand et al., 2017; Voelcker et al., 2025), while value equivalence preserves Bellman updates for selected function-policy families (Grimm et al., 2020). Homomorphism, state abstraction,

and bisimulation preserve compatible decision structure (Ravindran and Barto, 2003; Li et al., 2006; Ferns et al., 2004); BS-MPC uses a bisimulation objective in latent control (Shimizu and Tomizuka, 2025).

Task-Sufficient Predictive Models. MIST-WM learns compact task-specific representations (Feng et al., 2026), TC-WM compresses visual features for planning (Fu et al., 2026), and World Models in Pieces identifies goal-relevant transitions (Lu et al., 2026). DRTS instead targets the plan-conditioned distribution over the full future regime trajectory. Order, dwell time, and critical-entry events are therefore explicit predictive targets. Unlike compatible bisimulation or homomorphism, DRTS need not preserve reward-transition equivalences beyond those required by the downstream regime trajectory.

3 Decision-Regime Trajectory Sufficiency

We define a task-relative regime abstraction, its induced trajectory distribution, and the corresponding planning sufficiency criterion.

3.1 Decision Regimes and Abstraction Residual

Let $x_t \in \mathcal{X}$ denote the planner's information state: a physical state in an MDP or a history/belief state under partial observability. A downstream decision problem is

$$D = (\mathcal{A}, G_D, \mathcal{C}_D, H, \mathcal{P}_D), \tag{1}$$

where $\mathcal{A}$ is the action or plan space, G_D is the trajectory objective, $\mathcal{C}_D$ denotes constraints, H is the planning horizon, and $\mathcal{P}_D$ fixes the downstream decision operator or planner family.

The operator induces a decision signature $\sigma_D(x)$, such as an active-set pattern, controller branch, or maneuver mode. We use a fixed deterministic tie-breaking rule whenever the downstream operator admits multiple equivalent branches. We define $x_i \sim_D x_j$ iff $\sigma_D(x_i) = \sigma_D(x_j)$ and let $\rho_D : \mathcal{X} \to \mathcal{R}_D$ map an information state to its equivalence class. For a trajectory $\tau = x_{t+1:t+H}$, write $R^D_{1:H} = \rho^H_D(\tau)$. The partition is therefore relative to D rather than an intrinsic labeling of state space.

Definition 1 (Regime Abstraction Residual)**.** *For a decision abstraction ρ and the regime-level readout $\bar{G}_D$ used by the planner, let $\mathcal{R}^H_D(A)$ denote regime sequences with nonempty trajectory support under plan A, and define (suppressing the readout in the notation)*

$$\eta_D(\rho) = \sup_{A,\, r \in \mathcal{R}^H_D(A)} \; \sup_{\tau : \rho^H(\tau) = r} \left| G_D(\tau, A) - \bar{G}_D(r, A) \right|. \tag{2}$$

For a fixed partition, the within-class midpoint minimizes this residual and recovers the half-range optimum. Small η_D means that trajectories merged into one regime sequence remain close in downstream utility under the actual planning readout.

3.2 Decision-Regime Trajectory Sufficiency

Let $P_T(\tau \mid x, A)$ be the environment trajectory distribution for candidate plan $A = a_{t:t+H-1}$. Its pushforward through the decision abstraction is

$$Q^D_T(\cdot \mid x, A) = (\rho^H_D)_{\#} P_T(\cdot \mid x, A). \tag{3}$$

For a learned model M, let Q^D_M denote the corresponding induced distribution.

Definition 2 (Decision-Regime Trajectory Sufficiency)**.** *M is (D, H, ϵ)-DRT-sufficient on planner-relevant pairs $\mathcal{X}_P \times \mathcal{A}_P$ if*

$$\sup_{x \in \mathcal{X}_P, A \in \mathcal{A}_P} D_{\mathrm{TV}}\big(Q^D_T(\cdot \mid x, A), Q^D_M(\cdot \mid x, A)\big) \le \epsilon. \tag{4}$$

Equal marginals $P(R_h)$ leave transition order, dwell time, and first entry into a critical regime unspecified.

Let $\mathcal{F}_D = \{f(R, A) : \|f\|_\infty \le 1\}$ denote bounded regime-measurable decision functionals.

Proposition 1 (Decision-Functional Characterization)**.** *For fixed* (x, A)*,*

$$2D_{\mathrm{TV}}(Q_T^D, Q_M^D) = \sup_{f \in \mathcal{F}_D} \left| \mathbb{E}_{Q_T^D}[f] - \mathbb{E}_{Q_M^D}[f] \right|. \tag{5}$$

Equation (5) covers bounded regime-entry, occupancy, and switch-time functionals. Full trajectory fidelity and compatible model-preserving abstractions imply DRTS under standard measurability conditions, while restricted value equivalence may induce a different regime-trajectory law. Appendix A gives the formal relations and counterexamples.

3.3 Consequence for Planning

Fix the regime readout $\bar{G}_D$ and assume it is uniformly bounded, $\|\bar{G}_D\|_\infty \le M < \infty$. By Definition 1,

$$|G_D(\tau, A) - \bar{G}_D(\rho_D^H(\tau), A)| \le \eta_D(\rho). \tag{6}$$

For a model-induced regime distribution Q_M^D, define $\hat{J}_M(A) = \mathbb{E}_{R \sim Q_M^D}[\bar{G}_D(R, A)]$ and the true objective $J_T(A) = \mathbb{E}_{\tau \sim P_T}[G_D(\tau, A)]$.

Proposition 2 (Approximate Planning Fidelity)**.** *If Eq.* (4) *holds with error* ϵ*, then*

$$|J_T(A) - \hat{J}_M(A)| \le 2M\epsilon + \eta_D(\rho). \tag{7}$$

Whenever the corresponding maximizers exist, let $A^* \in \arg\max_A J_T(A)$ *and* $\hat{A} \in \arg\max_A \hat{J}_M(A)$*. Then*

$$J_T(A^*) - J_T(\hat{A}) \le 4M\epsilon + 2\eta_D(\rho). \tag{8}$$

The bound holds for continuous plan spaces; a finite candidate bank also preserves its best plan when the true gap exceeds $4M\epsilon + 2\eta_D(\rho)$. For constrained tasks, Proposition 2 applies to the scalarized planning objective used by the downstream planner; feasibility-related structure is additionally encoded in the regime signature and evaluated through false-safe and violation metrics. The proof is in Appendix A. BoundaryWorld approximates this predictive target from latent rollouts.

4 BoundaryWorld

4.1 Overview

BoundaryWorld retains the latent world-model backbone and the same CEM proposal/refinement mechanics used by the controlled methods. Its change lies in the supervision applied to counterfactual rollouts. Figure 2 summarizes training and planning. For each training branch, the fixed downstream operator provides the target regime trajectory and the latent rollout is trained to predict it. At deployment, oracle labels are removed and candidate plans are ranked with method-specific model scores under the shared CEM search.

4.2 Oracle-Induced Supervision and Joint Regime-Trajectory Learning

The fixed decision operator $\mathcal{P}_D$ converts each training information state into a decision signature; the regime label is then given by the fixed map ρ_D,

$$x_t \xrightarrow{\mathcal{P}_D} \sigma_D(x_t), \qquad r_t^* = \rho_D(x_t). \tag{9}$$

Safety-Gymnasium combines the downstream control branch with active safety constraints; MetaDrive combines the maneuver branch with collision, lane, and safety-distance constraints. Appendix B specifies the signatures and granularity variants. All controlled methods derive labels from the same replay or shared counterfactual branches.

Let $z_t = E_\phi(h_t)$ and let the stochastic latent dynamics induce

$$P_\theta(\hat{z}_{t+1:t+H} \mid z_t, A). \tag{10}$$

The rollout-conditioned regime head q_ψ—the *decision lens* in Figure 2—reads the predicted rollout,

$$q_\psi(r_h \mid r_{<h}, \hat{z}_{t+1:t+h}, A_{<h}), \tag{11}$$

Conditioning on the predicted rollout makes regime prediction depend on the latent dynamics. Training uses teacher forcing for $r^*_{<h}$ and deployment is autoregressive. Gradients from $\mathcal{L}_{\mathrm{DRT}}$ pass through q_ψ into E_ϕ and P_θ; post-hoc and no-dynamics-gradient variants stop this path.

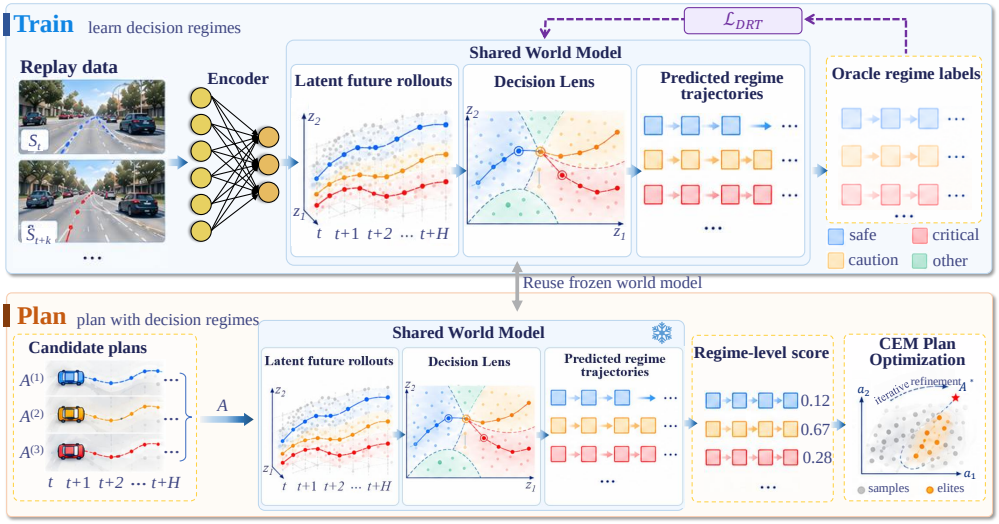

Figure 2: BoundaryWorld trains latent rollouts with joint regime-trajectory supervision and scores CEM candidates without oracle labels; colors denote schematic regimes induced by D.

4.3 Planner-Relevant Counterfactual Coverage

CEM may evaluate plans outside the behavior-policy support, so training samples a planner-relevant intervention distribution

$$(x, A) \sim \mu_{\text{plan}}, \qquad A^{(1)}, \dots, A^{(K_A)} \sim \mu_{\text{plan}}(\cdot \mid x), \tag{12}$$

from replay actions and proposals drawn from a shared reference-planner buffer that is frozen before the controlled comparison. When simulator branching is available, every method reuses the same proposals and branch budget. Counterfactual targets are obtained by forking the simulator from the same saved state and rolling out each proposed action sequence.

4.4 Practical Distribution Estimation and Planning

For a candidate plan A, the model induces the exact regime-trajectory law

$$\widetilde{Q}^{D}_{\theta,\psi}(R \mid z_t, A) = \int \prod_{h=1}^{H} q_\psi(r_h \mid r_{<h}, \hat{z}_{t+1:t+h}, A_{<h})\, dP_\theta(\hat{z}_{t+1:t+H} \mid z_t, A). \tag{13}$$

With K_z stochastic latent rollouts, we use the Monte-Carlo approximation

$$\widehat{Q}^{D}_{\theta,\psi}(R \mid z_t, A) = \frac{1}{K_z} \sum_{k=1}^{K_z} \prod_{h=1}^{H} q_\psi(r_h \mid r_{<h}, \hat{z}^{(k)}_{t+1:t+h}, A_{<h}), \quad \hat{z}^{(k)}_{t+1:t+H} \sim P_\theta(\cdot \mid z_t, A). \tag{14}$$

Training minimizes the finite-sample mixture likelihood

$$\mathcal{L}_{\text{DRT}} = -\log \widehat{Q}^{D}_{\theta,\psi}(R^{D,*}_{1:H} \mid z_t, A), \qquad \mathcal{L}_{\text{BW}} = \mathcal{L}_{\text{WM}} + \lambda \mathcal{L}_{\text{DRT}}. \tag{15}$$

For the exact law in Eq. (13), population cross-entropy is $H(Q^D_T) + D_{\text{KL}}(Q^D_T \| \widetilde{Q}^D_{\theta,\psi})$; Pinsker then gives $D_{\text{TV}}(Q^D_T, \widetilde{Q}^D_{\theta,\psi}) \le \sqrt{D_{\text{KL}}/2}$. Training and planning use the finite-K_z estimator in Eq. (14). A post-hoc head is limited to decision information already present in a frozen latent model; BoundaryWorld propagates $\mathcal{L}_{\text{DRT}}$ through reparameterized latent rollouts into the encoder and dynamics.

Planning scores are derived from the decision problem D through the same regime-measurable readout that defines the residual in Section 3. The experiments use

$$\bar{G}_D(R, A) = \sum_{h=1}^{H} \gamma^{h-1} \left[u_D(r_h, a_h) - \lambda_c c_D(r_h) - \lambda_a \|a_h\|_2^2 \right], \tag{16}$$

Table 1: Safety-Gymnasium results under matched data, capacity, counterfactual budget, and planner.

Method	Seq. NLL↓	DRT-JS↓	Crit. F1↑	Switch Err.↓	False Safe↓	Plan Flip↓	Regret↓	Viol.↓	Return↑
Base-WM	1.185 ± 0.061	0.176 ± 0.010	0.709 ± 0.011	1.920 ± 0.088	0.125 ± 0.019	0.286 ± 0.024	0.338 ± 0.031	0.152 ± 0.018	74.1 ± 1.1
Value-Aware-WM	1.045 ± 0.039	0.161 ± 0.010	0.756 ± 0.016	1.540 ± 0.121	0.109 ± 0.014	0.229 ± 0.013	0.228 ± 0.028	0.134 ± 0.012	78.4 ± 1.5
BS-MPC-style	0.978 ± 0.057	0.134 ± 0.006	0.771 ± 0.010	1.490 ± 0.073	0.094 ± 0.012	0.239 ± 0.021	0.281 ± 0.019	0.116 ± 0.019	79.2 ± 0.9
Cost-Supervised-WM	1.022 ± 0.046	0.153 ± 0.009	0.783 ± 0.018	1.620 ± 0.116	0.077 ± 0.016	0.252 ± 0.017	0.276 ± 0.035	0.056 ± 0.013	78.5 ± 1.4
Marginal-Regime-WM	0.858 ± 0.051	0.117 ± 0.009	0.815 ± 0.008	1.140 ± 0.083	0.075 ± 0.006	0.184 ± 0.022	0.205 ± 0.015	0.088 ± 0.015	82.0 ± 1.2
Direct-DRT	0.742 ± 0.027	0.099 ± 0.005	0.849 ± 0.015	1.060 ± 0.089	0.067 ± 0.010	0.176 ± 0.014	0.181 ± 0.026	0.084 ± 0.010	82.7 ± 1.4
BoundaryWorld	0.632 ± 0.045	0.078 ± 0.009	0.879 ± 0.013	0.750 ± 0.064	0.046 ± 0.009	0.131 ± 0.016	0.146 ± 0.018	0.061 ± 0.014	85.3 ± 0.9

where u_D and c_D are deterministic regime-level utility and constraint terms induced by the same downstream reward/constraint specification used to define D; their environment-specific construction is given in Appendix B. The held-out residual $\hat{\eta}_D$ is computed for this exact readout, so Proposition 2 applies to the scoring function used in the experiments. Candidate plans receive

$$\hat{J}(A) = \mathbb{E}_{R \sim \widehat{Q}^D_{\theta,\psi}} [\bar{G}_D(R, A)] \tag{17}$$

and are ranked by a shared CEM planner. In practice we estimate Eq. (17) with the same K_z samples used for Eq. (14): each latent rollout generates one autoregressive regime trajectory $R^{(k)}$, and the candidate score is $K_z^{-1} \sum_{k=1}^{K_z} \bar{G}_D(R^{(k)}, A)$. This avoids enumerating the exponentially large regime-sequence support. Base-WM, Value-Aware-WM, BS-MPC-style, and Cost-Supervised-WM receive a post-hoc DRT head fitted after the latent model is frozen. Marginal-Regime-WM trains the same head with independent per-horizon labels, whereas BoundaryWorld jointly optimizes the autoregressive head, encoder, and dynamics with $\mathcal{L}_{\mathrm{DRT}}$. Direct-DRT predicts $P(R_{1:H} \mid x, A)$ without latent dynamics. All variants use the same regime annotations and scoring rule.

4.5 Training and Deployment

Training samples $(x, A) \sim \mu_{\mathrm{plan}}$, rolls the latent model for H steps, obtains the target regime trajectory from the fixed operator, and updates the encoder, dynamics, and head with Eq. (15). At deployment, CEM proposes K_A candidates; K_z latent rollouts per candidate estimate Eq. (14), Eq. (17) ranks them, and the first action is executed before replanning. No oracle labels or simulator branches are queried online. Regime scoring adds $O(K_A K_z H)$ head evaluations beyond the shared latent-rollout cost.

5 Experiments

5.1 Evaluation Protocol

The evaluation covers five axes: predictive versus decision fidelity, sensitivity to decision-critical perturbations, long-horizon behavior, predictive capacity, and the abstraction–prediction trade-off induced by regime granularity. An analytic Diagnostic MDP illustrates the state-error/decision-error mismatch; quantitative benchmark results use Safety-Gymnasium (Ji et al., 2023) and MetaDrive (Li et al., 2023). Controlled methods share replay data, latent capacity, counterfactual budget, horizon, five paired seeds, and CEM proposal/refinement settings.

Baselines are Base-WM, a VAML-style Value-Aware-WM (Farahmand et al., 2017), BS-MPC-style (Shimizu and Tomizuka, 2025), Cost-Supervised-WM, Marginal-Regime-WM, and Direct-DRT. The first four use post-hoc DRT heads on frozen latent representations; Marginal-Regime-WM trains with per-horizon regime labels; BoundaryWorld propagates the joint trajectory loss through the encoder and dynamics. Appendix B gives the controlled objectives and state-probe protocol.

Sequence NLL is the length-normalized autoregressive loss $-H^{-1} \sum_{h=1}^{H} \log q_\psi(r_h \mid r_{<h}, \cdot)$; DRT-JS is Jensen–Shannon divergence between matched regime-trajectory distributions. Crit. F1 scores oracle-defined critical entries, Switch Err. is first-switch timing error in planning steps, and False Safe counts plans predicted safe that violate a simulator constraint. Aggregate Value Error averages $|\hat{J}(A) - J_T(A)|$ over the common candidate bank; Plan Flip marks disagreement with its simulator-best candidate and regret is $J_T(A^*) - J_T(\hat{A})$. Main tables use $H = 20$ and five paired seeds; Appendix B gives full metric and statistical definitions.

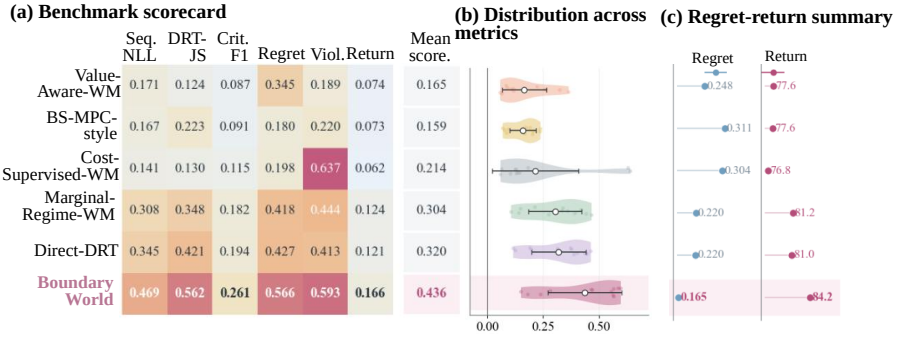

Figure 3: Normalized benchmark summary across predictive, decision, and control metrics. Panel (b) summarizes the spread across six normalized metric values rather than repeated stochastic samples; their unweighted mean is used only as a visual summary. Exact values appear in Tables 1–2.

Table 2: MetaDrive results under matched model, intervention, rollout, and CEM budgets.

Method	Seq. NLL↓	DRT-JS↓	Crit. F1↑	Switch Err.↓	False Safe↓	Plan Flip↓	Regret↓	Viol.↓	Return↑
Base-WM	1.410 ± 0.074	0.208 ± 0.014	0.664 ± 0.021	2.250 ± 0.118	0.164 ± 0.025	0.335 ± 0.031	0.422 ± 0.041	0.201 ± 0.018	70.4 ± 1.9
Value-Aware-WM	1.095 ± 0.043	0.174 ± 0.012	0.736 ± 0.011	1.620 ± 0.147	0.124 ± 0.018	0.248 ± 0.019	0.268 ± 0.023	0.149 ± 0.022	76.8 ± 1.2
BS-MPC-style	1.185 ± 0.065	0.165 ± 0.008	0.727 ± 0.017	1.920 ± 0.096	0.136 ± 0.015	0.279 ± 0.027	0.341 ± 0.036	0.160 ± 0.014	75.9 ± 1.6
Cost-Supervised-WM	1.205 ± 0.052	0.181 ± 0.013	0.748 ± 0.019	1.870 ± 0.159	0.096 ± 0.020	0.302 ± 0.029	0.332 ± 0.027	0.072 ± 0.019	75.0 ± 1.8
Marginal-Regime-WM	0.930 ± 0.058	0.133 ± 0.010	0.806 ± 0.009	1.490 ± 0.082	0.086 ± 0.014	0.205 ± 0.016	0.235 ± 0.031	0.107 ± 0.012	80.4 ± 1.1
Direct-DRT	0.965 ± 0.038	0.124 ± 0.007	0.791 ± 0.018	1.230 ± 0.124	0.094 ± 0.009	0.224 ± 0.029	0.258 ± 0.024	0.125 ± 0.016	79.3 ± 2.0
BoundaryWorld	0.745 ± 0.055	0.090 ± 0.010	0.852 ± 0.014	0.960 ± 0.078	0.058 ± 0.012	0.149 ± 0.018	0.184 ± 0.026	0.083 ± 0.013	83.2 ± 1.3

Table 3: Cross-method association between predictive metrics and planning failures; partial correlations control for horizon and task/scenario.

Metric	Safety ρ_{regret}	Safety ρ_{FS}	Drive ρ_{regret}	Drive ρ_{FS}	Partial ρ_{regret}	Partial ρ_{FS}
State Error	0.301	0.259	0.344	0.318	0.205	0.182
Aggregate Value Error	0.548	0.492	0.559	0.438	0.451	0.397
Sequence NLL	0.786	0.705	0.821	0.769	0.702	0.656
DRT-JS	0.842	0.812	0.874	0.819	0.773	0.731
Switch-Time Error	0.815	0.836	0.798	0.861	0.721	0.758

5.2 Prediction Fidelity and Planning Reliability

The Diagnostic MDP uses $x = (n, c)$, where the high-variance nuisance coordinate n dominates state error and the low-amplitude coordinate c controls a decision switch at $c = 0$. A model can predict n accurately yet bias c near the switch, while a second model may discard nuisance detail and preserve the decision regime. Figure 1 qualitatively illustrates the same mismatch.

Across both domains, BoundaryWorld attains the lowest DRT-JS and planning regret: 0.078/0.146 in Safety-Gymnasium and 0.090/0.184 in MetaDrive. Cost-Supervised-WM gives the lowest violation rate, but BoundaryWorld has lower plan-flip and regret and higher return (Tables 1–2). Figure 3 summarizes the normalized cross-metric pattern.

DRT-JS shows a substantially higher cross-method correlation with regret than State Error in Safety-Gymnasium (0.842 vs. 0.301) and MetaDrive (0.874 vs. 0.344), while Switch-Time Error tracks false-safe failures most closely (Table 3). At $H = 40$, BoundaryWorld reaches regret 0.182/0.230 versus 0.306/0.301 for the strongest alternatives among the horizon-sweep baselines in Safety/Drive (Table 4); Appendix C gives the corresponding DRT-JS values.

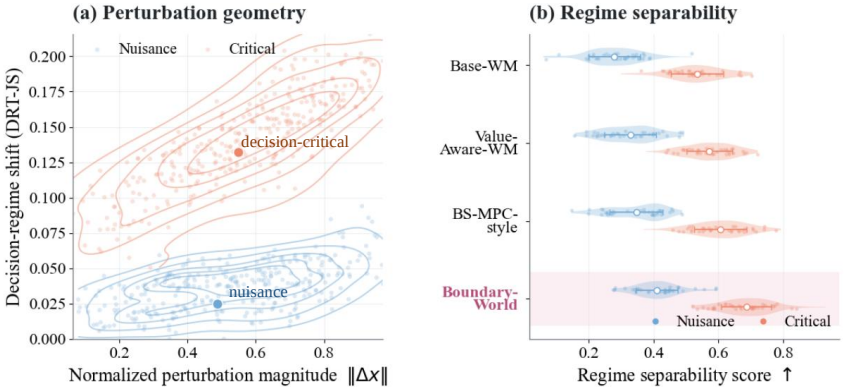

Figure 4: Paired nuisance and decision-critical perturbations used to inspect regime-distribution sensitivity and latent separation.

Table 4: Long-horizon planning in Safety-Gymnasium and MetaDrive under a fixed candidate-plan budget.

Env.	Method	Seq. NLL↓				Regret↓				FS@40↓	Viol.@40↓
		H=5	10	20	40	H=5	10	20	40		
Safety	Base-WM	0.714	0.804	1.185	1.693	0.131	0.159	0.338	0.444	0.188	0.221
	Value-Aware-WM	0.421	0.786	1.045	1.685	0.084	0.149	0.228	0.306	0.158	0.184
	BS-MPC-style	0.448	0.649	0.978	1.542	0.082	0.134	0.281	0.351	0.151	0.167
	Marginal-Regime-WM	0.454	0.625	0.858	1.423	0.055	0.123	0.205	0.325	0.124	0.139
	Direct-DRT	0.318	0.568	0.742	1.301	0.094	0.147	0.181	0.395	0.107	0.123
	BoundaryWorld	0.314	0.413	0.632	0.791	0.064	0.075	0.146	0.182	0.071	0.082
Drive	Base-WM	0.790	0.889	1.410	2.488	0.226	0.260	0.422	0.733	0.249	0.281
	Value-Aware-WM	0.710	0.817	1.095	1.747	0.170	0.217	0.268	0.453	0.196	0.229
	BS-MPC-style	0.535	0.913	1.185	1.661	0.128	0.216	0.341	0.426	0.202	0.218
	Marginal-Regime-WM	0.399	0.696	0.930	1.263	0.088	0.140	0.235	0.301	0.156	0.181
	Direct-DRT	0.452	0.528	0.965	1.603	0.105	0.120	0.258	0.460	0.179	0.194
	BoundaryWorld	0.402	0.495	0.745	1.059	0.095	0.084	0.184	0.230	0.095	0.109

5.3 Decision-Critical Representation

Paired nuisance and decision-critical perturbations probe whether the learned representation distinguishes visual change from changes that alter the oracle decision signature. Figure 4 visualizes this contrast; the fixed aggregate evaluation slice described in Appendix B.2 is used for the ablations below.

Marginal regime supervision recovers most critical-entry accuracy (0.866 vs. 0.870) but not the planning gains of joint trajectory supervision (regret 0.206 vs. 0.152); no-dynamics-gradient and head-bypass variants rise to 0.259 and 0.265. These ablations support the importance of joint temporal supervision and gradient flow through predicted rollouts (Table 5).

5.4 Abstraction Granularity and Predictive Capacity

Under the 512-to-32 latent-capacity sweep, BoundaryWorld retains higher return and lower DRT-JS/regret than Base-WM and BS-MPC-style (Appendix Table A4). Granularity shows the complementary trade-off: coarse signatures are easiest to predict but have the largest abstraction residual, while fine signatures reduce $\hat{\eta}_D$ at the cost of harder trajectory prediction. The medium partition gives the lowest regret and highest return (Table 6; Figure 5); Appendix Table A5 gives domain-specific values.

5.5 Discussion and Scope

DRTS is task-relative: changing D can change the regime map, and training requires a planner or constraint structure from which regime signatures are derived. Direct-DRT remains a competitive

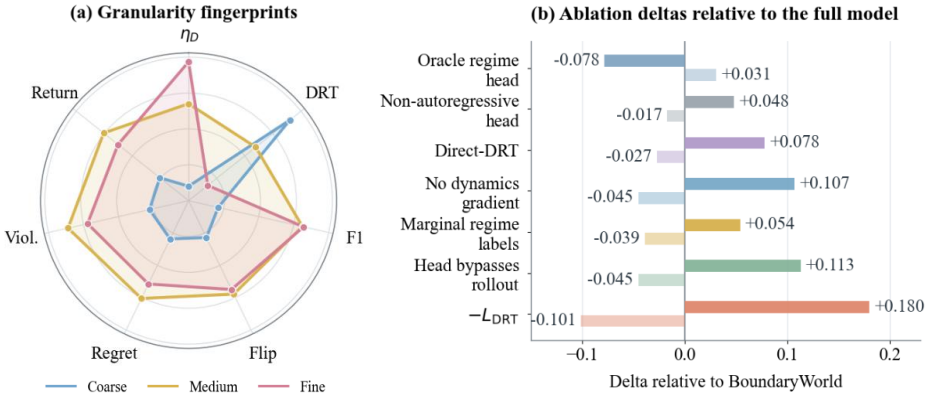

Figure 5: Granularity and mechanism diagnostics. Radar axes are normalized and direction-aligned; panel (b) compares mechanism variants, Direct-DRT, and the oracle upper bound using native ΔRegret and ΔReturn/100. Exact values are tabulated.

Table 5: Mechanism study, direct alternative, and oracle upper bound for regime-trajectory learning.

Variant	Seq. NLL↓	DRT-JS↓	Switch Err.↓	Crit. F1↑	False Safe↓	Plan Flip↓	Regret↓	Viol.↓	Return↑
BoundaryWorld	0.671	0.081	0.860	0.870	0.052	0.138	0.152	0.065	84.5
$-\mathcal{L}_{\mathrm{DRT}}$	1.248	0.189	1.870	0.716	0.126	0.276	0.332	0.154	74.4
Marginal regime labels	0.979	0.128	1.537	0.866	0.093	0.197	0.206	0.096	80.6
Head bypasses rollout	0.747	0.174	1.598	0.792	0.102	0.216	0.265	0.111	80.0
No dynamics gradient	0.949	0.109	1.122	0.853	0.062	0.216	0.259	0.106	80.0
Non-autoregressive head	0.710	0.129	1.556	0.804	0.053	0.179	0.200	0.079	82.8
Direct-DRT	0.720	0.089	0.997	0.863	0.079	0.216	0.230	0.090	81.8
Oracle regime head (upper bound)	0.423	0.035	0.455	0.947	0.021	0.069	0.074	0.041	87.6

Table 6: Regime-granularity trade-off using support-aware residual estimates.

Signature	$\hat{\eta}_D$ ↓	DRT-JS↓	Crit. F1↑	Plan Flip↓	Regret↓	Viol.↓	Return↑
Coarse	0.261	0.068	0.832	0.163	0.187	0.090	82.3
Medium	0.147	0.081	0.870	0.139	0.152	0.065	84.5
Fine	0.089	0.098	0.871	0.141	0.160	0.071	83.9

fixed-task alternative. BoundaryWorld retains latent dynamics to provide compositional transition structure across unseen action combinations and longer rollouts, while supporting state-dependent objectives beyond a fixed regime readout (Appendix Table A7). Regime design and model learning are therefore coupled through the abstraction–prediction trade-off.

6 Conclusion and Future Work

BoundaryWorld targets future distinctions that affect downstream plans by learning action-conditioned decision-regime trajectories. The abstraction residual and DRTS error separate the two sources of planning error, while joint trajectory supervision shapes latent rollouts under the same CEM search mechanics. Across both benchmark domains, stronger regime-trajectory fidelity coincides with lower long-horizon planning regret, including under constrained latent capacity. Learning a single abstraction that adapts across broader or nonstationary decision problems remains open.

Reproducibility Statement. The paper specifies the decision abstraction, objectives, controlled baselines, evaluation metrics, and statistical protocol. Appendix B details regime construction, shared data and budgets, metrics, and controlled objectives; Appendix C reports expanded task-, scenario-, horizon-, capacity-, ablation-, and shift-level results; and Appendix D gives model architecture, hyperparameters, and statistical settings. Anonymous supplementary code provides the core BoundaryWorld model, training, and CEM planning implementation.

AI Use Statement. Large language models were used for literature organization, language editing, LaTeX drafting, and consistency checks.

REFERENCES

Yuntao Chen, Yuqi Wang, and Zhaoxiang Zhang. DrivingGPT: Unifying driving world modeling and planning with multi-modal autoregressive transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 26890–26900, 2025. doi: 10.1109/ICCV51701.2025.02496.

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*, 2018.

Junha Chun, Youngjoon Jeong, and Taesup Kim. Sparse imagination for efficient visual world model planning. In *International Conference on Learning Representations*, 2026.

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*, pages 1486–1494, 2017.

Fan Feng, Yujia Zheng, Minghao Fu, Yongqiang Chen, Guangyi Chen, Kevin Patrick Murphy, Biwei Huang, and Kun Zhang. Learning task-sufficient world models by synergizing agentic exploration and structured modeling. In *International Conference on Machine Learning*, 2026.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 162–169, 2004.

Minghao Fu, Fan Feng, Nicklas Hansen, and Biwei Huang. Back to parsimonious latents: Learning task-centric world models from visual foundations. *arXiv preprint arXiv:2605.25620*, 2026.

Christopher Grimm, Andre Barreto, Satinder P. Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, 2019.

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

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

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*, pages 8387–8406, 2022.

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, Datasets and Benchmarks Track*, 2023.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics*, pages 531–539, 2006.

Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. MetaDrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3461–3475, 2023. doi: 10.1109/TPAMI.2022.3190471.

Yikai Lu, Yifei Wu, Xinyu Lu, and Tongxin Li. World models in pieces: Structural certification for general agents. In *International Conference on Machine Learning*, 2026.

Zhao-Han Peng, Shaohui Li, Zhi Li, Shulan Ruan, Yu Liu, and You He. From observations to events: Event-aware world models for reinforcement learning. In *International Conference on Learning Representations*, 2026.

Balaraman Ravindran and Andrew G. Barto. SMDP homomorphisms: An algebraic approach to abstraction in Semi-Markov decision processes. In *International Joint Conference on Artificial Intelligence*, 2003.

Yutaka Shimizu and Masayoshi Tomizuka. Bisimulation metric for model predictive control. In *International Conference on Learning Representations*, 2025.

Claas A. Voelcker, Anastasiia Pedan, Arash Ahmadian, Romina Abachi, Igor Gilitschenski, and Amir-Massoud Farahmand. Calibrated value-aware model learning with probabilistic environment models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 61745–61768, 2025.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-In-World: World models in a closed-loop world. In *International Conference on Learning Representations*, 2026a.

Weipu Zhang, Adam Jelley, Trevor McInroe, Amos Storkey, and Gang Wang. Object-centric world models from few-shot annotations for sample-efficient reinforcement learning. In *International Conference on Learning Representations*, 2026b.

Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Xueyang Zhang, Yida Wang, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, Wenjun Mei, and Xingang Wang. DriveDreamer4D: World models are effective data machines for 4D driving scene representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12015–12026, 2025.

A Proofs and Additional Theory

A.1 Proof of Proposition 1

For probability measures P and Q, total variation is $D_{\mathrm{TV}}(P,Q) = \frac{1}{2}\|P-Q\|_1$. The dual representation of the signed measure $P-Q$ gives

$$\sup_{\|f\|_\infty \le 1} |\mathbb{E}_P[f] - \mathbb{E}_Q[f]| = \|P-Q\|_1 = 2D_{\mathrm{TV}}(P,Q). \tag{18}$$

Applying this identity to $P = Q_T^D(\cdot \mid x, A)$ and $Q = Q_M^D(\cdot \mid x, A)$ proves Eq. (5). Because f is any bounded measurable function of the regime sequence, the same statement covers entry events, occupancy, switch time, and bounded risk functionals.

A.2 Full Fidelity and Decision-Regime Sufficiency

Let $\Phi = \rho_D^H$ be the measurable map from trajectories to regime trajectories. The data-processing inequality for total variation gives

$$D_{\mathrm{TV}}(\Phi_{\#}P_T, \Phi_{\#}P_M) \le D_{\mathrm{TV}}(P_T, P_M). \tag{19}$$

Hence full trajectory-distribution fidelity implies DRTS. The converse fails in a simple nuisance-variable construction. Let the state be (r, n), where r is the decision regime and n is a nuisance variable. Two models can induce the same distribution of $r_{1:H}$ for every candidate plan while assigning disjoint distributions to $n_{1:H}$. Their DRT discrepancy is zero although their full trajectory TV distance is positive.

A.3 Relation to Compatible Bisimulation or Homomorphism

Suppose an abstraction map ϕ preserves the reward/transition structure required by a compatible bisimulation or MDP homomorphism and ρ_D is measurable with respect to ϕ, i.e., $\rho_D = \tilde{\rho}_D \circ \phi$. Equality of the abstract trajectory law then implies equality after the further pushforward through $\tilde{\rho}_D^H$, so the induced regime-trajectory distributions coincide. The reverse implication is not required. If two models share the same regime trajectories but differ in reward or transition detail inside a regime, they can satisfy DRTS while violating the conditions needed for bisimulation or a reward-preserving homomorphism.

A.4 Value Equivalence and DRTS Are Distinct

Consider a fixed action and one selected value function with terminal values $V(s_S) = 0$, $V(s_M) = 1$, and $V(s_C) = 2$. Model M_1 transitions to s_S and s_C with probability $1/2$ each; model M_2 transitions to s_M with probability one. With equal immediate reward and discount, both models produce the same Bellman update for the restricted family $\{V\}$ because $\mathbb{E}_{M_1}[V] = \mathbb{E}_{M_2}[V] = 1$. They are therefore value equivalent for that restricted function/action family. If $\rho_D(s_S) = \mathrm{Safe}$, $\rho_D(s_M) = \mathrm{Moderate}$, and $\rho_D(s_C) = \mathrm{Catastrophic}$, their regime distributions differ. Conversely, two models can agree on the regime distribution while differing on a value function that varies inside one regime. The two criteria are therefore not interchangeable in general.

A.5 Proof of Proposition 2

For any planner-relevant plan A, insert the true regime-level expectation between $J_T(A)$ and $\hat{J}_M(A)$:

$$|J_T(A) - \hat{J}_M(A)| \le \left|\mathbb{E}_{P_T}[G_D(\tau, A) - \bar{G}_D(\rho_D^H(\tau), A)]\right| \tag{20}$$

$$+ \left|\mathbb{E}_{Q_T^D}[\bar{G}_D] - \mathbb{E}_{Q_M^D}[\bar{G}_D]\right|. \tag{21}$$

The first term is at most $\eta_D(\rho)$ by Eq. (6). Since $|\bar{G}_D| \le M$, the second term is at most $2M D_{\mathrm{TV}}(Q_T^D, Q_M^D) \le 2M\epsilon$. Thus $|J_T(A) - \hat{J}_M(A)| \le \delta$ with $\delta = 2M\epsilon + \eta_D(\rho)$. For $A^* \in \arg\max J_T$ and $\hat{A} \in \arg\max \hat{J}_M$,

$$J_T(A^*) - J_T(\hat{A}) \le [J_T(A^*) - \hat{J}_M(A^*)] + [\hat{J}_M(\hat{A}) - J_T(\hat{A})] \tag{22}$$

$$\le 2\delta = 4M\epsilon + 2\eta_D(\rho), \tag{23}$$

where $\hat{J}_M(\hat{A}) \ge \hat{J}_M(A^*)$ removes the middle comparison term. If a finite candidate set has true best-versus-second-best gap greater than 2δ, the same uniform approximation argument preserves the maximizing candidate.

B Experimental Protocols and Decision-Regime Construction

B.1 Decision-Regime Construction

The fixed downstream decision problem induces the regime map. In the Diagnostic MDP, $\rho_D(x)$ is the sign-defined decision mode of the critical coordinate. In Safety-Gymnasium, the signature

combines the branch selected by the fixed downstream planner with the binary pattern of active environment safety constraints; Appendix C denotes the three fixed benchmark tasks by S1–S3. In MetaDrive, the signature combines the planner maneuver (keep, brake, left change, or right change) with active collision-risk, lane-boundary, and safety-distance constraints; Sparse, Dense, and Critical refer to the fixed scenario strata used throughout the evaluation. Coarse signatures keep only the decision branch, medium signatures add whether any safety constraint is active, and fine signatures retain the full active-constraint pattern. Ambiguous branches use the deterministic tie-breaking rule of Section 3, and residual estimation uses only cells with nonempty held-out support.

Benchmark identities and scenario strata. The three Safety-Gymnasium tasks used in Table A1 are S1 = `SafetyPointGoal1-v0`, S2 = `SafetyPointButton1-v0`, and S3 = `SafetyPointPush1-v0`. MetaDrive scenarios are assigned to the Sparse, Dense, and Critical strata using the local non-ego vehicle count n_{30} within 30 m and the minimum predicted time-to-collision $\mathrm{TTC}_{\min}$: Critical if $\mathrm{TTC}_{\min} \leq 2.0$ s; otherwise Dense if $n_{30} \geq 8$ or $\mathrm{TTC}_{\min} \leq 4.0$ s; and Sparse otherwise. These task identities and stratification rules are fixed before test evaluation and are shared by all methods.

The empirical residual $\hat{\eta}_D$ is estimated on held-out trajectories within the same reachable regime sequence using the exact readout in Eq. (16). The planning terms $u_D(r, a)$ and $c_D(r)$ are deterministic lookups derived from the downstream reward and active-constraint specification associated with each regime signature; they are fixed before model training. To avoid a downward bias when fine partitions have little support, regime cells below a minimum sample count are merged with their parent signature or excluded from residual estimation according to a validation-fixed rule. Support counts are reported in Table A5; bootstrap intervals are estimated for residual-analysis checks.

B.2 Planner-Relevant Data and Common Budgets

Replay trajectories are shared across all controlled methods. Candidate plans used for counterfactual supervision follow the distribution μ_{plan} in Eq. (12). Planner proposals are generated once by a shared reference planner and stored in a frozen proposal buffer. When simulator branching is available, every method receives the same proposal sequences and branch-query budget; each counterfactual target is obtained by forking the simulator from the same saved state. The fixed map ρ_D is used to produce training labels and is unavailable during deployment. Marginal-Regime-WM and other regime-supervised controls receive the same annotations.

For planning-regret evaluation, each frozen evaluation state uses a common candidate bank. Every candidate is rolled out K_{eval} times in the true simulator to estimate its expected objective. The simulator-best candidate defines the reference plan for regret and Plan Flip. The shared candidate bank makes regret and Plan Flip comparable across methods. False-safe uncertainty is estimated at the episode level with bootstrap or binomial confidence intervals. Aggregate analyses without an explicit environment/task/scenario column use the same fixed evaluation slice and weighting across methods, including the mechanism, capacity, and shift studies.

Aggregate evaluation slice. Unless a table reports an explicit environment, task, or scenario column, the aggregate analyses use a fixed held-out slice containing 600 Safety-Gymnasium states and 600 MetaDrive states from the five frozen final-evaluation seeds $\{100, 101, 102, 103, 104\}$, disjoint from the training and validation seeds. The slice uses equal task/scenario weighting (200 states from each of S1–S3 and 200 from each of Sparse/Dense/Critical) and the main planning horizon $H = 20$. The same evaluation states, candidate banks, and weights are used for every compared method.

B.3 Distributional and State-Fidelity Metrics

Sequence NLL is reported per planning step, $-H^{-1} \sum_{h=1}^{H} \log q_\psi(r_h \mid r_{<h}, \cdot)$, so horizon sweeps are not inflated mechanically by sequence length. DRT-JS is Jensen–Shannon divergence between matched empirical regime-trajectory distributions. Crit. F1 treats oracle-defined critical regime-entry events as positives; Switch Err. is absolute first-switch timing error in planning steps. False Safe counts selected plans predicted safe by the model but violating an environment safety constraint under simulator rollout. Aggregate Value Error is the candidate-bank mean absolute error $|\hat{J}(A) - J_T(A)|$, Plan Flip marks disagreement with the simulator-best candidate, and regret is $J_T(A^*) - J_T(\hat{A})$.

DRT-JS is computed from repeated rollouts on matched (x, A) pairs with a fixed smoothing rule and is treated as a secondary distributional diagnostic.

State fidelity is measured with one shared frozen probe g_{probe} trained to map every method's latent state to the same simulator-state vector $s_t \in \mathbb{R}^{d_s}$, with $\hat{s}_{t+h} = g_{\text{probe}}(z_{t+h})$. The reported State Error is

$$E_{\text{state}} = \frac{1}{H}\sum_{h=1}^{H}\left[\frac{1}{d_s}\sum_{j=1}^{d_s}\left(\frac{\hat{s}_{t+h,j} - s_{t+h,j}}{\sigma_j^{\text{tr}} + 10^{-8}}\right)^2\right]^{1/2}. \tag{24}$$

The probe architecture and training split are identical across methods. Simulator-variable normalization, temporal aggregation, and dimensional weighting use training-split standard deviations σ_j^{tr} for each continuous simulator variable, wrapped angular differences before normalization, equal weight over state dimensions and rollout steps, and equal macro-weight over the three task/scenario strata and five paired seeds, isolating latent-state fidelity from decoder architecture.

Perturbation diagnostics. For Fig. 4, let Δx denote the perturbation expressed in the same d_s simulator-state coordinates used by the state-fidelity probe. The normalized perturbation magnitude and regime-separability score are respectively

$$m(\Delta x) = \left[\frac{1}{d_s}\sum_{j=1}^{d_s}\left(\frac{\Delta x_j}{\sigma_j^{\text{tr}} + 10^{-8}}\right)^2\right]^{1/2}, \qquad S_{\text{sep}} = \frac{\bar{d}_{\text{between}} - \bar{d}_{\text{within}}}{\bar{d}_{\text{between}} + \bar{d}_{\text{within}} + 10^{-8}}. \tag{25}$$

Both are evaluated on matched nuisance and decision-critical perturbation pairs using training-split state standard deviations for $m(\Delta x)$ and validation-whitened latent coordinates for the distance statistics, where $\bar{d}_{\text{within}}$ is the mean latent distance for matched pairs that preserve the oracle regime and $\bar{d}_{\text{between}}$ is the mean distance for matched pairs that change it. These diagnostics are descriptive and are not used to train the model.

Correlation analyses use method-by-seed-by-task/scenario-by-horizon observations. We report within-domain Spearman correlations and pooled partial correlations controlling for horizon and task/scenario identity; the main-text values are interpreted as cross-method associations rather than causal effects. Correlation uncertainty is estimated with clustered bootstrap resampling at the seed–task/scenario level.

B.4 Controlled Baseline Objectives

All controlled world-model methods share the same encoder, stochastic latent dynamics, training data, and state-predictive base objective. Let $s_t \in \mathbb{R}^{d_s}$ denote the simulator-state variables used for state-fidelity evaluation, and let

$$\bar{s}_{t,j} = \frac{s_{t,j} - \mu_j^{\text{tr}}}{\sigma_j^{\text{tr}} + 10^{-8}}$$

denote their training-set standardization after the preprocessing described in Section B.3. Starting from $z_t = E_\phi(x_t)$, the stochastic dynamics generate a reparameterized rollout $\hat{z}_{t+h} \sim p_\theta(\cdot \mid \hat{z}_{t+h-1}, a_{t+h-1})$ with $\hat{z}_t = z_t$.

The shared latent-dynamics objective is

$$\mathcal{L}_{\text{dyn}} = -\frac{1}{H}\sum_{h=1}^{H}\log p_\theta(\text{sg}[E_\phi(x_{t+h})] \mid \hat{z}_{t+h-1}, a_{t+h-1}). \tag{26}$$

To anchor the learned representation to the underlying state variables, a training-only decoder D_ω reconstructs standardized simulator state from both the encoded current state and predicted future latents:

$$\mathcal{L}_{\text{state}} = \frac{1}{(H+1)d_s}\sum_{h=0}^{H}\|D_\omega(\tilde{z}_{t+h}) - \bar{s}_{t+h}\|_2^2, \qquad \tilde{z}_t = z_t, \quad \tilde{z}_{t+h} = \hat{z}_{t+h}\ (h \geq 1). \tag{27}$$

The common base world-model objective is therefore

$$\mathcal{L}_{\text{WM}} = \mathcal{L}_{\text{dyn}} + \beta_x \mathcal{L}_{\text{state}}. \tag{28}$$

Gradients from $\mathcal{L}_{\text{state}}$ update D_ω, E_ϕ, and P_θ through $\tilde{z}_{t:t+H}$, whereas the target branch in $\mathcal{L}_{\text{dyn}}$ remains stop-gradient. Thus the common base objective anchors the latent representation to task-agnostic simulator-state information without using decision-regime labels. The decoder D_ω is used only during training and is discarded for planning. It is distinct from the frozen probe g_{probe} used to report State Error in Section B.3. Base-WM optimizes only $\mathcal{L}_{\text{WM}}$.

The controlled auxiliary objectives are

$$\mathcal{L}_{\text{VA}} = \mathcal{L}_{\text{WM}} + \lambda_V \mathcal{L}_V, \qquad \mathcal{L}_V = \mathbb{E}_{(x,A)\sim\mu_{\text{plan}}}\left[\left(\hat{J}_\theta(A) - J_T(A)\right)^2\right], \tag{29}$$

$$\mathcal{L}_{\text{BS}} = \mathcal{L}_{\text{WM}} + \lambda_B \mathcal{L}_{\text{bisim}}, \tag{30}$$

$$\begin{aligned}\mathcal{L}_{\text{bisim}} = \mathbb{E}_{i,j,a}\Big[&\|z_i - z_j\|_1 - |\mathfrak{r}_i - \mathfrak{r}_j| \\ &- \gamma W_2(P_\theta(\cdot \mid z_i, a), P_\theta(\cdot \mid z_j, a))\Big]^2.\end{aligned} \tag{31}$$

$$\mathcal{L}_{\text{Cost}} = \mathcal{L}_{\text{WM}} + \lambda_C \mathcal{L}_{\text{cost}}, \qquad \mathcal{L}_{\text{cost}} = \frac{1}{H}\sum_{h=1}^{H}\left[(\hat{c}_{t+h} - c_{t+h})^2 + \text{BCE}(\hat{v}_{t+h}, v_{t+h})\right]. \tag{32}$$

Here $\mathfrak{r}_i$ denotes the scalar environment reward at state sample i, distinguishing it from the regime label r_h used throughout the main text. Marginal-Regime-WM augments $\mathcal{L}_{\text{WM}}$ with independent per-horizon regime cross-entropies using the same oracle labels. Direct-DRT predicts the joint regime sequence directly from (x, A) and does not use latent dynamics. For Base-WM, Value-Aware-WM, BS-MPC-style, and Cost-Supervised-WM, the DRT readout is fitted after the latent model is frozen. Marginal-Regime-WM includes per-horizon regime supervision during world-model training, whereas BoundaryWorld jointly optimizes the autoregressive DRT head, encoder, and latent dynamics through $\mathcal{L}_{\text{DRT}}$. All objective weights are selected on validation seeds and then frozen for the controlled comparison.

C Full Numerical Results

The appendix reports expanded domain-, horizon-, capacity-, abstraction-, and critical-state results at finer resolution than the main text. Tables A1 and A2 disaggregate the two main domains; Tables A3 and A4 stress temporal horizon and latent capacity; Tables A5 and A6 examine abstraction and mechanism; and Tables A7 and A8 probe counterfactual transfer and low-margin states. The subsections below group results by the scientific question they address.

C.1 Task- and Scenario-Level Results

Table A1 reports Safety-Gymnasium task-level results across distinct constraint geometries. Table A2 stratifies MetaDrive by traffic criticality and shows how performance changes as small predictive differences become more consequential for the planner. The same metrics and controlled methods are used in both tables.

Table A1: Safety-Gymnasium task-level decision and planning results for the controlled methods.

Task	Method	DRT-JS↓	Crit. F1↑	Switch Err.↓	False Safe↓	Plan Flip↓	Regret↓	Viol.↓	Return↑
S1	Base-WM	0.159	0.741	1.606	0.114	0.254	0.266	0.114	78.1
	Value-Aware-WM	0.141	0.777	1.353	0.076	0.211	0.152	0.110	80.7
	BS-MPC-style	0.125	0.782	1.328	0.074	0.219	0.181	0.102	82.5
	Cost-Supervised-WM	0.142	0.809	1.430	0.055	0.248	0.257	0.028	83.0
	Marginal-Regime-WM	0.090	0.846	0.823	0.066	0.137	0.158	0.075	85.5
	Direct-DRT	0.081	0.891	0.856	0.040	0.130	0.104	0.072	87.0
	BoundaryWorld	0.064	0.898	0.505	0.017	0.094	0.101	0.039	89.1
S2	Base-WM	0.184	0.704	1.934	0.118	0.307	0.354	0.163	72.3
	Value-Aware-WM	0.162	0.750	1.524	0.110	0.204	0.232	0.133	77.6
	BS-MPC-style	0.141	0.785	1.360	0.087	0.236	0.289	0.111	79.7
	Cost-Supervised-WM	0.157	0.780	1.550	0.074	0.243	0.279	0.060	77.2
	Marginal-Regime-WM	0.123	0.829	1.119	0.072	0.180	0.206	0.092	82.5
	Direct-DRT	0.101	0.841	0.949	0.072	0.173	0.148	0.076	82.3
	BoundaryWorld	0.084	0.863	0.784	0.065	0.151	0.149	0.061	85.6
S3	Base-WM	0.185	0.682	2.220	0.143	0.297	0.394	0.179	71.9
	Value-Aware-WM	0.180	0.741	1.743	0.141	0.272	0.300	0.159	76.9
	BS-MPC-style	0.136	0.746	1.782	0.121	0.262	0.373	0.135	75.4
	Cost-Supervised-WM	0.160	0.760	1.880	0.102	0.265	0.292	0.080	75.3
	Marginal-Regime-WM	0.138	0.770	1.478	0.087	0.235	0.251	0.097	78.0
	Direct-DRT	0.115	0.815	1.375	0.089	0.225	0.291	0.104	78.8
	BoundaryWorld	0.086	0.876	0.961	0.056	0.148	0.188	0.083	81.2

Table A2: MetaDrive scenario-level decision and planning results across increasingly decision-critical traffic conditions.

Scenario	Method	DRT-JS↓	Crit. F1↑	Switch Err.↓	False Safe↓	Plan Flip↓	Regret↓	Viol.↓	Return↑
Sparse	Base-WM	0.180	0.718	1.973	0.145	0.310	0.334	0.169	76.6
	Value-Aware-WM	0.152	0.770	1.383	0.093	0.229	0.196	0.117	80.8
	BS-MPC-style	0.148	0.755	1.649	0.112	0.265	0.250	0.149	81.6
	Cost-Supervised-WM	0.155	0.768	1.518	0.077	0.269	0.271	0.041	80.9
	Marginal-Regime-WM	0.110	0.849	1.103	0.067	0.162	0.185	0.083	84.2
	Direct-DRT	0.088	0.855	0.784	0.055	0.136	0.145	0.090	86.2
	BoundaryWorld	0.067	0.875	0.605	0.027	0.105	0.109	0.050	87.4
Dense	Base-WM	0.202	0.656	2.187	0.152	0.339	0.405	0.193	69.4
	Value-Aware-WM	0.179	0.751	1.686	0.114	0.230	0.269	0.144	77.2
	BS-MPC-style	0.162	0.730	1.859	0.119	0.253	0.334	0.144	76.3
	Cost-Supervised-WM	0.177	0.731	1.806	0.083	0.298	0.334	0.075	73.7
	Marginal-Regime-WM	0.137	0.812	1.605	0.073	0.200	0.235	0.121	81.4
	Direct-DRT	0.124	0.802	1.111	0.075	0.220	0.208	0.126	79.8
	BoundaryWorld	0.091	0.850	0.991	0.063	0.160	0.167	0.084	84.8
Critical	Base-WM	0.242	0.618	2.590	0.195	0.356	0.527	0.241	65.2
	Value-Aware-WM	0.191	0.687	1.791	0.165	0.285	0.339	0.186	72.4
	BS-MPC-style	0.185	0.696	2.252	0.177	0.319	0.439	0.187	69.8
	Cost-Supervised-WM	0.211	0.745	2.286	0.128	0.339	0.391	0.100	70.4
	Marginal-Regime-WM	0.152	0.757	1.762	0.118	0.253	0.285	0.117	75.6
	Direct-DRT	0.160	0.716	1.795	0.152	0.316	0.421	0.159	71.9
	BoundaryWorld	0.112	0.831	1.284	0.084	0.182	0.276	0.115	77.4

C.2 Horizon and Capacity Stress Tests

Longer rollouts expose the temporal component of DRTS beyond one-step regime classification. Table A3 provides both sequence NLL and DRT-JS across horizons, together with regret and long-horizon safety outcomes. Table A4 reduces latent dimension for Base-WM, BS-MPC-style, and BoundaryWorld on the same fixed aggregate evaluation slice for every method; the sweep measures how trajectory-level decision supervision behaves as predictive capacity becomes scarce.

Table A3: Complete horizon sweep for regime-trajectory fidelity and downstream planning reliability.

Env.	Method	Seq. NLL↓					DRT-JS↓					Regret↓					FS@40↓	Viol.@40↓
		1	5	10	20	40	1	5	10	20	40	1	5	10	20	40		
Safety	Base-WM	0.348	0.714	0.804	1.185	1.693	0.047	0.091	0.106	0.176	0.246	0.060	0.131	0.159	0.338	0.444	0.188	0.221
Safety	Value-Aware-WM	0.326	0.421	0.786	1.045	1.685	0.045	0.059	0.147	0.161	0.247	0.049	0.084	0.149	0.228	0.306	0.158	0.184
Safety	BS-MPC-style	0.296	0.448	0.649	0.978	1.542	0.039	0.072	0.094	0.134	0.217	0.055	0.082	0.134	0.281	0.351	0.151	0.167
Safety	Marginal-Regime-WM	0.245	0.454	0.625	0.858	1.423	0.030	0.058	0.091	0.117	0.196	0.034	0.055	0.123	0.205	0.325	0.124	0.139
Safety	Direct-DRT	0.176	0.318	0.568	0.742	1.301	0.020	0.054	0.062	0.099	0.212	0.024	0.094	0.147	0.181	0.395	0.107	0.123
Safety	**BoundaryWorld**	0.197	0.314	0.413	0.632	0.791	0.024	0.044	0.063	0.078	0.106	0.029	0.064	0.075	0.146	0.182	0.071	0.082
Drive	Base-WM	0.421	0.790	0.889	1.410	2.488	0.061	0.100	0.144	0.208	0.365	0.081	0.226	0.260	0.422	0.733	0.249	0.281
Drive	Value-Aware-WM	0.346	0.710	0.817	1.095	1.747	0.046	0.109	0.120	0.174	0.209	0.054	0.170	0.217	0.268	0.453	0.196	0.229
Drive	BS-MPC-style	0.371	0.535	0.913	1.185	1.661	0.050	0.067	0.121	0.165	0.251	0.064	0.128	0.216	0.341	0.426	0.202	0.218
Drive	Marginal-Regime-WM	0.280	0.399	0.696	0.930	1.263	0.036	0.066	0.096	0.133	0.193	0.043	0.088	0.140	0.235	0.301	0.156	0.181
Drive	Direct-DRT	0.209	0.452	0.528	0.965	1.603	0.025	0.042	0.073	0.124	0.236	0.035	0.105	0.120	0.258	0.460	0.179	0.194
Drive	**BoundaryWorld**	0.231	0.402	0.495	0.745	1.059	0.029	0.047	0.071	0.090	0.136	0.039	0.095	0.084	0.184	0.230	0.095	0.109

Table A4: Latent-capacity sweep for Base-WM, BS-MPC-style, and BoundaryWorld.

Method	Dim.	State Err.↓	Seq. NLL↓	DRT-JS↓	Crit. F1↑	Plan Flip↓	Regret↓	Viol.↓	Return↑
Base-WM	512	0.073	1.146	0.165	0.718	0.278	0.321	0.153	74.6
	256	0.069	1.119	0.160	0.728	0.269	0.312	0.149	75.2
	128	0.095	1.283	0.190	0.692	0.305	0.368	0.170	72.7
	64	0.131	1.447	0.209	0.671	0.339	0.405	0.178	70.9
	32	0.180	1.672	0.243	0.629	0.394	0.501	0.222	65.7
BS-MPC-style	512	0.079	0.951	0.130	0.788	0.216	0.247	0.113	79.8
	256	0.085	0.929	0.129	0.790	0.215	0.241	0.113	79.9
	128	0.101	1.063	0.153	0.755	0.246	0.279	0.134	78.0
	64	0.136	1.214	0.171	0.731	0.286	0.304	0.146	75.8
	32	0.187	1.352	0.191	0.695	0.305	0.369	0.164	72.3
BoundaryWorld	512	0.086	0.589	0.067	0.895	0.119	0.131	0.056	86.0
	256	0.096	0.601	0.074	0.884	0.129	0.134	0.060	85.6
	128	0.108	0.677	0.082	0.869	0.139	0.153	0.066	84.4
	64	0.141	0.738	0.086	0.852	0.153	0.159	0.070	83.1
	32	0.195	0.776	0.099	0.821	0.166	0.181	0.079	81.0

C.3 Abstraction and Mechanism

Table A5 examines the abstraction side of the regret bound. Coarse, medium, and fine signatures are compared using held-out support-aware residual estimates, sequence fidelity, and downstream regret. Table A6 then isolates the mechanism on the same fixed aggregate evaluation slice: joint versus marginal trajectory supervision, dependence on predicted latent rollouts, gradient flow into dynamics, oracle-head error, and the necessity of latent dynamics. Figures 4–5 provide the perturbation and granularity/mechanism visualizations.

Table A5: Regime-granularity sweep with support-aware abstraction-residual estimation.

Env.	Signature	Min supp.	$\hat{\eta}_D \downarrow$	DRT-JS↓	Crit. F1↑	Switch Err.↓	Plan Flip↓	Regret↓	Viol.↓	Return↑
Safety	Coarse	417	0.241	0.064	0.845	0.940	0.147	0.164	0.078	83.4
Safety	Medium	223	0.134	0.075	0.882	0.780	0.126	0.133	0.059	85.5
Safety	Fine	97	0.082	0.091	0.879	0.810	0.123	0.140	0.064	84.9
Drive	Coarse	506	0.281	0.072	0.818	1.110	0.179	0.209	0.102	81.2
Drive	Medium	263	0.159	0.086	0.858	0.940	0.151	0.170	0.071	83.4
Drive	Fine	113	0.095	0.105	0.862	0.930	0.158	0.180	0.078	82.9

Table A6: Mechanism study for sequence structure, latent rollout dependence, gradient flow, and regime supervision.

Variant	State Err.↓	Seq. NLL↓	DRT-JS↓	Crit. F1↑	Switch Err.↓	False Safe↓	Plan Flip↓	Regret↓	Viol.↓	Return↑
BoundaryWorld	0.108	0.671	0.081	0.870	0.860	0.052	0.138	0.152	0.065	84.5
$-\mathcal{L}_{\mathrm{DRT}}$	0.095	1.248	0.189	0.716	1.870	0.126	0.276	0.332	0.154	74.4
Marginal regime labels	0.104	0.979	0.128	0.866	1.537	0.093	0.197	0.206	0.096	80.6
Head bypasses rollout	0.099	0.747	0.174	0.792	1.598	0.102	0.216	0.265	0.111	80.0
No dynamics gradient	0.097	0.949	0.109	0.853	1.122	0.062	0.216	0.259	0.106	80.0
Oracle regime head (upper bound)	0.109	0.423	0.035	0.947	0.455	0.021	0.069	0.074	0.041	87.6
Direct-DRT	N/A	0.720	0.089	0.863	0.997	0.079	0.216	0.230	0.090	81.8
Non-autoregressive head	0.106	0.710	0.129	0.804	1.556	0.053	0.179	0.200	0.079	82.8

C.4 Counterfactual Transfer and Critical States

Table A7 compares Direct-DRT with BoundaryWorld on the fixed aggregate evaluation slice under unseen action composition, longer rollouts, and nearby perturbations of the downstream decision problem. Reward and constraint shifts modify the task-specific specification used to define D, and oracle evaluation labels are recomputed under the shifted problem. Under these shifted decision problems, the learned encoder, latent dynamics, and regime head remain frozen; only the task specification used by the oracle and the regime-measurable planning readout is changed, so no model parameters are updated on shifted test conditions. These rows quantify sensitivity to local changes in D, while the action-composition and horizon rows probe predictive extrapolation. Table A8 stratifies the same aggregate evaluation states by decision-margin quantile. It includes Base-WM, Value-Aware-WM, BS-MPC-style, and BoundaryWorld, placing low-margin behavior against both state- and value-oriented alternatives.

Table A7: Counterfactual action/horizon stress tests and sensitivity to nearby reward and constraint shifts.

Shift	Method	Seq. NLL↓	DRT-JS↓	Crit. F1↑	Plan Flip↓	Regret↓	Viol.↓	Return↑
Unseen action composition	Direct-DRT	1.080	0.154	0.759	0.251	0.302	0.133	77.1
	BoundaryWorld	0.780	0.104	0.814	0.201	0.232	0.113	80.1
Longer rollout horizon	Direct-DRT	1.460	0.188	0.721	0.289	0.356	0.157	75.2
	BoundaryWorld	0.940	0.128	0.791	0.189	0.211	0.127	80.7
Reward $+10\%$	Direct-DRT	0.700	0.086	0.857	0.156	0.184	0.081	83.0
	BoundaryWorld	0.730	0.089	0.851	0.151	0.174	0.076	83.4
Reward -10%	Direct-DRT	0.920	0.125	0.804	0.211	0.243	0.109	79.9
	BoundaryWorld	0.870	0.115	0.809	0.181	0.163	0.094	82.7
Constraint $+20\%$	Direct-DRT	1.010	0.144	0.781	0.229	0.269	0.127	78.7
	BoundaryWorld	0.920	0.129	0.801	0.199	0.229	0.067	80.5

Table A8: Critical-state stratification by decision-margin quantile.

Margin bin	Method	State Err.↓	DRT-JS↓	Crit. F1↑	False Safe↓	Plan Flip↓	Regret↓	Viol.↓
Q1: lowest	Base-WM	0.092	0.251	0.601	0.224	0.401	0.518	0.259
	Value-Aware-WM	0.088	0.221	0.644	0.195	0.358	0.456	0.229
	BS-MPC-style	0.094	0.201	0.677	0.174	0.333	0.423	0.207
	BoundaryWorld	0.103	0.108	0.820	0.081	0.184	0.206	0.101
Q2	Base-WM	0.086	0.213	0.680	0.166	0.318	0.408	0.188
	Value-Aware-WM	0.093	0.185	0.697	0.126	0.292	0.327	0.164
	BS-MPC-style	0.090	0.168	0.742	0.127	0.267	0.346	0.149
	BoundaryWorld	0.100	0.088	0.862	0.057	0.157	0.163	0.073
Q3	Base-WM	0.094	0.158	0.720	0.099	0.253	0.267	0.135
	Value-Aware-WM	0.090	0.153	0.773	0.133	0.221	0.284	0.125
	BS-MPC-style	0.098	0.128	0.786	0.077	0.199	0.209	0.153
	BoundaryWorld	0.107	0.092	0.844	0.065	0.108	0.126	0.062
Q4: highest	Base-WM	0.089	0.121	0.815	0.067	0.170	0.192	0.089
	Value-Aware-WM	0.096	0.099	0.829	0.050	0.146	0.158	0.069
	BS-MPC-style	0.095	0.101	0.853	0.052	0.142	0.154	0.075
	BoundaryWorld	0.104	0.060	0.918	0.026	0.082	0.084	0.039

D Implementation and Statistical Reporting

All principal simulator comparisons use five paired training/test seeds. Rare-event outcomes such as false-safe rate use episode-level bootstrap confidence intervals in addition to seed-level summaries. Hyperparameters are selected on validation seeds and frozen before the test suite is evaluated. Table A9 summarizes protocol invariants shared across the controlled comparison.

World-model architecture. All controlled world-model methods use the same encoder, stochastic latent-dynamics model, and training-only state decoder. The encoder is a three-layer MLP with hidden widths 256–256 and a linear latent projection, with main-run latent dimension 128. The stochastic transition $P_\theta(z_{t+1} \mid z_t, a_t)$ is a diagonal-Gaussian residual model whose mean increment and clipped log-standard-deviation are predicted by a two-layer 256-unit MLP from $[z_t, a_t]$. The state decoder D_ω is a two-layer MLP with hidden width 256 and a linear output layer over the standardized simulator-state variables. It is optimized only through $\mathcal{L}_{\text{state}}$ and is discarded after world-model training. The autoregressive DRT head q_ψ uses a one-layer GRU with hidden size 128 followed by a linear softmax head over regime labels at each step. Shared activation, normalization, and regularization choices are SiLU activations, LayerNorm after hidden layers, no dropout, and no explicit weight decay. The training-only decoder D_ω is distinct from the frozen state probe g_{probe} used for State Error evaluation.

Table A9: Protocol invariants shared across the controlled comparison.

Setting	Value
Paired seeds	5
Main planning horizon	$H = 20$
Horizon sweep	$\{1, 5, 10, 20, 40\}$
Latent-capacity sweep	$\{32, 64, 128, 256, 512\}$
Counterfactual proposals	Shared frozen reference-planner buffer
CEM population / elite fraction	512 / 12.5% (64 elites)
K_z / K_{eval}	16 / 32
Replay size / training steps / optimizer	1.0×10^6 / 5.0×10^5 / Adam (3×10^{-4})

Per-seed outputs support the state-probe, clustered-correlation, regime-construction, and residual analyses. Run-level optimization settings are shared across the controlled comparison. The remaining run-level settings not listed in Table A9 are batch size 256, CEM refinement iterations 5, base

state-reconstruction weight $\beta_x = 1.0$, and auxiliary-loss weights $\lambda = 1.0$, $\lambda_V = 0.5$, $\lambda_B = 0.1$, and $\lambda_C = 0.5$ in the final runs.