How Far Should an Agent Imagine? Decision-Stable Planning with Learned World Models

Anonymous authors
Paper under double-blind review

Abstract

Model-based reinforcement learning uses learned world models to evaluate future consequences, but deeper imagination couples delayed decision information with accumulating model error. Existing adaptive-horizon methods commonly use predictive uncertainty or reliability to regulate model use. The same uncertainty level can imply different planning needs when action margins or reward delays differ. We introduce MarginPlan, which selects planning depth according to decision stability. For each root action, it constructs calibrated return intervals over plausible dynamics models and certifies stability when the preferred action's lower return bound exceeds the upper bounds of every alternative. If the certificate remains open, a continuation predictor estimates whether another imagination block will reduce action ambiguity enough to offset added interval uncertainty and model-call cost. Controlled experiments independently vary uncertainty, action-value margins, delayed evidence, error correlation, and candidate-set size, including uncertainty-scale-matched pairs and approximately computation-matched baselines. MarginPlan makes planning depth depend on whether additional lookahead can change the action decision, jointly accounting for predictive uncertainty and decision geometry.

Keywords: model-based reinforcement learning; world models; adaptive planning; uncertainty quantification; decision stability

1 Introduction

Learned world models let an agent evaluate consequences before acting. They support visual control through latent imagination (Hafner et al., 2025), scalable continuous control (Hansen et al., 2024), offline world-action pretraining (Cheng et al., 2025), and zero-shot physical planning (Zhou et al., 2025). Their planning horizon remains consequential: longer rollouts can reveal delayed evidence, while additional model steps can amplify approximation error (Ramasubramanian et al., 2025).

Recent work improves both model prediction and model use. Transformer world models extend temporal representation (Burchi and Timofte, 2025), time-aware models vary prediction scale (Nhu et al., 2025), latent-action models improve transfer (Gao et al., 2025), and continual planners update models online (Liu et al., 2025). MACURA adapts rollout use from model uncertainty (Frauenknecht et al., 2024), while dynamic-horizon value expansion selects horizons from reliability estimates (Wang et al., 2024). Bayesian offline planning (Jeong et al., 2025) and robust value-aware modeling (Qiao et al., 2026) incorporate uncertainty into planning or model learning. These approaches motivate adaptive model use, but they do not make action-ranking stability the horizon criterion.

Predictive uncertainty does not determine a decision by itself. Consider two states with the same return uncertainty. When the leading action is separated by a wide margin, the decision can remain unchanged across plausible return perturbations; when the top actions are nearly tied, the same uncertainty can leave the ranking unresolved. Delayed reward creates a second source of depth dependence: a low-uncertainty shallow rollout may still omit decisive evidence. The useful horizon therefore depends jointly on uncertainty, action geometry, and delayed information (Voelcker et al., 2025; Elmachtoub and Grigas, 2021; Wilder et al., 2019).

Confidence-based decision stopping provides a useful formal analogy. Best-arm identification allocates additional samples until the preferred arm is separated with sufficient evidence (Garivier

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.

and Kaufmann, 2016). In world-model planning, depth changes the information source itself: further propagation can reveal delayed reward structure while accumulating model error. Robust search incorporates uncertainty into action evaluation (Dam et al., 2025), and value expansion exhibits a horizon-dependent bias–variance tradeoff (Feinberg et al., 2018). Planning depth therefore controls how far an imperfect model should be propagated before selecting an action.

We propose MARGINPLAN, a decision-stable adaptive planner. For each root action, a common continuation policy generates imagined futures under a family of plausible world models. We apply split-conformal calibration to a maximum normalized residual taken jointly over candidate actions and all checked horizons (Lei et al., 2018; Romano et al., 2019). If the current leader's lower bound exceeds every competitor's upper bound, its identity is invariant over the represented uncertainty set and planning stops. This certificate uses uncertainty only through its effect on action ordering.

When the certificate remains open, MARGINPLAN predicts whether one more imagination block will reduce action ambiguity. A lightweight continuation head is trained from deeper rollouts but uses only statistics available at the current depth at deployment. The planner continues when predicted ambiguity reduction exceeds the increase in interval uncertainty plus model-call cost. This creates a state-dependent rule that can stop early under moderate uncertainty when the action margin is wide, yet continue under smaller uncertainty when the decision remains unresolved.

We test this mechanism with shared-prediction controlled experiments that vary action margin, model uncertainty, evidence delay, correlation in model error, and candidate-set size independently. The design includes uncertainty-scale-matched controlled pairs, approximately computation-matched adaptive baselines, calibration and false-certification measurements, component ablations, and scalability tests. Our contributions are:

1. **We formulate adaptive imagination as decision stability.** The stopping certificate depends on return uncertainty and action separation jointly, and its guarantee is stated under simultaneous action-by-horizon coverage so it remains valid for adaptive stopping.
2. **We introduce decision-aware continuation.** A learned predictor estimates the benefit of deeper imagination from current-depth information, and continuation trades predicted ambiguity reduction against added uncertainty and computation.
3. **We isolate these effects with matched controlled tests.** A fixed injected uncertainty scale is paired with different margins or evidence delays, and all adaptive rules share the same return predictions so that differences primarily reflect the horizon-selection rule rather than the underlying return predictor.

2 RELATED WORK

World models and adaptive horizons. Latent dynamics models established planning from pixels (Hafner et al., 2019), recurrent imagination made long synthetic rollouts practical (Ha and Schmidhuber, 2018; Hafner et al., 2020), and MuZero learned decision-oriented latent dynamics (Schrittwieser et al., 2020). TD-MPC and TD-MPC2 couple compact latent models with trajectory optimization (Hansen et al., 2022; 2024). MVE and STEVE analyze or mix rollout horizons (Feinberg et al., 2018; Buckman et al., 2018); MBPO limits synthetic rollout length (Janner et al., 2019); MACURA and DMVE adapt model use from uncertainty or reliability (Frauenknecht et al., 2024; Wang et al., 2024). **Decision-aware uncertainty.** Ensembles estimate epistemic spread (Lakshminarayanan et al., 2017; Chua et al., 2018), while calibration aligns predictive intervals with observed errors (Kuleshov et al., 2018; Romano et al., 2019). Decision-focused learning instead judges errors through downstream choices (Wilder et al., 2019; Elmachtoub and Grigas, 2021), and recent value-aware models connect model error to control value (Voelcker et al., 2025; Qiao et al., 2026). **Confidence-based stopping.** Fixed-confidence best-arm identification stops when evidence separates a preferred arm from competitors (Garivier and Kaufmann, 2016). MARGINPLAN adopts interval separation for action identification while treating rollout depth as a coupled source of information and model error. The continuation rule evaluates this depth-dependent tradeoff directly. Unlike uncertainty- or reliability-based horizon rules, MARGINPLAN does not ask whether the model is sufficiently confident in isolation. It asks whether the preferred root action is invariant under calibrated return uncertainty; when it is not, the planner estimates whether one additional rollout block is expected to reduce decision ambiguity enough to justify its added uncertainty and computation.

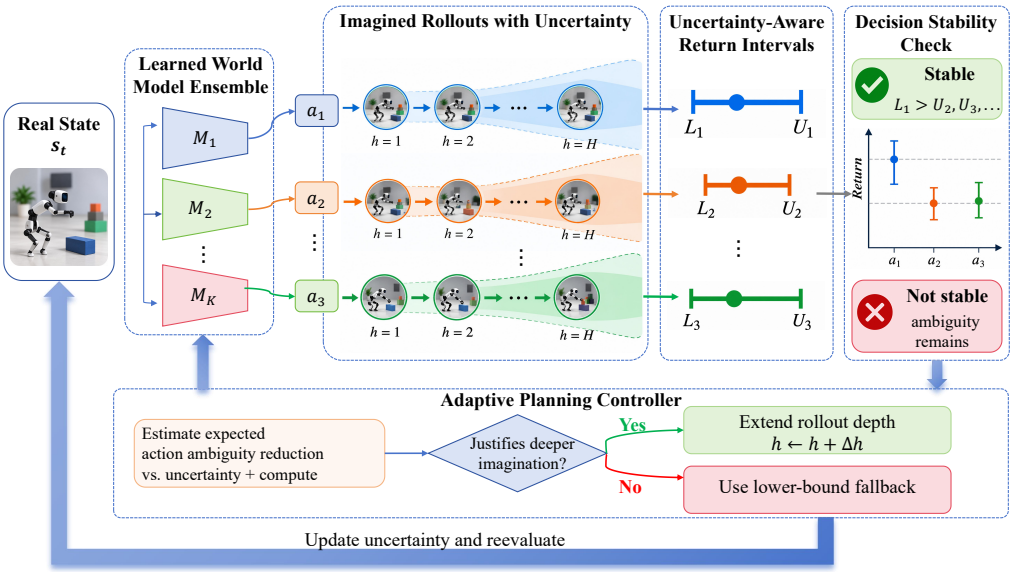


Figure 1: MarginPlan forms calibrated return intervals, certifies stable action rankings, and extends unresolved rollouts only when predicted ambiguity reduction justifies added interval uncertainty and model-call cost.

3 Problem Setup

We consider a discounted MDP $\mathcal{P} = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$ (Bellman, 1957; Sutton and Barto, 2018). At state s_t, the planner evaluates a finite root-action set $\mathcal{C}_t \subseteq \mathcal{A}$. The first imagined action is fixed to candidate $a \in \mathcal{C}_t$; all later imagined actions follow the same continuation policy π. A family of M plausible world models $\mathcal{M}_t = \{\theta^{(m)}\}_{m=1}^{M}$ represents epistemic uncertainty. At depth h, model m yields

$$G_h^{(m)}(s_t, a; \pi) = \sum_{k=0}^{h-1} \gamma^k \hat{r}_{t+k}^{(m)} + \gamma^h \hat{V}_\psi^{(m)}(\hat{s}_{t+h}^{(m)}), \tag{1}$$

where $a_t = a$, $a_{t+k} \sim \pi(\cdot \mid \hat{s}_{t+k}^{(m)})$ for $k > 0$, and $\hat{V}_\psi^{(m)}$ is the terminal value estimate associated with model sample m. The goal is to choose both a_t and a state-dependent depth $h_t \leq H_{\max}$ while limiting further model propagation when additional imagination is not expected to improve the action decision.

4 MarginPlan: Decision-Stable Planning

4.1 Overview

Figure 1 summarizes the planning loop. Candidate root actions share a continuation policy and are propagated through the same world-model family. Return samples form calibrated intervals; interval dominance yields a stability certificate, while unresolved decisions enter the adaptive controller. The controller either extends cached rollout prefixes by Δ_h steps or executes the lower-bound fallback action when another block is not worthwhile. Compute accounting counts only newly generated model transitions.

4.2 Simultaneous Return Intervals

For each action and depth, let $\bar{G}_h(a)$ and $s_h(a)$ be the ensemble mean and standard deviation of $G_h^{(m)}(a)$. We calibrate the same h-step bootstrapped return that appears in the stopping rule. For

held-out decision i, define the normalized maximum residual

$$R_i = \max_{a \in \mathcal{C}_i,\, h \in \mathcal{H}} \frac{|Q_i(a) - \bar{G}_{i,h}(a)|}{s_{i,h}(a) + \epsilon}, \qquad \mathcal{H} = \{1, 1 + \Delta_h, \ldots, H_{\max}\}, \tag{2}$$

where $Q_i(a)$ is the held-out environment return target. Let $q_{1-\alpha}$ be the finite-sample conformal quantile of $\{R_i\}$. With $N_{\text{cal}} = 10{,}000$ and target coverage 0.90, each seed uses order-statistic rank $\lceil (N_{\text{cal}} + 1)(1 - \alpha) \rceil = 9001$; reported q values are means of the five seed-specific multipliers. We use

$$L_h(a) = \bar{G}_h(a) - q_{1-\alpha}(s_h(a) + \epsilon), \quad U_h(a) = \bar{G}_h(a) + q_{1-\alpha}(s_h(a) + \epsilon). \tag{3}$$

Under exchangeable calibration and test planning decisions, this construction targets simultaneous coverage over candidates and all checked horizons. Here a planning decision includes the procedure that constructs $\mathcal{C}_i$; calibration and test decisions use the same candidate-set construction, with conditional or stratified calibration when that construction changes. The maximum over (a, h) is crucial because the algorithm itself chooses the stopping horizon.

4.3 Decision-Stability Certificate

Let $a_h^\star = \arg\max_a \bar{G}_h(a)$ and define the signed stability margin

$$\Gamma_h = L_h(a_h^\star) - \max_{b \neq a_h^\star} U_h(b). \tag{4}$$

Planning is certified when $\Gamma_h > 0$. The certificate can therefore close under nontrivial uncertainty if the action margin is sufficiently wide, while remaining open at low uncertainty when two actions are nearly tied.

Proposition 1 (adaptive decision stability). Let $\tau \in \mathcal{H}$ be any stopping time produced from the interval sequence, define $\mathsf{Cert} = \{\Gamma_\tau > 0\}$ and $\mathsf{Wrong} = \{a_\tau^\star \notin \arg\max_a Q(a)\}$, and suppose

$$\Pr\left(\bigcap_{h \in \mathcal{H}} \bigcap_{a \in \mathcal{C}_t} \{Q(a) \in [L_h(a), U_h(a)]\} \right) \geq 1 - \alpha. \tag{5}$$

Then the probability of issuing an incorrect positive certificate is bounded by

$$\Pr(\mathsf{Cert} \cap \mathsf{Wrong}) \leq \alpha. \tag{6}$$

Equation (6) controls the marginal false-certificate probability. Conditional certificate precision is evaluated empirically in Section 5.

Proof sketch. On the simultaneous event in Eq. (5), coverage holds at every horizon the algorithm could select. At the realized τ, any positive certificate gives $Q(a_\tau^\star) \geq L_\tau(a_\tau^\star) > U_\tau(b) \geq Q(b)$ for every competitor b. Hence $\mathsf{Cert} \cap \mathsf{Wrong}$ can occur only when the simultaneous coverage event fails, whose probability is at most α. A full statement is given in Appendix B. □

4.4 Value of Deeper Imagination

When $\Gamma_h \leq 0$, define action ambiguity and average interval width as

$$A_h = \left[\max_{b \neq a_h^\star} U_h(b) - L_h(a_h^\star) \right]_+, \qquad W_h = |\mathcal{C}_t|^{-1} \sum_a (U_h(a) - L_h(a)). \tag{7}$$

A lightweight predictor f_ω maps current-depth statistics $\phi_h = (h/H_{\max}, A_h, W_h, \Gamma_h$, top-two mean gap, dispersion summaries) to $(\widehat{A}_{h+\Delta_h}, \widehat{W}_{h+\Delta_h})$. During training, actual deeper rollouts provide targets through

$$\mathcal{L}_{\text{cont}} = (\widehat{A}_{h+\Delta_h} - A_{h+\Delta_h})^2 + (\widehat{W}_{h+\Delta_h} - W_{h+\Delta_h})^2 + \xi \|\omega\|_2^2, \tag{8}$$

where ξ is the ridge coefficient. At deployment, f_ω receives only quantities available at depth h. We define

$$I_h = [A_h - \widehat{A}_{h+\Delta_h}]_+ + \eta [W_h - \widehat{W}_{h+\Delta_h}]_+, \tag{9}$$

$$C_h = [\widehat{W}_{h+\Delta_h} - W_h]_+, \tag{10}$$

Algorithm 1 Decision-stable adaptive imagination

Input: $s_t, \mathcal{C}_t, \mathcal{M}_t, \pi, H_{\max}, \Delta_h$

initialize and cache one rollout prefix per (a, m)
for $h = 1, 1 + \Delta_h, \ldots, H_{\max}$ **do**
  extend cached prefixes to h and compute $[L_h(a), U_h(a)]$
  $a_h^\star \leftarrow \arg\max_a \bar{G}_h(a)$; compute Γ_h
  if $\Gamma_h > 0$ **then**
    return $a_h^\star, h$
  end if
  if $h + \Delta_h > H_{\max}$ **then**
    return $\arg\max_a L_h(a), h$
  end if
  predict $(\widehat{A}_{h+\Delta_h}, \widehat{W}_{h+\Delta_h})$ from ϕ_h
  if Eq. (11) fails **then**
    return $\arg\max_a L_h(a), h$
  end if
end for
return $\arg\max_a L_{H_{\max}}(a), H_{\max}$

with incremental model calls $\Delta N_h = M|\mathcal{C}_t|\Delta_h$. The rollout continues when

$$I_h > \lambda C_h + \kappa_{\text{comp}} \Delta N_h, \tag{11}$$

where κ_{comp} converts one model call to the normalized return scale. Otherwise the planner executes $\arg\max_a L_h(a)$. The $\eta = 0$ sensitivity in Appendix Table 14 tests whether explicit interval shrinkage contributes beyond ambiguity itself.

5 Experiments

5.1 Protocol and Questions

We use a shared-prediction controlled testbed to isolate horizon selection. Each root action has a known full return and an evidence delay. Shallow imagination retains terminal ambiguity about delayed effects; deeper imagination reduces that ambiguity while adding transition-model error. We vary the top-two gap $g \in [0.04, 0.50]$, uncertainty scale $\sigma \in [0.04, 0.34]$, delay, cross-member Gaussian-error correlation, and $|\mathcal{C}_t|$. Unless noted, $M = 24$, $H_{\max} = 12$, $\Delta_h = 1$, and the default cross-member Gaussian-error correlation is $\rho = 0.50$. All horizon rules consume the same cached return predictions. We ask: **Q1** does the selected depth depend on decision geometry at a fixed injected uncertainty scale; **Q2** does the gain persist at approximately matched computation; **Q3** which components are necessary; and **Q4** does certification remain reliable as uncertainty structure or candidate-set construction changes?

Metrics and statistical protocol. For controlled decision i, let $\mathcal{A}_i^\star$ denote the set of candidates maximizing the full-return target Q_i, and let $\hat{a}_i$ be the executed action. We report

$$\text{Acc} = N^{-1} \sum_i \mathbf{1}\{\hat{a}_i \in \mathcal{A}_i^\star\},$$

$$\text{Regret} = N^{-1} \sum_i [\max_a Q_i(a) - Q_i(\hat{a}_i)].$$

Returns use the simulator's normalized scale; ties within 10^{-8} are counted as correct and have zero regret. For certification, $\text{FalseCert} = \Pr(\mathsf{Cert} \cap \mathsf{Wrong})$ and $\text{Precision} = 1 - \text{FalseCert}/\Pr(\mathsf{Cert})$ when the certification rate is nonzero. Controlled results use five paired seeds; per seed we use 60,000 training decisions, 10,000 calibration decisions, and 20,000 disjoint test decisions. The split-conformal order statistic is therefore computed from $N_{\text{cal}} = 10{,}000$ decisions per seed. Table means are over the same five paired seeds; Appendix Table 11 reports seed-level variability for accuracy, regret, and selected depth in the main comparison.

Adaptive baselines. All baselines share cached rollouts and are tuned only on the held-out validation grid. We define

$$u_h = |\mathcal{C}_t|^{-1} \sum_a s_h(a),$$

$$r_h = |\mathcal{C}_t|^{-1} \sum_a \frac{|\bar{G}_h(a) - \bar{G}_{h-\Delta_h}(a)|}{s_h(a) + s_{h-\Delta_h}(a) + \epsilon}.$$

Uncertainty-stop terminates when $u_h < \tau_u = 0.112$. Reliability-stop is first eligible at $h = 1 + \Delta_h$, once both h and $h - \Delta_h$ statistics exist, and then terminates when $r_h < \tau_r = 0.46$. Consistency-stop terminates once the ensemble-mean argmax remains unchanged for $K_c = 3$ consecutive checked depths. None uses calibrated action-separation intervals for stopping. For the approximately computation-matched comparison, each rule is tuned on validation data toward $B = 576$ model calls; the selected settings are reported in Appendix Table 8 and are then applied to the test split.

Across the shared-prediction benchmark, fixed horizons trace the expected compute–error trade-off, while the adaptive baselines respond to uncertainty, return reliability, or action-identity persistence. MARGINPLAN attains an overall accuracy of .944 with .014 regret and 518 model calls. More importantly, its selected depth changes from 3.6 on wide-gap decisions to 7.5 on narrow-gap decisions, whereas the uncertainty-only rule remains near six steps in both subsets. This pattern is consistent with planning depth depending on decision geometry in addition to predictive uncertainty (Table 1 and Figure 2).

Table 1: Shared-prediction results overall and under wide- and narrow-margin decisions.

	Overall				Wide gap $g \geq 0.30$			Narrow gap $g \leq 0.12$		
Method	Acc.	Regret	Depth	Calls	Acc.	Depth	Calls	Acc.	Depth	Calls
Fixed-2	.842	.050	2.0	192	.944	2.0	192	.731	2.0	192
Fixed-4	.889	.027	4.0	384	.963	4.0	384	.806	4.0	384
Fixed-8	.931	.017	8.0	768	.974	8.0	768	.884	8.0	768
Fixed-12	.918	.029	12.0	1152	.968	12.0	1152	.861	12.0	1152
Uncertainty-stop	.920	.024	5.9	566	.949	5.8	557	.876	6.0	576
Reliability-stop	.923	.019	5.6	538	.952	5.3	509	.881	5.7	547
Consistency-stop	.930	.022	5.7	547	.957	4.8	461	.889	6.5	624
MARGINPLAN	.944	.014	5.4	518	.969	3.6	346	.895	7.5	720

5.2 DECISION GEOMETRY AND MATCHED COMPUTE

Holding the injected uncertainty scale fixed reveals a strong dependence on decision geometry. At $\sigma = .16$ and delay 7, narrowing the action gap from .42 to .08 reduces the certification rate from .744 to .286 and increases selected depth from 2.9 to 8.0. Holding the gap at .16 instead, delaying evidence from 3 to 10 steps reduces certification from .612 to .331 and increases depth from 3.8 to 7.2. These paired comparisons separate the need for deeper planning from the magnitude of the injected uncertainty scale itself (Table 2 and Figure 3).

Table 2: Uncertainty-scale-matched pairs isolate margin and evidence-delay effects.

Case	σ	Gap	Delay	Acc.	Regret	Depth	Calls	Cert.
Wide margin	.16	.42	7	.978	.006	2.9	278	.744
Narrow margin	.16	.08	7	.889	.017	8.0	768	.286
Early evidence	.16	.16	3	.961	.010	3.8	365	.612
Delayed evidence	.16	.16	10	.909	.019	7.2	691	.331

The improvement is not explained by additional computation. After validation tuning toward $B = 576$ model calls, all realized budgets remain within 19 calls (3.3%) of the target. At this approximately matched compute level, MARGINPLAN reaches .947 accuracy with .013 regret, compared with .925–.934 accuracy and .017–.024 regret for the adaptive baselines (Table 3).

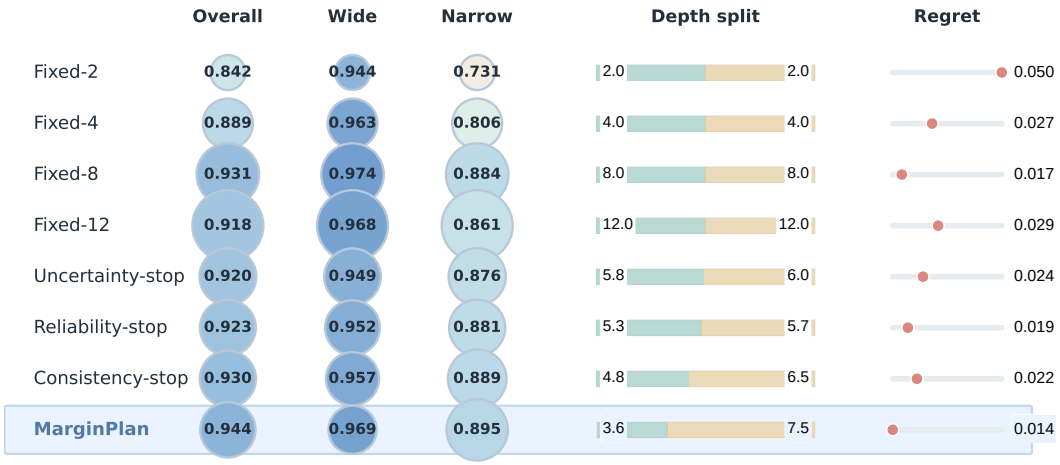


Figure 2: Shared-prediction comparison across overall, wide-gap, and narrow-gap settings. Bubble labels show accuracy, bubble area reflects model calls, split bars show selected depth, and red markers show overall regret.

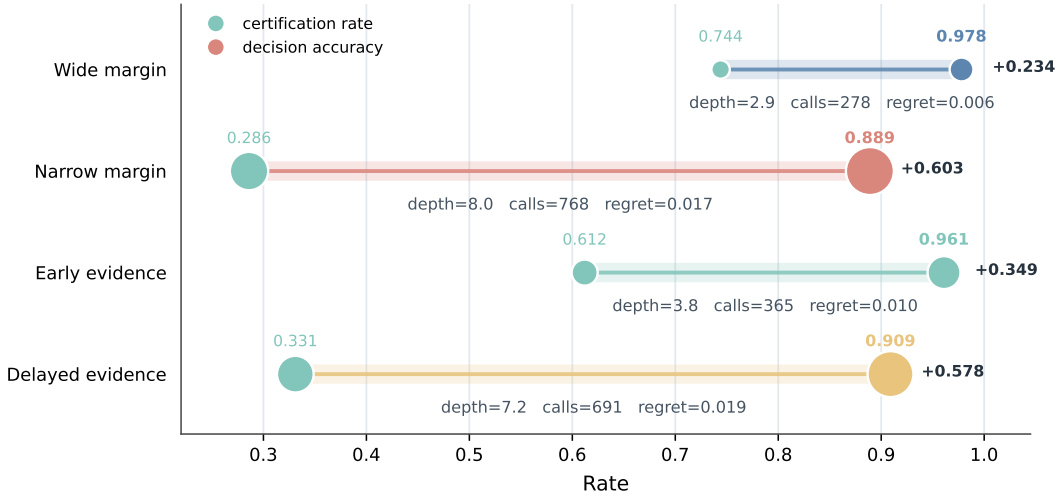


Figure 3: Uncertainty-scale-matched comparisons. Markers show certification and accuracy; size reflects calls, with depth and regret annotated. The annotated difference is accuracy minus certification rate. Narrow margins and delayed evidence increase planning depth.

Table 3: Approximately computation-matched comparison after validation-budget tuning.

Method	Acc.	Regret	Depth	Calls	ΔCalls
Fixed-6	.913	.031	6.0	576	0
Uncertainty-stop	.925	.020	5.9	566	-10
Reliability-stop	.928	.024	6.1	586	+10
Consistency-stop	.934	.017	5.8	557	-19
MarginPlan	.947	.013	6.0	576	0

5.3 Reliability, Ablations, and Scalability

Ablations separate the roles of calibration, certification, and continuation. Removing the ambiguity-gain term reduces accuracy from .944 to .923, while disabling the certificate increases mean depth from 5.4 to 6.3 and model calls from 518 to 605. Removing the width-increase cost pushes planning to a mean depth of 7.2 and 691 calls. These changes show that the certificate and continuation

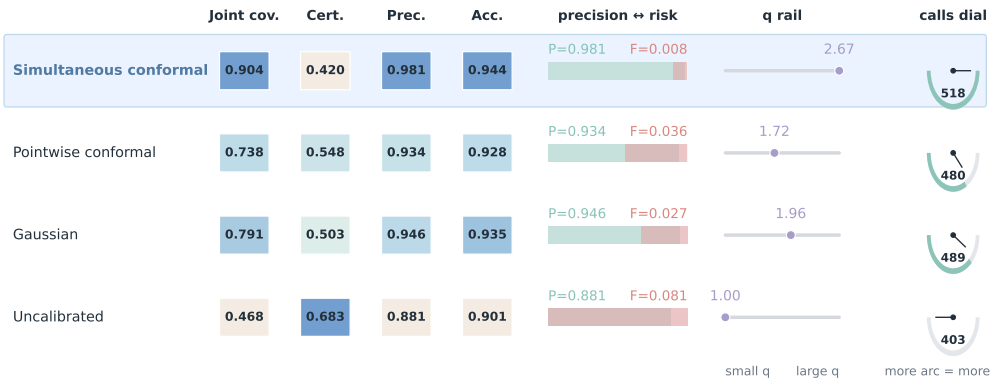


Figure 4: Calibration and certification across interval schemes. Tiles summarize coverage, certification, precision, accuracy, false-certification risk, calibration multiplier, and model calls. In the precision-risk bar, P denotes conditional precision and F denotes the marginal false-certification rate.

terms regulate complementary parts of the decision–compute trade-off (Table 4). The "Uncalibrated ($q = 1$)" variant uses the raw interval multiplier reported in Table 5; Pointwise calibration uses the corresponding pointwise conformal intervals. "No certificate" disables the $\Gamma_h > 0$ early exit, "No ambiguity gain" removes the first term of I_h, and "No width-increase cost" sets $\lambda = 0$ while retaining the model-call penalty. "Mean-only stop" uses the validation-tuned mean-gap convergence test defined in Appendix A, with τ_m reported in Table 7. The continuation predictor is evaluated in Appendix C; the $\eta = 0$ variant is reported in Appendix Table 14.

Table 4: Component ablations of calibration, certification, and continuation.

Variant	Acc.	Regret	Depth	Calls
Full MarginPlan	.944	.014	5.4	518
Uncalibrated ($q = 1$)	.901	.033	4.2	403
Pointwise calibration	.928	.020	5.0	480
No certificate	.938	.018	6.3	605
No ambiguity gain	.923	.026	5.0	480
No width-increase cost	.941	.022	7.2	691
Mean-only stop	.910	.031	4.4	422

Calibration quality directly controls the reliability of early certification. Simultaneous conformal calibration attains .904 joint coverage with a marginal false-certification rate of .008, whereas pointwise, Gaussian, and uncalibrated intervals provide lower joint coverage and more aggressive certification; the uncalibrated variant reaches a false-certification rate of .081 (Figure 4 and Table 5). False cert. denotes $\Pr(\mathsf{Cert} \cap \mathsf{Wrong})$, while Precision is the empirical accuracy among certified decisions. Conformal multipliers are estimated from their designated calibration splits; Gaussian and uncalibrated variants use the fixed multipliers shown in the q column. Across the five paired seeds, the simultaneous-conformal row has seed s.d. $(.006, .014, .002, .005, .008, 27)$ for (Joint cov., Cert., False cert., Prec., Acc., Calls), respectively.

Table 5: Interval calibration and certificate reliability under correlated model error.

Scheme	q	Joint cov.	Cert.	False cert.	Prec.	Acc.	Calls
Simultaneous conformal	2.67	.904	.420	.008	.981	.944	518
Pointwise conformal	1.72	.738	.548	.036	.934	.928	480
Gaussian	1.96	.791	.503	.027	.946	.935	489
Uncalibrated	1.00	.468	.683	.081	.881	.901	403

Larger candidate sets make decision separation progressively harder and increase planning cost. As $|\mathcal{C}_t|$ grows from 2 to 32, MarginPlan accuracy decreases from .964 to .889 while mean depth rises from 4.3 to 7.9; the corresponding model-call count grows from 206 to 6067. Because the calibration

score maximizes over candidates, each scaling condition is calibrated under its corresponding candidate-generation rule. Appendix Tables 13 and 15 further evaluate certificate behavior under correlation shift and candidate-set scaling; in the latter, certification decreases with candidate-set size while joint coverage remains close to the target (Table 6).

Table 6: Candidate-set scaling for MarginPlan and the uncertainty- and reliability-based adaptive baselines.

	MarginPlan				Uncertainty-stop				Reliability-stop			
$\lvert\mathcal{C}_t\rvert$	Acc.	Regret	Depth	Calls	Acc.	Regret	Depth	Calls	Acc.	Regret	Depth	Calls
2	.964	.009	4.3	206	.941	.020	5.6	269	.945	.016	5.2	250
4	.944	.014	5.4	518	.920	.024	5.9	566	.923	.019	5.6	538
8	.927	.024	6.2	1190	.901	.035	6.3	1210	.905	.028	6.0	1152
16	.909	.026	7.0	2688	.879	.038	6.5	2496	.883	.043	6.4	2458
32	.889	.041	7.9	6067	.853	.061	6.8	5222	.859	.050	6.7	5146

6 Conclusion and Future Work

We presented MarginPlan, an adaptive world-model planner that selects imagination depth from decision stability. Simultaneous action-by-horizon calibration provides finite-sample marginal false-certificate control under the stated exchangeability and coverage assumptions, while the continuation predictor allocates additional model propagation to unresolved decisions expected to benefit from deeper evidence. Uncertainty-scale-matched tests, budget-controlled comparisons, calibration studies, ablations, and candidate-set scaling show how the selected horizon changes with action geometry as well as uncertainty. Appendix G further evaluates the horizon-selection rule under a fixed learned model. Extending the criterion to larger latent planners and continuous candidate optimization is a natural next direction.

Reproducibility Statement

Appendix A provides the experimental protocol and hyperparameters, and Appendix B gives the adaptive-stopping guarantee. The remaining appendices report additional diagnostics and robustness analyses. The accompanying anonymous code archive contains the core implementation of calibration, continuation prediction, and the MARGINPLAN stopping rule.

AI Use Statement

Generative AI tools were used solely for language editing and polishing.

References

Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957.

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, 2018.

Maxime Burchi and Radu Timofte. Learning transformer-based world models with contrastive predictive coding. In *International Conference on Learning Representations*, 2025.

Jie Cheng, Ruixi Qiao, Yingwei Ma, Binhua Li, Gang Xiong, Qinghai Miao, Yongbin Li, and Yisheng Lv. Scaling offline model-based RL via jointly-optimized world-action model pretraining. In *International Conference on Learning Representations*, 2025.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Tuan Quang Dam, Kishan Panaganti, Brahim Driss, and Adam Wierman. Online robust reinforcement learning through monte-carlo planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 12314–12343, 2025.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2021. doi: 10.1287/mnsc.2020.3922.

Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I. Jordan, Joseph E. Gonzalez, and Sergey Levine. Model-based value estimation for efficient model-free reinforcement learning. *arXiv preprint arXiv:1803.00101*, 2018. URL https://arxiv.org/abs/1803.00101.

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*, pages 13973–14005, 2024.

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. AdaWorld: Learning adaptable world models with latent actions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 18744–18771, 2025.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027. PMLR, 2016.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2555–2565, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

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

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. PMLR, 2022. URL https://proceedings.mlr.press/v162/hansen22a.html.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jihwan Jeong, Xiaoyu Wang, Jingmin Wang, Scott Sanner, and Pascal Poupart. Reflect-then-plan: Offline model-based planning through a doubly bayesian lens. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 27113–27133, 2025.

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*, pages 2796–2804. PMLR, 2018. URL https://proceedings.mlr.press/v80/kuleshov18a.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jing Lei, Max GŚell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Zichen Liu, Guoji Fu, Chao Du, Wee Sun Lee, and Min Lin. Continual reinforcement learning by planning with online world models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 38397–38423, 2025.

Anh N Nhu, Sanghyun Son, and Ming Lin. Time-aware world model for adaptive prediction and control. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 46265–46287, 2025.

Zhongjian Qiao, Jiafei Lyu, Boxiang Lyu, Yao Shu, Siyang Gao, and Shuang Qiu. Model-based offline RL via robust value-aware model learning with implicitly differentiable adaptive weighting. In *International Conference on Learning Representations*, 2026.

Shrinivas Ramasubramanian, Benjamin Freed, Alexandre Capone, and Jeff G. Schneider. Improving model-based reinforcement learning by converging to flatter minima. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3834.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588: 604–609, 2020. doi: 10.1038/s41586-020-03051-4.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, second edition, 2018.

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.

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.

Junjie Wang, Qichao Zhang, and Dongbin Zhao. Dynamic-horizon model-based value estimation with latent imagination. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7): 8812–8825, 2024. doi: 10.1109/TNNLS.2022.3215788.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 79115–79135, 2025.

A Additional Experimental Protocol

Appendix A specifies the controlled simulator, calibration protocol, and budget-matched settings. Appendices B–G provide the adaptive-stopping guarantee, continuation-predictor diagnostics, coverage and robustness analyses, continuation-rule sensitivity, candidate-set scaling, and the learned-world-model evaluation. Controlled tables use the same evaluation split and paired seeds unless the candidate-set construction changes; those conditions use matching calibration data.

The controlled task represents a single planning decision with delayed evidence. Candidate actions are indexed so that a_1 and a_2 form the top-two pair. We set $Q_i(a_1) = 1$ and $Q_i(a_2) = 1 - g$; for $j \geq 3$, the remaining values are sampled as $Q_i(a_j) \sim \mathrm{Unif}[0.45, \min(0.78, 1 - g)]$. Because $g \leq 0.50$, this interval is nonempty, and $Q_i(a_j) < Q_i(a_2)$ almost surely, so g is the realized top-two return gap. The leading action receives evidence delay d, the runner-up receives $\max(1, d - 2)$, and remaining delays are sampled uniformly from $\{1, \ldots, d - 1\}$. Unless a diagnostic fixes a factor, training, calibration, and test decisions sample $g \sim \mathrm{Unif}[0.04, 0.50]$, $\sigma \sim \mathrm{Unif}[0.04, 0.34]$, and $d \sim \mathrm{Unif}\{3, \ldots, 10\}$. At imagination depth h, the unrevealed fraction is

$$u_{a,h} = \frac{(d_a - h)_+}{d_a},$$

and the ensemble-error scale is

$$\sigma_{a,h} = \sigma \left(0.18 + 0.65 u_{a,h} + 0.025 h^{1.25}\right).$$

For ensemble member m, the controlled return prediction is generated as

$$\widetilde{G}_{i,h}^{(m)}(a) = Q_i(a) + \sigma_{a,h}\left(\sqrt{\rho}\, z_{i,a}^{\mathrm{sh}} + \sqrt{1 - \rho}\, z_{i,a}^{(m)}\right) + u_{a,h} b_{i,a} + \left(\frac{h}{H_{\max}}\right)^{1.4} c_{i,a}, \tag{12}$$

where $z_{i,a}^{\mathrm{sh}}$ and $z_{i,a}^{(m)}$ are independent standard normal variables, with $z_{i,a}^{\mathrm{sh}}$ shared by all ensemble members for action a. For fixed (i, a, h), the zero-mean Gaussian component has pairwise cross-member correlation ρ. The same shared and member-specific latent draws are reused across checked horizons, while $\sigma_{a,h}$ changes with depth, preserving cross-horizon dependence. The terms $b_{i,a} \sim \mathcal{N}(0, (0.55\sigma)^2)$ and $c_{i,a} \sim \mathcal{N}(0, (0.06\sigma)^2)$ model missing-evidence bias and smaller depth-dependent compounding error, respectively. The planner receives only the ensemble predictions; $Q_i(a)$ is used for calibration targets and evaluation.

The continuation predictor is a two-output linear ridge regressor that minimizes Eq. (8) on the current-depth feature vector in Section 4.4. Nonconstant features are standardized using the training split, the intercept is unregularized, and $\xi = 2 \times 10^{-3}$. One training example is formed at each checked $h < H_{\max}$, with $(A_{h+\Delta_h}, W_{h+\Delta_h})$ from the realized deeper rollout as the target. The predictor is fitted once per seed on training decisions and is then fixed for calibration and test decisions. Table 7 lists the shared configuration. The Mean-only ablation in Table 4 stops when the ensemble-mean argmax is unchanged across two consecutive checks and the change in the top-two mean-return gap is below τ_m; τ_m is selected on validation data and then fixed.

Table 7: Controlled-test and calibration settings used throughout the paper.

Quantity	Value
Candidate actions (default)	4
Ensemble size M	24
Maximum depth $H_{\max}$	12
Rollout increment Δ_h	1
Gap sweep g	$[0.04, 0.50]$
Uncertainty sweep	$[0.04, 0.34]$
Default evidence delay	7
Default cross-member Gaussian correlation ρ	0.50
Interval stabilizer ϵ	10^{-3}
Continuation ridge coefficient ξ	2×10^{-3}
Training decisions / seed	60,000
Calibration decisions / seed N_{cal}	10,000
Test decisions / seed N_{test}	20,000
Paired random seeds	5
Uncertainty threshold τ_u	0.112
Reliability threshold τ_r	0.46
Consistency patience K_c	3
Mean-gap tolerance τ_m (ablation only)	0.018
Calibration target $1 - \alpha$	0.90
Mean simultaneous quantile $q_{1-\alpha}$	2.67
Mean pointwise quantile	1.72
$\lambda, \eta, \kappa_{\mathrm{comp}}$	$0.65, 0.35, 2 \times 10^{-5}$

The budget-matched experiment in Table 3 uses a separate validation sweep around $B = 576$ calls. The validation objective first minimizes $|\text{Calls} - B|$ and then breaks ties by validation regret; the selected settings are then applied to the common test split. Table 8 records the settings needed to reproduce that operating point. For Consistency-stop, $h_{\min}$ is the earliest depth at which the patience rule may terminate.

Table 8: Validation-selected settings for the approximately computation-matched comparison.

Method	Budget-matched setting
Fixed-6	$h = 6$
Uncertainty-stop	$\tau_u = 0.109$
Reliability-stop	$\tau_r = 0.40$
Consistency-stop	$K_c = 3,\ h_{\min} = 2$
MARGINPLAN	$\lambda = 0.48,\ \eta = 0.35,\ \kappa_{\text{comp}} = 2\times10^{-5}$

B ADAPTIVE-STOPPING GUARANTEE

For a held-out decision i, Eq. (2) takes the maximum normalized residual over every candidate action and every horizon that Algorithm 1 may inspect. A split-conformal quantile of these maxima jointly calibrates the entire finite collection over actions and horizons. Under exchangeability of the complete planning decision—including the candidate-set construction—the new decision satisfies the simultaneous event in Eq. (5) with the target finite-sample coverage, up to the standard split-conformal quantile discretization. If $|\mathcal{C}_t|$ or the candidate-generation rule changes, calibration is repeated under that construction or performed conditionally/stratified.

Condition on that simultaneous event. Because the event holds for all $h \in \mathcal{H}$, it also holds at the data-dependent stopping time τ. If $\Gamma_\tau > 0$, then for every $b \neq a^\star_\tau$,

$$Q(a^\star_\tau) \geq L_\tau(a^\star_\tau) > U_\tau(b) \geq Q(b).$$

Thus, on the simultaneous event, every positive certificate is valid within the candidate set. Consequently $\mathsf{Cert} \cap \mathsf{Wrong}$ is a subset of the simultaneous-coverage failure event, which proves Eq. (6). The bound controls the marginal event $\mathsf{Cert} \cap \mathsf{Wrong}$; conditional certificate precision is reported empirically. Stops triggered by Eq. (11) are recorded as lower-bound fallbacks and evaluated through decision accuracy and regret.

C CONTINUATION PREDICTOR DIAGNOSTICS

The continuation predictor is trained only from training-side deeper rollouts. Its inputs are current-depth statistics; targets are the ambiguity and average interval width one block deeper. Table 9 reports predictor errors for both targets on a fixed held-out set containing all checked depths and shared across deployment settings.

Table 9: Held-out continuation-predictor error over five paired seeds (mean $\pm$ s.d.).

Evaluation subset	MAE_A	MAE_W
All checked depths	$.024 \pm .003$	$.033 \pm .004$
Open-certificate states	$.028 \pm .004$	$.037 \pm .005$
Low-margin quartile	$.031 \pm .005$	$.040 \pm .006$

Table 10 separates certified stops, lower-bound fallbacks, and depth-cap terminations, and reports their rate, accuracy, depth, and calls.

Table 10: Stopping-mode statistics for MarginPlan on the main controlled grid.

Stop reason	Rate	Accuracy	Depth	Calls
Certified stable	.420	.981	3.5	336
Lower-bound fallback	.500	.936	5.93	569
Depth cap	.080	.800	12.0	1152

Table 11: Seed-level variability for the overall shared-prediction comparison (five paired seeds). Fixed-horizon depth is deterministic; adaptive depth varies across decisions and seeds.

Method	Accuracy	Regret	Mean depth
Fixed-2	$.842 \pm .011$	$.050 \pm .004$	$2.00 \pm .00$
Fixed-4	$.889 \pm .014$	$.027 \pm .006$	$4.00 \pm .00$
Fixed-8	$.931 \pm .007$	$.017 \pm .002$	$8.00 \pm .00$
Fixed-12	$.918 \pm .012$	$.029 \pm .007$	$12.00 \pm .00$
Uncertainty-stop	$.920 \pm .009$	$.024 \pm .003$	5.9 ± 0.5
Reliability-stop	$.923 \pm .013$	$.019 \pm .005$	5.6 ± 0.4
Consistency-stop	$.930 \pm .010$	$.022 \pm .004$	5.7 ± 0.6
MARGINPLAN	$.944 \pm .008$	$.014 \pm .004$	5.4 ± 0.4

D COVERAGE AND CERTIFICATE RELIABILITY

Increasing the target coverage produces the expected conservatism–compute trade-off: raising the target from .80 to .95 reduces certification from .503 to .351 and false certification from .015 to .004, while model calls increase from 475 to 568 (Table 12). Each target uses its corresponding calibration quantile on the same evaluation split and seed set; the .90 row uses the main-experiment calibration.

Table 12: Target coverage controls interval width, certification frequency, and computation.

Target	q	Joint cov.	Cert.	False cert.	Acc.	Regret	Calls
.80	2.18	.812	.503	.015	.934	.020	475
.90	2.67	.904	.420	.008	.944	.014	518
.95	3.21	.953	.351	.004	.948	.013	568

Shared model error can reduce ensemble disagreement when multiple members move together. Table 13 varies test-time cross-member Gaussian-error correlation while holding each seed-specific multiplier fixed at the value calibrated for $\rho = 0.50$; $q = 2.67$ is the mean of the five multipliers. The resulting coverage shift quantifies sensitivity to the cross-member Gaussian-error correlation. Each row reports joint coverage, certification, false certification, accuracy, regret, and calls.

Table 13: Robustness to cross-member Gaussian correlation in world-model errors.

Corr.	Joint cov.	Cert.	False cert.	Acc.	Regret	Calls
.00	.919	.458	.005	.952	.010	491
.25	.912	.441	.006	.949	.015	503
.50	.904	.420	.008	.944	.014	518
.75	.879	.391	.014	.934	.022	547
.90	.842	.354	.024	.920	.028	584

E CONTINUATION-RULE SENSITIVITY

Table 14 uses a single fixed evaluation split and seed set for every row. The first row gives the default configuration; each subsequent row changes one parameter from $(\lambda, \eta, \kappa_{\text{comp}}) = (0.65, 0.35, 2 \times 10^{-5})$. The one-at-a-time comparison uses a common predictor and evaluation draws across rows; the fixed predictor MAEs are reported separately in Table 9. The $\eta = 0$ row makes the interval-shrinkage diagnostic explicit.

Table 14: One-at-a-time sensitivity of MarginPlan on one fixed evaluation split and seed set.

Setting	λ	η	κ_{comp}	Acc.	Regret	Depth	Calls
Default	.65	.35	2×10^{-5}	.944	.014	5.4	518
λ low-1	.25	.35	2×10^{-5}	.936	.022	6.7	643
λ low-2	.50	.35	2×10^{-5}	.947	.013	5.9	566
λ high-1	.80	.35	2×10^{-5}	.939	.017	4.9	470
λ high-2	1.00	.35	2×10^{-5}	.928	.026	4.5	432
$\eta = 0$	.65	.00	2×10^{-5}	.931	.024	4.8	461
η low	.65	.15	2×10^{-5}	.938	.018	5.1	490
η high-1	.65	.60	2×10^{-5}	.946	.015	5.9	566
η high-2	.65	1.00	2×10^{-5}	.939	.019	6.5	624
κ zero	.65	.35	0	.939	.020	6.4	614
κ low	.65	.35	10^{-5}	.945	.016	5.9	566
κ high-1	.65	.35	4×10^{-5}	.937	.018	4.8	461
κ high-2	.65	.35	8×10^{-5}	.916	.030	4.0	384

F Candidate-Set Scaling

Increasing the candidate set introduces more competing return intervals and makes certification harder. Table 15 gives the complete measurements, including calls, simultaneous joint coverage, and certification statistics. Each candidate-set size uses calibration data generated by the corresponding candidate-set construction. For Uncertainty-stop and Reliability-stop, Cert. and False cert. are post-hoc diagnostics evaluated at each baseline's selected horizon.

Table 15: Detailed candidate-set scaling results.

$\lvert\mathcal{C}_t\rvert$	Method	Acc.	Regret	Depth	Calls	Joint cov.	Cert.	False cert.
2	MarginPlan	.964	.009	4.3	206	.907	.558	.005
2	Uncertainty-stop	.941	.020	5.6	269	.907	.489	.011
2	Reliability-stop	.945	.016	5.2	250	.907	.516	.009
4	MarginPlan	.944	.014	5.4	518	.904	.420	.008
4	Uncertainty-stop	.920	.024	5.9	566	.904	.361	.016
4	Reliability-stop	.923	.019	5.6	538	.904	.389	.014
8	MarginPlan	.927	.024	6.2	1190	.899	.326	.010
8	Uncertainty-stop	.901	.035	6.3	1210	.899	.268	.020
8	Reliability-stop	.905	.028	6.0	1152	.899	.299	.017
16	MarginPlan	.909	.026	7.0	2688	.906	.247	.013
16	Uncertainty-stop	.879	.038	6.5	2496	.906	.192	.024
16	Reliability-stop	.883	.043	6.4	2458	.906	.218	.021
32	MarginPlan	.889	.041	7.9	6067	.898	.174	.015
32	Uncertainty-stop	.853	.061	6.8	5222	.898	.128	.029
32	Reliability-stop	.859	.050	6.7	5146	.898	.149	.027

G Planner-Interface Evaluation with a Learned World Model

This study evaluates horizon selection on Walker Walk, Cheetah Run, Finger Spin, and Hopper Hop from the DeepMind Control Suite (Tassa et al., 2018). To isolate horizon selection, all methods use the same trained checkpoint, actor, replay data, continuation policy, candidate bank, and planner-side predictive samples. The planner uses the interface in Eq. (1) with $M = 24$ predictive samples per root action.

At each saved simulator state, the finite candidate bank contains four root actions: the actor mean and three Gaussian perturbations clipped to the environment action bounds. After the root action, every candidate follows the same continuation policy π. With four candidates and $M = 24$, one additional imagination step costs 96 model transitions, so reported calls equal mean selected depth times 96 up to rounding.

Calibration and evaluation are task-specific. Held-out simulator states are separated from the states used to fit the task checkpoint. For each calibration state and root candidate, the environment is branched from the saved state, executes the candidate action once, and then follows π to episode termination; that branch return defines $Q_i(a)$. The simultaneous multiplier is calibrated from these branch-return targets independently for each task. The comparison uses a fixed horizon, uncertainty-

based stopping, action-persistence stopping, and MARGINPLAN under the same planner interface. Table 16 reports five-seed return means with standard deviations together with mean selected depth and model calls.

Table 16: Planner-interface evaluation of selected horizon rules with one fixed learned model.

Task	Horizon rule	Return ↑	Depth ↓	Calls ↓
Walker Walk	Fixed	958.4 ± 13.8	12.0	1152
Walker Walk	Uncertainty-stop	956.9 ± 16.2	6.1	586
Walker Walk	Consistency-stop	968.1 ± 11.6	5.6	538
Walker Walk	MARGINPLAN	966.5 ± 12.4	5.2	499
Cheetah Run	Fixed	731.5 ± 21.7	12.0	1152
Cheetah Run	Uncertainty-stop	736.8 ± 23.4	6.8	653
Cheetah Run	Consistency-stop	740.2 ± 20.1	6.3	605
Cheetah Run	MARGINPLAN	747.6 ± 18.9	6.0	576
Finger Spin	Fixed	934.8 ± 14.2	12.0	1152
Finger Spin	Uncertainty-stop	920.6 ± 18.7	5.5	528
Finger Spin	Consistency-stop	932.1 ± 15.6	5.0	480
Finger Spin	MARGINPLAN	931.4 ± 16.1	4.6	442
Hopper Hop	Fixed	276.4 ± 19.5	12.0	1152
Hopper Hop	Uncertainty-stop	272.9 ± 21.0	7.1	682
Hopper Hop	Consistency-stop	283.7 ± 18.8	6.6	634
Hopper Hop	MARGINPLAN	289.8 ± 17.4	6.4	614