A Notation Table

Table 2: Summary of Notation

Symbol	Description
$\mathcal{S}$	Training dataset of transitions (fixed replay buffer)
s, s_0, s_t	State, initial state, state at time t
a, a_t	Action, action at time t
θ	Parameters of the world-model
d	Dimensionality of model parameters θ
$\mathcal{L}_{\mathcal{S}}(\theta)$	Empirical loss of the model on dataset $\mathcal{S}$ with parameters θ
$L(\theta; \pi)$	True (population) one-step model error under policy π
$\hat{L}(\theta; \pi)$	Empirical one-step model error under policy π (on samples from d^π)
ρ	Radius of the neighborhood for sharpness calculation / SAM perturbation
ϵ	Perturbation vector for model parameters
$R_\rho^{(1)}(\theta)$	First-order sharpness of training loss $\mathcal{L}_{\mathcal{S}}(\theta)$
$R_\rho^{(1)}(\theta; \pi)$	First-order sharpness of true model error $L(\theta; \pi)$ (theory context)
$L_{\mathcal{S}}^{\text{SAM}}(\theta)$	Sharpness-Aware Minimization objective function
λ	Weight decay coefficient (for SAM or general regularization)
M	True MDP (Markov Decision Process)
$\hat{M}_\theta$	Learned model of the MDP, parameterized by θ
$P(\cdot\|s, a)$	True transition probability function
$\hat{P}_\theta(\cdot\|s, a)$	Learned transition probability function
$r(s, a)$	True reward function
$\hat{r}_\theta(s, a)$	Learned reward function (if part of the model)
π	A policy $\pi : \mathcal{S} \to \mathcal{A}$
π^*	Optimal policy in the true environment M
$\hat{\pi}^*$	Optimal policy in the learned model $\hat{M}_\theta$
$V^\pi(s_0)$	True expected discounted return of policy π from s_0
$\hat{V}^\pi(s_0)$	Expected discounted return of policy π estimated by model $\hat{M}_\theta$
γ	Discount factor
$d^\pi(s, a)$	Discounted state-action occupancy measure for policy π
n	Number of samples (e.g., to estimate $\hat{L}(\theta; \pi)$ or size of $\mathcal{S}$)
δ	Confidence parameter (e.g., $1 - \delta$ probability)
C	Concentrability coefficient: $\max_{(s,a)} \frac{d^{\pi^*}(s,a)}{d^{\hat{\pi}^*}(s,a)}$
M_{loss} (in bounds)	Upper bound on the per-sample one-step model error value
$\Omega(d, n, \rho, \delta)$	Model complexity/capacity term in generalization bounds

B Proofs of Theoretical Results

In this section, we provide the proofs for Thm. 1 and Thm. 2 presented in Sec. 4. The proofs leverage standard results from learning theory and model-based reinforcement learning.

B.1 Preliminaries for Proofs

We first recall two key results that will be used in the proofs.

Simulation Lemma. For any policy π, discount factor γ, and a world-model $\hat{M}_\theta$ whose one-step population model error under π is $L(\theta; \pi) = \mathbb{E}_{(s,a)\sim d^\pi}[\|P(\cdot|s, a) - \hat{P}_\theta(\cdot|s, a)\|_1]$, the difference between the true value $V^\pi(s_0)$ and the model-estimated value $\hat{V}^\pi(s_0)$ is bounded (assuming rewards $r_t \in [0, R_{max}]$):

$$|V^\pi(s_0) - \hat{V}^\pi(s_0)| \leq \frac{\gamma R_{max}}{(1-\gamma)^2} L(\theta; \pi). \tag{4}$$

If rewards are part of the model error (e.g., bounded prediction error ϵ_r per step), the term $L(\theta;\pi)$ in the bound effectively incorporates both transition and reward errors. For simplicity and consistency with the main text where model error $\hat{L}$ directly contributes to the value gap scaled by $\gamma/(1-\gamma)^2$, we'll use a form of the simulation lemma where We define $L(\theta;\pi) := \mathbb{E}_{(s,a)\sim d^\pi}[\text{err}(s,a)]$ leads to:

$$|V^\pi(s_0) - \hat{V}^\pi(s_0)| \leq \frac{\gamma}{(1-\gamma)^2}\mathbb{E}_{(s,a)\sim d^\pi}[\text{err}(s,a)]. \tag{5}$$

We will denote $\mathbb{E}_{(s,a)\sim d^\pi}[\text{err}(s,a)]$ as $L(\theta;\pi)$, representing the true one-step prediction error of the model under policy π.

Bound on Population Model Error. Given an empirical model error $\hat{L}(\theta;\pi)$ calculated on a training set of size n, its first-order sharpness $R_\rho^{(1)}(\theta;\pi)$ (of the true model error landscape $L(\theta;\pi)$), a maximum per-sample loss M_{loss}, and model parameters $\theta \in \mathbb{R}^d$, the true population model error $L(\theta;\pi)$ can be bounded. Specifically, with probability at least $1-\delta$:

$$L(\theta;\pi) \lesssim \hat{L}(\theta;\pi) + R_\rho^{(1)}(\theta;\pi) + \sqrt{\frac{M_{loss}}{n}} + \Omega(d,n,\rho,\delta), \tag{6}$$

where $\Omega(d,n,\rho,\delta)$ is a model complexity term. The $\lesssim$ indicates an approximation or bound that may hide constants. This type of bound often arises from PAC-Bayesian analysis or uniform convergence arguments applied to sharpness-aware contexts.

B.2 Proof of Thm. 1 (Return-Estimation Gap)

Theorem 3 (Return-Estimation Gap Restated, cf. Thm. 1)**.** *For any policy π and discount factor $\gamma \in (0,1)$, with probability at least $1-\delta$ over n i.i.d. training samples for the model error estimation:*

$$\left|V^\pi(s_0) - \hat{V}^\pi(s_0)\right| \;\leq\; \frac{\gamma}{(1-\gamma)^2}\left[\hat{L}(\theta;\pi) + R_\rho^{(1)}(\theta;\pi) + \sqrt{\tfrac{M_{loss}}{n}} + \Omega(d,n,\rho,\delta)\right].$$

Proof. We start with the Simulation Lemma (Eq. 5):

$$|V^\pi(s_0) - \hat{V}^\pi(s_0)| \leq \frac{\gamma}{(1-\gamma)^2}L(\theta;\pi).$$

Now, we substitute the upper bound for the population model error $L(\theta;\pi)$ from Equation 6. This bound holds with probability at least $1-\delta$:

$$L(\theta;\pi) \lesssim \hat{L}(\theta;\pi) + R_\rho^{(1)}(\theta;\pi) + \sqrt{\frac{M_{loss}}{n}} + \Omega(d,n,\rho,\delta).$$

Plugging this into the simulation lemma bound, we get:

$$\left|V^\pi(s_0) - \hat{V}^\pi(s_0)\right| \;\leq\; \frac{\gamma}{(1-\gamma)^2}\left[\hat{L}(\theta;\pi) + R_\rho^{(1)}(\theta;\pi) + \sqrt{\tfrac{M_{loss}}{n}} + \Omega(d,n,\rho,\delta)\right].$$

This completes the proof. The term $\hat{L}(\theta;\pi)$ is the empirical one-step model error under policy π, and $R_\rho^{(1)}(\theta;\pi)$ is the first-order sharpness of the true model error landscape $L(\theta;\pi)$. □

B.3 Proof of Theorem 2 (Performance Gap)

Theorem 4 (Performance Gap Restated, cf. Theorem 2)**.** *Let π^* be the optimal policy in the true environment and $\hat{\pi}^*$ be the policy that is optimal according to the learned model $\hat{M}_\theta$. If $C = \max_{(s,a)} \frac{d^{\pi^*}(s,a)}{d^{\hat{\pi}^*}(s,a)}$ is the concentrability coefficient measuring the distribution mismatch between π^* and $\hat{\pi}^*$:*

$$\left|V^{\pi^*}(s_0) - V^{\hat{\pi}^*}(s_0)\right| \leq \frac{\gamma(1+C)}{(1-\gamma)^2}\left[\hat{L}(\theta;\hat{\pi}^*) + R_\rho^{(1)}(\theta;\hat{\pi}^*) + \sqrt{\tfrac{M_{loss}}{n}} + \Omega(d,n,\rho,\delta)\right].$$

Proof. Since π^* is the optimal policy in the true environment M, we have $V^{\pi^*}(s_0) \geq V^{\hat{\pi}^*}(s_0)$. Therefore, the performance gap is:

$$|V^{\pi^*}(s_0) - V^{\hat{\pi}^*}(s_0)| = V^{\pi^*}(s_0) - V^{\hat{\pi}^*}(s_0). \tag{7}$$

We can decompose this difference as:

$$V^{\pi^*}(s_0) - V^{\hat{\pi}^*}(s_0) = [V^{\pi^*}(s_0) - \hat{V}^{\pi^*}(s_0)] + [\hat{V}^{\pi^*}(s_0) - \hat{V}^{\hat{\pi}^*}(s_0)] + [\hat{V}^{\hat{\pi}^*}(s_0) - V^{\hat{\pi}^*}(s_0)]. \tag{8}$$

Applying the Simulation Lemma (Equation 5) to the first and third terms:

$$|V^{\pi^*}(s_0) - \hat{V}^{\pi^*}(s_0)| \leq \frac{\gamma}{(1-\gamma)^2} L(\theta; \pi^*) \tag{9}$$

$$|\hat{V}^{\hat{\pi}^*}(s_0) - V^{\hat{\pi}^*}(s_0)| \leq \frac{\gamma}{(1-\gamma)^2} L(\theta; \hat{\pi}^*) \tag{10}$$

Since $\hat{\pi}^*$ is the optimal policy in the learned model $\hat{M}_\theta$, we have $\hat{V}^{\hat{\pi}^*}(s_0) \geq \hat{V}^{\pi^*}(s_0)$. Therefore, the middle term in Equation 8 is non-positive:

$$\hat{V}^{\pi^*}(s_0) - \hat{V}^{\hat{\pi}^*}(s_0) \leq 0. \tag{11}$$

Substituting these into Equation 8 and using the bounds from 9 and 10:

$$\begin{aligned} V^{\pi^*}(s_0) - V^{\hat{\pi}^*}(s_0) &\leq \frac{\gamma}{(1-\gamma)^2} L(\theta; \pi^*) + 0 + \frac{\gamma}{(1-\gamma)^2} L(\theta; \hat{\pi}^*) \\ &= \frac{\gamma}{(1-\gamma)^2} [L(\theta; \pi^*) + L(\theta; \hat{\pi}^*)]. \end{aligned} \tag{12}$$

Using the definition of the concentrability coefficient $C = \max_{(s,a)} \frac{d^{\pi^*}(s,a)}{d^{\hat{\pi}^*}(s,a)}$, we can bound the model error under π^* in terms of the model error under $\hat{\pi}^*$. Assuming the model error is an expectation over the state-action distribution, $L(\theta; \pi) = \mathbb{E}_{(s,a)\sim d^\pi}[\text{err}(s, a; \theta)]$:

$$L(\theta; \pi^*) = \sum_{s,a} d^{\pi^*}(s,a)\text{err}(s,a;\theta) \leq \sum_{s,a} C \cdot d^{\hat{\pi}^*}(s,a)\text{err}(s,a;\theta) = C \cdot L(\theta; \hat{\pi}^*).$$

Substituting this into Equation 12:

$$\begin{aligned} V^{\pi^*}(s_0) - V^{\hat{\pi}^*}(s_0) &\leq \frac{\gamma}{(1-\gamma)^2} [C \cdot L(\theta; \hat{\pi}^*) + L(\theta; \hat{\pi}^*)] \\ &= \frac{\gamma(1+C)}{(1-\gamma)^2} L(\theta; \hat{\pi}^*). \end{aligned} \tag{13}$$

Finally, we substitute the upper bound for the population model error $L(\theta; \hat{\pi}^*)$ from Equation 6, which holds with probability at least $1 - \delta$:

$$L(\theta; \hat{\pi}^*) \lesssim \hat{L}(\theta; \hat{\pi}^*) + R_\rho^{(1)}(\theta; \hat{\pi}^*) + \sqrt{\frac{M_{loss}}{n}} + \Omega(d, n, \rho, \delta).$$

Plugging this into Equation 13 yields the final result:

$$\left|V^{\pi^*}(s_0) - V^{\hat{\pi}^*}(s_0)\right| \leq \frac{\gamma(1+C)}{(1-\gamma)^2} \left[\hat{L}(\theta; \hat{\pi}^*) + R_\rho^{(1)}(\theta; \hat{\pi}^*) + \sqrt{\tfrac{M_{loss}}{n}} + \Omega(d, n, \rho, \delta)\right].$$

This completes the proof. The terms $\hat{L}(\theta; \hat{\pi}^*)$ and $R_\rho^{(1)}(\theta; \hat{\pi}^*)$ refer to the empirical one-step model error and the first-order sharpness of the true model error landscape $L(\theta; \hat{\pi}^*)$, evaluated under the model-optimal policy $\hat{\pi}^*$. □

C On the Correlation Between Loss Sharpness and Model Lipschitz Continuity

In the context of model-based reinforcement learning, the smoothness of the learned dynamics model $\hat{M}_\theta$ (parameterized by θ) can be crucial for robust planning and generalization. Lipschitz continuity

is a common measure of such smoothness. Here, we briefly explore a tangential connection between the sharpness of the model's training loss and the Lipschitz properties of the model. While our main theoretical results (Theorems 1 and 2) directly incorporate first-order sharpness $R_\rho^{(1)}$, understanding its relationship with other smoothness measures like Lipschitz constants can provide additional intuition.

Let $\hat{f}_\theta(x)$ represent a component of our learned model (e.g., the next-state prediction function, where $x = (s, a)$ is a state-action pair) and $f(x)$ be the corresponding true environment dynamics. The model is trained to minimize a loss, commonly mean squared error (MSE) for continuous state predictions:

$$\mathcal{L}(\theta) = \mathbb{E}_{x \sim P_x}\left[(f(x) - \hat{f}_\theta(x))^2\right],$$

where P_x is the distribution of training data.

[Lipschitz Continuity] A function $g : \mathcal{X} \to \mathcal{Y}$ (where $\mathcal{Y}$ could be $\mathbb{R}^k$) is Lipschitz continuous with constant L_g if for all $x_1, x_2 \in \mathcal{X}$:

$$\|g(x_1) - g(x_2)\|_\mathcal{Y} \leq L_g \|x_1 - x_2\|_\mathcal{X}.$$

[Second-Order Sharpness (Spectral Norm of Hessian)] A common measure of sharpness related to the curvature of the loss landscape $\mathcal{L}(\theta)$ at a point θ is the spectral norm of its Hessian matrix:

$$S_2(\theta) = \|\nabla_\theta^2 \mathcal{L}(\theta)\|_2.$$

While our main text focuses on first-order sharpness $R_\rho^{(1)}(\theta)$ (Eq. 1), $S_2(\theta)$ provides another perspective on local curvature. (Note: A related concept, $R_\rho^{(2)}(\theta)$, could involve maximizing $\lambda_{\max}(\nabla_{\theta\theta}^2 \mathcal{L}(\theta + \epsilon))$ over a neighborhood, but $S_2(\theta)$ is the local Hessian norm at θ.)

We can show a relationship between this sharpness measure and the properties of $\hat{f}_\theta$. The Hessian of the MSE loss $\mathcal{L}(\theta)$ is:

$$\begin{aligned}
\nabla_\theta^2 \mathcal{L}(\theta) &= \nabla_\theta \left(\mathbb{E}_{x \sim P_x}\left[-2(f(x) - \hat{f}_\theta(x))\nabla_\theta \hat{f}_\theta(x)\right]\right) \\
&= 2\mathbb{E}_{x \sim P_x}\left[\nabla_\theta \hat{f}_\theta(x)\left(\nabla_\theta \hat{f}_\theta(x)\right)^\top - (f(x) - \hat{f}_\theta(x))\nabla_\theta^2 \hat{f}_\theta(x)\right].
\end{aligned}$$

Taking spectral norms and using the triangle inequality:

$$\begin{aligned}
S_2(\theta) = \|\nabla_\theta^2 \mathcal{L}(\theta)\|_2 &\leq 2\left\|\mathbb{E}_{x \sim P_x}\left[\nabla_\theta \hat{f}_\theta(x)\left(\nabla_\theta \hat{f}_\theta(x)\right)^\top\right]\right\|_2 \\
&\quad + 2\left\|\mathbb{E}_{x \sim P_x}\left[(f(x) - \hat{f}_\theta(x))\nabla_\theta^2 \hat{f}_\theta(x)\right]\right\|_2 \\
&\leq 2\mathbb{E}_{x \sim P_x}\left[\|\nabla_\theta \hat{f}_\theta(x)\|_2^2\right] + 2\sup_x |f(x) - \hat{f}_\theta(x)| \cdot \mathbb{E}_{x \sim P_x}\left[\|\nabla_\theta^2 \hat{f}_\theta(x)\|_2\right].
\end{aligned} \tag{14}$$

The term $\|\nabla_\theta \hat{f}_\theta(x)\|_2$ represents the sensitivity of the model's output to changes in its parameters θ at a given input x. The term $\|\nabla_\theta^2 \hat{f}_\theta(x)\|_2$ represents the curvature of the model function $\hat{f}_\theta$ with respect to its parameters.

If the model $\hat{f}_\theta(x)$ itself is "smooth" with respect to its inputs x (i.e., has a small Lipschitz constant $L_{\hat{f}_x} = \sup_x \|\nabla_x \hat{f}_\theta(x)\|_2$) and also with respect to its parameters (i.e., $\|\nabla_\theta \hat{f}_\theta(x)\|_2$ and $\|\nabla_\theta^2 \hat{f}_\theta(x)\|_2$ are bounded, perhaps by constants $L_{\hat{f}_\theta}$ and $H_{\hat{f}_\theta}$ respectively), then the sharpness $S_2(\theta)$ tends to be smaller, especially if the model error $|f(x) - \hat{f}_\theta(x)|$ is also small.

For instance, if we assume that for our parameterization, $\sup_x \|\nabla_\theta \hat{f}_\theta(x)\|_2 \leq K_1$ and $\sup_x \|\nabla_\theta^2 \hat{f}_\theta(x)\|_2 \leq K_2$, then:

$$S_2(\theta) \leq 2K_1^2 + 2\|f - \hat{f}_\theta\|_\infty K_2.$$

This inequality suggests that models whose output is less sensitive to parameter changes (smaller K_1, K_2) and that fit the data well (small $\|f - \hat{f}_\theta\|_\infty$) tend to reside in regions of lower second-order sharpness.

While this provides a connection to a specific type of sharpness ($S_2(\theta)$), the first-order sharpness $R_\rho^{(1)}$ targeted by SAM is related (as discussed in Section 4, $R_\rho^{(1)}$ is an indicator of local flatness). Methods that explicitly regularize the Lipschitz constant of the learned model (e.g., [2]) aim to enforce smoothness directly. SAM, by seeking flat minima of the training loss, indirectly promotes solutions where the loss (and thus often the model predictions) do not change drastically with small parameter perturbations. This implies a form of robustness that is conceptually related to having a small Lipschitz constant with respect to parameters, although SAM achieves this through a different mechanism than direct Lipschitz regularization. Exploring the interplay between SAM-induced flatness and explicit Lipschitz regularization of the world-model could be an interesting direction for future research, potentially leading to even more robust and generalizable models.

D t-statistic tests on HumanoidBench

We assess whether adding SAM improves return using one-tailed paired t-tests across n=4 seeds per task (null: no gain; alternative: *SAM* > *baseline*). Negative t indicates higher mean return with SAM under this convention.

Table 3: One-tailed paired t-tests: TD-MPC2+SAM vs. TD-MPC2 on HumanoidBench (4 seeds). Significance at α=0.05 in **bold**. The overall p-value using Fisher's method is $p = 0.0059$

Task	t-statistic	p-value	Significant?
`balance_hard`	-5.83	$\mathbf{0.002}$	**Yes**
`balance_simple`	-8.31	$\mathbf{0.002}$	**Yes**
`stair`	-1.64	0.170	No
`walk`	-1.69	0.184	No
`run`	-3.78	$\mathbf{0.032}$	**Yes**
`stand`	-1.87	0.148	No
`sit_simple`	-1.77	0.128	No
`sit_hard`	-2.20	0.108	No
`crawl`	1.69	0.160	No
`pole`	-3.19	$\mathbf{0.024}$	**Yes**
`hurdle`	-1.05	0.356	No

E SAM radius (ρ) ablation

We sweep ρ and observe a unimodal response: too small under-regularizes, too large over-perturbs, with a broad optimum in the middle.

E.1 Suite-level choices used in the paper

Table 4: ρ values used for main results.

Suite / Task set	ρ	Rationale
HumanoidBench (most tasks)	1.25×10^{-3}	Best overall trade-off across tasks
`sit_simple`, `sit_hard`	5.0×10^{-3}	Stabilizes sit tasks
High-DoF DMC	2.5×10^{-5}	Prefers smaller steps in state-based control
Atari-100k (TWISTER)	1.0×10^{-3}	Robust on pixel-based discrete control

E.2 Per-task sensitivity examples

F SAM on different TD-MPC2 components

Unless otherwise noted, SAM is applied only to the *dynamics* loss. Ablating the attachment point on `humanoid_run` (2M steps; 4 seeds):

Table 5: Return on `humanoid_run` vs. ρ settings (mean $\pm$ SEM over 4 seeds). Settings are ordered from smaller→larger ρ; **best** in bold.

Setting (ρ)	1×10^{-3}	5×10^{-6}	5×10^{-7}	1×10^{-7}	5×10^{-8}	1×10^{-8}
Episode return	348 ± 28	445 ± 56	$\mathbf{492 \pm 34}$	454 ± 17	439 ± 21	414 ± 21

Table 6: Score Bank Heist vs. ρ settings (mean $\pm$ SEM over 5 seeds).

Setting (ρ)	1×10^{-1}	1×10^{-2}	1×10^{-3}	1×10^{-4}	1×10^{-5}
Score	281 ± 277	714 ± 311	$\mathbf{886 \pm 445}$	785 ± 277	781 ± 213

Table 7: Where to apply SAM in TD-MPC2 (episode return; mean $\pm$ SEM).

Variant	Apply SAM to	Episode return
TD-MPC2 (baseline)	—	301 ± 11
TD-MPC2 + SAM (ours)	**Dynamics**	$\mathbf{484 \pm 19}$
TD-MPC2 + SAM	Reward & Value	10 ± 3
TD-MPC2 + SAM	Policy	463 ± 49

G Compute overhead

Vanilla SAM adds one inner ascent and one outer descent per model update. In our implementation this is a $\sim$**1.7**$\times$ training-time multiplier at fixed batch sizes, with modest memory overhead; environment interaction (sample complexity) is unchanged. When wall-clock is tight, apply SAM intermittently (every kth model update), anneal the inner step frequency late in training, or reuse cached activations to reduce the second backward pass cost.

H Max Hessian eigenvalue

We approximate the leading eigenvalue(s) of the Hessian of the *dynamics* loss via power iteration on replay mini-batches after training. Lower values indicate flatter minima.

Table 8: Max Hessian eigenvalue $\lambda_{\max}$ (arbitrary units; same scale across rows). Lower is flatter.

Task (2M steps)	TD-MPC2	TD-MPC2 + SAM
`humanoid_run`	141.6	**99.5**
`humanoid_walk`	92.1	**82.3**
`dog_run`	80.3	**46.5**
`dog_walk`	69.5	**40.1**
`dog_trot`	53.8	**31.6**

I HumanoidBench Results

Table 9: HumanoidBench returns (mean ± SEM over 4 seeds) for TD-MPC2 and TD-MPC2+SAM.

Metric	balance_hard	balance_simple	stair	walk	run	stand	sit_simple	sit_hard	crawl	pole	hurdle
TD-MPC2 Episode Reward	48 ± 6	28 ± 8	66 ± 6	644 ± 281	67 ± 8	639 ± 240	515 ± 187	508 ± 298	896 ± 53	207 ± 35	51 ± 12
TD-MPC2 w/ SAM Episode Reward	82 ± 10	145 ± 27	77 ± 12	885 ± 44	302 ± 124	870 ± 58	773 ± 223	843 ± 64	846 ± 26	273 ± 22	71 ± 36

Table 10: High-DoF DMC results at 2M environment steps (mean ± SEM over n=4 seeds). Bold is better.

Method	`humanoid_run`	`humanoid_walk`	`dog_run`	`dog_walk`	`dog_trot`
TD-MPC2	301 ± 11	883 ± 13	428 ± 39	887 ± 46	891 ± 21
TD-MPC2 w/ SAM	$\mathbf{484 \pm 19}$	$\mathbf{901 \pm 4}$	$\mathbf{552 \pm 17}$	$\mathbf{957 \pm 6}$	$\mathbf{920 \pm 14}$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim that applying SAM to the world-model in MBRL leads to flatter minima and improved downstream policy performance, supported by theory and experiments on HumanoidBench with TD-MPC2. These claims are reflected in the paper's theoretical analysis (Section 4) and experimental results (Section 5).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations, such as the current scope of empirical evaluation (primarily TD-MPC2 on HumanoidBench) and the task-dependent nature of the SAM hyperparameter ρ, are discussed in the Experiments (Section 5) and Conclusion (Section 7).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be

used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The main theoretical results (Theorems 1 and 2) are presented with their assumptions in Section 4. Detailed proofs are provided in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper details the experimental setup in Section 5, including the benchmark (HumanoidBench), base algorithm (TD-MPC2), SAM integration, and key hyperparameters like ρ values and training steps. Further details on TD-MPC2 hyperparameters are referenced to its original publication.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is provided with the submission/

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 describes the environment, tasks, algorithm modifications, training duration, seeds, and key SAM hyperparameters. Further details for TD-MPC2 are referenced.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The learning curves in Figure 4, 2 report mean returns over 4 seeds with shaded regions representing the standard error of the mean (SEM), as stated in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper does not currently detail the specific compute resources (e.g., GPU type, number of GPUs, approximate training time per run). This information would be added to an appendix if the paper is accepted or if required by submission guidelines.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: The research focuses on algorithmic improvements in simulated environments and does not involve human subjects, sensitive data, or applications with immediate ethical concerns that would violate the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: The paper currently focuses on the technical contributions. A discussion of broader impacts (e.g., improved robot capabilities, potential misuse of advanced AI) is not included but could be added.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The research involves training reinforcement learning agents in simulated environments. The models produced are specific to these simulations and do not pose a high direct risk for misuse in the manner of large pre-trained generative models or scraped datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites TD-MPC2, HumanoidBench, and SAM, which are the primary existing assets. HumanoidBench is based on MuJoCo. Licenses for these assets are typically permissive for academic research (e.g., Apache 2.0 for TD-MPC2 and HumanoidBench, MuJoCo is now open-source).

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper primarily introduces a modification to an existing algorithm (TD-MPC2 by integrating SAM) and evaluates it on an existing benchmark (HumanoidBench). No new datasets or standalone software packages are introduced as primary contributions. If code is released, it will be documented.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

Justification: The research does not involve crowdsourcing or human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The research does not involve human subjects, so IRB approval is not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Large language models were not used as a core component of the proposed methodology or experiments. Any LLM usage was restricted to assisting with writing, editing, or formatting, which does not impact the scientific contributions of the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.