L4DC 2026 · LEARNING FOR DYNAMICS & CONTROL

TD-M(PC)²: Improving Temporal-Difference MPC Through Policy Constraint

Haotian Lin[1] · **Pengcheng Wang**[2] · **Jeff Schneider**[1] · **Guanya Shi**[1]
[1] Carnegie Mellon University · [2] University of California, Berkeley
Carnegie Mellon University UC Berkeley

37
evaluation tasks
14 HB + 23 DMC

>100%
HB aggregate gain
vs TD-MPC2; Reach excluded

<10 LoC
policy-side change
same actor architecture

MAIN CLAIM **MPC exploration is useful; the bottleneck is bootstrapping the actor outside planner-data support. TD-M(PC)² constrains exploitation while leaving the planner free.**

Abstract

Simplified from the paper abstract.

Plan-based MBRL combines sampling-based MPC with learned value and policy priors, but value learning is trained almost entirely on planner-collected data. We identify a structural mismatch between the planner policy used for exploration and the nominal policy evaluated by the value function, causing persistent value overestimation in high-dimensional tasks.
TD-M(PC)² adds a soft conservative policy update that keeps the learned actor close to replay-buffer behavior while preserving model-based exploration. Across HumanoidBench and DMControl, it reduces value bias and improves performance, with the largest gains on 61-DoF humanoid locomotion.

Problem	Fix	Evidence
π_H vs π mismatch	actor-side β log μ	2,159% bias + 37 tasks

Analysis

Why value bias persists in plan-based MBRL.

Mechanism: planner data are generated by π_H, but TD targets evaluate the nominal actor π. When π disagrees with π_H, Q(s',π(s')) is queried outside replay support.

1 MISMATCH planner π_H fills replay; actor π is evaluated → **2 OOD BOOTSTRAP** Bellman target queries Q(s',π(s')) → **3 ACCUMULATION** ε_k amplifies; bias does not self-correct

THEOREM 1: PLANNER PERFORMANCE BOUND

$$\limsup_{k\to\infty} |V^* - V^{\pi_{H,k}}| \le \limsup_{k\to\infty} \frac{2}{1-\gamma^H}\left[C(\epsilon_{m,k}, H, \gamma) + \frac{\epsilon_{p,k}}{2} + \frac{\gamma^H(1+\gamma^2)}{(1-\gamma)^2}\epsilon_k\right]$$

THEOREM 2: ERROR ACCUMULATION

$$\delta_k \le \frac{1}{1-\gamma^H}\left[2C(\epsilon_{m,k-1}, H, \gamma) + \epsilon_{p,k-1} + (1+\gamma^H)\delta_{k-1} + \frac{2\gamma(1+\gamma^{H-1})}{1-\gamma}\epsilon_{k-1}\right]$$

DISTRIBUTION SHIFT LOWER BOUND

$$D_{\mathrm{TV}}\big(p^{\pi}(s,a) \,\|\, p^{\pi'}(s,a)\big) \ge \tfrac{1-\gamma}{2R_{\max}}\left|J^{\pi} - J^{\pi'}\right|$$

Interpretation. H-step planning reduces dependence on terminal value error, but the error itself grows when actor queries leave planner-data support.

Empirical signature. Value bias scales with action dimension: Dog-trot reaches 231%, h1hand-run reaches 2,159%, and h1hand-slide reaches 746% under vanilla TD-MPC2.

Method

Conservative exploitation; unchanged model-based exploration.

TD-M(PC)² adopts a soft state-conditional divergence constraint to the replay behavior policy μ, implemented using the stored planner Gaussian. The method changes the actor update, not the planner.

CONSTRAINED POLICY IMPROVEMENT

$$\pi_{k+1} \in \arg\max_{\pi}\ \mathbb{E}_{s\sim d_{\mathcal{D}}}\big[\mathbb{E}_{a\sim\pi(\cdot|s)} Q_k(s,a) - \beta D(\pi(\cdot|s)\|\mu(\cdot|s))\big]$$

IMPLEMENTED ACTOR LOSS

$$\mathcal{L}_\pi = -\mathbb{E}_{s,\mu\sim\mathcal{B},\ a\sim\pi}\left[\frac{Q(s,a)}{S_q} + \frac{\beta \log \mu(a|s)}{S_q} - \alpha \log \pi(a|s)\right]$$

Store in replay	No density model
planner mean/std with each transition	reuse μ ≈ π_H; <10 LoC
Actor-side only	**Optimistic conservatism**
MPC search, model, critic target unchanged	entropy term + β threshold via S_q

LOSS SCALE

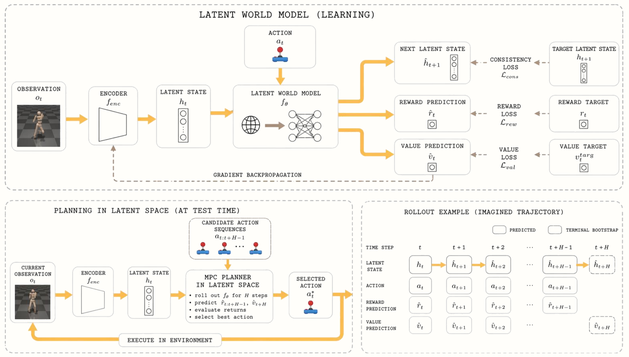
$$S_q = \mathrm{EMA}(\max\{\mathrm{Per}(Q, 95) - \mathrm{Per}(Q, 5), 1\}, \xi)$$

Why it works. The constraint is applied only during actor improvement. The replay planner still explores with MPC, while the actor used in TD targets is discouraged from querying high-value unsupported actions. This attacks OOD bootstrapping without adding a density model or changing the world model, critic target, or planner objective.

Constrain exploitation; preserve exploration.

Architecture

TD-MPC latent world model learning, test-time MPC planning, and imagined rollouts.



1 LEARN A LATENT WORLD MODEL
Encode **o_t → h_t**; predict next latent, reward, and value. Consistency, reward, and value losses train the encoder and model jointly.

2 PLAN INSIDE THE MODEL
MPC samples action sequences and rolls out **f_θ** for H steps.
Score = imagined rewards + terminal value; execute the first action.

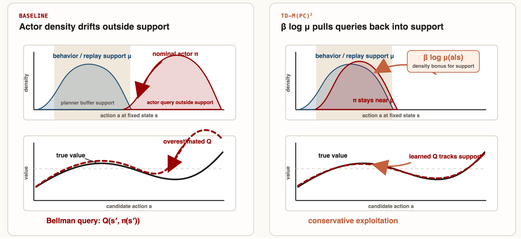
3 WHERE BIAS ENTERS
Replay follows planner **π_H**, but TD targets query actor **π**. If π leaves support, bootstrapped value becomes biased.

TD-M(PC)² intervention: keep the MBRL stack unchanged; constrain actor-side exploitation.

Stored support	Actor update	Unchanged planner
$\mu_H(a \mid s) = \mathcal{N}(a;\, \bar{a}_H(s), \Sigma_H(s))$	$Q_\theta(s,a) + \beta \log \mu_H(a \mid s)$	$f_\theta,\ r_\theta,\ Q_\theta,\ \pi_H$ unchanged
store planner Gaussian parameters with replay	penalize unsupported high-Q actions	MPC search and model learning stay intact

Mitigate Value Overestimation

β log μ pulls actor queries back into replay support.



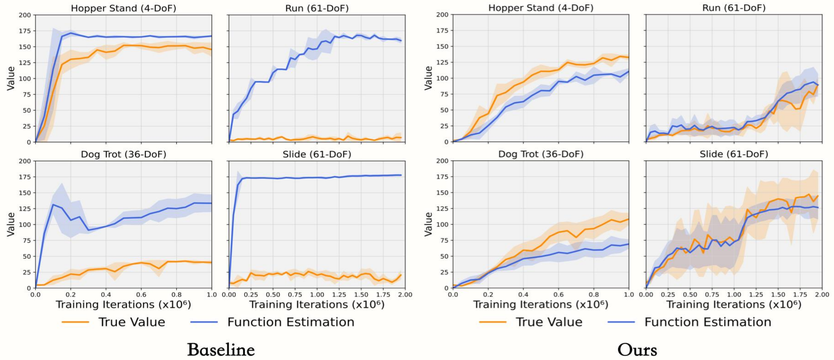
Failure
actor density drifts beyond planner replay support; Bellman queries hit weakly covered tails

Constraint
β log μ_H(a|s) is a soft support prior during actor improvement, not during MPC search

Effect
TD targets remain near data-rich actions, so value bias stops amplifying across updates

Value-bias diagnostic

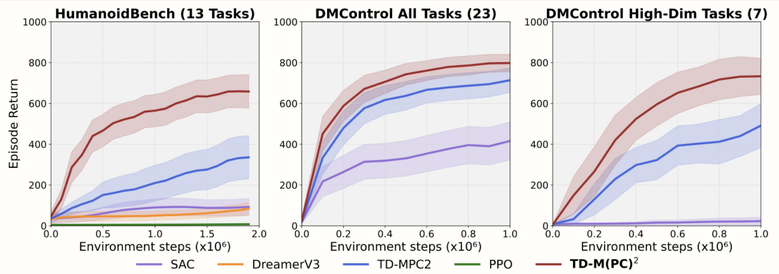
Initial-state Monte-Carlo return under nominal actor versus critic estimate.



Measured as Vhat(s0) minus Monte-Carlo return under nominal π; TD-M(PC)² moves high-DoF estimates closer to Monte-Carlo value.

Average performance

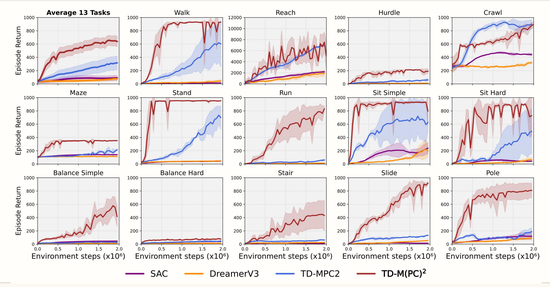
Main-paper aggregate across HumanoidBench and DMControl.



13	23	7	
HumanoidBench avg	DMControl tasks	high-dim DMC	**Takeaway.** All aggregate suites improve; largest gains are on high-dimensional control.

HumanoidBench per-task curves

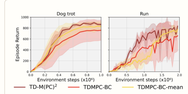
Complete h1hand-v0 locomotion results.



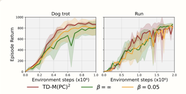
Full HB evidence. The gain is not a single-task artifact.
Hard 61-DoF tasks: run, slide, stair, pole.
Reach excluded due to reward scale.
Pattern: support matters most when value errors drive large whole-body motions.

Ablations

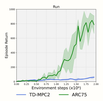
Constraint mechanism, sensitivity, and planner-side checks.



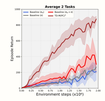
ACTOR OBJECTIVE
BC-style actor losses help, but the planner-likelihood term gives the cleaner high-DoF improvement.



β SENSITIVITY
Performance is stable across β; too-small β weakens support, while moderate β is robust.



PLANNER REG.
Planner-side regularization removes useful exploration and does not solve actor-side bootstrapping.



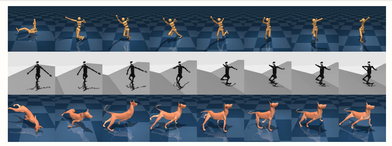
EXPLORATION
Broad MPC exploration stays intact; the constraint acts during actor improvement only.

Mechanism check: constrain exploitation without removing model-based exploration.

Qualitative rollouts

Stable whole-body behavior on high-dimensional tasks.

h1hand-run · 61-DoF
h1hand-slide · 61-DoF
dog-run · 36-DoF



Paper, code, videos, and reproducibility material
https://darthutopian.github.io/tdmpc_square/