How Many Samples Should a Planner Draw? The Answer Depends on a Convention Nobody Reports

Anonymous authors
Paper under double-blind review

Abstract

Sampling-based planners (CEM, MPPI) are the deployment-time engine of modern world-model agents, and a practitioner's first question about them is how to spend planner compute: draw more candidate action sequences, or refine them over more iterations? We show that the first half of this question is not well posed as usually asked. Implementations differ in whether the elite set is a fixed *fraction* of the population (mbrl-lib's default, used by PETS and PlaNet) or a fixed *count* (TD-MPC2's convention), and this choice, usually inherited from a library default, **can reverse the sign of the measured effect of drawing more samples**. On identical PlaNet models, seeds and episodes, increasing the population $40\times$ changes return by -63 under one convention and $+246$ under the other (both $p = 0.002$); the other two PlaNet tasks point the same way, while on PETS the convention changes the effect's size (significantly only on Ant) but not its sign. Instrumenting the optimizer gives a post-hoc mechanism: the elite count sets how concentrated the refit proposal is, so tying it to the population *widens* the proposal as sampling grows (PETS dispersion $0.127 \to 0.219$), while a fixed count *tightens* it ($0.167 \to 0.149$). Sample-count comparisons are therefore not commensurable unless the elite convention is stated. The second half of the question holds the population fixed, so the convention does not enter: across released TD-MPC2 agents on 14 tasks in two domains and two further agent families, *refinement iterations* are the more reliable way to convert planner compute into return. They recover full-planner performance on weak-prior high-dimensional bodies and on pixel PlaNet, where width does not (e.g. $+534$, $p = 0.002$, where the best width-only configuration still falls 232 short), with 90% of the gain arriving between the second and fourth iteration on all four TD-MPC2 tasks swept (4-D to 38-D); on state-based PETS one well-configured wide pass matches them, and on a poorly fit model they hurt. How much refinement a task needs tracks the prior-to-task gap, a contrast carried by three weak-prior high-dimensional bodies. Every sweep was preregistered with dated amendments; of the twenty predictions tested, four held, twelve failed, one held in part and three could not be established, and we report all of them.

1 Introduction

A world model is only as useful as the level of abstraction at which it is asked to predict and at which its predictions are used. Modern world-model agents fix both choices at once: a latent model predicts one control step ahead, and a sampling-based planner (CEM, MPPI) spends its deployment budget re-deriving, from those one-step predictions, an action the agent's own policy prior has already proposed. Underneath the usual engineering question of how to spend that budget sits a question about the representation: how much of the planner's work is really the model's? If the prior already encodes what search would find, planning is redundant; if it does not, planning is the agent. That boundary can be measured, because an agent like TD-MPC2 (Hansen et al., 2024) selects every action by sampling-based model-predictive control over a learned latent model, and the same weights can be run with the planner on, cheap, or off.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

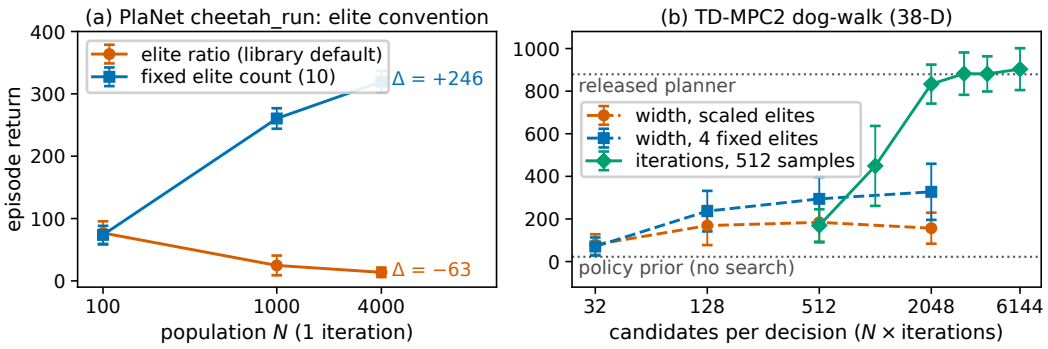

Figure 1: The two halves of the question, each on one measured cell. (a) How many samples: on identical PlaNet models and eval seeds, widening the population at one iteration lowers return when the elite set is a fixed fraction and raises it when it is a fixed count (Table 2; mean ± sample SD over 10 paired episodes). (b) Where to spend compute: on a released 38-D TD-MPC2 agent, width at one iteration stays far below the released planner under either elite rule, while refinement iterations at 512 samples reach it (mean ± SD over 3 released seeds, 10 episodes each). Cells chosen as the abstract's PlaNet cell and the one high-dimensional body run under both elite rules and the iterations sweep; all other cells are in Tables 1–2 and Fig. 2.

This paper measures that boundary on its most concrete available instance: the anatomy of planner compute. A sampling planner has two levers — *parallel* width (candidate sequences per refinement step) and *sequential* depth (refinement iterations) — and how the value of search distributes over them, and over tasks, has not been measured across a released agent population. Every sweep's predictions were preregistered before its numbers existed, all agents are official released checkpoints, and one arm of our central sweep is configuration-identical to the released planner, so the harness validates itself in situ. Read through the question above, the answer on the TD-MPC2 release is that the policy prior plus a few refinement steps capture almost all of what the full planner delivers, that adding width at one iteration captures little of it where search matters, and that how much refinement a task needs tracks how well the prior fits the task. Before width can be measured at all, one setting, usually left implicit, must be fixed: whether the elite set scales with the population (§8.1). We conjecture that this makes the value of planning a property of the representation — of the abstraction at which the model predicts — rather than of search width, and we return to that reading in §10.

The findings form two nested surprises. (i) Search value is not a perturbation: on high-dimensional bodies released agents barely function without it. (ii) The scaling axis is not the obvious one — on those bodies widening the search buys little, while deepening it recovers the full planner. "Extra search buys little" holds only on low-dimensional tasks.

Contributions.

1. **A confound in width sweeps**: the elite convention significantly changes the sample-count effect in four of six cells and reverses its sign on all three PlaNet tasks, with a post-hoc mechanism measured inside the optimizers of two agent families (elite-set breadth controls proposal concentration). Sample-count comparisons require reporting the convention.
2. **An anatomy of what does convert compute into return**: refinement iterations, with 90% recovery between iterations 2 and 4 on every task swept across a $10\times$ dimensionality range, even though the per-iteration rate differs.
3. **The demand side**: how much refinement a task needs tracks the prior-to-task gap, with action dimensionality as its observable proxy ($\rho = +0.55$, $p = 0.009$ across 22 released tasks; with 22 tasks the two variables' partial correlations cannot be told apart).
4. **Preregistration with reported failures**: twelve of twenty preregistered predictions failed and three could not be established (Table 3), including a shared contraction rate across agent families.

2 RELATED WORK

Budget-efficient optimizers. iCEM (Pinneri et al., 2020) makes CEM real-time (colored noise, elite memory), reporting 2.7–22× sample savings. That work engineers a better optimizer; we measure where search value comes from across a released population, holding the optimizer fixed. **The role of planning.** Hamrick et al. (2021) found, in MuZero-style discrete domains, that planning matters mainly for learning. Our regime is complementary — continuous control, sampling MPC, deployment-time value on released weights — where planning can be nearly the whole agent. **Test-time compute for control.** ELASTIC (Li et al., 2026) adaptively allocates sequential vs. parallel inference compute for diffusion *policies*; Nguyen et al. (2026) sweep CEM budgets on low-dimensional tasks to *amortize* planning away. Neither provides the dimensionality contrast or the axis decomposition; both motivate it. **Prior–planner interaction.** TD-M(PC)^2 (Lin et al., 2025) and KL-adaptive priors (Serra-Gomez et al., 2026) repair the prior mismatch we quantify across an entire release. **Optimizer exploitation.** That strong optimizers exploit model error is documented (Zhan et al., 2022; Boney et al., 2019); we identify one knob that controls it. **Agents.** We study TD-MPC2 on DMControl (Tassa et al., 2018) and MyoSuite (Caggiano et al., 2022), and PETS (Chua et al., 2018) and PlaNet (Hafner et al., 2019) in MBRL-Lib (Pineda et al., 2021), whose elite set is a population fraction; we know of no study isolating this setting as a confound in width sweeps.

3 METHOD

Agents: a population study by design. Official single-task TD-MPC2 checkpoints (5M class), 22 DMControl tasks × 3 released seeds, 1-D to 38-D. Studying *released* agents is deliberate: the population is fixed before any hypothesis, every result is verifiable from public artifacts, and the findings describe the ecosystem practitioners deploy. Other agent families (§8) test dependence on this population; the planner-free boundary condition in §8 bounds the toggle's regime.

The toggle. Search value is measured by flipping the planner at deployment on identical weights: *full* (MPPI, 512 samples, 64 elites, 6 iterations — the released configuration), *cheap* (32/4/1), and *none* (policy prior). Ten episodes per arm, eval seeds paired across arms. We write $\mathrm{having} = \mathrm{cheap} - \mathrm{none}$ and $\mathrm{depth} = \mathrm{full} - \mathrm{cheap}$.

Statistics. Correlations: permutation tests (20,000 resamples). Paired arm contrasts: exact sign-flip tests over 10 eval seeds (smallest two-sided p is 0.002); p is uncorrected unless marked Holm. Seed-level bootstrap for fitted constants; preregistrations with dated amendments reproduced in App. D.

Caveat carried up front. These checkpoints trained *with* the planner, so the toggle measures deployment-time value for search-trained agents. The measurement is identical across tasks, so cross-task structure is unaffected; absolute values do not transfer to planner-free training.

4 SEARCH VALUE AND DIMENSIONALITY

Three preregistered predictions (that having tracks the policy-to-ceiling gap, that depth is small everywhere and shrinks with dimensionality, that sparse tasks need planners more) all *failed*; the second failed with the opposite sign, and significantly: $\mathrm{depth} \sim \mathrm{action\ dim}$ gives $\rho = +0.55$, $p = 0.009$ across the 22 tasks.

Dimensionality or prior competence? The account of §7 says the mechanistic variable is the proposal-to-target gap, with dimensionality merely its proxy. Writing $\mathrm{prior\ gap} = (1000 - R_{\mathrm{prior}})/1000$, across 22 tasks: depth ∼ dimensionality $\rho = +0.55$ ($p = 0.009$); depth ∼ prior gap $\rho = +0.59$ ($p = 0.005$); dimensionality ∼ prior gap $\rho = +0.51$ ($p = 0.017$). Partialling: depth∼dim|gap $= +0.36$, depth∼gap|dim $= +0.43$. The variables are genuinely entangled; the prior gap has the larger partial, as the account requires, but a task bootstrap puts the difference of partials at $[-0.97, 0.74]$, so the data do not separate the two. We therefore state the phenomenon as a **prior-competence × dimensionality interaction**, keeping dimensionality as the practitioner-facing proxy because it is known before training.

How robust is the correlation, and where does it live? Four checks on the depth–dimensionality correlation, all from the same per-checkpoint returns. (i) Leave one task out: ρ stays between $+0.49$ and $+0.64$. (ii) Each released seed on its own: $\rho = +0.57$ ($p = 0.007$), $+0.56$ ($p = 0.008$)

	return ↑			planner value	
task (action dim)	none	cheap	full	having	depth
high-dimensional, weak prior					
humanoid-walk (21)†	3±1	9±2	896±15	+6±1	**+887**±13
dog-walk (38)	22±13	77±52	879±134	+54±40	**+803**±186
dog-run (38)	54±10	77±18	774±87	+22±10	**+697**±105
low-dimensional					
walker-walk (6)	430±47	836±84	988±1	+405±114	+152±84
hopper-hop (4)	351±163	531±72	492±65	+180±93	−38±10
acrobot-swingup (1)	635±18	570±28	627±4	−65±10	+56±24

Table 1: Planner value decomposition, representative tasks (full 22-task table in App. A): mean ± sample SD over the 3 released seeds (10 paired episodes each). Bold: depth exceeds the policy's own return by an order of magnitude. †seeds 1–2; seed 3 is a defective released checkpoint (App. E).

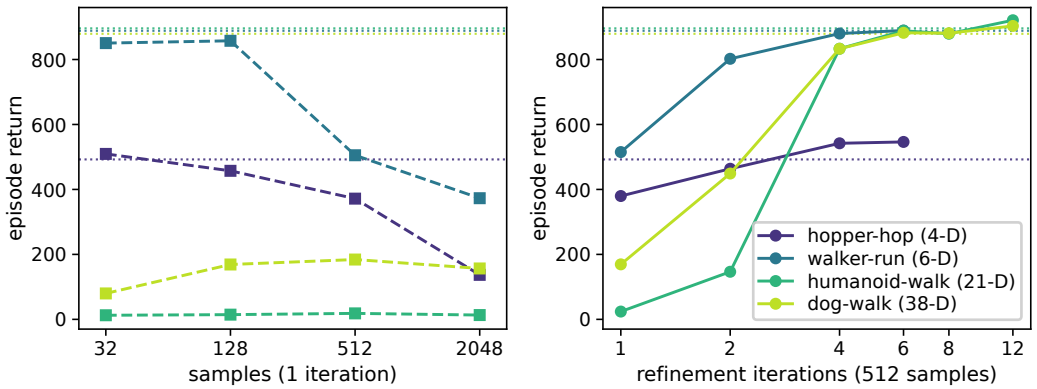

Figure 2: The two axes (dotted: released full-MPPI references). Left: samples at one iteration with elites scaled as $\max(4, N/8)$; the high-dimensional bodies stay far below the full planner and the low-dimensional ones decline (elite dilution, §5). Right: iterations at 512 samples recover full-planner performance by 4–6 on every task.

and $+0.45$ ($p = 0.042$; 21 tasks without the defective checkpoint). (iii) Measurement noise, the sensitivity check our preregistration required before any cross-task claim: per-task evaluation noise (mean per-episode standard deviation over arms and checkpoints) does not track depth ($\rho = +0.11$, $p = 0.61$), and depth expressed in noise units still rises with dimensionality ($\rho = +0.57$, $p = 0.006$). having fails the same check ($\rho = +0.57$ with noise, $p = 0.006$), so we draw no conclusion from it. (iv) Restricted to the 19 tasks of at most 6-D, the correlation is $+0.30$ ($p = 0.20$), and the magnitude $|\text{depth}|$ does track noise ($\rho = +0.49$, $p = 0.022$): the small low-dimensional depths are partly noise. On the three high-dimensional bodies depth is 29–54 noise units, against under 2 for the largest low-dimensional effects (walker-walk, cheetah-run). The effect is therefore a contrast between weak-prior high-dimensional bodies and everything else, not a smooth dose–response in dimensionality, and we state it that way.

5 Sequential beats parallel

We decompose search depth into its axes on four tasks spanning the range (4-D to 38-D), 3 seeds (humanoid: seeds 1–2), paired episodes (Fig. 2).

Parallel. Sweeping $\{32, 128, 512, 2048\}$ samples at one iteration: humanoid is inert at 13–18 (against ∼890 for the full planner) and dog-walk peaks at 184 (against 879); neither high-dimensional body approaches full planning at any width. On the low-dimensional tasks, whose priors are already competent, one wide pass does come close: walker-run reaches 858 at 128 samples, against 888 for the full planner.

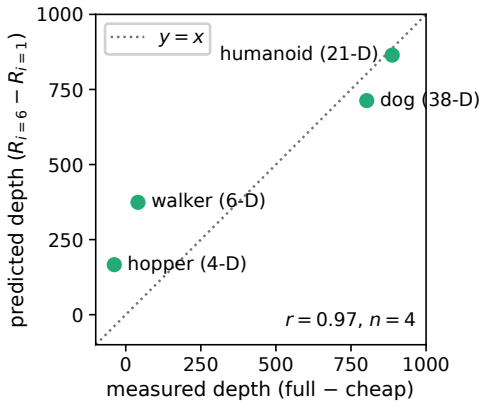

Figure 3: Search depth measured directly (released budgets) against the value predicted from the iterations sweep alone ($n = 4$). The two agree on the high-dimensional bodies; the low-dimensional points carry the elite-dilution artifact discussed in the text.

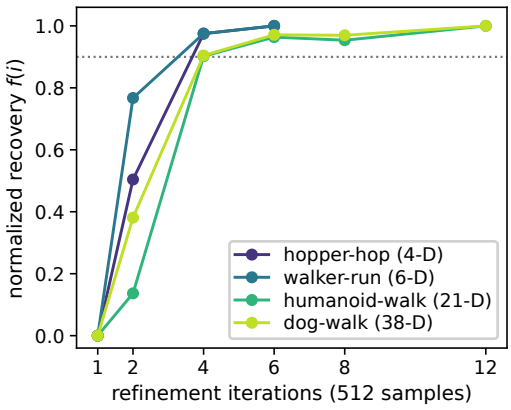

Figure 4: Normalized recovery $f(i)$ at 512 samples (§7): every task crosses 90% (dotted) between i=2 and i=4 across a 10× dimensionality range, at different rates.

Sequential. Sweeping iterations at 512 samples: every task recovers its reference and saturates by 4–6 (dog-walk $169 \to 882$ vs. reference 879; Spearman over the 12 seed-level points $\rho = +0.89$, $p < 0.001$). The i=6 arm is configuration-identical to the released planner and reproduces it within noise — an in-situ harness validation.

Elite dilution, not width. A preregistered control explains the declining samples curves: the original sweep scaled elites with width ($m = \max(4, N/8)$), confounding the two. At 4 *fixed* elites the decline vanishes (hopper stays within 484–542 across 64×; dog rises monotonically, 71 to 327, against 80 to 157 with scaled elites, still far below its 882 via iterations). The harm was elite-set dilution, not width; the central claims never depended on it.

6 The Bridge: Refinement Reproduces the Depth Column

If refinement is what depth provides, Table 1's depth column should be reproducible from the iterations sweep alone via $\widehat{\text{depth}} = R_{i=6} - R_{i=1}$. Where the effect lives it is: within 3–11% on the high-dimensional bodies (humanoid 887 measured vs. 864 predicted; dog 803 vs. 713), and $r = 0.97$ over the four swept tasks, a correlation those two points carry. On the low-dimensional tasks the prediction overshoots (hopper −38 vs. 167; walker-run 41 vs. 374), because $R_{i=1}$ at 512 samples carries the elite-dilution loss above.

7 A Refinement Account, and Measuring It Inside the Planner

MPPI/CEM approximate an exponentially tilted target $p^*(a) \propto q(a)e^{\eta \hat{J}(a)}$ by iteratively refitting a proposal q (Williams et al., 2017). Three propositions follow (full statements, assumptions and where they fail: App. C).

P1 (contraction). Under smoothed moment projection with constant α, the natural-parameter gap contracts geometrically, so normalized recovery is $f(i) = 1 - (1 - \alpha)^{i-1}$ — a property of the optimizer, not the task. **P2 (width cannot substitute).** A single iteration is self-normalized importance sampling, whose effective sample size decays as $e^{-\mathrm{KL}(p^* \| q)}$ (Chatterjee & Diaconis, 2018): when the prior gap is large, no affordable width suffices. **P3 (elite dilution).** With fixed elite count the max-of-N bias grows only as $\sqrt{\log N}$; with $m \propto N$ the refit is increasingly set by mediocre candidates — predicting declines exactly under the scaled schedule, as observed.

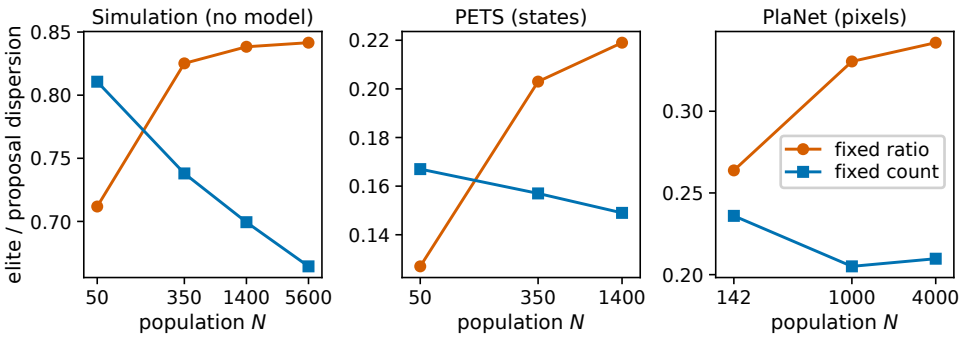

Figure 5: P4 across three settings: a pure order-statistics simulation with no model and no learning (left), a trained PETS planner over an ensemble MLP with state observations (centre), and a trained PlaNet planner over a recurrent latent model with pixel observations (right). In all three, tying the elite set to the population keeps the refit proposal broad as sampling grows, while a fixed elite count sharpens it: the post-hoc mechanism we propose for the sign reversal in returns, consistent across both architectures measured.

P4 (the elite convention sets proposal concentration). *Assumptions.* Candidates are sampled i.i.d. from the current proposal and scored with additive noise; the next proposal is refit to the top-m (the *elites*). *Statement.* Under a fixed elite *ratio* $m = \rho N$, the elites are all candidates above the $(1-\rho)$ quantile, so as N grows their empirical dispersion converges upward to the conditional dispersion of that tail region — a fixed positive constant. Under a fixed elite *count*, the top-m of N are the m most extreme order statistics and concentrate toward the maximum, so their dispersion decreases in N. P4 is post hoc: our preregistered prediction had the opposite sign (Table 3, row 20). *Consequence.* Growing the population sharpens the proposal under a fixed count and leaves it broad (indeed broader, approaching the asymptote from below) under a fixed ratio. The sign of "does drawing more samples help?" therefore depends on the convention, not only on the agent or the task. Fig. 5 shows the prediction in a pure order-statistics simulation with no model and no learning, beside the same quantity measured inside two trained planners from different families (an ensemble MLP over states, and a recurrent latent model over pixels); the shapes agree in all three, though PlaNet's fixed-count dispersion flattens above $N{=}1000$ (0.205 vs. 0.210). This also connects to classical CEM analysis, where convergence requires the elite quantile to shrink or to be smoothed — the fixed-ratio convention never shrinks it.

What holds, and what does not. The saturation point is similar: every measured task curve crosses 90% recovery between $i{=}2$ and $i{=}4$ (linear interpolation $i_{90} \in [3.3, 4.0]$; Fig. 4), though geometric fits place it at 2.6–6.0; the grid sets the narrow range. The *rate* is not: per-task α spans 0.37–0.77 and seed-bootstrap intervals for the extremes are disjoint ($[0.76, 0.77]$ vs. $[0.37, 0.37]$), so we reject a universal contraction constant. What generalizes is the budget (4–6 iterations), not the rate.

Instrumenting the planner. P1's constant can also be measured directly. Recording the CEM proposal mean at each refinement step on a trained PETS agent (50 replans, all fitted) gives a contraction of $c = 0.265$ (IQR 0.243–0.283) against a *nominal* step of 0.9 — a $3.4\times$ gap. This is P1's exact-moments assumption failing in the predicted direction: the elite mean is itself noisy, so most of each nominal step buys estimation noise. It also explains why several iterations are needed when one update nominally almost jumps to the target.

8 Generality

Out-of-family domain (MyoSuite). All ten released MyoSuite hand tasks (39-D overactuated muscle control, 3 seeds; Table 5 in App. B): iterations dominate width on **9 of 10** tasks (median per-task Spearman over seed-level points: $+0.69$ vs. -0.41). Single-shot width can be destructive here (obj-hold falls $757 \to -1710$ at 2048 samples, restored to 766 by iterations), though this sweep used the scaled-elite schedule of §5, so the collapse includes elite dilution. MyoSuite also *bounds* the contrast: all ten bodies are 39-D, yet on four of them (key-turn, pose, obj-hold, reach-hard) the full plan-

			width, elite *ratio*		width, elite *count*		iterations	
family	task	dim	Δ return	p	Δ return	p	Δ return	p
PETS	Hopper	3	$+39 \pm 43$	0.006	$+88 \pm 82$	0.016	$\mathbf{+89} \pm 36$	0.002
(states)	Walker2d	6	-4 ± 81	0.889	-56 ± 98	0.105	-107 ± 90	0.014
	Ant	8	$\mathbf{+1273} \pm 719$	0.002	$+87 \pm 359$	0.439	-265 ± 750	0.273
PlaNet	cheetah_run	6	$\mathbf{-63} \pm 22$	0.002	$\mathbf{+246} \pm 21$	0.002	$\mathbf{+534} \pm 15$	0.002
(pixels)	walker_walk	6	-18 ± 29	0.088	$\mathbf{+78} \pm 53$	0.002	$\mathbf{+118} \pm 83$	0.002
	finger_spin	2	-19 ± 16	0.012	$\mathbf{+153} \pm 40$	0.002	$\mathbf{+511} \pm 184$	0.004

Table 2: Agent-family sweeps. Width arms compare the largest against the smallest population at one refinement iteration, under both elite conventions; iteration arms compare the most against the fewest iterations at fixed population. Δ: mean paired difference in episode return ± sample SD of the 10 paired differences. p: exact paired sign-flip test over 10 eval seeds, uncorrected; bold: significant after Holm correction over the 18 tests ($p_{\text{Holm}} < 0.05$).

ner moves return by only -96 to $+14$ over the prior, while on pen-twirl it doubles it ($2204 \rightarrow 4712$). Dimensionality alone does not set search value, which a dimensionality-only account gets wrong.

8.1 Agent families: PETS and PlaNet

We repeat the sweeps on agents from a different codebase and different model classes: PETS (CEM over a probabilistic ensemble, *no policy prior*) on three gym-MuJoCo tasks, and PlaNet (CEM over a recurrent latent model, *pixel* observations) on three DMControl tasks. Models were trained by us; sweeps use 10 paired episodes per arm with the planner RNG re-seeded identically per arm (Table 2; per-arm returns in Table 6, App. B); we report all cells.

The convention changes the answer, and reverses it on PlaNet. Under the elite-*ratio* convention, widening the search *lowers* return on all three PlaNet tasks (significantly on two; walker_walk $p = 0.088$); under a fixed elite *count*, on the same models and seeds, it is significantly *helpful* on all three. With a Holm correction over the table's 18 tests the count-side gains all survive, but the ratio-side decline only on cheetah_run. On PETS the discrepancy is large but not a reversal — Ant shows $+1273$ under a ratio against $+87$ under a count. A paired test of the per-episode difference between the two width effects is significant after Holm correction over the six cells on all three PlaNet tasks and on Ant ($p_{\text{Holm}} = 0.012$ each), but not on Hopper ($p = 0.13$) or Walker2d ($p = 0.037$, 0.074 Holm). We therefore state the claim in the form the data supports: **the elite convention significantly changes the measured effect of population size in four of six cells, and reverses its sign in three (all PlaNet), with a family-dependent direction.** Width sweeps are not comparable across implementations that differ in this choice.

Refinement dominates where search helps at all, but not universally. Iterations produce large significant gains on four of six cells (up to $+534$ on a pixel RSSM), no significant effect on one (Ant, whose returns are large and highly variable), and a *loss* on one ($p = 0.014$; 0.096 with Holm) — PETS Walker2d, the agent whose model validation error is two orders of magnitude worse than its siblings' (0.157 vs 0.0015). Where width is configured well (fixed elite count), iterations still dominate it on PlaNet by 1.5–3.3×, but merely match it on PETS Hopper ($+88$ vs $+89$): with a 3-D action space and an accurate model, one well-configured wide pass is enough.

Why refinement can hurt (supporting; the phenomenon is not ours). That a strong optimizer exploits an inaccurate model is well documented (Zhan et al., 2022; Boney et al., 2019); we add a direct measurement of it (Fig. 6). Recording the value the planner *predicts* for the plan it executes, alongside the return actually obtained, the optimism gap is flat across an eightfold change in refinement on the accurate-model agent ($+5.8 \rightarrow +5.8$, validation error 0.0015) and grows sevenfold on the inaccurate-model agent ($+1.6 \rightarrow +11.0$, validation error 0.157), where predicted value climbs $47.7 \rightarrow 58.9$ while realized return moves only $46.2 \rightarrow 47.9$. Growing optimism is not by itself the signature of a bad model: on PlaNet cheetah_run the gap also rises ($+0.75 \rightarrow +4.67$ from one to ten iterations), but realized reward rises with it ($8.3 \rightarrow 14.2$ per 12-step horizon). What separates PETS Walker2d is that realized return does not follow the prediction up — refinement is then spent on model error, which is why the same operation is worth $+89$ on one agent and -107 on the other. A

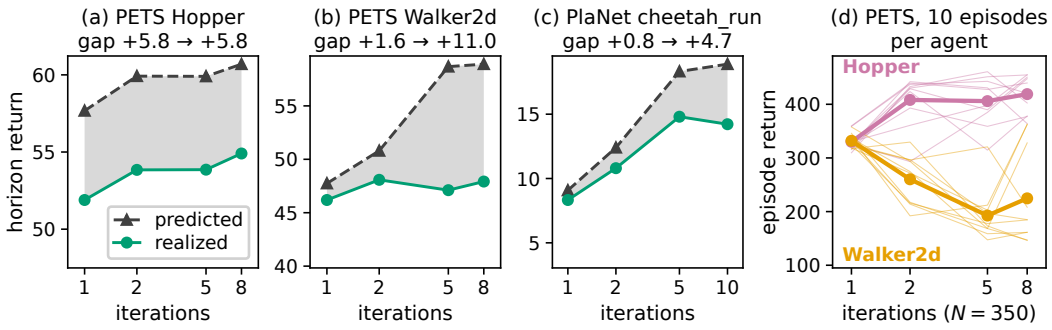

Figure 6: Case study: where refinement's extra predicted value goes. (a–c) The planner's predicted return for the plan it executes (dashed) against the reward it then obtains over the same horizon (solid; shaded: optimism gap), means over 40 replans (PETS, horizon 30) or 30 replans (PlaNet, horizon 12); the probe logs means only. (d) All 10 paired eval episodes (thin) and their mean (thick) for the two PETS agents in the iterations arm of Table 2. The PETS pair is the one named in the preregistered prediction (ledger row 19); cheetah_run is the abstract's PlaNet cell. In (a) and (c) realized reward follows the prediction up; on the poorly fit Walker2d model (b) extra refinement turns into optimism, and its episodes fall (d).

preregistered within-run version (early vs. late snapshots of one PETS Hopper run: $+100$ vs. $+71$, $p = 0.086$) is inconclusive, since validation error barely moved over the window (ledger row 15).

What this costs us. Two preregistered predictions do not survive contact with these agents: that width at one iteration never reaches the training configuration (it holds on all three PlaNet tasks and fails on all three PETS tasks; on Walker2d and Ant the training configuration scores 193 and 761 against 384 and 1293 for the best one-iteration width), and that the iterations effect is positive wherever the model is usable (false on Walker2d, and unresolved on Ant). Both are in Table 3.

A boundary condition (planner-free training). On our own TD-MPC2 agents trained with the planner *disabled*, attaching a planner at deployment is net *negative* on walker-run at every checkpoint (-5 to -40), and hopper-hop barely trains at all without search (prior return 35 at 400k steps, mean of two seeds, one of which scores 0). Such policies have already absorbed what search would supply. We therefore scope the paper explicitly: it measures search value for agents whose policies were *shaped by search during training* — the regime every released checkpoint belongs to.

9 The Preregistration Ledger

Every sweep in this paper was preceded by a written prediction and a decision rule. Table 3 lists all twenty predictions the sweeps tested, grouped by sweep, with the outcome as measured; failures are not set apart from successes. Two patterns matter for how the rest of the paper should be read. First, most failures are of our *mechanistic* guesses (coverage, collapse, a shared contraction rate), while the *measurements* they were meant to explain replicated; the paper's claims are therefore stated at the level of the measurements. Second, three outcomes (rows 9, 15, 16) are recorded as not established because the test we specified could not decide them.

10 Practical Rules and Limitations

Rules. (1) Allocate sequential compute: 4–6 refinement iterations at moderate width beat every single-shot width we ran on DMControl TD-MPC2 and PlaNet, and on 9 of 10 MyoSuite tasks (on PETS one wide pass can match them). (2) Budget by prior competence, using action dimensionality as its observable proxy: below $\sim$6-D a 32-sample planner captures most value; above $\sim$20-D with a weak prior the planner *is* the agent. (3) On TD-MPC2 and PlaNet, keep the elite count fixed when widening, and always report the convention. (4) Audit released checkpoints: one of 66 was defective in a way invisible to aggregate reporting.

	prediction (sweep)	measured	verdict
1	having rises with the prior gap (22 tasks)	$\rho = +0.31$, $p = 0.17$	failed
2	depth small everywhere; if anything falls with action dim	rises: $\rho = +0.55$, $p = 0.009$	failed
3	sparse tasks gain more from having at matched gap	sparse residual -64, $p = 0.89$	failed
4	hopper, walker saturate by 32–128 samples (1 iteration)	decline: hopper $509 \to 137$, walker $858 \to 373$	failed
5	humanoid, dog rise monotonically with samples	humanoid inert (13–18); dog $80 \to 184 \to 157$	failed
6	humanoid, dog rise steeply with iterations	humanoid $24 \to 888$, dog $169 \to 882$	held
7	hopper, walker flat or falling in iterations	rise: hopper $380 \to 546$, walker $515 \to 889$	failed
8	humanoid keeps rising to i=12; dog stays on its plateau	humanoid 921, dog 903 at i=12	held
9	PETS/PlaNet: return flat in population at training iterations	3-level rank test cannot reject; paired test significant in 5 of 6 cells	not established
10	PETS/PlaNet recovery is geometric with TD-MPC2's $\alpha \in [0.37, 0.77]$, RMSE ≤ 0.15	criteria met in 2 of 6 cells	failed
11	PETS refinement need at least that of TD-MPC2's weak-prior tasks	0.21, -0.48, -0.28 vs. 0.97 (humanoid)	failed
12	width at one iteration never reaches the training configuration	holds on 3/3 PlaNet, fails on 3/3 PETS	failed
13	fixed elite count removes the width decline, leaving width flat and below iterations	decline removed; width rises on 3/3 PlaNet; equals iterations on PETS Hopper	in part
14	iterations help wherever the model is usable	Walker2d -107 ($p = 0.014$); Ant n.s.	failed
15	the iterations effect grows with model quality within one run	$+100$ early vs. $+71$ late, $p = 0.086$; no quality gradient	not established
16	iterations effect ordered Hopper $>$ Ant $>$ Walker2d by model error	fails on absolute, holds on relative scale; Ant $p = 0.27$	not established
17	measured in-planner contraction c well below nominal 0.9	$c = 0.265$	held
18	c lies in the return-derived range 0.37–0.77	0.265	failed
19	optimism gap widens with iterations on Walker2d, not on Hopper	$+1.6 \to +11.0$ vs. $+5.8 \to +5.8$	held
20	fixed ratio *shrinks* proposal dispersion; fixed count flat or rising	reversed: $0.127 \to 0.219$ vs. $0.167 \to 0.149$	failed

Table 3: All preregistered predictions tested by the sweeps in this paper, grouped by sweep. Rows 1–3: 22-task toggle; 4–5: samples sweep; 6–8: iterations sweeps; 9–16: agent-family sweeps and controls; 17–18: contraction probe; 19: optimism probe; 20: dispersion probe (Fig. 5).

What this says about the model. Read as a statement about the representation rather than the search, the results line up by how well the policy prior already fits the task. Where the prior is nearly blind, the planner supplies almost the whole return (humanoid-walk, 3 to 896); where it is competent, a fraction (walker-walk, 430 to 988); where it is strong, little or nothing (hopper-hop; four of ten MyoSuite hands; our planner-free-trained agents). Across the 22 released tasks the prior gap predicts that residual at least as well as dimensionality does ($\rho = +0.59$ vs. $+0.55$; partials $+0.43$ vs. $+0.36$, not separable), and what the planner recovers it recovers by refining the prior's proposal a few times (90% by the fourth iteration), not by sampling around it more widely, which on the weak-prior bodies stays far below the full planner at every width we ran. On this reading the lever is the representation: search pays back, one refinement at a time, what the prior failed to encode. We leave open at which timescale a model should predict so that the residual is small — a choice the options framework (Sutton et al., 1999) and temporally abstract world models (Kim et al., 2019; Hafner et al., 2022) make by design, and which the dynamics' structural information (Li & Pan, 2016) might inform.

Limitations. Deployment-time scope (train-time search value is not measured); returns-based tasks without early termination; 3 released seeds; the refinement account assembles known results, one of its predictions (a universal rate) is refuted, and P4 is post hoc; the elite-convention result rests on six cells from two families (one reverses significantly on both sides after Holm correction), and we did not survey how often papers state the convention; the PETS dispersion probe (12 replans per cell) is reported at three-decimal precision without per-replan intervals, and the PlaNet probe uses N=142 and 10 iterations, not the return sweep's 100 and 1; and PlaNet evaluations are not bit-reproducible even with the planner RNG re-seeded (identical configurations differ by up to 116 points, Table 6), so its smaller effects (walker_walk) should be read against that floor.

REPRODUCIBILITY STATEMENT

All checkpoints are public releases. Harnesses, per-episode data, and preregistration documents with dated amendments are in the supplementary material; the configuration-identity replication doubles as an end-to-end harness audit.

AI USE STATEMENT

We used a generative AI assistant (Claude, Anthropic), under the authors' direction, to give feedback on experimental design, to implement analysis and plotting code, to help interpret results, to draft and edit the text, to create and edit figures and tables, to identify and summarise related literature, and to format references. We did *not* use generative AI tools to generate synthetic data reported as measurements, or for translation. We have reviewed all AI-assisted work: every reported number was checked by the authors against the logged experimental outputs, and every reference was checked against its published record. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

REFERENCES

N. Hansen, H. Su, X. Wang. TD-MPC2: Scalable, robust world models for continuous control. ICLR, 2024.

C. Pinneri et al. Sample-efficient cross-entropy method for real-time planning. CoRL, 2020.

J. Hamrick et al. On the role of planning in model-based deep RL. ICLR, 2021.

A. Z. Li, G. Swamy, Y. Bisk, A. Bajcsy. ELASTIC: Efficiently learning to adaptively scale test-time compute for generative control policies. arXiv:2606.31132, 2026.

H. Nguyen, X. Xu, X. Huang. Latent geometry beyond search: amortizing planning in world models. arXiv:2605.08732, 2026.

H. Lin, P. Wang, J. Schneider, G. Shi. TD-M(PC)^2: Improving temporal difference MPC through policy constraint. arXiv:2502.03550, 2025.

Á. Serra-Gomez, D. Jarne Ornia, D. Tirumala, T. Moerland. A KL-regularization framework for learning to plan with adaptive priors. ICML, 2026. arXiv:2510.04280.

X. Zhan et al. Model-based offline planning with trajectory pruning. IJCAI, 2022.

R. Boney et al. Regularizing trajectory optimization with denoising autoencoders. NeurIPS, 2019.

G. Williams et al. Information theoretic MPC for model-based reinforcement learning. ICRA, 2017.

S. Chatterjee, P. Diaconis. The sample size required in importance sampling. Ann. Appl. Probab., 2018.

R. S. Sutton, D. Precup, S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence, 112(1–2):181–211, 1999.

T. Kim, S. Ahn, Y. Bengio. Variational temporal abstraction. NeurIPS, 2019.

D. Hafner, K.-H. Lee, I. Fischer, P. Abbeel. Deep hierarchical planning from pixels. NeurIPS, 2022.

A. Li, Y. Pan. Structural information and dynamical complexity of networks. IEEE Trans. Inf. Theory, 62(6):3290–3339, 2016.

Y. Tassa et al. DeepMind Control Suite. arXiv:1801.00690, 2018.

V. Caggiano, H. Wang, G. Durandau, M. Sartori, V. Kumar. MyoSuite: A contact-rich simulation suite for musculoskeletal motor control. L4DC, 2022. arXiv:2205.13600.

K. Chua, R. Calandra, R. McAllister, S. Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. NeurIPS, 2018.

D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, J. Davidson. Learning latent dynamics for planning from pixels. ICML, 2019.

L. Pineda, B. Amos, A. Zhang, N. O. Lambert, R. Calandra. MBRL-Lib: A modular library for model-based reinforcement learning. arXiv:2104.10159, 2021.

A FULL 22-TASK DECOMPOSITION

task	dim	none	cheap	full	having	depth
acrobot-swingup	1	635	570	627	-65	+56
cartpole-balance	1	1000	1000	999	+0	-1
cartpole-balance-sparse	1	993	1000	1000	+7	+0
cartpole-swingup	1	648	632	676	-16	+44
cartpole-swingup-sparse	1	253	282	325	+29	+43
pendulum-swingup	1	566	564	520	-1	-44
cup-catch	2	981	980	980	-1	+0
cup-spin	2	849	849	849	-0	-0
finger-spin	2	951	987	987	+35	+0
finger-turn-hard	2	895	972	980	+78	+7
reacher-easy	2	983	984	987	+1	+3
reacher-hard	2	980	979	984	-1	+5
hopper-hop	4	351	531	492	+180	-38
hopper-stand	4	675	915	867	+240	-48
fish-swim	5	640	797	820	+157	+23
cheetah-run	6	407	719	885	+312	+166
walker-run	6	803	848	888	+45	+41
walker-stand	6	819	944	996	+126	+51
walker-walk	6	430	836	988	+405	+152
humanoid-walk	21	3	6	598	+4	+591
dog-run	38	54	77	774	+22	+697
dog-walk	38	22	77	879	+54	+803

Table 4: All 22 tasks (means over 3 released seeds $\times$ 10 paired episodes). humanoid-walk includes the defective seed-3 checkpoint here for completeness; see App. E.

B DETAIL TABLES FOR THE GENERALITY SWEEPS

	width sweep (1 iteration, scaled elites)				iterations sweep (512 samples, 64 elites)			
task (39-D)	prior	N=32	N=512	N=2048	prior	i=1	i=2	i=6
key-turn	1172	927	502	595	1172	485	951	1076
key-turn-hard	624	828	807	759	786	553	886	988
obj-hold	757	755	−140	−1710	757	−216	757	766
obj-hold-hard	−687	−526	−953	−1906	−341	−1266	−468	471
pen-twirl	2204	2851	2464	2531	2204	2464	3922	4712
pen-twirl-hard	2678	2900	2232	2405	3255	2972	3613	4422
pose	684	658	578	577	684	589	694	698
pose-hard	−43	−134	−205	−255	−116	−188	−70	−29
reach	452	540	591	295	452	575	758	759
reach-hard	681	700	725	686	701	710	755	707

Table 5: MyoSuite, all ten released hand tasks (mean of 3 seeds $\times$ 10 paired episodes). Each sweep evaluates the prior in its own run; the two prior columns differ where the prior's own returns are noisy. i=6 at 512 samples is the released planner.

			width, elite ratio			width, elite count			iterations			
family	task	train	N_1	N_2	N_3	N_1	N_2	N_3	$i{=}1$	$i{=}2$	$i{=}5$	last
PETS	Hopper	406	290	330	329	277	400	365	330	408	406	419
PETS	Walker2d	193	336	332	332	384	339	328	332	260	193	225
PETS	Ant	761	20	1204	1293	299	433	386	1204	822	761	939
PlaNet	cheetah_run	551	77	25	14	74	260	320	18	273	431	552
PlaNet	walker_walk	202	61	53	44	56	104	134	50	59	212	169
PlaNet	finger_spin	398	23	5	5	19	111	172	3	230	485	514

Table 6: Mean returns behind Table 2 (10 paired episodes per cell). PETS: $N \in \{50, 350, 1400\}$, training configuration 350×5 iterations, last $= 8$ iterations at $N{=}350$. PlaNet: $N \in \{100, 1000, 4000\}$, training configuration 1000×10, last $= 10$ at $N{=}1000$. Width arms use one iteration; the count arms hold 10 elites; the iteration arms use the library's elite ratio. On PETS, $i{=}1$ and $i{=}5$ reproduce the ratio arm at N_2 and the training configuration exactly. On PlaNet they do not: identical configurations evaluated in different arms differ by up to 116 points (finger_spin training configuration, 398 vs. 514; one episode scores 0 in one repeat and 574 in the other), none significantly (paired $p \geq 0.33$). This residual nondeterminism is the noise floor against which the PlaNet effects in Table 2 should be read.

C Statements and assumptions for P1–P3

We restate P1–P3 with their assumptions, sketch why each holds, and — because the assumptions are idealizations — state explicitly where our measurements agree and where they do not. None of the mathematics is new; the contribution is the assembly and the tests.

P1 (geometric contraction; assumptions and sketch). *Assumptions.* (i) The proposal q_k lies in an exponential family with natural parameter η_k (Gaussian, in every planner we study); (ii) the per-iteration update is the smoothed information projection $\eta_{k+1} = (1-\alpha)\,\eta_k + \alpha\,\eta_k^*$, where η_k^* are the natural parameters matching the moments of the reweighted (or elite) population and $\alpha \in (0, 1]$ is the smoothing constant; (iii) the moment estimate is unbiased, i.e. the sample size is large enough that $\eta_k^* \approx \eta^*$. *Sketch.* Under (i)–(iii), $\eta_k - \eta^* = (1-\alpha)^k(\eta_0 - \eta^*)$, so the parameter gap contracts geometrically at a rate that does not depend on its initial magnitude. If return is locally monotone in the gap, the *normalized* recovery curve is $f(i) = 1 - (1-\alpha)^{\,i-1}$ — identical across tasks up to α, which is a property of the optimizer, not of the task. This is the collapse of Fig. 4. *Where reality departs.* Assumption (iii) fails when the initial gap is very large: early iterations then estimate η_k^* from a population containing almost no high-return sequences, and the effective contraction stalls until the proposal reaches the mode. This predicts a sigmoidal, not geometric, curve exactly in the weak-prior high-dimensional regime — observed on humanoid-walk (§7). MPPI's exponentially weighted mean is a soft-elite instance of the same update, with an effective α set jointly by temperature and population; we therefore fit α rather than reading it from the configuration.

P2 (width cannot substitute; assumptions and sketch). *Assumptions.* A single-iteration search is self-normalized importance sampling of the tilted target p^* under the proposal q, with unbiased return estimates. *Sketch.* Chatterjee & Diaconis (2018) show that the sample size required for self-normalized importance sampling to be accurate scales as $\exp(\mathrm{KL}(p^* \| q))$ (equivalently, ESS/N decays exponentially in the divergence). Since the tilted target concentrates on near-optimal action sequences and q is the policy prior, the exponent grows with prior–task mismatch — which, across our suites, grows with action dimensionality. Hence any affordable N leaves the single-iteration search far from p^*: *returns are flat-to-weakly-rising in width and never reach the full planner*. Observed at both extremes of our range (§5). *Where reality departs.* Our return estimates are *not* unbiased — they come from a learned model and value function — which is precisely the door P3 walks through.

P3 (elite dilution, not width; assumptions and sketch). *Assumptions.* Scores are the true objective plus zero-mean noise of scale σ (model and value error); m elites are selected from N candidates and their mean defines the next proposal. *Sketch.* For *fixed* m, the optimism induced by selecting the top-m of N noisy scores grows only as $\sigma\sqrt{2\log N}$, so widening the search is nearly free: the selected set tracks the true optimum with slowly growing bias. If instead the elite count

scales with width, $m = \rho N$, the selected set is a fixed *quantile*: its mean is pulled toward the bulk of the population, so as N grows the refit proposal is increasingly determined by mediocre candidates. *Prediction: declines with width appear when $m \propto N$ and vanish at fixed m*, which is exactly the preregistered control of §5 (hopper: flat $505 \to 525$ at fixed m, versus an apparent monotone decline under the scaled schedule). *Where reality departs.* The noise is neither zero-mean nor independent across candidates — correlated model error is the well-documented failure mode of learned-model planning (Zhan et al., 2022; Boney et al., 2019); our statement should be read as identifying *which knob* controls the effect, not as a complete account of its magnitude.

P4 (the elite convention sets proposal concentration). *Assumptions.* Candidates are sampled i.i.d. from the current proposal and scored with additive noise; the next proposal is refit to the top-m (the *elites*). *Statement.* Under a fixed elite *ratio* $m = \rho N$, the elites are all candidates above the $(1 - \rho)$ quantile, so as N grows their empirical dispersion converges upward to the conditional dispersion of that tail region — a fixed positive constant. Under a fixed elite *count*, the top-m of N are the m most extreme order statistics and concentrate toward the maximum, so their dispersion decreases in N. *Consequence.* Growing the population sharpens the proposal under a fixed count and leaves it broad (indeed broader, approaching the asymptote from below) under a fixed ratio. The sign of "does drawing more samples help?" therefore depends on the convention, not only on the agent or the task. Fig. 5 shows the prediction in a pure order-statistics simulation with no model and no learning, beside the same quantity measured inside two trained planners (PETS and PlaNet); the shapes agree in all three. This also connects to classical CEM analysis, where convergence requires the elite quantile to shrink or to be smoothed — the fixed-ratio convention never shrinks it.

D PREREGISTRATIONS

The preregistration documents and their dated amendments are provided verbatim in the supplementary material. Each sweep's predictions and decision rules were committed before the corresponding numbers existed; every protocol change (including the three corrections reported in the main text) is a dated amendment committed before the affected results.

E DEFECTIVE CHECKPOINT REPORT

The released humanoid-walk checkpoint for seed 3 returns 0.5–1.8 under every arm (prior; 32–2048 samples; 1–12 iterations) across four independent sweeps; we recommend exclusion and report it upstream.