

Phase 4b Final Benchmark Statistical Report

BENCHMARK COMPLETE

Pre-Registered Hypothesis Testing (Wilcoxon + Holm-Bonferroni)

Benchmark Specification

Evaluated Controllers: SAC, TD3, TAD3
Final Milestone: 1.5M steps
Bootstrap Resamples: 10,000 ($\alpha = 0.05$)
Completed At: N/A

Hypothesis Family Verdict

Comparisons Tested: 12 pairs
Statistically Significant: 4 / 12 ($p < 0.05$)
Correction Method: Holm-Bonferroni Step-Down
Unit of Analysis: Training Run Seed (\$n=10\$)

1. Descriptive Summary at Final Milestone

Gate metrics (*Completion*, *Success*) and unconditional error (*RMSE all*) are computed over **every** episode. *RMSE completed* is conditional on surviving the episode, over the # Q seeds that produced one — it is the quantity the signed-rank family tests, and it is not comparable across controllers whose completion rates differ sharply.

Controller	Condition	Completion	Success	RMSE all	RMSE completed	# Q
SAC	no_disturbance	1.000	1.000	0.0055 m	0.0055 m	10/10
TD3	no_disturbance	0.967	0.967	0.0417 m	0.0157 m	10/10
TAD3	no_disturbance	0.900	0.800	0.2787 m	0.1884 m	9/10
FLATPID	no_disturbance	1.000	1.000	0.0235 m	0.0235 m	10/10
SAC	composite	1.000	0.833	0.1322 m	0.1322 m	10/10
TD3	composite	0.767	0.333	0.3482 m	0.2832 m	10/10
TAD3	composite	0.733	0.167	0.5511 m	0.4336 m	9/10
FLATPID	composite	0.133	0.000	0.7700 m	0.4259 m	3/10

2. Statistical Hypothesis Testing Scorecard (12-Comparison Family)

Paired Wilcoxon signed-rank test across 10 benchmark seeds with Holm-Bonferroni familywise error correction at the final milestone.

3 of 12 comparisons are underpowered. Completion gating leaves too few paired seeds for the exact signed-rank test to reach $\alpha = 0.05$ at any effect size (floor 2^{1-n}). These are reported as *Underpowered*, not as evidence of no difference — read them against Section 1.

Comparison (A vs B)	Condition	# pairs	Mean Difference	95% Bootstrap CI	p-value	Outcome
SAC vs TD3	no_disturbance	10	-0.0101 m	[-0.018, -0.004]	1.3672e-02	✗ Not Sig.
SAC vs TAD3	no_disturbance	9	-0.1834 m	[-0.511, -0.015]	3.9062e-03	✓ Significant
SAC vs FLATPID	no_disturbance	10	-0.0180 m	[-0.020, -0.016]	1.9531e-03	✓ Significant

TD3 vs TAD3	no_disturbance	9	-0.1718 m	[-0.503, 0.002]	1.6406e-01	✗ Not Sig.
TD3 vs FLATPID	no_disturbance	10	-0.0078 m	[-0.013, 0.000]	1.0547e-01	✗ Not Sig.
TAD3 vs FLATPID	no_disturbance	9	0.1649 m	[-0.003, 0.493]	4.9609e-01	✗ Not Sig.
SAC vs TD3	composite	10	-0.1510 m	[-0.228, -0.084]	3.9062e-03	✓ Significant
SAC vs TAD3	composite	9	-0.3051 m	[-0.578, -0.137]	3.9062e-03	✓ Significant
SAC vs FLATPID	composite	3	-0.3240 m	[-0.378, -0.268]	2.5000e-01	⚠ Underpowered (floor 0.250)
TD3 vs TAD3	composite	9	-0.1372 m	[-0.448, 0.060]	7.3438e-01	✗ Not Sig.
TD3 vs FLATPID	composite	3	-0.2061 m	[-0.250, -0.124]	2.5000e-01	⚠ Underpowered (floor 0.250)
TAD3 vs FLATPID	composite	3	-0.1458 m	[-0.224, -0.098]	2.5000e-01	⚠ Underpowered (floor 0.250)

3. Milestone Learning Curves & Sample Efficiency

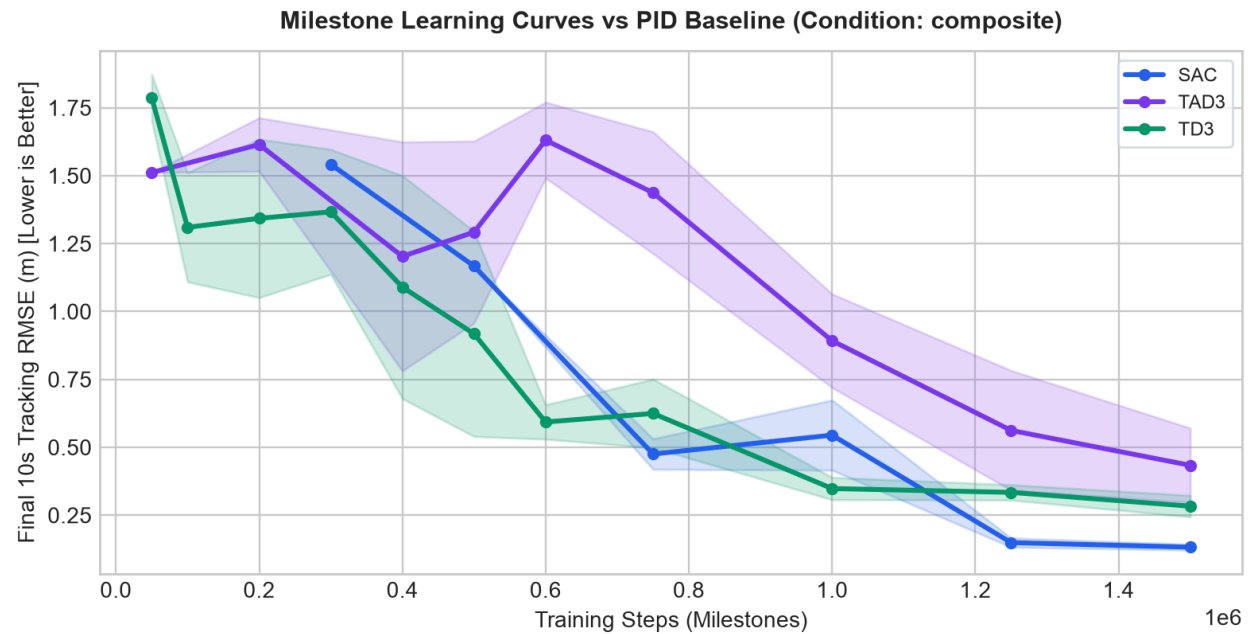


Figure 1: Progression of position tracking RMSE across training milestone checkpoints (0 to 1M steps) with ± 1 SE ribbons compared against the Classical PID baseline.

4. Pairwise Effect Sizes & Confidence Intervals

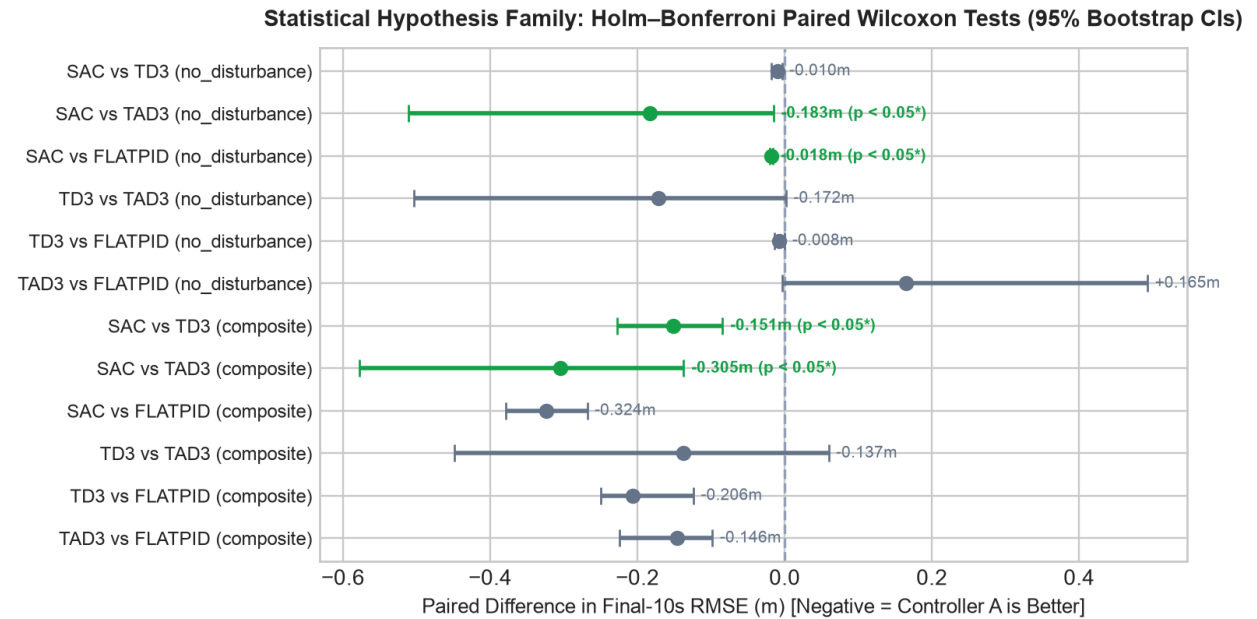


Figure 2: Paired difference in final-10s tracking error ($A - B$) with 10,000-resample bootstrap 95% confidence intervals and Holm-Bonferroni significance indicators.

5. Headline Performance Dimensions

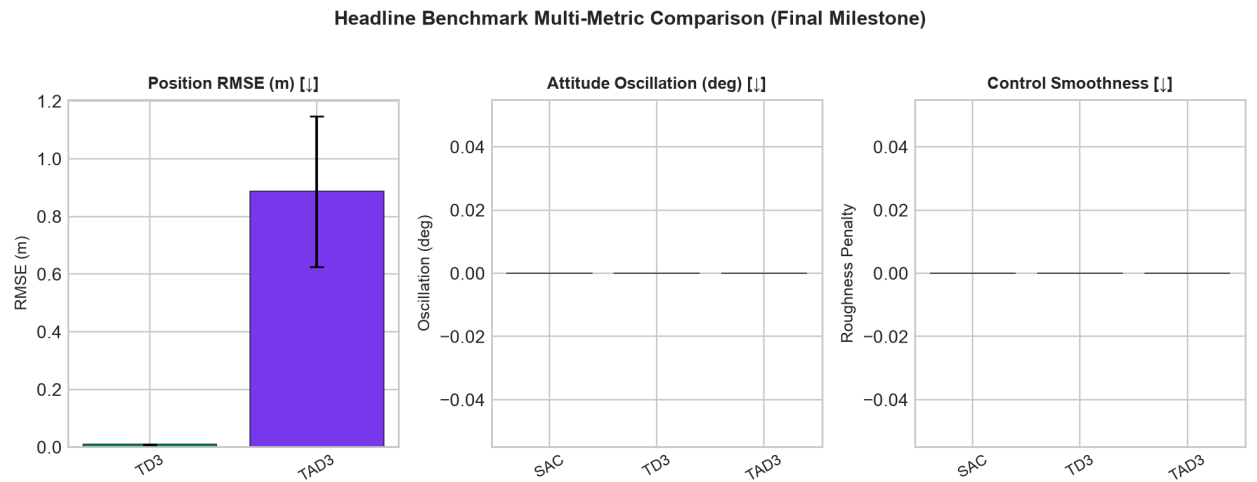


Figure 3: Multi-metric headline comparison at final milestone across tracking error, attitude oscillation, and control smoothness.