

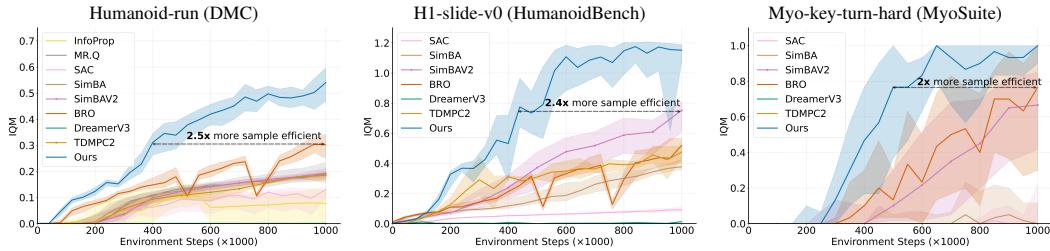
000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 WIMLE: UNCERTAINTY-AWARE WORLD MODELS WITH IMLE FOR SAMPLE-EFFICIENT CONTINUOUS CONTROL

006
007
Anonymous authors
008
Paper under double-blind review
009
010
011
012

ABSTRACT

013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
Model-based reinforcement learning promises strong sample efficiency but often underperforms in practice due to compounding model error, unimodal world models that average over multi-modal dynamics, and overconfident predictions that bias learning. We introduce WIMLE, a model-based method that extends Implicit Maximum Likelihood Estimation (IMLE) to the model-based RL framework to learn stochastic, multi-modal world models without iterative sampling and to estimate predictive uncertainty via ensembles and latent sampling. During training, WIMLE weights each synthetic transition by its predicted confidence, preserving useful model rollouts while attenuating bias from uncertain predictions and enabling stable learning. Across 40 continuous-control tasks spanning DeepMind Control, MyoSuite, and HumanoidBench, WIMLE achieves superior sample efficiency and competitive or better asymptotic performance than strong model-free and model-based baselines. Notably, on the challenging Humanoid-run task, WIMLE improves sample efficiency by over 50% relative to the strongest competitor, and on HumanoidBench it solves 8 of 14 tasks (versus 4 for BRO and 5 for SimBAv2). These results highlight the value of IMLE-based multi-modality and uncertainty-aware weighting for stable model-based RL.

1 INTRODUCTION



040
041
042
043
044
045
046
047
048
049
050
051
052
053
Figure 1: Sample efficiency on challenging tasks from each benchmark suite. WIMLE achieves superior sample efficiency and asymptotic performance over strong model-free and model-based baselines. Y-axes show interquartile mean. Shaded areas indicate 95% confidence intervals.

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
Reinforcement learning has become a powerful framework for solving complex decision-making problems across diverse domains such as autonomous control (Kiumarsi et al., 2018), strategic game playing (Hosu & Rebedea, 2016), and natural language processing (Lambert, 2025; Cetina et al., 2021). However, a significant challenge in RL is the need for a substantial number of interactions with the environment to learn a good policy. Without a simulator, learning requires real-world trials, which are costly, slow, and risky (Chen et al., 2022; Weng et al., 2023; Hessel et al., 2018; Schulman et al., 2017b).

101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
10010
10011
10012
10013
10014
10015
10016
10017
10018
10019
10020
10021
10022
10023
10024
10025
10026
10027
10028
10029
10030
10031
10032
10033
10034
10035
10036
10037
10038
10039
10040
10041
10042
10043
10044
10045
10046
10047
10048
10049
10050
10051
10052
10053
10054
10055
10056
10057
10058
10059
10060
10061
10062
10063
10064
10065
10066
10067
10068
10069
10070
10071
10072
10073
10074
10075
10076
10077
10078
10079
10080
10081
10082
10083
10084
10085
10086
10087
10088
10089
10090
10091
10092
10093
10094
10095
10096
10097
10098
10099
100100
100101
100102
100103
100104
100105
100106
100107
100108
100109
100110
100111
100112
100113
100114
100115
100116
100117
100118
100119
100120
100121
100122
100123
100124
100125
100126
100127
100128
100129
100130
100131
100132
100133
100134
100135
100136
100137
100138
100139
100140
100141
100142
100143
100144
100145
100146
100147
100148
100149
100150
100151
100152
100153
100154
100155
100156
100157
100158
100159
100160
100161
100162
100163
100164
100165
100166
100167
100168
100169
100170
100171
100172
100173
100174
100175
100176
100177
100178
100179
100180
100181
100182
100183
100184
100185
100186
100187
100188
100189
100190
100191
100192
100193
100194
100195
100196
100197
100198
100199
100200
100201
100202
100203
100204
100205
100206
100207
100208
100209
100210
100211
100212
100213
100214
100215
100216
100217
100218
100219
100220
100221
100222
100223
100224
100225
100226
100227
100228
100229
100230
100231
100232
100233
100234
100235
100236
100237
100238
100239
100240
100241
100242
100243
100244
100245
100246
100247
100248
100249
100250
100251
100252
100253
100254
100255
100256
100257
100258
100259
100260
100261
100262
100263
100264
100265
100266
100267
100268
100269
100270
100271
100272
100273
100274
100275
100276
100277
100278
100279
100280
100281
100282
100283
100284
100285
100286
100287
100288
100289
100290
100291
100292
100293
100294
100295
100296
100297
100298
100299
100300
100301
100302
100303
100304
100305
100306
100307
100308
100309
100310
100311
100312
100313
100314
100315
100316
100317
100318
100319
100320
100321
100322
100323
100324
100325
100326
100327
100328
100329
100330
100331
100332
100333
100334
100335
100336
100337
100338
100339
100340
100341
100342
100343
100344
100345
100346
100347
100348
100349
100350
100351
100352
100353
100354
100355
100356
100357
100358
100359
100360
100361
100362
100363
100364
100365
100366
100367
100368
100369
100370
100371
100372
100373
100374
100375
100376
100377
100378
100379
100380
100381
100382
100383
100384
100385
100386
100387
100388
100389
100390
100391
100392
100393
100394
100395
100396
100397
100398
100399
100400
100401
100402
100403
100404
100405
100406
100407
100408
100409
100410
100411
100412
100413
100414
100415
100416
100417
100418
100419
100420
100421
100422
100423
100424
100425
100426
100427
100428
100429
100430
100431
100432
100433
100434
100435
100436
100437
100438
100439
100440
100441
100442
100443
100444
100445
100446
100447
100448
100449
100450
100451
100452
100453
100454
100455
100456
100457
100458
100459
100460
100461
100462
100463
100464
100465
100466
100467
100468
100469
100470
100471
100472
100473
100474
100475
100476
100477
100478
100479
100480
100481
100482
100483
100484
100485
100486
100487
100488
100489
100490
100491
100492
100493
100494
100495
100496
100497
100498
100499
100500
100501
100502
100503
100504
100505
100506
100507
100508
100509
100510
100511
100512
100513
100514
100515
100516
100517
100518
100519
100520
100521
100522
100523
100524
100525
100526
100527
100528
100529
100530
100531
100532
100533
100534
100535
100536
100537
100538
100539
100540
100541
100542
100543
100544
100545
100546
100547
100548
100549
100550
100551
100552
100553
100554
100555
100556
100557
100558
100559
100560
100561
100562
100563
100564
100565
100566
100567
100568
100569
100570
100571
100572
100573
100574
100575
100576
100577
100578
100579
100580
100581
100582
100583
100584
100585
100586
100587
100588
100589
100590
100591
100592
100593
100594
100595
100596
100597
100598
100599
100600
100601
100602
100603
100604
100605
100606
100607
100608
100609
100610
100611
100612
100613
100614
100615
100616
100617
100618
100619
100620
100621
100622
100623
100624
100625
100626
100627
100628
100629
100630
100631
100632
100633
100634
100635
100636
100637
100638
100639
100640
100641
100642
100643
100644
100645
100646
100647
100648
100649
100650
100651
100652
100653
100654
100655
100656
100657
100658
10065

054 generating synthetic rollouts that augment training data for policy learning (Janner et al., 2019; Ha &
 055 Schmidhuber, 2018; Hafner et al., 2020; Clavera et al., 2020) and (2) planning by simulating future
 056 trajectories to guide action selection (Zhu et al., 2023; Frauenknecht et al., 2025; Janner et al., 2019;
 057 Lowrey et al., 2019; Hafner et al., 2019; Argenson & Dulac-Arnold, 2021). In this work, we focus
 058 on the former.

059 Historically, MBRL has struggled to surpass strong model-free baselines, largely because com-
 060 pounding rollout errors bias training and mislead the policy (Janner et al., 2019; Xiao et al., 2019;
 061 Talvitie, 2017; Frauenknecht et al., 2025; Venkatraman et al., 2015; Asadi et al., 2018b;a). We at-
 062 tribute this to two key issues: (1) standard predictive models struggle when the same state-action
 063 pair yields different, conflicting supervision due to partial observability, contact-rich dynamics, or
 064 inherent stochasticity (Kurutach et al., 2018); and (2) a lack of uncertainty awareness in model
 065 predictions (Frauenknecht et al., 2025), which leads to overconfidence in regions with complex dy-
 066 namics or limited data. Despite attempts to address these issues (Janner et al., 2019; Zhu et al., 2023;
 067 Frauenknecht et al., 2025; Somalwar et al., 2025; Hansen et al., 2022), MBRL methods have yet to
 068 consistently outperform strong model-free baselines in practice (Nauman et al., 2024; Lee et al.,
 069 2025b).

070 To address these issues, we propose WIMLE (World models with **IMLE**)—an uncertainty-aware
 071 model-based RL approach. We integrate IMLE (Li & Malik, 2018), a mode-covering generative
 072 model with demonstrated success in low-data regimes (Aghabozorgi et al., 2023; Vashist et al.,
 073 2024), into the MBRL framework. This allows us to learn world models that handle different,
 074 conflicting supervision and from which we extract predictive uncertainty estimates. We incorporate
 075 these uncertainty estimates into the RL objective to prevent overconfident predictions from biasing
 076 learning. To the best of our knowledge, this is the first work to extend IMLE for uncertainty-aware
 077 world models in MBRL.

078 We evaluate WIMLE on 40 tasks across DMC, HumanoidBench, and MyoSuite. WIMLE deliv-
 079 ers considerable gains in sample efficiency and asymptotic performance over strong model-free and
 080 model-based baselines. Notably, on the notoriously challenging Humanoid-run task, WIMLE im-
 081 proves the sample efficiency of the most competitive method by over 50%. On HumanoidBench,
 082 WIMLE successfully solves 8 of 14 tasks, compared to 4 for BRO and 5 for SimbaV2 (Figure 10).
 083 Across suites, Figure 1 shows one example task per benchmark, each showing more than 50%
 084 sample-efficiency improvement for WIMLE over the strongest competing method.

085 2 PRELIMINARIES

086 2.1 RL

087 We consider an infinite-horizon discounted Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ (Bell-
 088 man, 1957) with initial state distribution ρ_0 . At time t , the agent observes $s_t \in \mathcal{S}$, selects $a_t \sim$
 089 $\pi_\phi(a | s_t)$, receives reward $r_t = r(s_t, a_t)$, and the environment transitions as $s_{t+1} \sim P(\cdot | s_t, a_t)$.
 090 The objective is to learn a policy that maximizes the expected discounted return

$$093 \quad J(\pi_\phi) = \mathbb{E}_{\tau \sim (\rho_0, P, \pi_\phi)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]. \quad (1)$$

094 In the continuous-control settings considered here, states and actions are real-valued. The discount
 095 factor satisfies $\gamma \in (0, 1)$. We denote the action-value function under policy π by $Q^\pi(s, a)$:

$$096 \quad Q^\pi(s, a) = \mathbb{E}_{\tau \sim (\pi, P)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \right]. \quad (2)$$

103 2.2 MODEL-BASED RL

104 In model-based RL, we learn a parametric world model with parameters θ that approximates the
 105 unknown environment transition dynamics $P(s_{t+1}, r_t | s_t, a_t)$ through a learned conditional distri-
 106 bution

$$107 \quad \hat{p}_\theta(s_{t+1}, r_t | s_t, a_t) \quad (3)$$

108 trained from limited environment interactions. A model rollout (prediction) of horizon H under a
 109 policy π_ϕ from s_0 is the sequence $\hat{\tau} = (s_0, a_0, r_0, s_1, \dots, s_H)$ generated by
 110

$$111 \quad a_t \sim \pi_\phi(\cdot | s_t), \quad (s_{t+1}, r_t) \sim \hat{p}_\theta(\cdot | s_t, a_t). \quad (4)$$

112 Such rollouts are commonly used for planning or to provide synthetic transitions for RL training
 113 (Janner et al., 2019; Zhu et al., 2023).

114 2.3 IMPLICIT MAXIMUM LIKELIHOOD ESTIMATION

116 Implicit Maximum Likelihood Estimation (IMLE) learns a latent-variable generator $g_\theta(z)$ that maps
 117 noise $z \sim \mathcal{N}(0, I)$ to data space. Given data $\{x_i\}_{i=1}^N$, the IMLE objective is:
 118

$$119 \quad \theta^* = \arg \min_{\theta} \mathbb{E}_{\{z_j\}_{j=1}^m} \sum_{i=1}^N \min_{1 \leq j \leq m} \|g_\theta(z_j) - x_i\|^2. \quad (5)$$

122 In practice, given current parameters θ , we realize this objective by drawing a pool of candidate
 123 latents $\{z_j\}_{j=1}^m$ i.i.d. from $\mathcal{N}(0, I)$ per data point and selecting the nearest generated sample; this
 124 step is gradient-free and fully parallelizable,

$$125 \quad z_i^* = \arg \min_{1 \leq j \leq m} \|g_\theta(z_j) - x_i\|^2, \quad (6)$$

127 and minimizing the resulting empirical loss using stochastic gradient descent.

$$129 \quad \theta \leftarrow \theta - \eta \nabla_\theta \frac{1}{|B|} \sum_{i \in B} \|g_\theta(z_i^*) - x_i\|^2. \quad (7)$$

131 Optimizing Eq. equation 5 yields maximum likelihood estimation (MLE) of θ and ensures mode
 132 coverage (Aghabozorgi et al., 2023): each data point is represented by at least one generated sample.
 133 In practice, IMLE is sample efficient and effective for modeling multi-modal distributions. Condi-
 134 tional IMLE $g_\theta(c, z)$ models multi-modal conditional distributions; we adopt this form in WIMLE.
 135 For a more detailed discussion of IMLE, we refer readers to (Li & Malik, 2018; Aghabozorgi et al.,
 136 2023).

137 3 WIMLE

140 WIMLE addresses the key limitations of traditional MBRL through three main components: (1)
 141 IMLE-trained stochastic world models that capture complex multi-modal transition dynamics, (2)
 142 predictive uncertainty estimation that reflects the model’s confidence in its predictions, and (3)
 143 uncertainty-weighted learning that scales the influence of synthetic data based on model confidence.
 144 We detail each component below.

145 3.1 IMLE WORLD MODEL

147 Recent MBRL approaches span autoregressive sequence models, latent-variable generators, diffu-
 148 sion models, and planning-centric objectives (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020;
 149 2021; Robine et al., 2023; Micheli et al., 2023; Zhang et al., 2023; Hansen et al., 2024; Huang et al.,
 150 2024). Diffusion models are effective but rely on iterative sampling, which limits their usage in
 151 the online RL setting where rollout throughput is critical (Huang et al., 2024; Karras et al., 2022).
 152 Despite progress, these methods often require substantial data and still struggle to consistently sur-
 153 pass strong model-free baselines (Nauman et al., 2024; Lee et al., 2025b). Moreover, simple and
 154 sample-efficient unimodal Gaussian world models underfit inherently multi-modal, complex dynam-
 155 ics in partially observable or contact-rich settings, exacerbating model bias and compounding errors
 156 (Janner et al., 2019; Zhu et al., 2023).

157 On the other hand, we leverage IMLE to learn transitions, a one-step generative method that—unlike
 158 diffusion models—avoids iterative sampling and enables fast online rollouts. In practice, IMLE
 159 yields strong rollout throughput; Figure 2 reports wall-clock time among model-based methods. We
 160 represent the world model as a conditional stochastic generator g_θ that maps a state-action pair and
 161 latent noise to the next outcome:

$$161 \quad (\tilde{s}_{t+1}, \tilde{r}_t) = g_\theta(s_t, a_t, z), \quad z \sim \mathcal{N}(0, I). \quad (8)$$

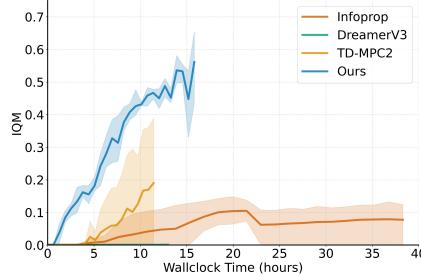


Figure 2: Wall-clock comparison among model-based methods (3 seeds) on a single NVIDIA L40S GPU for the humanoid-run task. Y-axis shows interquartile mean; shaded areas indicate 95% confidence intervals.

Here, the latent variable z induces a distribution over next outcomes for the same state–action pair, capturing inherent stochasticity and multi-modality in the dynamics.

IMLE Training Procedure. Given a dataset of transitions $\{(s_t, a_t, r_t, s_{t+1})\}_{t=1}^N$, we form targets $y_i = [r_t, s_{t+1}]$ and train g_θ using the IMLE objective. The training proceeds in two alternating steps:

Assignment Step: For each data point y_i , we sample m candidate latents $\{z_j\}_{j=1}^m$ and assign the nearest candidate that minimizes the prediction error:

$$z_i^* = \arg \min_{1 \leq j \leq m} \|g_\theta(s_i, a_i, z_j) - y_i\|^2. \quad (9)$$

This assignment step is computationally efficient as it requires no gradient computation, is fully parallelizable across data points, and typically uses small values of m (e.g., 5-10) in conditional IMLE settings.

Update Step: We then perform gradient descent on the empirical loss using the assigned latents:

$$\theta \leftarrow \theta - \eta \nabla_\theta \frac{1}{|B|} \sum_{i \in B} \|g_\theta(s_i, a_i, z_i^*) - y_i\|^2, \quad (10)$$

where B is a minibatch of indices and $\eta > 0$ is the learning rate.

This procedure ensures *mode coverage* by matching each data point to at least one generated sample, avoiding collapse to a single mean prediction. In contrast, a standard Gaussian regression model trained with least squares (Janner et al., 2019) predicts the conditional mean; in multi-modal settings this falls between modes and produces averaged, often implausible next states—known as regression to the mean (Galton, 1886; Barnett et al., 2005)—that compound over rollouts. IMLE’s per-sample latent assignment avoids this averaging and yields sharper, mode-consistent predictions (Aghabozorgi et al., 2023; Vashist et al., 2024).

Inference and Rollouts. After training, we generate rollouts following the procedure described in Section 2.2. Multi-step rollouts of horizon H are generated by initializing from a real state s_0 and iteratively applying: $a_t \sim \pi_\phi(\cdot | s_t)$, $z_t \sim \mathcal{N}(0, I)$, $(s_{t+1}, r_t) = g_\theta(s_t, a_t, z_t)$ for $t = 0, \dots, H-1$.

3.2 UNCERTAINTY ESTIMATION

Reliable uncertainty estimation is crucial for deciding when to trust model predictions. We therefore compute a predictive uncertainty for each synthetic transition and use it to reweight the RL objective. Each transition’s contribution is scaled by its estimated confidence. This preserves useful rollouts and reduces bias from uncertain predictions without changing the underlying algorithm. Alternative integrations exist. For example, Infoprop (Frauenknecht et al., 2025) computes an information-theoretic corruption measure and uses it to truncate rollouts during generation. In contrast, we integrate uncertainty directly into the learning objective via confidence weights.

We use a single predictive uncertainty measure $\sigma(s, a)$ that reflects the model’s confidence in its next-step prediction. Concretely, we maintain an ensemble of K IMLE world models (see Section 3.6.1) and, for each model, draw m latent samples, yielding predictions:

$$\{g_{\theta_k}(s, a, z_j)\}_{k=1..K, j=1..m} \quad (11)$$

We define $\sigma(s, a)$ as the standard deviation across these predictions—a direct measure of model agreement (see Algorithm 1, lines 13–14):

$$\sigma(s, a) = \text{std}_{k,j}[g_{\theta_k}(s, a, z_j)] \quad (12)$$

In practice, we compute per-dimension standard deviations for the predicted reward and next state and average them to obtain a single scalar for the transition. $\sigma(s, a)$ decreases when models agree and increases when predictive uncertainty is high (e.g., limited data or complex dynamics). By the law of total variance we can decompose $\sigma^2(s, a)$ into epistemic and aleatoric components,

$$\sigma^2(s, a) = \underbrace{\text{Var}_k [\mathbb{E}_z [g_{\theta_k}(s, a, z)]]}_{\text{ensemble / epistemic}} + \underbrace{\mathbb{E}_k [\text{Var}_z [g_{\theta_k}(s, a, z)]]}_{\text{latent / aleatoric}}, \quad (13)$$

Following recent insights on uncertainty from Smith et al. (2024), this total predictive variance is the Bayes risk of acting under a squared-error loss.

3.3 UNCERTAINTY-WEIGHTED LEARNING

Having defined a single predictive uncertainty $\sigma(s, a)$, we now describe how it enters learning. The key idea is simple: weight each synthetic transition by the model’s confidence so reliable predictions contribute more and uncertain ones less. Because uncertainty typically grows with rollout horizon due to error accumulation, later steps in the rollout receive smaller weights, naturally down-weighting distant, noisier predictions.

To choose these weights, we invoke a standard heteroscedastic regression result (formalized in Section 3.4): when estimating a scalar quantity from independent noisy observations with different noise variances, the unique minimum-variance linear unbiased estimator weights each observation inversely to its variance. We use this principle as guidance and map each predictive variance $\sigma(s, a)$ to a bounded, inverse-variance weight

$$w(s, a) = \frac{1}{\sigma(s, a) + 1}, \quad (14)$$

which preserves the inverse-variance ordering and keeps $w \in (0, 1]$, avoiding exploding gradients.

During rollout generation, we compute per-transition weights $w_i = w(s_i, a_i)$ for each synthetic transition (s_i, a_i, r_i, s'_i) using the predictive uncertainty defined above. We incorporate these weights into the RL objective by modifying the temporal difference (TD) loss:

$$\mathcal{L}_{\text{critic}} = \mathbb{E}_{(s_i, a_i, r_i, s'_i) \sim \mathcal{D}} [w_i \cdot \delta_i^2], \quad (15)$$

where $\delta_i = r_i + \gamma Q_{\phi}(s_{i+1}, a_{i+1}) - Q_{\phi}(s_i, a_i)$ is the TD error for transition i , with $a_{i+1} \sim \pi_{\phi}(\cdot | s_{i+1})$, Q_{ϕ} is a parameterized Q-function, and w_i is the corresponding uncertainty weight. For real environment data, we simply use $w_i = 1$.

This approach lets us leverage synthetic rollouts while keeping TD updates well-conditioned: high-variance, uncertain transitions have a smaller impact on the stochastic gradients, while the Bellman fixed point remains unchanged, so we tame noisy (including purely stochastic) updates without discarding or biasing them.

3.4 THEORETICAL ANALYSIS

To make the effect of our weighting more concrete, we provide a theoretical analysis of its properties. First, we establish that multiplying each squared Bellman error $(y - Q(s, a))^2$ by a positive weight $w(s, a)$ preserves the Bellman fixed point. Second, in a tractable setting where the critic is linear in a feature representation, we demonstrate that choosing weights inversely proportional to the (total) target noise variance minimizes the variance of the learned parameters, thereby improving convergence rate and sample efficiency. The lemmas below formalize these results: the first shows that any strictly positive reweighting leaves the Bellman target Q^* unchanged, and the second shows that, in the linear-critic regime, inverse-variance weighting $w_i \propto 1/\sigma_i^2$ yields the minimum-covariance linear unbiased estimator. Full proofs are provided in Appendix B.

Lemma (Positive weights preserve the Bellman target). Let $y = r + \gamma V(s')$ denote the one-step Bellman target for a value function V with conditional mean $\mu(s, a) = \mathbb{E}[y | s, a]$. Given a strictly positive weight function $w(s, a) > 0$, the corresponding population weighted Bellman squared loss is

$$\mathcal{L}_w(Q) = \mathbb{E}_{(s, a) \sim d^{\pi}} \mathbb{E}_{(r, s') \sim P(\cdot | s, a)} [w(s, a) (y - Q(s, a))^2]. \quad (16)$$

Then the unique minimizer of $\mathcal{L}_w(Q)$ over all action-value functions Q is $Q^*(s, a) = \mu(s, a)$ for all (s, a) , i.e., any strictly positive reweighting leaves the Bellman fixed point unchanged and only re-emphasizes different regions of (s, a) . A full proof is given in Appendix B.

270 **Lemma (Linear critics, inverse-variance weighting).** Consider the setting where (1) the action-
 271 value function is linear in some feature representation $\phi(s, a) \in \mathbb{R}^d$, i.e., $Q_\theta(s, a) = \phi(s, a)^\top \theta$
 272 for parameters $\theta \in \mathbb{R}^d$, and (2) for a batch of m transitions $x_i = (s_i, a_i)$ the TD targets satisfy
 273 $y_i = \mu(x_i) + \varepsilon_i$ with $\mathbb{E}[\varepsilon_i | x_i] = 0$, $\text{Var}(\varepsilon_i | x_i) = \sigma_i^2$, and independent noise terms ε_i . Among
 274 all linear unbiased estimators of θ , the inverse-variance choice $w_i \propto 1/\sigma_i^2$ yields the minimum
 275 covariance matrix for $\hat{\theta}$. A proof follows the classical Gauss–Markov theorem and is provided in
 276 Appendix B.

277 *Proof sketch.* Stacking targets into a vector y and features into a design matrix Φ , minimizing
 278 $\sum_i w_i (y_i - Q_\theta(x_i))^2$ yields the weighted least-squares estimator $\hat{\theta}_w = (\Phi^\top W \Phi)^{-1} \Phi^\top W y$ with
 279 $W = \text{diag}(w_i)$. Choosing W proportional to the inverse noise covariance $\Sigma^{-1} = \text{diag}(1/\sigma_i^2)$
 280 yields an estimator whose covariance $\text{Cov}(\hat{\theta}_w)$ is minimal among all linear unbiased estimators by
 281 the Gauss–Markov theorem. We refer to Appendix B for a full derivation.

283 **Implications for WIMLE.** Here σ_i^2 is the total predictive variance at (s_i, a_i) , which can include
 284 both epistemic and aleatoric components (Eq. 13). The lemma above shows that, in the linear-critic
 285 regime, weighting each transition inversely to this total variance minimizes the covariance of the
 286 learned parameters, independently of the source of the noise. Lower parameter variance means
 287 fewer samples are needed to reach a given accuracy, i.e., inverse-variance weighting is provably
 288 more sample-efficient in this regime. Combined with the Bellman fixed-point lemma above (proved
 289 in Appendix B), this means our choice of $w(s, a)$ shrinks update variance without ever changing
 290 the Bellman solution—even in the limit of a perfect world model where all uncertainty (and thus
 291 down-weighted noise) is purely aleatoric.

293 3.5 ALGORITHM

295 Algorithm 1 presents the overall WIMLE procedure. For a more complete implementation of the
 296 algorithm, including training frequencies and hyperparameters, see Algorithm 3 in Appendix A.

298 **Algorithm 1** WIMLE: World Models with Implicit Maximum Likelihood Estimation

299 1: **Input:** Rollout horizon H , ensemble size K , number of rollouts M , number of latent codes m
 300 2: {Red text indicates steps that fundamentally differ from MBPO.}
 301 3: initialize ensemble world models $\{g_{\theta_k}\}_{k=1}^K$, environment and model datasets \mathcal{D}_{env} , $\mathcal{D}_{\text{model}}$
 302 4: **for** environment steps **do**
 303 5: Collect environment transitions using π_ϕ ; add to \mathcal{D}_{env}
 304 6: // **IMLE World Model Training**
 305 7: Train ensemble $\{g_{\theta_k}\}_{k=1}^K$ in parallel on bootstrap samples of \mathcal{D}_{env} using **IMLE** (Eqs. 9, 10)
 306 8: **for** M model rollouts **do**
 307 9: Sample starting state s_0 from \mathcal{D}_{env}
 308 10: **for** $t = 0$ to $H - 1$ **do**
 309 11: $a_t \sim \pi_\phi(\cdot | s_t)$
 310 12: Sample m latents $\{z_j\}_{j=1}^m \sim \mathcal{N}(0, I)$
 311 13: Generate predictions $\{(\tilde{s}_{t+1}, \tilde{r}_t)_{k,j}\}_{k=1,j=1}^{K,m} = \{g_{\theta_k}(s_t, a_t, z_j)\}_{k=1,j=1}^{K,m}$ from all ensemble
 312 members
 313 14: Compute predictive uncertainty: $\sigma_t = \text{std}_{k,j}[g_{\theta_k}(s_t, a_t, z_j)]$ {aggregated over ensem-
 314 bles and latents}
 315 15: Set weight $w_t = 1/(\sigma_t + 1)$
 316 16: Add weighted transition $(s_t, a_t, r_t, s_{t+1}, w_t)$ to $\mathcal{D}_{\text{model}}$
 317 17: // **Uncertainty-Weighted Policy Learning**
 318 18: Sample batch from $\mathcal{D}_{\text{env}} \cup \mathcal{D}_{\text{model}}$ (real data has $w = 1$)
 319 19: Update policy using **weighted** RL objective:
 320 20: $\mathcal{L} = \mathbb{E}_{(s, a, r, s', w) \sim \text{batch}}[w \cdot \ell_{\text{RL}}(s, a, r, s')]$

321
 322 The algorithm maintains the underlying RL method as a black box through the weighted loss func-
 323 tion \mathcal{L} , where ℓ_{RL} represents any standard RL objective (e.g., TD error for critics, policy gradient for
 actors). The key insight is that uncertainty weights w_t automatically scale the contribution of each

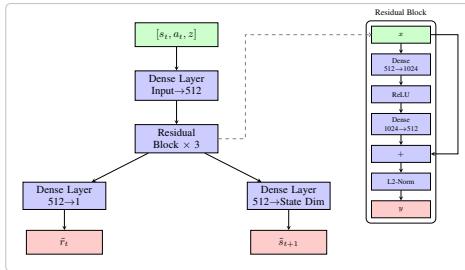


Figure 3: WIMLE world model architecture.

synthetic transition—high-confidence predictions receive higher weights while uncertain predictions contribute proportionally less.

3.6 DESIGN CHOICES

3.6.1 TRAINING

Model Rollouts. We experiment with synthetic rollouts using horizons up to $H = 8$. All rollouts are initialized from real environment states sampled uniformly from the environment dataset \mathcal{D}_{env} , following standard practice in model-based RL to ensure rollouts start from the data distribution. We select task-specific rollout horizons through empirical experimentation as described in Appendix D.1.

RL Training. We use Soft Actor-Critic (SAC) (Haarnoja et al., 2018) with distributional Q-learning (Bellemare et al., 2017) as our underlying RL algorithm. Following recent work that has demonstrated the effectiveness of distributional RL for continuous control (Nauman et al., 2024; Lee et al., 2025b; Dabney et al., 2018a), we specifically adapt quantile Q-learning (Dabney et al., 2018b; Nauman et al., 2024).

Ensemble Training. We train an ensemble of $K = 7$ IMLE world models in parallel to improve predictive uncertainty estimation and calibration (Section 3.3). Each ensemble member is initialized with different random parameters and trained on bootstrap samples of the environment data. The parallel training of ensemble members is computationally efficient and scales well with available compute resources, enabling reliable predictive uncertainty without significant computational overhead.

3.6.2 ARCHITECTURE

Figure 3 illustrates the WIMLE world model architecture. The network takes as input state s_t , action a_t , and latent variable z , followed by a dense layer that maps to a 512-dimensional hidden representation. The core of the architecture consists of three residual blocks, each containing dense layers with ReLU activations and L2 normalization. Following recent findings by Lee et al. (2025b), we employ L2 normalization within the residual blocks, which has been shown to improve stability and performance in RL settings. The network outputs separate predictions for rewards and next states through dedicated dense heads.

4 EXPERIMENTS

We evaluate WIMLE across diverse continuous-control benchmarks—DeepMind Control Suite (including Dog and Humanoid), MyoSuite, and HumanoidBench (Tassa et al., 2018; Caggiano et al., 2022; Sferrazza et al., 2024). Across 40 tasks spanning locomotion and dexterous manipulation with high-dimensional state/action spaces and sparse rewards, we compare against strong model-free and model-based methods, including MR.Q, PPO, SAC, Simba, SimbaV2, BRO, TD-MPC2, and DreamerV3 (Fujimoto et al., 2025; Schulman et al., 2017a; Haarnoja et al., 2018; Lee et al., 2025a;b; Nauman et al., 2024; Hansen et al., 2024; Hafner et al., 2023), and present per-benchmark results. Through our experiments, we aim to answer: (i) How does WIMLE compare to strong model-free and model-based methods? (ii) How does IMLE-based multi-modality in the world model affect results compared to standard unimodal Gaussian models? (iii) How do uncertainty estimates evolve during training, and how do they affect performance?

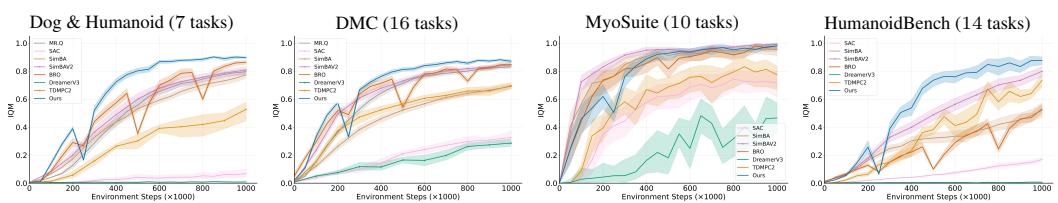
378
379

4.1 EXPERIMENTAL SETUP

380 All experiments are run for 1M environment steps with 10 random seeds unless otherwise specified.
 381 We report the interquartile mean (IQM) and 95% confidence intervals computed with RLiable (Agar-
 382 wal et al., 2021), using stratified bootstrap across tasks and seeds. Following BRO and SimbaV2
 383 (Nauman et al., 2024; Lee et al., 2025b), we aggregate normalized scores per BRO/SimbaV2 proto-
 384 col (DMC [0,1], MyoSuite success, HumanoidBench success-normalized). Where official baseline
 385 results are available, we report the authors’ numbers; otherwise, we run public implementations with
 386 their recommended settings. We provide full details about the experimental setup, hyperparameters,
 387 and baselines in Section D and E of the appendix.

388

4.2 COMPARISON TO BASELINES

389
390
391
392
393
394
395

396 Figure 4: Aggregate results across benchmarks. WIMLE outperforms strong model-free and model-
 397 based baselines overall. Gains are most pronounced on the challenging Dog & Humanoid subset,
 398 where it achieves superior sample efficiency and asymptotic performance. On MyoSuite, it performs
 399 asymptotically on par with strong baselines that are already near the maximum score (1.0), and on
 400 HumanoidBench it significantly outperforms the baselines, solving 8/14 tasks versus BRO 4 and
 401 SimbaV2 5. Y-axes show interquartile mean; shaded areas denote 95% confidence intervals.

402
403
404
405
406
407
408
409
410
411
412
413
414

We summarize aggregate performance across benchmarks in Figure 4 and provide detailed per-task results in Section C of the appendix. WIMLE consistently leads among strong model-free and model-based methods on Dog & Humanoid, the full DMC suite, and HumanoidBench, while performing asymptotically on par with strong MyoSuite baselines that are already close to the maximum score (1.0). Notably, gains are largest on the high-dimensional and challenging Dog & Humanoid tasks (Dog: $|\mathcal{S}|=223$, $|\mathcal{A}|=38$; Humanoid: $|\mathcal{S}|=67$, $|\mathcal{A}|=24$). On HumanoidBench, WIMLE significantly outperforms baselines, solving 8 of 14 tasks versus BRO 4 and SimbaV2 5 (Figure 10). We summarize performance across timesteps in Section E.1, where WIMLE performs best or competitively across most evaluations. We attribute these improvements to IMLE-driven multi-modality in the world model and uncertainty-weighted learning that scales the influence of synthetic rollouts by model confidence, mitigating bias from overconfident predictions while preserving useful signal, which we discuss in more detail in the next section. Per-task performance is reported in Figures 7, 8, 9 and 10.

415
416

4.3 METHOD ANALYSIS

417
418

We analyze how uncertainty-aware weighting and multi-modal dynamics modeling impact performance and how predictive uncertainty evolves during training.

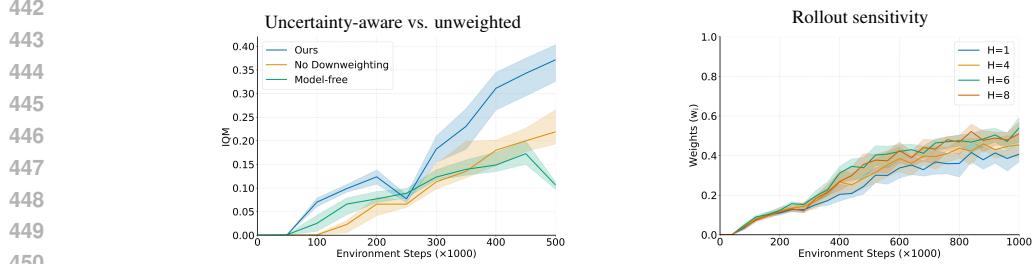
419
420
421
422
423
424
425
426
427
428
429

Effect of uncertainty-aware weighting Figure 5 (Left) compares WIMLE with uncertainty-aware weighting to an unweighted variant that is *identical in every respect except that all per-transition weights are fixed to $w_i=1.0$* . The unweighted curve lags and can even underperform a strong model-free baseline early on, indicating that ignoring predictive uncertainty significantly biases learning and hinders performance. Figure 5 (Right) studies rollout sensitivity. Increasing the model rollout horizon from $H=1$ to $H=4$ to $H=6$ improves performance, and extending to $H=8$ maintains performance rather than showing the severe degradation typically observed in model-based methods when increasing rollout length (Janner et al., 2019). This improved stability at longer horizons demonstrates that uncertainty-aware weighting reduces the model bias typically introduced by longer horizon errors, enabling us to leverage longer synthetic rollouts without considerable performance degradation.

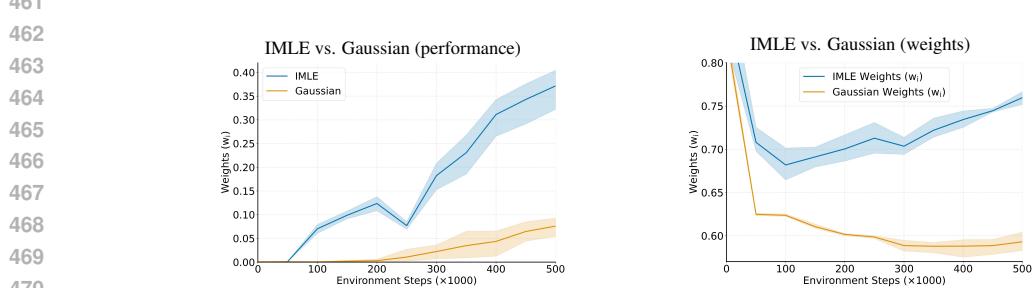
430
431

Impact of IMLE-based multi-modality Figure 6 (Left) contrasts WIMLE (IMLE world model) with an otherwise identical unimodal Gaussian world model (MBPO-style; (Janner et al., 2019)),

432 with both variants using uncertainty-aware weighting. The IMLE variant significantly outperforms
 433 the Gaussian, underscoring the value of modeling multi-modal transition dynamics for uncertainty
 434 estimation in complex, contact-rich control. Figure 6 (Right) shows how weights evolve. During
 435 a brief warm-up with limited environment samples and training, both models are uncalibrated and
 436 weights can appear transiently high; as data accumulates and the estimators calibrate, weights drop
 437 to reflect high uncertainty and low confidence. As training progresses and more data are collected,
 438 IMLE’s weights increase to reflect higher confidence in the predictions, whereas the Gaussian’s
 439 remain relatively flat, indicating limited calibration. Together, these results show that multi-modal
 440 modeling improves both performance and the quality of uncertainty estimates, reducing the risk of
 441 overconfident, biased predictions misleading the policy.



451 Figure 5: Uncertainty-aware weighting reduces model bias and enables stable training at longer
 452 horizons on Humanoid-run. **Left:** Uncertainty-aware WIMLE compared to an unweighted variant
 453 that is identical except all per-transition weights are fixed to $w_i = 1.0$ and a model-free variant
 454 that is identical except that it does not use the model; the unweighted curve lags and can even
 455 underperform the model-free variant early on, indicating that ignoring uncertainty will bias learning
 456 and hinder performance. **Right:** Rollout ablation ($H = 1, 4, 6, 8$) for WIMLE: increasing the
 457 model rollout horizon from $H=1$ to $H=4$ to $H=6$ improves performance, and extending to $H=8$
 458 does not substantially degrade performance, suggesting that uncertainty-aware weighting mitigates
 459 harm from error accumulation at longer horizons. All variants use the same SAC backbone and
 460 distributional critics; only the ablated components differ. All plots are on DMC’s Humanoid-run
 461 task with 5 seeds.



471 Figure 6: Multi-modality strengthens model-based learning. **Left:** WIMLE (IMLE world model)
 472 versus an otherwise identical unimodal Gaussian world model (MBPO-style; (Janner et al., 2019);
 473 both use uncertainty-aware weighting): the IMLE variant significantly outperforms the Gaussian,
 474 highlighting the value of multi-modal modeling and IMLE’s efficacy. **Right:** Weight dynamics:
 475 After a brief warm-up phase, IMLE’s weights are lower when uncertainty is high and increase as
 476 training progresses and more data are collected, reflecting growing model confidence; the unimodal
 477 Gaussian fails to capture this evolution, yielding relatively flat weights over time. All plots are on
 478 Humanoid-run.

479 5 RELATED WORK

481 **Model-free RL.** Foundational model-free methods such as PPO (Schulman et al., 2017a) and
 482 SAC (Haarnoja et al., 2018) remain strong references for continuous control. Recent advances focus
 483 on scaling and regularization: BRO (Nauman et al., 2024) scales critic networks to 5M parameters
 484 with strong regularization and optimistic exploration, achieving state-of-the-art performance. Simba
 485 (Lee et al., 2025a) introduces an architecture that embeds simplicity bias through running statistics
 486 normalization, residual feedforward blocks, and layer normalization, enabling effective parameter

486 scaling; SimbaV2 (Lee et al., 2025b) further constrains feature and weight norms via hyperspherical
 487 normalization. Contemporary work like MR.Q (Fujimoto et al., 2025) explores improved value
 488 estimation for better sample efficiency. Collectively, these methods provide strong model-free base-
 489 lines.
 490

491 **Model-based RL.** Model-based RL methods learn world models to improve sample efficiency
 492 via synthetic rollouts and planning. DreamerV3 (Hafner et al., 2023) learns a latent world model
 493 and achieves strong performance in continuous control with large-scale training. MBPO (Janner
 494 et al., 2019) uses short model-generated rollouts branched from real data to avoid model exploitation
 495 while maintaining sample efficiency. TD-MPC2 (Hansen et al., 2024) learns implicit world models
 496 through joint-embedding prediction and performs local trajectory optimization in latent space for
 497 scalable multi-task learning. STORM (Zhang et al., 2023) combines Transformer-based sequence
 498 modeling with categorical VAEs for efficient world model learning in visual domains. Diffusion-
 499 based world models generate trajectories via iterative denoising and incur high inference cost, which
 500 hinders online RL (Janner et al., 2022; Ajay et al., 2023; He et al., 2023). Despite these algorithmic
 501 advances, model-based methods have struggled to consistently surpass recent model-free approaches
 502 like BRO (Nauman et al., 2024) and SimbaV2 (Lee et al., 2025b).
 503

504 **Model Bias.** Model bias and error accumulation remain fundamental challenges in MBRL. Tra-
 505 jectory models (Asadi et al., 2019; Lambert et al., 2021) address the compounding-error problem by
 506 learning multi-step models that directly predict outcomes of action sequences, avoiding the accumu-
 507 lation of one-step prediction errors. Self-correcting models (Talvitie, 2017) train models to correct
 508 themselves when producing errors. Infoprop (Frauenknecht et al., 2025) integrates uncertainty by
 509 truncating rollouts using information-theoretic corruption measures, but is not competitive on com-
 510 plex, high-dimensional tasks such as Humanoid-run (see Figure 1). In contrast, we estimate a single
 511 predictive uncertainty and weight each synthetic transition accordingly, integrating this directly into
 512 the learning objective to preserve useful synthetic data while reducing the influence of uncertain
 513 predictions. This yields state-of-the-art results on challenging tasks (Figures 1 and 4).
 514

515 **Implicit Maximum Likelihood Estimation.** IMLE (Li & Malik, 2018) trains implicit generative
 516 models by minimizing the expected distance from each data point to its nearest generated sam-
 517 ple, avoiding mode-collapse and GAN (Goodfellow et al., 2014) training issues. Adaptive IMLE
 518 (Aghabozorgi et al., 2023) extends this with adaptive thresholding and curriculum learning for bet-
 519 ter few-shot performance. These methods demonstrate that likelihood-based objectives can achieve
 520 good sample quality without adversarial training on low-data settings.
 521

522 6 LIMITATIONS AND FUTURE WORK

523 WIMLE uses world models solely to generate synthetic rollouts. Other uses, such as planning
 524 with the model or integrating the model into policy-gradient formulations, remain unexplored here.
 525 Future work should evaluate WIMLE in these settings. Our experiments use proprioceptive state
 526 observations only. Extending WIMLE to image-based control is an important direction, especially
 527 since IMLE has been shown to be effective in few-shot image synthesis (Aghabozorgi et al., 2023;
 528 Vashist et al., 2024). Finally, similar to MBPO (Janner et al., 2019) and POMP (Zhu et al., 2023),
 529 the rollout horizon is a task-dependent hyperparameter. Learning to adapt the horizon online based
 530 on model confidence is a promising avenue for future research.
 531

532 7 CONCLUSION

533 WIMLE advances model-based reinforcement learning by extending IMLE to learn stochastic,
 534 multi-modal world models and by weighting synthetic data with predictive confidence. This re-
 535 duces model bias and stabilizes learning while retaining the benefits of synthetic rollouts. Across
 536 40 continuous-control tasks in DMC, MyoSuite, and HumanoidBench, WIMLE achieves superior
 537 sample efficiency and competitive or higher asymptotic performance than strong model-free and
 538 model-based baselines. Gains are largest on challenging Dog and Humanoid locomotion tasks. On
 539 HumanoidBench, WIMLE significantly outperforms baselines and solves 8 of 14 tasks. The ap-
 540 proach integrates cleanly with standard RL objectives and scales with compute through ensembles
 541 and parallel latent sampling. We hope these results renew interest in practical world models for
 542 challenging continuous control.
 543

540 REFERENCES
541

542 Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare.
543 Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information*
544 *processing systems*, 34:29304–29320, 2021.

545 Mehran Aghabozorgi, Shichong Peng, and Ke Li. Adaptive IMLE for few-shot pretraining-free
546 generative modelling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt,
547 Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML*
548 *2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning*
549 *Research*, pp. 248–264. PMLR, 2023. URL <https://proceedings.mlr.press/v202/aghabozorgi23a.html>.

550 Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is
551 conditional generative modeling all you need for decision-making? *International Conference on*
552 *Learning Representations*, 2023.

553 Arthur Argenson and Gabriel Dulac-Arnold. Model-based offline planning. *ArXiv*, abs/2008.05556,
554 2021.

555 Kavosh Asadi, Evan Cater, Dipendra Misra, and Michael L. Littman. Towards a Simple Approach
556 to Multi-step Model-based Reinforcement Learning. *arXiv*, 2018a.

557 Kavosh Asadi, Dipendra Misra, and Michael L. Littman. Lipschitz Continuity in Model-based
558 Reinforcement Learning. *arXiv*, 2018b.

559 Kavosh Asadi, Dipendra Misra, Seungchan Kim, and Michel L. Littman. Combating the
560 Compounding-Error Problem with a Multi-step Model. *arXiv*, 2019.

561 Adrian G Barnett, Jolieke C Van Der Pols, and Annette J Dobson. Regression to the mean: what it
562 is and how to deal with it. *International journal of epidemiology*, 34(1):215–220, 2005.

563 Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforce-
564 ment learning. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th Interna-
565 tional Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*,
566 volume 70 of *Proceedings of Machine Learning Research*, pp. 449–458. PMLR, 2017. URL
567 <http://proceedings.mlr.press/v70/bellemare17a.html>.

568 Richard Bellman. A markovian decision process. *Indiana Univ. Math. J.*, 6:679–684, 1957. ISSN
569 0022-2518.

570 Vittorio Caggiano, Huawei Wang, Guillaume Durandau, Massimo Sartori, and Vikash Kumar.
571 Myosuite: A contact-rich simulation suite for musculoskeletal motor control. In Roya Firoozi,
572 Negar Mehr, Esen Yel, Rika Antonova, Jeannette Bohg, Mac Schwager, and Mykel J. Kochen-
573 derfer (eds.), *Learning for Dynamics and Control Conference, L4DC 2022, 23-24 June 2022,
574 Stanford University, Stanford, CA, USA*, volume 168 of *Proceedings of Machine Learning*
575 *Research*, pp. 492–507. PMLR, 2022. URL <https://proceedings.mlr.press/v168/caggiano22a.html>.

576 Víctor Uc Cetina, Nicolás Navarro-Guerrero, Ana Martín-González, Cornelius Weber, and Stefan
577 Wermter. Survey on reinforcement learning for language processing. *Artificial Intelligence Re-
578 view*, 56:1543–1575, 2021. URL [https://api.semanticscholar.org/CorpusID:
579 233210638](https://api.semanticscholar.org/CorpusID:233210638).

580 Jie Chen, Jian Sun, and Gang Wang. From unmanned systems to autonomous intelligent systems.
581 *Engineering*, 12:16–19, 2022.

582 Ignasi Clavera, Yao Fu, and P. Abbeel. Model-augmented actor-critic: Backpropagating through
583 paths. *ArXiv*, abs/2005.08068, 2020.

584 Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for dis-
585 tributional reinforcement learning, 2018a. URL <https://arxiv.org/abs/1806.06923>.

594 Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement
 595 learning with quantile regression. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Pro-
 596 ceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th
 597 innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Ed-
 598 ucational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, Febru-
 599 ary 2-7, 2018*, pp. 2892–2901. AAAI Press, 2018b. doi: 10.1609/AAAI.V32I1.11791. URL
 600 <https://doi.org/10.1609/aaai.v32i1.11791>.

601 Bernd Frauenknecht, Devdutt Subhasish, Friedrich Solowjow, and Sebastian Trimpe. On rollouts
 602 in model-based reinforcement learning. In *The Thirteenth International Conference on Learn-
 603 ing Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL
 604 <https://openreview.net/forum?id=Uh5GRmLvt>.

605 Scott Fujimoto, Pierluca D’Oro, Amy Zhang, Yuandong Tian, and Michael Rabbat. Towards
 606 general-purpose model-free reinforcement learning. *arXiv preprint arXiv:2501.16142*, 2025.

607 Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropo-
 608 logical Institute of Great Britain and Ireland*, 15:246–263, 1886.

609 Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,
 610 Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL <https://arxiv.org/abs/1406.2661>.

611 David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances
 612 in Neural Information Processing Systems 31*, pp. 2451–2463. Curran Associates, Inc., 2018.

613 Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy
 614 maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer G. Dy and
 615 Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning,
 616 ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings
 617 of Machine Learning Research*, pp. 1856–1865. PMLR, 2018. URL <http://proceedings.mlr.press/v80/haarnoja18b.html>.

618 Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James
 619 Davidson. Learning latent dynamics for planning from pixels. In *International Conference on
 620 Machine Learning*, pp. 2555–2565, 2019.

621 Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning
 622 behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.
 623 URL <https://openreview.net/forum?id=S1lOTC4tDS>.

624 Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with
 625 discrete world models. In *International Conference on Learning Representations*, 2021. URL
 626 <https://openreview.net/forum?id=0aabwyZbOu>.

627 Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains
 628 through world models. *arXiv preprint arXiv:2301.04104*, 2023.

629 Nicklas Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predic-
 630 tive control. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu,
 631 and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23
 632 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Re-
 633 search*, pp. 8387–8406. PMLR, 2022. URL <https://proceedings.mlr.press/v162/hansen22a.html>.

634 Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: scalable, robust world models for contin-
 635 uous control. In *The Twelfth International Conference on Learning Representations, ICLR 2024,
 636 Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=Oxh5CstDJU>.

637 Haoran He, Chenjia Bai, Kang Xu, Zhuoran Yang, Weinan Zhang, Dong Wang, Bin Zhao, and Xue-
 638 long Li. Diffusion model is an effective planner and data synthesizer for multi-task reinforcement
 639 learning. *Advances in Neural Information Processing Systems*, 2023.

648 Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan
 649 Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in
 650 deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*,
 651 volume 32, 2018.

652 Ionel-Alexandru Hosu and Traian Rebedea. Playing atari games with deep reinforcement learning
 653 and human checkpoint replay. *CoRR*, abs/1607.05077, 2016. URL <http://arxiv.org/abs/1607.05077>.

654 Renming Huang, Yunqiang Pei, Guoqing Wang, Yangming Zhang, Yang Yang, Peng Wang, and
 655 Hengtao Shen. Diffusion models as optimizers for efficient planning in offline rl, 2024. URL
 656 <https://arxiv.org/abs/2407.16142>.

657 Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model:
 658 Model-based policy optimization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer,
 659 Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural
 660 Information Processing Systems 32: Annual Conference on Neural Information Processing
 661 Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 12498–12509,
 662 2019. URL [https://proceedings.neurips.cc/paper/2019/
 663 hash/5faf461eff3099671ad63c6f3f094f7f-Abstract.html](https://proceedings.neurips.cc/paper/2019/hash/5faf461eff3099671ad63c6f3f094f7f-Abstract.html).

664 Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for
 665 flexible behavior synthesis. *International Conference on Machine Learning*, 2022.

666 Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-
 667 based generative models, 2022. URL <https://arxiv.org/abs/2206.00364>.

668 Bahare Kiumarsi, Kyriakos G. Vamvoudakis, Hamidreza Modares, and Frank L. Lewis. Optimal
 669 and autonomous control using reinforcement learning: A survey. *IEEE Transactions
 670 on Neural Networks and Learning Systems*, 29:2042–2062, 2018. URL <https://api.semanticscholar.org/CorpusID:21709652>.

671 Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. Model-ensemble
 672 trust-region policy optimization. In *6th International Conference on Learning Representations,
 673 ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.
 674 OpenReview.net, 2018. URL <https://openreview.net/forum?id=SJJinbWRZ>.

675 Nathan Lambert. Reinforcement learning from human feedback, 2025. URL <https://arxiv.org/abs/2504.12501>.

676 Nathan O. Lambert, Albert Wilcox, Howard Zhang, Kristofer S. J. Pister, and Roberto Calandra.
 677 Learning Accurate Long-term Dynamics for Model-based Reinforcement Learning. *IEEE Conf
 678 on Decision and Control*, 2021.

679 Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations
 680 for reinforcement learning. In *International Conference on Machine Learning*, pp. 5639–
 681 5650. PMLR, 2020.

682 Hojoon Lee, Dongyoon Hwang, Donghu Kim, Hyunseung Kim, Jun Jet Tai, Kaushik Subramanian,
 683 Peter R. Wurman, Jaegul Choo, Peter Stone, and Takuma Seno. Simba: Simplicity bias for scaling
 684 up parameters in deep reinforcement learning. In *The Thirteenth International Conference on
 685 Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a.
 686 URL <https://openreview.net/forum?id=jXLiDKsuDo>.

687 Hojoon Lee, Youngdo Lee, Takuma Seno, Donghu Kim, Peter Stone, and Jaegul Choo. Hyper-
 688 spherical normalization for scalable deep reinforcement learning. *CoRR*, abs/2502.15280, 2025b.
 689 doi: 10.48550/ARXIV.2502.15280. URL <https://doi.org/10.48550/arXiv.2502.15280>.

690 Ke Li and Jitendra Malik. Implicit maximum likelihood estimation. *CoRR*, abs/1809.09087, 2018.
 691 URL <http://arxiv.org/abs/1809.09087>.

702 Kendall Lowrey, Aravind Rajeswaran, Sham M. Kakade, Emanuel Todorov, and Igor Mordatch.
 703 Plan online, learn offline: Efficient learning and exploration via model-based control. *ArXiv*,
 704 abs/1811.01848, 2019.

705

706 Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world mod-
 707 els. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Ki-
 708 gali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL [https://openreview.net/](https://openreview.net/forum?id=vhFu1Acb0xb)
 709 forum?id=vhFu1Acb0xb.

710 Michal Nauman, Mateusz Ostaszewski, Krzysztof Jankowski, Piotr Milos, and Marek Cy-
 711 gan. Bigger, regularized, optimistic: scaling for compute and sample efficient con-
 712 tinuous control. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela
 713 Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neu-
 714 ral Information Processing Systems 38: Annual Conference on Neural Information Pro-
 715 cessing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15,
 716 2024*. URL http://papers.nips.cc/paper_files/paper/2024/hash/cd3b5d2ed967e906af24b33d6a356cac-Abstract-Conference.html.

717

718 Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world mod-
 719 els are happy with 100k interactions. In *The Eleventh International Conference on Learning
 720 Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL
 721 <https://openreview.net/forum?id=TdBaDGCPjly>.

722

723 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
 724 optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017a.

725

726 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
 727 optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017b.

728

729 Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. Humanoid-
 730 bench: Simulated humanoid benchmark for whole-body locomotion and manipulation. In Dana
 731 Kulic, Gentiane Venture, Kostas E. Bekris, and Enrique Coronado (eds.), *Robotics: Science and
 732 Systems XX, Delft, The Netherlands, July 15-19, 2024*, 2024. doi: 10.15607/RSS.2024.XX.061.
 733 URL <https://doi.org/10.15607/RSS.2024.XX.061>.

734

735 Freddie Bickford Smith, Jannik Kossen, Eleanor Trollope, Mark van der Wilk, Adam Foster, and
 736 Tom Rainforth. Rethinking aleatoric and epistemic uncertainty. *CoRR*, abs/2412.20892, 2024.
 737 doi: 10.48550/ARXIV.2412.20892. URL <https://doi.org/10.48550/arXiv.2412.20892>.

738

739 Anne Somalwar, Bruce D. Lee, George J. Pappas, and Nikolai Matni. Learning with imperfect
 740 models: When multi-step prediction mitigates compounding error. *CoRR*, abs/2504.01766, 2025.
 741 doi: 10.48550/ARXIV.2504.01766. URL <https://doi.org/10.48550/arXiv.2504.01766>.

742

743 Erik Talvitie. Self-correcting models for model-based reinforcement learning. In Satinder Singh
 744 and Shaul Markovitch (eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intel-
 745 ligence, February 4-9, 2017, San Francisco, California, USA*, pp. 2597–2603. AAAI Press, 2017.
 746 doi: 10.1609/AAAI.V31I1.10850. URL <https://doi.org/10.1609/aaai.v31i1.10850>.

747

748 Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David
 749 Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy P. Lillicrap, and Martin
 750 A. Riedmiller. Deepmind control suite. *CoRR*, abs/1801.00690, 2018. URL <http://arxiv.org/abs/1801.00690>.

751

752 Chirag Vashist, Shichong Peng, and Ke Li. Rejection sampling IMLE: designing priors for better
 753 few-shot image synthesis. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten
 754 Sattler, and Gü̈l Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan,
 755 Italy, September 29-October 4, 2024, Proceedings, Part XXI*, volume 15079 of *Lecture Notes in
 Computer Science*, pp. 441–456. Springer, 2024. doi: 10.1007/978-3-031-72664-4_25. URL
https://doi.org/10.1007/978-3-031-72664-4_25.

756 Arun Venkatraman, Martial Hebert, and J.. Bagnell. Improving Multi-Step Prediction of Learned
757 Time Series Models. *AAAI*, 2015.

758

759 Boxi Weng, Jian Sun, Gao Huang, Fang Deng, Gang Wang, and Jie Chen. Competitive meta-
760 learning. *IEEE/CAA Journal of Automatica Sinica*, 10(9):1902–1904, 2023.

761 Chenjun Xiao, Yifan Wu, Chen Ma, Dale Schuurmans, and Martin Müller. Learning to combat
762 compounding-error in model-based reinforcement learning. *CoRR*, abs/1912.11206, 2019. URL
763 <http://arxiv.org/abs/1912.11206>.

764

765 Weirui Ye, Shaohuai Liu, Thanard Kurutach, Pieter Abbeel, and Yang Gao. Mastering atari games
766 with limited data. *Advances in Neural Information Processing Systems*, 34:25476–25488, 2021.

767

768 Weipu Zhang, Gang Wang, Jian Sun, Yetian Yuan, and Gao Huang. STORM: efficient
769 stochastic transformer based world models for reinforcement learning. In Alice Oh, Tris-
770 stan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Ad-
771 vances in Neural Information Processing Systems 36: Annual Conference on Neural Infor-
772 mation Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16,
773 2023*. URL http://papers.nips.cc/paper_files/paper/2023/hash/5647763d4245b23e6a1cb0a8947b38c9-Abstract-Conference.html.

774

775 Jinhua Zhu, Yue Wang, Lijun Wu, Tao Qin, Wengang Zhou, Tie-Yan Liu, and Houqiang Li. Making
776 better decision by directly planning in continuous control. In *The Eleventh International Confer-
777 ence on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net,
778 2023. URL <https://openreview.net/forum?id=r8Mu7idxyF>.

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810 AUTHOR STATEMENT: USE OF LANGUAGE MODELS
811812 We used large language models to help polish writing and improve clarity. All ideas, methods, ex-
813 periments, and analyses were created and verified by the authors. Any suggested text was reviewed
814 and edited by the authors for accuracy and originality.
815816 A ALGORITHM DETAILS
817818 Algorithm 3 provides the detailed implementation of WIMLE, including training frequencies, batch
819 sizes, and other practical considerations omitted from the main algorithm for clarity. The IMLE
820 training procedure is detailed in Algorithm 2. Hyperparameters are provided in Section D.
821822 **Algorithm 2** IMLE World Model Training
823

 824 1: **Input:** Environment dataset \mathcal{D}_{env} , ensemble $\{g_{\theta_k}\}_{k=1}^K$, number of latent codes m , number of
825 updates U , learning rate η
 826 2: **for** $u = 1$ to U **do**
 827 3: Sample minibatch $\{(s_i, a_i, r_i, s_{i+1})\}_{i \in B}$ with replacement from \mathcal{D}_{env}
 828 4: Form targets $y_i = [r_i, s_{i+1}]$ for all $i \in B$
 829 5: **// Assignment Step (Eq. 9)**
 830 6: Sample m candidate latents $\{z_j\}_{j=1}^m \sim \mathcal{N}(0, I)$
 831 7: **for** $k = 1$ to K in parallel **do**
 832 8: $z_{i,k}^* = \arg \min_{1 \leq j \leq m} \|g_{\theta_k}(s_i, a_i, z_j) - y_i\|^2$ for all $i \in B$
 833 9: **// Update Step (Eq. 10)**
 834 10: **for** $k = 1$ to K in parallel **do**
 835 11: $\theta_k \leftarrow \theta_k - \eta \nabla_{\theta_k} \frac{1}{|B|} \sum_{i \in B} \|g_{\theta_k}(s_i, a_i, z_{i,k}^*) - y_i\|^2$

836
837 **Algorithm 3** WIMLE: Detailed Implementation
838

 839 1: **Input:** Rollout horizon H , ensemble size $K = 7$, batch size B , model training frequency
840 train_freq , number of latent codes m , number of model updates U
 841 2: Initialize policy π_ϕ , ensemble of IMLE world models $\{g_{\theta_k}\}_{k=1}^K$, environment dataset \mathcal{D}_{env} ,
842 model dataset $\mathcal{D}_{\text{model}}$
 843 3: **for** environment steps **do**
 844 4: Collect environment transition using π_ϕ ; add to \mathcal{D}_{env}
 845 5: **if** step mod $\text{train_freq} = 0$ **then**
 846 6: **// IMLE World Model Training**
 847 7: Train ensemble $\{g_{\theta_k}\}_{k=1}^K$ in parallel using Algorithm 2
 848 8: **// Uncertainty-Aware Rollout Generation**
 849 9: Clear $\mathcal{D}_{\text{model}}$
 850 10: Sample batch of starting states $\{s_0^{(i)}\}_{i=1}^B$ with replacement from \mathcal{D}_{env}
 851 11: **for** $t = 0$ to $H - 1$ **do**
 852 12: $a_t^{(i)} \sim \pi_\phi(\cdot | s_t^{(i)})$ for all $i \in \{1, \dots, B\}$
 853 13: Sample m latents $\{z_j\}_{j=1}^m \sim \mathcal{N}(0, I)$
 854 14: Generate predictions $\{g_{\theta_k}(s_t^{(i)}, a_t^{(i)}, z_j)\}_{k=1, j=1}^{K, m}$ from all ensemble members for all i
 855 15: Compute predictive uncertainty: $\sigma_t^{(i)} = \text{std}_{k,j} [g_{\theta_k}(s_t^{(i)}, a_t^{(i)}, z_j)]$ {aggregated over en-
856 sembles and latents}
 857 16: Set weight $w_t^{(i)} = 1/(\sigma_t^{(i)} + 1)$
 858 17: Select transitions $(s_{t+1}^{(i)}, r_t^{(i)})$ from predictions
 859 18: Add weighted transitions $\{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)}, s_{t+1}^{(i)}, w_t^{(i)})\}_{i=1}^B$ to $\mathcal{D}_{\text{model}}$
 860 19: **// Uncertainty-Weighted Policy Learning**
 861 20: Sample batch from $\mathcal{D}_{\text{env}} \cup \mathcal{D}_{\text{model}}$ (real data has $w = 1$)
 862 21: Update policy using weighted RL objective:
 863 22: $\mathcal{L} = \mathbb{E}_{(s, a, r, s', w) \sim \text{batch}} [w \cdot \ell_{\text{RL}}(s, a, r, s')]$

864 B THEORETICAL ANALYSIS OF RISK-WEIGHTED BELLMAN ESTIMATION
865866 **Bellman fixed point under positive reweighting.** Fix a policy π and a target value function V .
867 Recall the one-step Bellman target

868
$$y = r + \gamma V(s'), \quad (17)$$

869

870 with conditional mean

871
$$\mu(s, a) = \mathbb{E}[y | s, a] = \mathbb{E}_{(r, s') \sim P(\cdot | s, a)}[r + \gamma V(s')]. \quad (18)$$

872

873 Given a strictly positive weight function $w : \mathcal{S} \times \mathcal{A} \rightarrow (0, \infty)$, the corresponding population
874 weighted Bellman regression loss for an arbitrary action-value function Q is
875

876
$$\mathcal{L}_w(Q) = \mathbb{E}_{(s, a) \sim d^\pi} \mathbb{E}_{(r, s') \sim P(\cdot | s, a)}[w(s, a) (y - Q(s, a))^2]. \quad (19)$$

877

878 **Lemma (Weights do not change the Bellman target).** Assume that $w(s, a) > 0$ for all (s, a) in the
879 support of d^π . Then the unique minimizer of $\mathcal{L}_w(Q)$ over all action-value functions $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$
880 is

881
$$Q^*(s, a) = \mu(s, a) \quad \text{for all } (s, a), \quad (20)$$

882

883 which is the same minimizer as for the unweighted objective (obtained by setting $w \equiv 1$).
884885 *Proof.* Fix (s, a) and consider the conditional risk as a function of a scalar $q \in \mathbb{R}$:

886
$$\ell_{s, a}(q) = \mathbb{E}[w(s, a) (y - q)^2 | s, a] = w(s, a) \mathbb{E}[(y - q)^2 | s, a]. \quad (21)$$

887 Since $w(s, a)$ is a strictly positive constant with respect to the inner expectation, it does not affect
888 the minimizer in q . The derivative of the unweighted term with respect to q is
889

890
$$\frac{\partial}{\partial q} \mathbb{E}[(y - q)^2 | s, a] = 2(q - \mathbb{E}[y | s, a]) = 2(q - \mu(s, a)), \quad (22)$$

891

892 which vanishes if and only if $q = \mu(s, a)$. Thus, for each fixed (s, a) , the unique minimizer of
893 $\ell_{s, a}(q)$ is $q = \mu(s, a)$, independently of the choice of $w(s, a) > 0$.
894895 The global objective $\mathcal{L}_w(Q)$ in Eq. 19 is the expectation of $\ell_{s, a}(Q(s, a))$ under d^π . For any action-
896 value function Q , we can write

897
$$\ell_{s, a}(Q(s, a)) - \ell_{s, a}(\mu(s, a)) = w(s, a) \mathbb{E}[(Q(s, a) - \mu(s, a))^2 | s, a] \geq 0, \quad (23)$$

898

899 with equality if and only if $Q(s, a) = \mu(s, a)$. Integrating this inequality with respect to d^π shows
900 that $\mathcal{L}_w(Q) \geq \mathcal{L}_w(\mu)$, with strict inequality whenever Q differs from μ on a set of positive d^π -
901 measure. Hence, $Q^* = \mu$ is the unique minimizer of \mathcal{L}_w , and this minimizer does not depend on the
902 particular choice of strictly positive weights. \square 903 **Proof of the linear-critic GLS lemma.** We restate the setting from the main text. Let $x_i = (s_i, a_i)$
904 denote m state-action pairs with feature vectors $\phi(x_i) \in \mathbb{R}^d$, and let y_i be the corresponding one-
905 step TD targets. Assume a linear value model
906

907
$$y_i = \phi(x_i)^\top \theta^* + \varepsilon_i, \quad (24)$$

908

909 where $\theta^* \in \mathbb{R}^d$ is the true parameter vector, the noise satisfies $\mathbb{E}[\varepsilon_i | x_i] = 0$, and the conditional
910 variances are $\text{Var}(\varepsilon_i | x_i) = \sigma_i^2$. Collect the targets into a vector $y \in \mathbb{R}^m$, the features into a design
911 matrix $\Phi \in \mathbb{R}^{m \times d}$ whose i -th row is $\phi(x_i)^\top$, and define the noise vector $\varepsilon = (\varepsilon_1, \dots, \varepsilon_m)^\top$. Then
912

913
$$y = \Phi \theta^* + \varepsilon, \quad (25)$$

914

915 with $\mathbb{E}[\varepsilon] = 0$ and covariance

916
$$\Sigma = \text{Cov}(\varepsilon) = \text{diag}(\sigma_1^2, \dots, \sigma_m^2). \quad (26)$$

917

918 We assume that Φ has full column rank so that all normal equations below are solvable.

918 **Linear unbiased estimators and GLS.** Consider linear estimators of θ^* of the form
 919

$$920 \quad \hat{\theta} = Ay, \quad (27)$$

921 for some matrix $A \in \mathbb{R}^{d \times m}$. Unbiasedness for all θ^* requires
 922

$$923 \quad \mathbb{E}[\hat{\theta}] = \mathbb{E}[A(\Phi\theta^* + \varepsilon)] = A\Phi\theta^* = \theta^* \quad \text{for all } \theta^*, \quad (28)$$

924 which is equivalent to the constraint
 925

$$926 \quad A\Phi = I_d. \quad (29)$$

927 Under this constraint, the covariance of $\hat{\theta}$ is
 928

$$929 \quad \text{Cov}(\hat{\theta}) = A \text{Cov}(y) A^\top = A\Sigma A^\top. \quad (30)$$

930 The generalized least-squares (GLS) estimator corresponds to the specific choice
 931

$$932 \quad A_{\text{GLS}} = (\Phi^\top \Sigma^{-1} \Phi)^{-1} \Phi^\top \Sigma^{-1}, \quad (31)$$

933 so that
 934

$$935 \quad \hat{\theta}_{\text{GLS}} = A_{\text{GLS}}y. \quad (32)$$

936 It is immediate that $A_{\text{GLS}}\Phi = I_d$, so $\hat{\theta}_{\text{GLS}}$ is linear and unbiased. Its covariance is
 937

$$938 \quad \text{Cov}(\hat{\theta}_{\text{GLS}}) = A_{\text{GLS}}\Sigma A_{\text{GLS}}^\top = (\Phi^\top \Sigma^{-1} \Phi)^{-1}. \quad (33)$$

940 **Optimality of GLS.** Let $\tilde{\theta} = Ay$ be any other linear unbiased estimator, so that $A\Phi = I_d$ by
 941 equation 29. Define
 942

$$943 \quad C = A - A_{\text{GLS}}. \quad (34)$$

944 Then
 945

$$946 \quad C\Phi = A\Phi - A_{\text{GLS}}\Phi = I_d - I_d = 0. \quad (35)$$

947 The covariance of $\tilde{\theta}$ can be expanded as
 948

$$949 \quad \text{Cov}(\tilde{\theta}) = A\Sigma A^\top \quad (36)$$

$$950 \quad = (A_{\text{GLS}} + C)\Sigma(A_{\text{GLS}} + C)^\top \quad (37)$$

$$951 \quad = A_{\text{GLS}}\Sigma A_{\text{GLS}}^\top + A_{\text{GLS}}\Sigma C^\top + C\Sigma A_{\text{GLS}}^\top + C\Sigma C^\top. \quad (38)$$

952 We now show that the cross terms vanish. Using the definition of A_{GLS} ,
 953

$$954 \quad A_{\text{GLS}}\Sigma = (\Phi^\top \Sigma^{-1} \Phi)^{-1} \Phi^\top \Sigma^{-1} \Sigma = (\Phi^\top \Sigma^{-1} \Phi)^{-1} \Phi^\top, \quad (39)$$

955 and similarly
 956

$$957 \quad \Sigma A_{\text{GLS}}^\top = \Sigma \Sigma^{-1} \Phi (\Phi^\top \Sigma^{-1} \Phi)^{-1} = \Phi (\Phi^\top \Sigma^{-1} \Phi)^{-1}. \quad (40)$$

958 Therefore,
 959

$$960 \quad A_{\text{GLS}}\Sigma C^\top = (\Phi^\top \Sigma^{-1} \Phi)^{-1} \Phi^\top C^\top = (\Phi^\top \Sigma^{-1} \Phi)^{-1} (C\Phi)^\top = 0, \quad (41)$$

$$961 \quad C\Sigma A_{\text{GLS}}^\top = C\Phi (\Phi^\top \Sigma^{-1} \Phi)^{-1} = 0, \quad (42)$$

962 where we used equation 35. Thus
 963

$$964 \quad \text{Cov}(\tilde{\theta}) = \text{Cov}(\hat{\theta}_{\text{GLS}}) + C\Sigma C^\top. \quad (43)$$

965 For any vector $v \in \mathbb{R}^d$,

$$966 \quad v^\top C\Sigma C^\top v = (C^\top v)^\top \Sigma (C^\top v) \geq 0, \quad (44)$$

967 because Σ is positive semidefinite. Hence $C\Sigma C^\top$ is positive semidefinite, and we have
 968

$$969 \quad \text{Cov}(\hat{\theta}_{\text{GLS}}) \preceq \text{Cov}(\tilde{\theta}), \quad (45)$$

970 with equality if and only if $C = 0$, i.e., $A = A_{\text{GLS}}$. This establishes that GLS has minimum
 971 covariance among all linear unbiased estimators.

972
973 **Connection to inverse-variance weighting.** In our setting, weighted least squares with per-
974 sample weights $w_i > 0$ corresponds to minimizing

$$975 \quad \sum_{i=1}^m w_i (y_i - \phi(x_i)^\top \theta)^2 = (y - \Phi\theta)^\top W(y - \Phi\theta), \quad (46)$$

976 where $W = \text{diag}(w_1, \dots, w_m)$. The normal equations yield the estimator

$$977 \quad \hat{\theta}_W = (\Phi^\top W \Phi)^{-1} \Phi^\top W y. \quad (47)$$

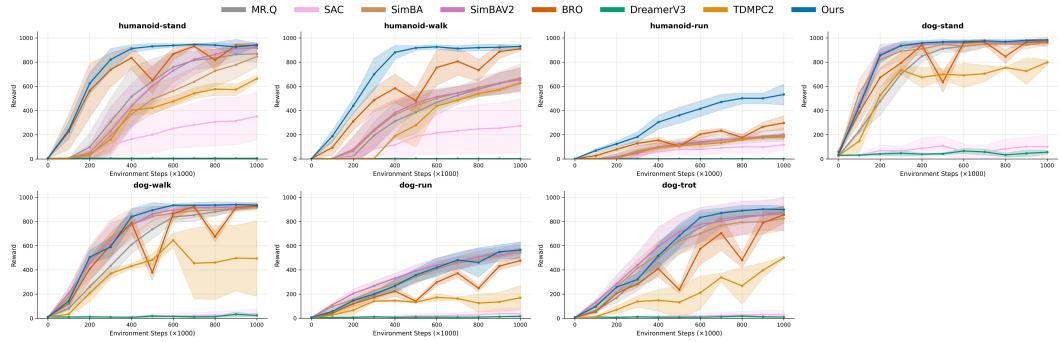
978 Choosing weights $w_i \propto 1/\sigma_i^2$ makes W proportional to Σ^{-1} , so that

$$979 \quad \hat{\theta}_W = (\Phi^\top \Sigma^{-1} \Phi)^{-1} \Phi^\top \Sigma^{-1} y = \hat{\theta}_{\text{GLS}}. \quad (48)$$

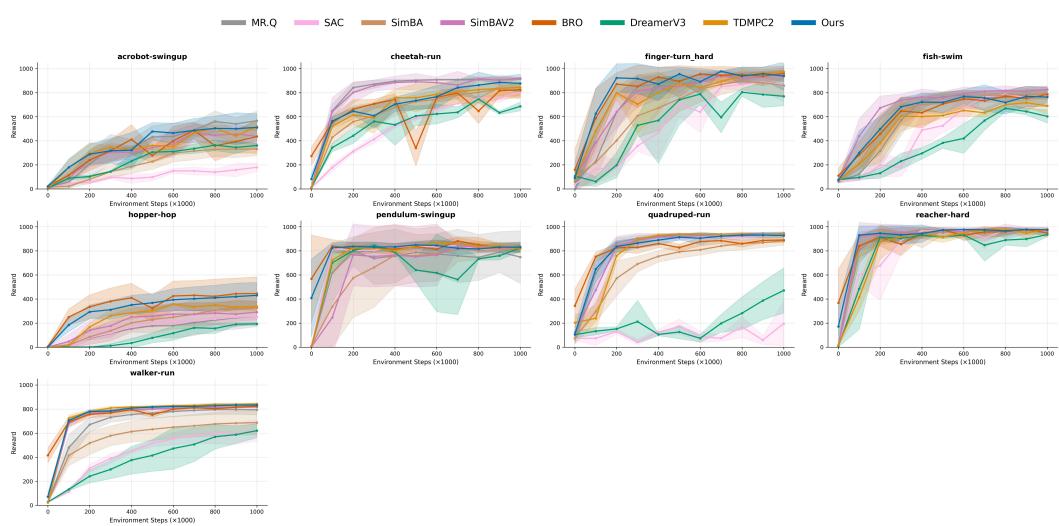
980 Thus inverse-variance weights $w_i \propto 1/\sigma_i^2$ recover the GLS estimator and, by the argument above,
981 minimize the covariance matrix among all linear unbiased estimators. \square

982 C PER-TASK RESULTS

983 We present detailed per-task performance results for WIMLE and other baselines across all benchmarks. The performance on each individual task is shown in Figures 7, 8, 9, and 10.



1000 Figure 7: Per-task results for high-dimensional Dog & Humanoid tasks from DeepMind Control Suite.
1001 We present the IQM of rewards and 95% confidence intervals for BRO and other baselines
1002 run for 1M steps.



1023 Figure 8: Per-task results for DeepMind Control Suite tasks with low-dimensional state/action
1024 spaces. We present the IQM of rewards and 95% confidence intervals for WIMLE and other base-
1025 lines run for 1M steps.

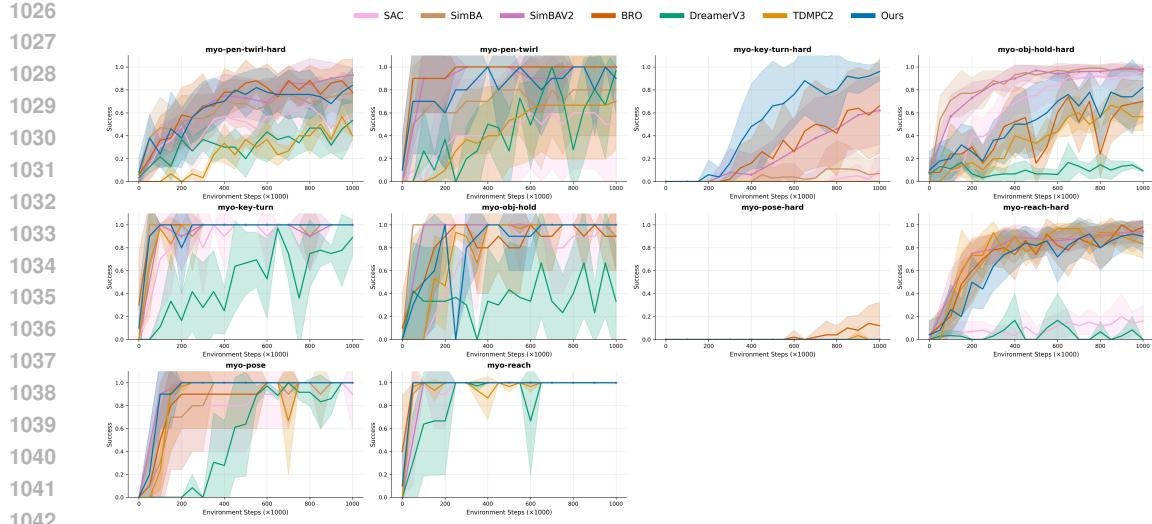


Figure 9: Per-task results for MyoSuite tasks. We present the IQM of success rate and 95% confidence intervals for WIMLE and other baselines run for 1M steps.

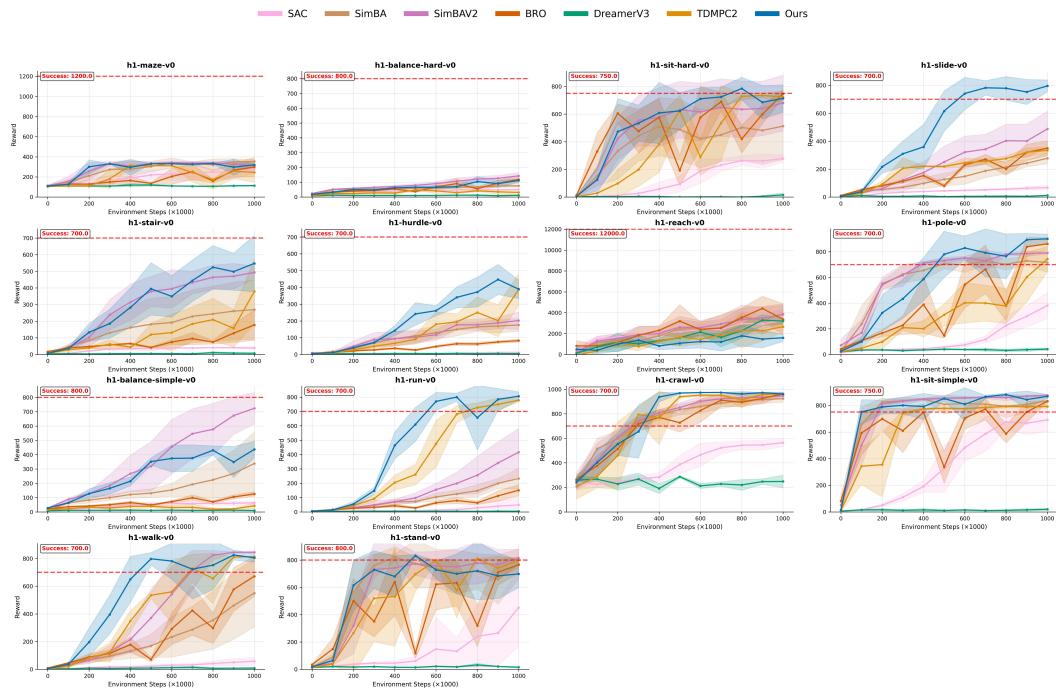


Figure 10: Per-task results for HumanoidBench tasks. We present the IQM of rewards and 95% confidence intervals for WIMLE and other baselines run for 1M steps. The red dashed line indicates the success threshold for each task.

D HYPERPARAMETERS

Table 1 lists the common hyperparameters used across all tasks. These parameters were selected through hyperparameter tuning based on standard practices in RL.

Based on our empirical evaluations, we found that increasing the model batch size to the maximum extent allowed by available GPU resources while proportionally decreasing the number of model

1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
Table 1: Common hyperparameters used across all tasks.

Parameter	Value
SAC Parameters	
Batch size	128
Actor learning rate	3×10^{-4}
Critic learning rate	3×10^{-4}
Number of quantiles	100
Updates per step	10
World Model Parameters	
Model learning rate	1×10^{-3}
Model batch size	512
Model updates	100
Number of latent codes	4
Model training frequency	1000
Number of rollouts	200
Number of ensembles	7

1105
1106 updates can achieve similar performance with improved training speed. When scaling the batch size
1107 in this manner, the model learning rate should be adjusted accordingly following standard Machine
1108 Learning practices.

D.1 ROLLOUT LENGTH SELECTION

1122
1123 We select task-specific rollout horizons H through experimentation. For easier tasks where base-
1124 lines already saturate near the maximum score (e.g., MyoSuite manipulation tasks), we start with
1125 short horizons ($H=1-2$) and increase only if performance benefits are observed, as longer horizons
1126 may still introduce slight performance degradation—though not to the extent seen in traditional
1127 MBRL methods. For harder tasks requiring longer-term planning (e.g., HumanoidBench, Dog &
1128 Humanoid), we begin with longer horizons ($H=8$) and decrease only if performance gains are seen
1129 empirically. However, we note that even simpler tasks may benefit from longer horizons in some
1130 cases, reflecting the task-specific nature of optimal rollout length. This selection balances the ben-
1131 efits of synthetic data augmentation with computational cost, as the marginal benefit of additional
1132 rollout steps diminishes beyond each task’s optimal horizon. We cap H at 8 to maintain rollout
1133 throughput and because we observe diminishing returns beyond task-specific optima; Tables 6, 7,
and 8 report the chosen H per task. An interesting future direction would be to dynamically adjust
rollout horizons based on the model’s uncertainty level, potentially allowing for adaptive rollout
lengths that scale with model confidence.

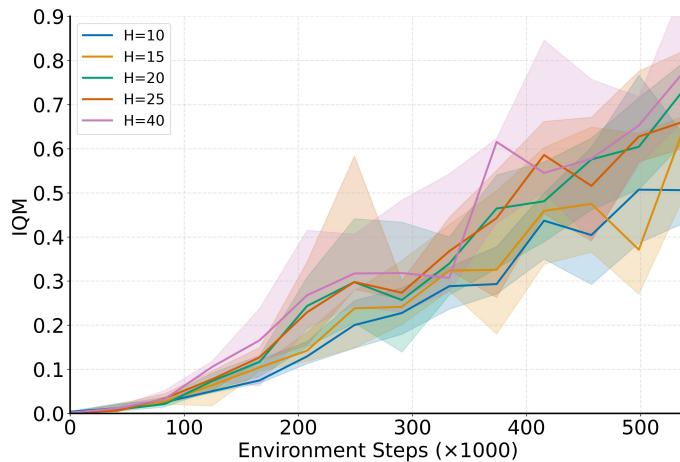


Figure 11: Rollout length sensitivity on h1-hurdle-v0 task of HumanoidBench up to $H = 40$. WIMLE remains stable at longer horizons without rollout scheduling.

E EXPERIMENT DETAILS

This section provides detailed descriptions of the benchmark environments used in our evaluation. We explain the task suites, their characteristics, and the normalization procedures used for fair comparison across different score scales. The following subsections describe each benchmark suite with complete task lists and their state/action dimensions.

E.1 DETAILED RESULTS

We present comprehensive IQM results for WIMLE and baseline methods across all benchmark suites at 100k, 200k, 500k, and 1M environment steps. The best performing method for each step count is highlighted in bold and the second best are underlined. WIMLE performs better across most evaluations.

Table 2: IQM results for DMC suite. Best scores are highlighted in bold, second best are underlined.

Method	100k	200k	500k	1M
MR.Q	0.153	0.362	0.714	0.830
SAC	0.037	0.082	0.210	0.326
SimBA	0.120	0.263	0.522	0.691
SimBAV2	0.235	0.495	<u>0.730</u>	0.845
BRO	<u>0.294</u>	<u>0.519</u>	0.542	<u>0.846</u>
DreamerV3	0.051	0.075	0.165	0.286
TD-MPC2	0.152	0.374	0.566	0.696
WIMLE	0.332	0.575	0.812	0.871

E.2 DEEPMIND CONTROL SUITE

DeepMind Control Suite (Tassa et al., 2018, DMC) is a standard continuous control benchmark encompassing locomotion and manipulation tasks with varying complexity. We evaluate 16 tasks from DMC, focusing on the most challenging locomotion tasks including Dog and Humanoid embodiments. All returns are normalized by dividing by 1000 to scale performance to [0,1]. The complete list of tasks with their observation and action dimensions is provided in Table 6.

1188 Table 3: IQM results for Dog & Humanoid suite. Best scores are highlighted in bold, second best
1189 are underlined.
1190

Method	100k	200k	500k	1M
MR.Q	0.042	0.127	0.557	0.796
SAC	0.007	0.008	0.043	0.069
SimBA	0.067	0.173	0.533	0.773
SimBAV2	0.082	0.200	<u>0.601</u>	0.808
BRO	<u>0.086</u>	<u>0.290</u>	0.355	<u>0.864</u>
DreamerV3	0.006	<u>0.006</u>	0.007	0.010
TD-MPC2	0.014	0.058	0.302	0.527
WIMLE	0.140	0.389	0.803	0.897

1200 Table 4: IQM results for MyoSuite. Best scores are highlighted in bold, second best are underlined.
1201

Method	100k	200k	500k	1M
SAC	0.038	0.350	0.622	0.714
SimBA	<u>0.566</u>	<u>0.728</u>	0.912	0.952
SimBAV2	0.724	0.830	0.956	0.990
BRO	0.440	0.736	0.816	<u>0.980</u>
DreamerV3	0.028	0.044	0.181	0.466
TD-MPC2	0.088	0.394	0.688	0.775
WIMLE	0.460	0.620	<u>0.928</u>	0.980

1211 Table 5: IQM results for HumanoidBench. Best scores are highlighted in bold, second best are
1212 underlined.
1213

Method	100k	200k	500k	1M
SAC	0.008	0.020	0.060	0.168
SimBA	0.070	0.164	0.322	0.521
SimBAV2	0.059	<u>0.179</u>	<u>0.488</u>	<u>0.799</u>
BRO	<u>0.064</u>	0.127	0.100	0.530
DreamerV3	0.003	0.003	0.005	0.007
TD-MPC2	0.023	0.064	0.382	0.734
WIMLE	0.056	0.258	0.735	0.876

1223 E.3 MYOSUITE

1225 MyoSuite (Caggiano et al., 2022) provides high-fidelity musculoskeletal simulations for dexterous
1226 manipulation tasks. We evaluate 10 tasks including both fixed-goal and randomized-goal (hard)
1227 settings. Performance is measured using success rates, which naturally scale to [0,1]. The complete
1228 list of tasks with their observation and action dimensions is provided in Table 7.
1229

1230 E.4 HUMANOIDBENCH

1232 HumanoidBench (Sferrazza et al., 2024) provides locomotion tasks for the UniTree H1 humanoid
1233 robot. We evaluate 14 tasks spanning balance, locomotion, and manipulation. For fair comparison
1234 across tasks with different score scales, all HumanoidBench scores are normalized using each task’s
1235 target success score and random score following the same procedure as in (Lee et al., 2025a;b):
1236

1238
$$\text{Success-Normalized}(x) := \frac{x - \text{random score}}{\text{Target success score} - \text{random score}}$$

1239

1240 The complete list of tasks with their observation and action dimensions is provided in Table 8, and
1241 the random scores and target success scores used for normalization are listed in Table 9.

1242 Table 6: **DMC Tasks.** Complete list of 16 DMC tasks evaluated, with state and action dimensions
 1243 and rollout lengths.

1245 Task	1246 State dim $ \mathcal{S} $	1247 Action dim $ \mathcal{A} $	1248 H
1249 acrobot-swingup	1250 6	1251 1	1252 8
1253 cheetah-run	1254 17	1255 6	1256 1
1257 finger-turn_hard	1258 12	1259 2	1260 1
1261 fish-swim	1262 24	1263 5	1264 8
1265 hopper-hop	1266 15	1267 4	1268 1
1269 pendulum-swingup	1270 3	1271 1	1272 1
1273 quadruped-run	1274 78	1275 12	1276 2
1277 reacher-hard	1278 6	1279 2	1280 1
1281 walker-run	1282 24	1283 6	1284 1
1285 humanoid-stand	1286 67	1287 24	1288 2
1289 humanoid-walk	1290 67	1291 24	1292 6
1293 humanoid-run	1294 67	1295 24	1296 6
1297 dog-stand	1298 223	1299 38	1300 6
1301 dog-walk	1302 223	1303 38	1304 4
1305 dog-run	1306 223	1307 38	1308 6
1309 dog-trot	1310 223	1311 38	1312 4

1262 Table 7: **MyoSuite Tasks.** Complete list of 10 MyoSuite tasks evaluated, with state and action
 1263 dimensions and rollout lengths.

1265 Task	1266 State dim $ \mathcal{S} $	1267 Action dim $ \mathcal{A} $	1268 H
1269 myo-key-turn	1270 93	1271 39	1272 6
1273 myo-key-turn-hard	1274 93	1275 39	1276 1
1277 myo-obj-hold	1278 91	1279 39	1280 4
1281 myo-obj-hold-hard	1282 91	1283 39	1284 1
1285 myo-pen-twirl	1286 83	1287 39	1288 2
1289 myo-pen-twirl-hard	1290 83	1291 39	1292 1
1293 myo-pose	1294 108	1295 39	1296 2
1297 myo-pose-hard	1298 108	1299 39	1300 1
1301 myo-reach	1302 115	1303 39	1304 4
1306 myo-reach-hard	1307 115	1308 39	1309 1

1276 Table 8: **HumanoidBench Tasks.** Complete list of 14 HumanoidBench tasks evaluated, with state
 1277 and action dimensions and rollout lengths.

1279 Task	1280 State dim $ \mathcal{S} $	1281 Action dim $ \mathcal{A} $	1282 H
1283 h1-balance-simple	1284 64	1285 19	1286 1
1287 h1-balance-hard	1288 77	1289 19	1290 6
1291 h1-crawl	1292 51	1293 19	1294 6
1296 h1-hurdle	1297 51	1298 19	1299 8
1301 h1-maze	1302 51	1303 19	1304 6
1306 h1-pole	1307 51	1308 19	1309 8
1311 h1-reach	1312 57	1313 19	1314 4
1316 h1-run	1317 51	1318 19	1319 8
1321 h1-slide-v0	1322 51	1323 19	1324 8
1326 h1-slide-v1	1327 51	1328 19	1329 8
1331 h1-sit-hard	1332 51	1333 19	1334 8
1336 h1-stair	1337 51	1338 19	1339 8
1341 h1-stand	1342 51	1343 19	1344 8
1346 h1-walk	1347 51	1348 19	1349 8

1296 Table 9: **HumanoidBench Normalization Scores.** Random scores and target success scores used
 1297 for normalization.

1299	Task	Random Score	Target Success Score
1300	h1-balance-simple	9.391	800
1301	h1-balance-hard	9.044	800
1302	h1-crawl	272.658	700
1303	h1-hurdle	2.214	700
1304	h1-maze	106.441	1200
1305	h1-pole	20.09	700
1306	h1-reach	260.302	12000
1307	h1-run	2.02	700
1308	h1-slide-v0	2.02	700
1309	h1-slide-v1	2.02	700
1310	h1-sit-hard	10.545	800
1311	h1-stair	2.214	700
1312	h1-stand	10.545	800
1313	h1-walk	2.377	700

1315 F ADDITIONAL RESULTS

1317 Figure 12 compares standard WIMLE (mixed real + imagined data) against an “imagined-only” vari-
 1318 ant whose critic and actor are trained exclusively on model-generated rollouts. Removing real trans-
 1319i-
 1320i-
 1321i-
 1322i-
 1323i-
 1324i-
 1325i-
 1326i-
 1327i-
 1328i-
 1329i-
 1330i-
 1331i-
 1332i-
 1333i-
 1334i-
 1335i-
 1336i-
 1337i-
 1338i-
 1339i-
 1340i-
 1341i-
 1342i-
 1343i-
 1344i-
 1345i-
 1346i-
 1347i-
 1348i-
 1349i-

Figure 12 compares standard WIMLE (mixed real + imagined data) against an “imagined-only” variant whose critic and actor are trained exclusively on model-generated rollouts. Removing real transitions does reduce performance slightly, but the imagined-only curve remains close to the original WIMLE results, underscoring that our synthetic trajectories are strong enough to sustain competitive learning.

Figure 14 and Table 10 illustrate the wall-clock efficiency of WIMLE. Despite the overhead of ensemble training, WIMLE’s superior sample efficiency results in a significantly lower total time-to-solution compared to baselines.

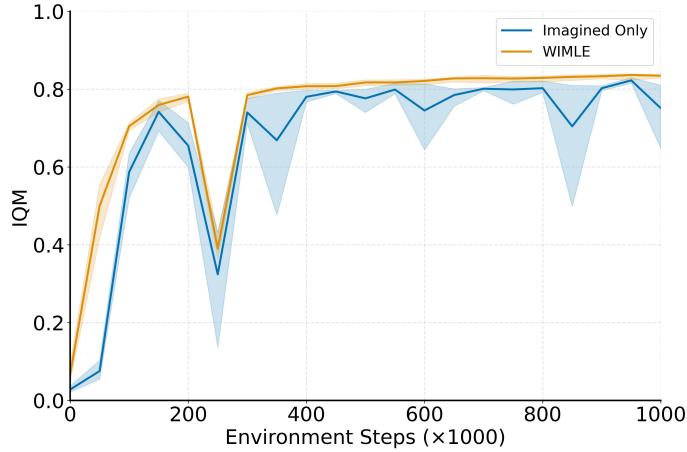


Figure 12: Imagined-only WIMLE (no real data in critic/actor updates) achieves performance comparable to the standard configuration, highlighting the strength of the generated rollouts.

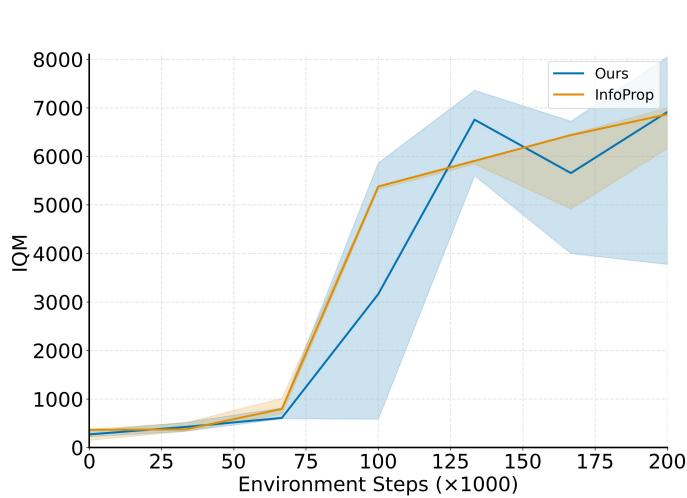


Figure 13: Humanoid (Mujoco) comparison against InfoProp for 4 seeds. Unlike InfoProp, which uses task-specific termination functions for synthetic rollouts, WIMLE runs with the generic setup used for our DMC experiments yet still matches or exceeds InfoProp.

Method	Step 300k		Step 500k		Step 1000k	
	Time (h)	IQM	Time (h)	IQM	Time (h)	IQM
Infoprop	11.49	0.044	19.15	0.102	38.30	0.077
DreamerV3	3.90	0.001	6.50	0.001	13.00	0.001
TD-MPC2	3.43	0.001	5.72	0.038	11.43	0.190
Ours	4.74	0.168	7.91	0.345	15.81	0.561

Table 10: Performance and time at different training milestones.

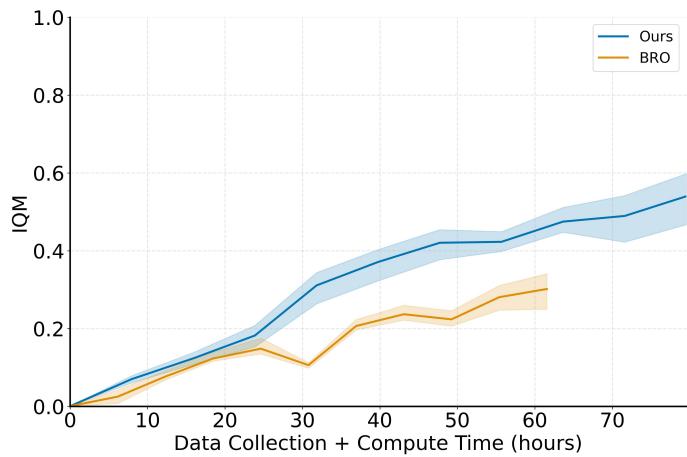


Figure 14: Projected total wall-clock time (algorithm compute + data collection) on Humanoid-run, assuming a 5Hz control rate. WIMLE reaches asymptotic performance significantly faster than BRO (the fastest model-free baseline) because its superior sample efficiency drastically reduces the time spent collecting real-world data, outweighing its higher per-update compute cost.