MAMBA Rebuttal Experiments

1 Additional algorithms

1.1 RL^2

RL^2 is a common meta-RL baseline; our additional experiments show that MAMBA is superior to RL^2 in all environments except EscapeRoom, where the difference is a bit in favor of RL^2 (although RL^2 used 5× more data than MAMBA). The Reacher environment is currently running but will be added to the final version of the paper. See Table 1.

1.2 DREAM

This baseline was suggested by **reviewer fKkT**. Since the experiments in DREAM deal with goal conditioned meta-RL tasks in GridWorld environments, we tested it in our Rooms environment. Our experiments in Rooms-3 show that DREAM fails to obtain any meaningful policy, and the meta-episode rewards remain negative (MAMBA, Dreamer, VariBAD and HyperX all managed to obtain returns greater than 100). Results from experiments with more rooms behave the same. This may result from bad hyper-parameter adaptation between the original paper's scenarios and Rooms; however, this is another advantage of using a Dreamer-based architecture as in MAMBA (same parameters fit different scenarios out-of-the-box).

2 Additional environments

2.1 Rooms7-8

We added more rooms to challenge MAMBA. We observe that MAMBA still manages to find a Bayes-optimal behavior and the drop in return is attributed to the exploration required to find the additional goals. We note that we are running the missing Dreamer-Vanilla baseline, results will be available for the final version of the paper. VariBAD or HyperX were not tested as they already failed to solve scenarios with less rooms. See Table 1.

2.2 PandaReacher

We tested MAMBA on a dynamically complex system with visual inputs and sparse rewards which requires a non-trivial Bayes-policy. In particular, we borrow the PandaReacher environment from [1]. In this environment a 7-degree-of-freedom (DoF) robotic arm is required to reach goals with the robot end-effector (i.e. the tip of the robot). The goals are hidden and must be discovered through exploration. The rewards are sparse, meaning only when the end-effector is within a short distance from the goal, a reward of 1 is obtained (otherwise the reward is 0). The observation is an image of size 84×84×4, and the action space is the direction the end-effector should go towards.

Since VariBAD and HyperX do not ingest visual inputs, similar to our visual Reacher experiment, we compared only between Dreamer-Vanilla, Dreamer-Tune, and MAMBA. The results after 6M steps for Dreamer-Vanilla, Dreamer-Tune, and MAMBA are 5.3, 14.5, and 133.3 correspondingly, demonstrating again that MAMBA is better than the other Dreamer versions. We also note that [1] reported returns of ∼135 in

	RL^2	VariBAD	HyperX	Dreamer-Vanilla	Dreamer-Tune	MAMBA (Ours)
HalfCircle:	239.4±3.1	218.0 ± 26.0	204.0 ± 19.9	**241.1 ± 9.9**	**239.7 ± 3.1**	**242.2 ± 7.9**
	∼10M	∼10M	∼15M	∼2M	∼2M	∼2M
Rooms-3:	108.0±31.7	**158.7 ± 2.6**	111.8 ± 20.2	137.5 ± 6.2	142.6 ± 4.1	**156.2 ± 1.7**
Rooms-4:	85.1±19.0	88.1 ± 39.2	14.1 ± 3.0	115.0 ± 6.7	119.8 ± 3.1	**136.7 ± 1.4**
Rooms-5:	72.4±36.8	0.9 ± 2.5	-15.8 ± 0.4	96.5 ± 2.5	93.9 ± 5.3	**113.1 ± 5.7**
Rooms-6:	64.9±15.9	-11.0 ± 0.9	-15.7 ± 0.7	69.5 ± 6.3	71.0 ± 6.3	**94.5 ± 1.2**
Rooms-7:	42.9±22.9	-	-	-	50.2±4.4	**73.2±2.9**
Rooms-8:	29.0±17.0	-	-	-	29.1±8.3	**55.6±3.0**
	∼100M	∼100M	∼100M	∼6M	∼6M	∼6M
Reacher-1:	-	473.7 ± 27.5	505.9 ± 36.0	552.0 ± 27.8	503.4 ± 69.0	**655.5 ± 12.3**
Reacher-2:	-	46.6 ± 33.2	30.0 ± 48.5	**247.4 ± 80.5**	**217.6 ± 64.3**	**285.8 ± 89.6**
Reacher-3:	-	0.2 ± 0.2	0.5 ± 0.2	183.6 ± 100.0	76.9 ± 80.5	**325.0 ± 47.0**
Reacher-4:	-	0.0 ± 0.0	-0.5 ± 1.1	0.4 ± 0.0	0.1 ± 0.2	**77.7 ± 61.1**
	-	∼150M	∼150M	∼10M	∼10M	∼10M
EscapeRoom:	79.9±4.4	70.7 ± 5.3	66.9 ± 6.5	68.2 ± 2.4	**73.2 ± 7.8**	**73.9 ± 3.1**
	∼20M	∼20M	∼20M	∼4M	∼4M	∼4M

Table 1: Total return comparison of VariBad, HyperX, Dreamer-Vanilla, Dreamer-Tune, and MAMBA on different meta environments. We report the final reward (mean ± std), and the number of time steps until convergence (below dashed line).

	VariBAD	HyperX	Dreamer-Vanilla	Dreamer-Tune	MAMBA
return:	1369.3 ± 75.3	-	2068.3 ± 156.7	2096.3 ± 79.8	**2405.9 ± 119.0**
steps:	100M	-	30M	30M	30M

Table 2: Comparison of VariBad, HyperX, Dreamer-Vanilla, Dreamer-Tune, and MAMBA in the Humanoid-Dir environment.

this environment (results were given as plots) after 20M environment steps compared to our 6M steps. We will report the performance after 20M steps for the final version.

2.3 Humanoid-Dir

First proposed in [2], in this environment the Humanoid MuJoCo task is initialized with a random target walking direction $\theta \in [0, 2\pi]$ on the 2D plane. Table 2 summarizes the results of this experiment, and we can see that again, using less environment steps (30M vs. 100M) the algorithms based on the Dreamer architecture perform better than VariBAD and HyperX. Among these, MAMBA is substantially better reaching the highest return.

2.4 HalfCircle-Wind

This is a modification of the HalfCircle environment, where a stochastic wind force is added at each step. Results in Table 3 show that the conclusions from HalfCircle also extend to this stochastic version.

	VariBAD	HyperX	Dreamer-Vanilla	Dreamer-Tune	MAMBA
return:	194.5± 44.7	177± 118.5	226.1±3.5	224.5±4.4	224.1±5.2
steps:	20M	20M	2M	2M	2M

Table 3: Comparison of VariBad, HyperX, Dreamer-Vanilla, Dreamer-Tune, and MAMBA on the HalfCircle-Wind environment.

References

[1] Era Choshen and Aviv Tamar. Contrabar: Contrastive bayes-adaptive deep rl. *arXiv preprint arXiv:2306.02418*, 2023.

[2] Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pages 5331–5340. PMLR, 2019.