9. Appendix

9.1. Environment Parameters

CartPoleSwingUp			Acrobot-v1		
Parameter	Train	Test Range	Parameter	Train	Test Range
Pole Length	0.6	$[0.2, 1.2]$	Link Length	1.0	$[1.0, 10.0]$
Pendulum Mass	0.5	$[1.0, 20]$	Link Mass	1.0	$[1.0, 10.0]$
Magnitude of Force	10	$[0.1, 2.0]$	Magnitude of Torque	1.0	$[0.1, 5.0]$

(a)

Quadruped			Walker		
Parameter	Train	Test Range	Parameter	Train	Test Range
Shin Length	0.25	$[0.1, 2.0]$	Thigh Length	0.225	$[0.1, 0.7]$
Joint Damping	30.0	$[20.0, 40.0]$	Joint Damping	0.1	$[0.1, 9.0]$
Contact Friction	1.4	$[0.5, 4.0]$	Contact Friction	0.7	$[0.02, 2.0]$

(b)

Table 2. (a) Parameter ranges for CartPoleSwingUp and Acrobot environments. Ranges are selected to be large enough so that failure modes exist for all model types. (b) Parameter ranges for the Quadruped and Walker environment.

9.2. Training Hyper-parameters

Table 3 provides the hyper-parameters used for training all models in each respective environment. Certain parameters do not apply to a particular model (i.e. dropout rate for non-dropout models). We evaluate over a range of hyperparameters for each baseline. For Dropout, we evaluated training with a dropout rate of 0.001, 0.01, 0.05, 0.1, 0.15, and 0.2. We evaluated with a maximum entropy coefficient of 0.001, 0.005, 0.01, 0.05, and 0.1. For Weight Decay, we evaluated training with a beta value of $1e-5$, $1e-4$, $1e-3$, $1e-2$, $1e-1$, $1e0$, and $1e1$. We evaluated VIB with a regularization value of $1e-4$, $1e-3$, $1e-2$, $1e-1$, $1e0$, $1e1$, and $1e2$.

Parameter	Environment			
	CartPole	Acrobot	Quadruped	Walker
Network size	[64,64,64]	[64,64.64]	[256,256,256]	[256,256,256]
Activation Function	ReLU	ReLU	ReLU	ReLU
Discount Factor	0.99	0.99	0.99	0.99
GAE λ	0.95	0.95	0.95	0.95
$\mathcal{L}_{\text{VF}}$ Coefficient	0.5	0.5	0.5	0.5
$\mathcal{L}_{\text{SNR}}$ Coefficient	20.0	20.0	10.0	5.0
Initial σ_i Range	[0.1, 0.2]	[0.1, 0.2]	[0.05, 0.2]	[0.05, 0.2]
Max SNR ($\Omega_{\text{Max SNR}}$)	10.0	10.0	10.0	10.0
Dropout Rate	0.1	0.2	0.01	0.001
Entropy Bonus Coefficient	0.01	0.05	0.01	0.005
Weight Decay Coefficient	1.0	1.0	0.1	0.01
VIB Coefficient	0.01	10.0	1.0	10.0

Table 3. Hyperparameters for each of the models presented in Table 1.

9.3. Model Comparison Heatmaps for CartpoleSwingUp

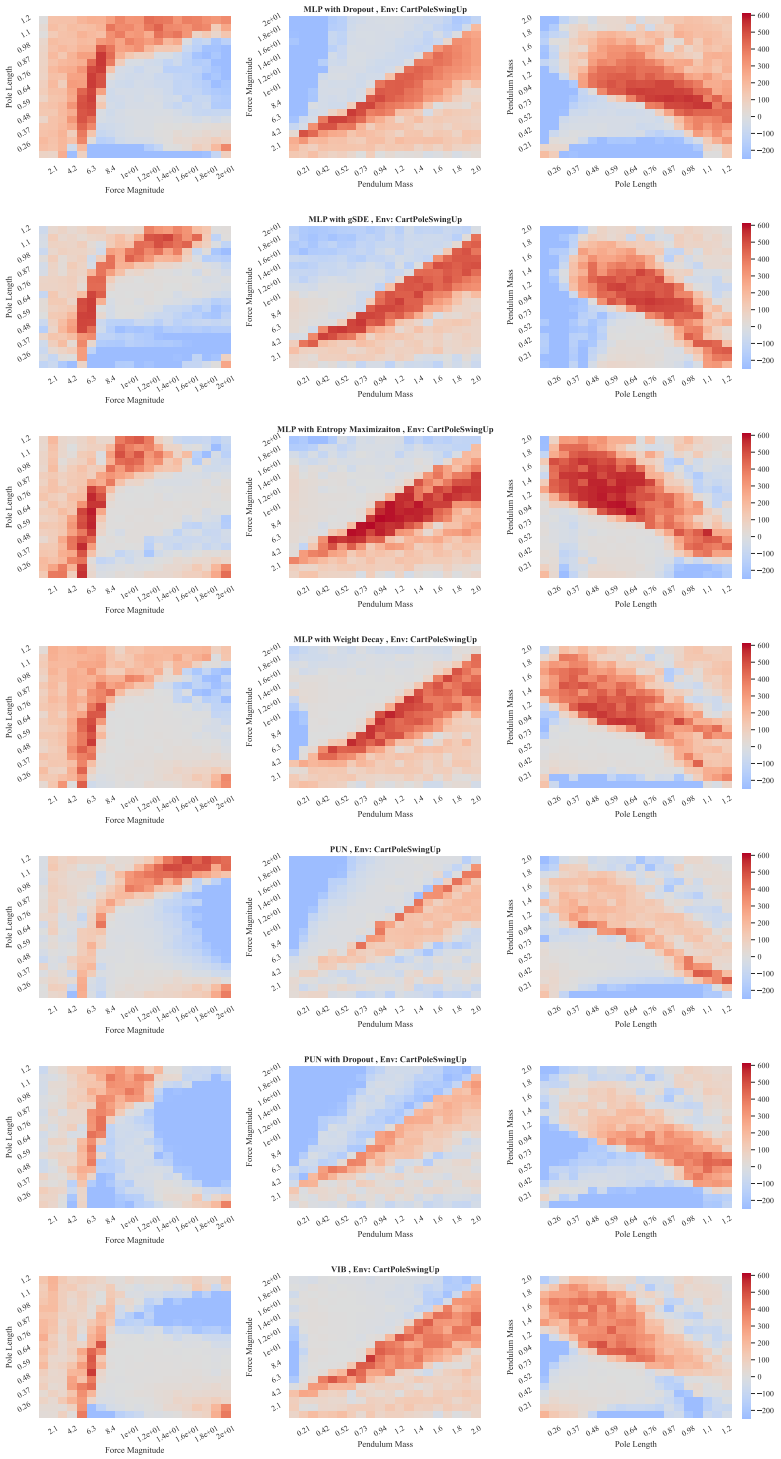

Figure 7. Heatmap of the difference in the ATP metric between each model and the MLP baseline for the mutated CartpoleSwingUp environments.

9.4. Model Comparison Heatmaps for Acrobot-v1

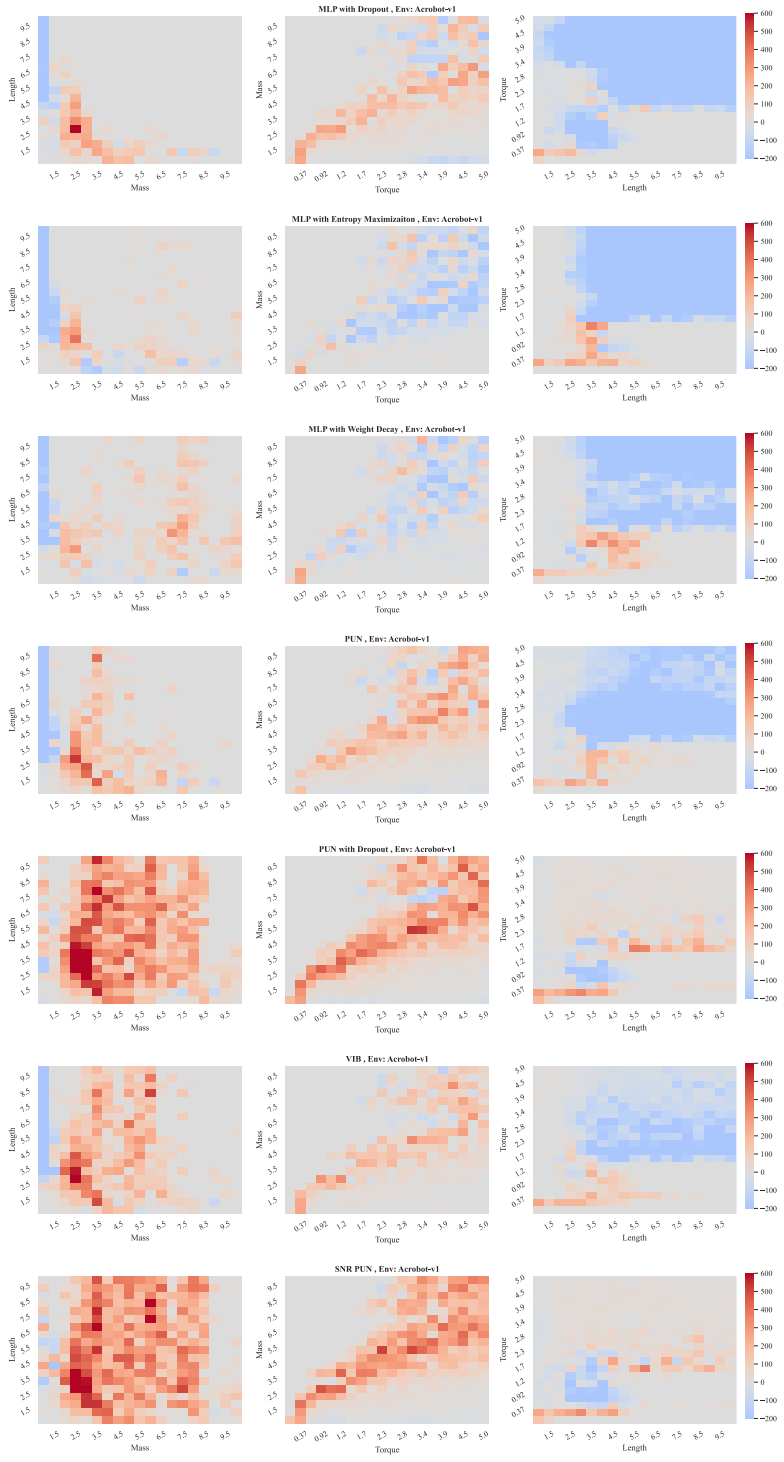

Figure 8. Heatmap of the difference in the ATP metric between each model and the MLP baseline for the mutated Acrobot-v1 environments.

9.5. Model Comparison Heatmaps for Quadruped

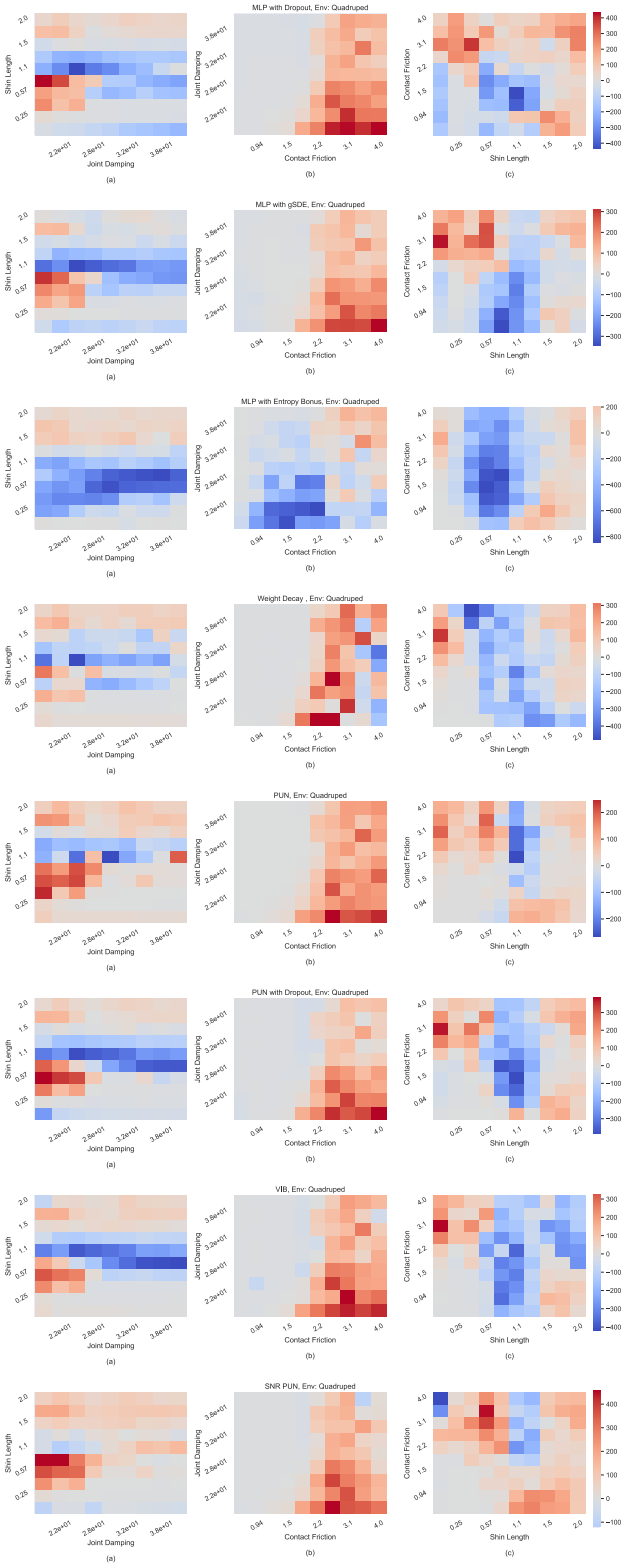

Figure 9. Heatmap of the difference in the ATP metric between each model and the MLP baseline for the mutated Quadruped environments.

9.6. Model Comparison Heatmaps for Walker

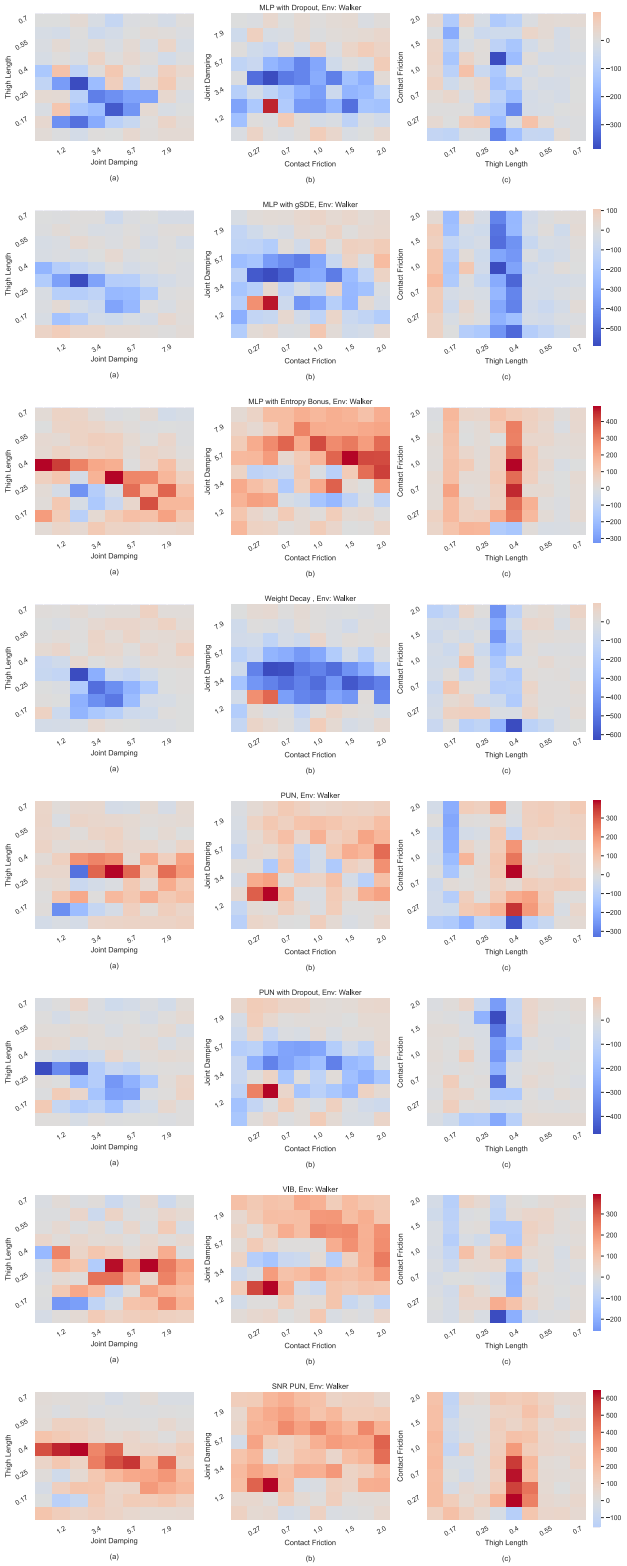

Figure 10. Heatmap of the difference in the ATP metric between each model and the MLP baseline for the mutated Walker environments.

9.7. Baseline Task Training Results for CartpoleSwingUp and Acrobot-v1

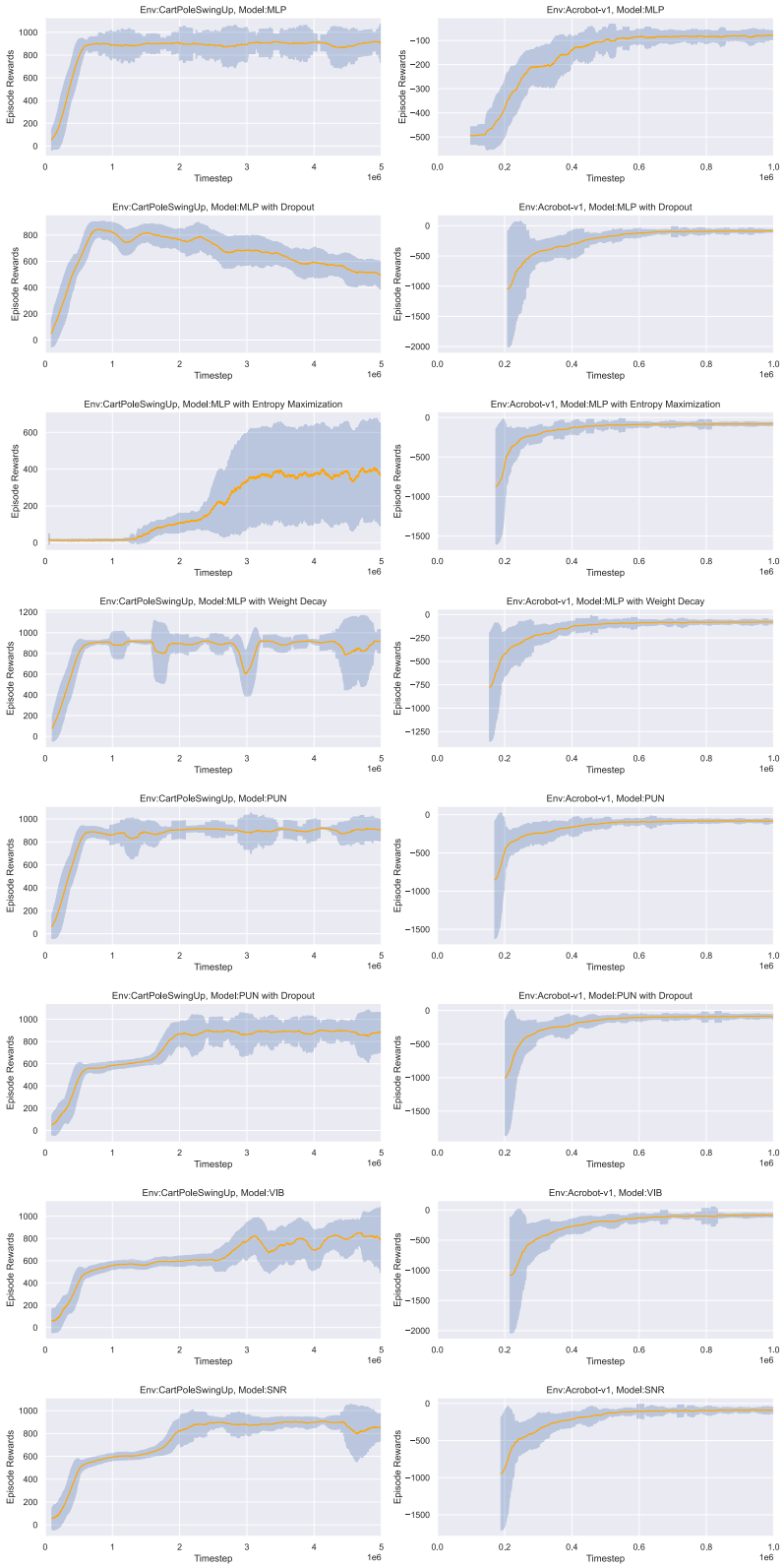

Figure 11. Training curves for each model on CartPoleSwingUp and Acrobot, on the respective baseline task. All models reach comparable performance on the training task.

9.8. Baseline Task Training Results for Two MujoCo Environments

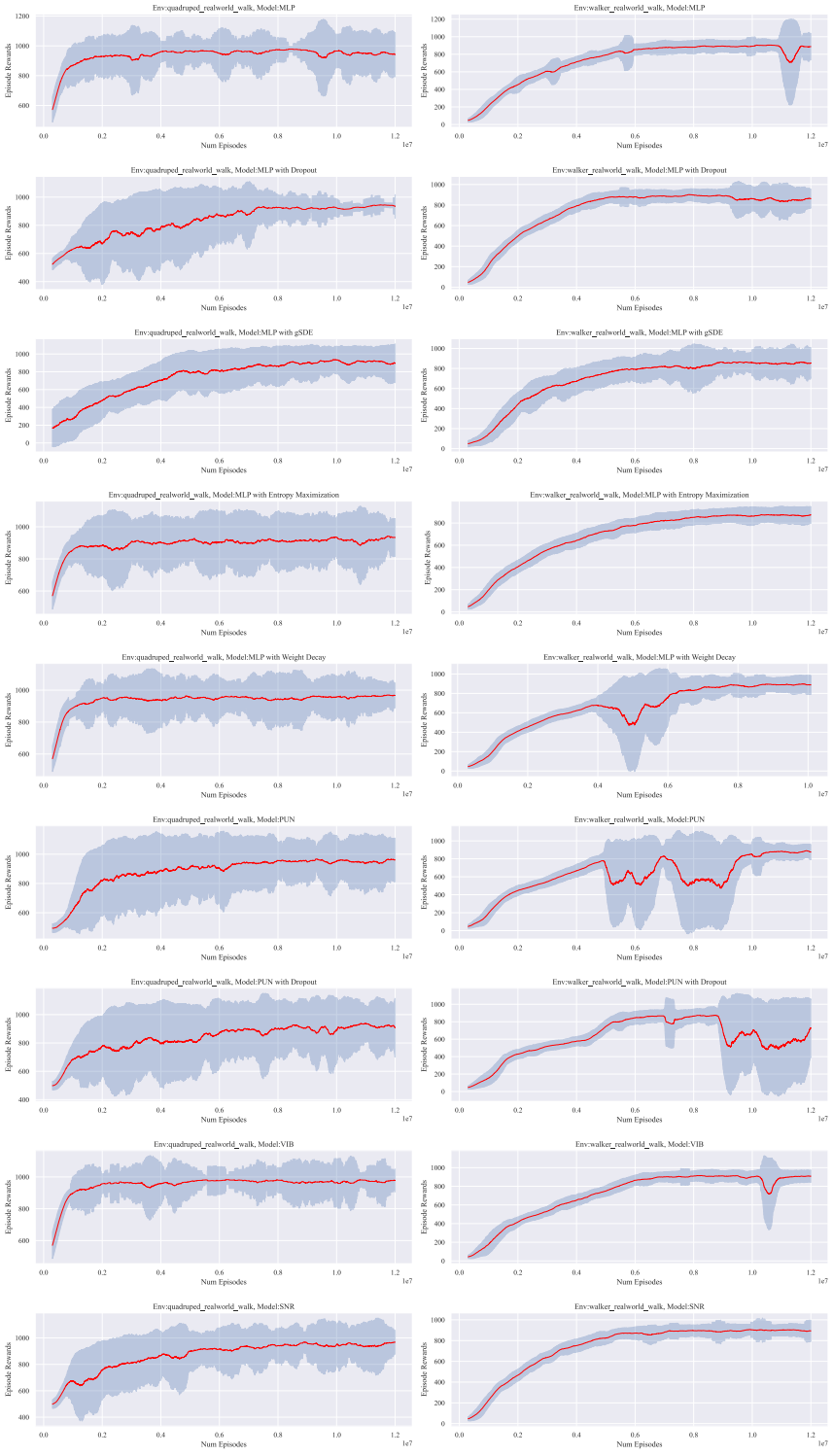

Figure 12. Training curves for each model and each MuJoCo environment on the respective baseline task. All models reach comparable performance on the training task.

9.9. Ablation Study

We provided an ablation study in the main text focused on our SNR term and evaluation resampling strategy. We provide the an extended ablation results for the new initialization, training resampling method, and deterministic critic here for completeness. We see that the largest negative impact comes from not resampling every time step during training, and the least impact comes from using standard initialization.

Model	Experiment	Cartpole	Acrobot	Quadruped	Walker
PUN	Standard Initialization	445.98 $\pm$ 23.47	211.83 $\pm$ 1.07	583.0 $\pm$ 9.77	149.14 $\pm$ 11.34
	No Resampling	235.10 $\pm$ 58.87	120.40 $\pm$ 31.71	173.22 $\pm$ 7.19	30.48 $\pm$ 4.51
SNR PUN	Standard Initialization	525.61 $\pm$ 25.46	400.82 $\pm$ 33.45	710.19 $\pm$ 7.77	131.47 $\pm$ 9.06
	No Resampling	303.45 $\pm$ 30.18	146.43 $\pm$ 2.02	244.19 $\pm$ 29.12	31.46 $\pm$ 3.11
	Shared Critic	409.29 $\pm$ 31.77	420.55 $\pm$ 33.13	497.23 $\pm$ 10.15	209.48 $\pm$ 23.46
SNR PUN	–	**670.48** $\pm$ 13.42	**421.77** $\pm$ 32.74	**750.47** $\pm$ 3.23	**310.96** $\pm$ 16.96

Table 4. Ablation studies results on using standard initialization, no parameter resampling, and shared critic function. The bottom row provides results for our full method.