

# Supplementary Materials

*The following content was not necessarily subject to peer review.*

## A Environments

All the environments are based on the *DeepMind Control Suite* (Tassa et al., 2018) and some adapted by (Touati et al., 2022).

- **Point-mass Maze:** a 2-dimensional continuous maze with four rooms. The states are 4-dimensional vectors encoding for positions and velocities of the point mass, and the actions are 2-dimensional vectors. Importantly, the initial position of the point-mass is always sampled from a uniform distribution over the spatial domain of the top-left room only. At test, we evaluate performance of agents on 20 goal-reaching tasks (5 goals in each room described by their (x,y) coordinates (see Fig. 3). This task is set as a goal-reaching task and hence we compute  $z_R$  at evaluation time by:  $z_R = B(s)$ .
- **Cheetah:** A 17 state-dimensional running planar biped consisting of positions and velocities of robot joints. Actions are 6-dimensional. We evaluate on 4 tasks: `walk`, `run`, `walk backward`, `run backward`. Rewards are linearly proportional to the achieved velocity up to the desired task velocity.
- **Walker:** A 24 state-dimensional planar walker consisting of positions and velocities of robot joints. Actions are 6-dimensional. We evaluate on 4 tasks: `stand`, `run`, `flip`. In the `stand` task reward is a combination of terms encouraging an upright torso and some minimal torso height. The `walk` and `run` task rewards include a component linearly proportional to the achieved velocity up to the desired task velocity. `flip` includes a component encouraging angular momentum.
- **Hopper:** A 15-dimensional planar one-legged hopper. Actions are 4 dimensional. We evaluate on 5 tasks: `stand`, `hop`, `flip`. In the `stand` the reward encourages a minimal torso height. In the `hop`, `hop backward` tasks the rewards have an additional term that is linearly proportional to the achieved velocity up to the desired task velocity. In the `flip`, `flip backward` includes a component encouraging angular momentum.
- **Quadruped** a four-leg spider navigating in 3D space. States and actions are 78 and 12 dimensional respectively. We evaluate on 4 tasks: `stand`, `walk`, `run`, `jump`. `stand` reward encourages an upright torso, `walk` and `run` have an additional term that is linearly proportional to the achieved velocity up to the desired task velocity. `jump` includes a term encouraging some minimal height of the center of mass.

## B Prior information on rewards

When dealing with high dimensionality environments, learning future probabilities for all states is very difficult and generally requires large  $d$  to accommodate for all possible rewards. In general, we are often interested in rewards that depend not on the full state but on a subset of it. If this is known in advance, the representation  $B$  can be trained on that part of the state only, with same theoretical guarantees (Appendix, Theorem 4 (Touati & Ollivier, 2021)). Hence, when knowing that the reward will be only a function of a subset of the state and action spaces  $G$ , we can leverage an environment-dependent feature map  $\varphi : S \times A \rightarrow G$ , and learn  $B(g)$  instead of  $B(s, a)$ , where  $g = \varphi(s, a)$ . Importantly, rewards can be arbitrary functions of  $g$ . This was also suggested in (Touati & Ollivier, 2021). In what follows, we list the feature maps that were used for the different environments.

- **Point-mass Maze:**  $\phi(s, a) = [x, y]$ .
- **Chetah:**  $\phi(s, a) = [v_x, L_y]$  where  $v_x$  is the velocity along the x-axis in the robot frame and  $L_x$  is the angular momentum about x-axis.

- **Walker:**  $\phi(s, a) = [v_x, torso_z, torso_{z_w}]$  where  $v_x$  is the horizontal velocity of the center of mass,  $torso_z$  is the height of the torso and  $torso_{z_w}$  is the projection from the z-axis of the torso to the z-axis of the world frame.
- **Hopper:**  $\phi(s, a) = [v_x, torso_{z,foot}]$  where  $v_x$  is the horizontal velocity of the center of mass and  $torso_{z,foot}$  is the height of the torso with respect to the foot.
- **Humanoid:**  $\phi(s, a) = [torso_z, v, torso_{z_w}]$  where  $torso_z$  is the height of the torso,  $v$  is the velocity of the center of mass in the local frame, and  $torso_{z_w}$  is the projection from the z-axis of the torso to the z-axis of the world frame.
- **Quadruped**  $\phi(s, a) = [v, torso_{z_w}]$  where  $v$  is the torso velocity vector in the local frame and  $torso_{z_w}$  is the projection from the z-axis of the torso to the z-axis of the world frame.

## C Hyperparameters

In Table 1 we summarize the hyperparameters used in our experiments. For a fair comparison, unless specified, we used the same parameters among all methods. Most of the parameters were adapted from (Touati et al., 2022).

Table 1: Hyperparameters.

Hyperparameter	Value
Optimizer	Adam (default hyperparameters)
Learning rate	$10^{-4}$
Batch size	256
Ratio gradient step/environment step	0.5
Z-dimension	50 (100 for maze)
Discount factor $\gamma$	0.98 (0.99 for maze)
Mix ratio for $z$ sampling	0.3
Momentum coefficient for target networks update	0.99
Number of reward labels for task inference	$10^4$
Number of ensemble members	5
Frequency of $z$ updates (training)	0.01

## D Additional experiments

### D.1 $F$ -uncertainty versus $Q$ -uncertainty exploration performance

As we have argued in Section 4,  $F^{\pi_z}$ -uncertainty and  $Q^{\pi_z}$ -uncertainty may lead to different exploration behaviors. In Fig. 5 we showed how, for a particular FB checkpoint from training on the Maze experiment, there is not a strong correlation signal ( $R^2$  score of 0.18) between the uncertainty of  $F^{\pi_z}$  to the uncertainty of  $Q^{\pi_z}$ . In the following experiment, we compare the performance of FBEE $^Q$  with a new ablation (FBEE $^F$ ), in which exploration is guided by the trace of the covariance of  $F^{\pi_z}$ . Specifically, we replace exploration as in Eq. (8) for:

$$\pi^E = \arg \max_{\pi_z} \text{tr}(\text{CoVar}[\mathbb{E}_{a \sim \pi_z(s)}[F(s, a, z)]] \quad \text{s.t.} \quad z \in \mathcal{Z}. \quad (9)$$

The results, shown in Fig. 6, indicate that these two exploration strategies lead to similar performance, suggesting that using predictive uncertainty in the  $F$ -representation to guide data collection is a viable alternative. Per task performance scores are shown in Fig. 8.

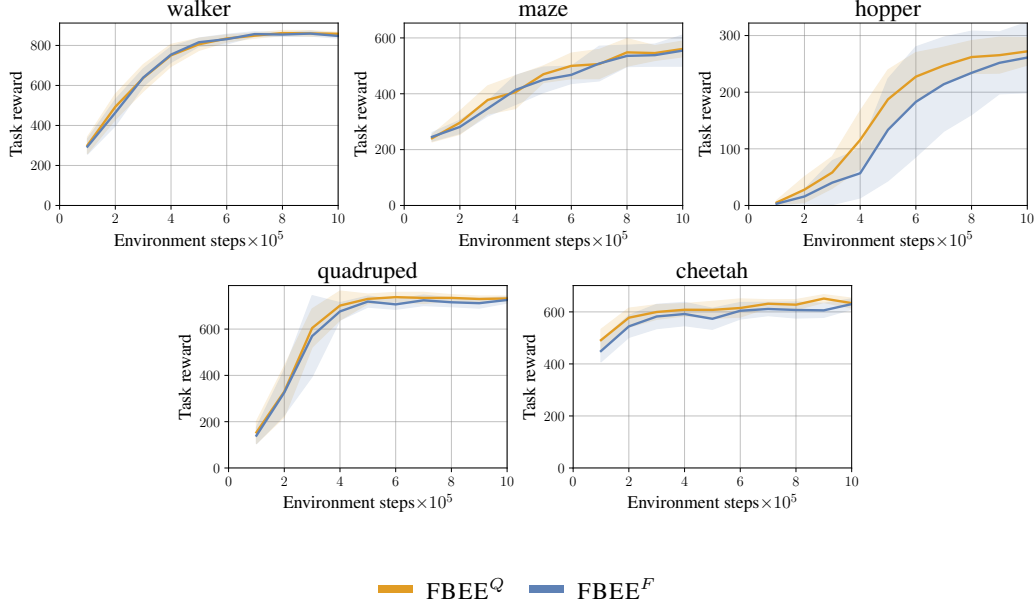


Figure 6: Scores comparison when using  $F$ -uncertainty versus  $Q$ -uncertainty exploration. Zero-shot scores averaged over different downstream task as number of environment samples increases. Metrics are averaged over 30 evaluation episodes and 10 independent random seeds. Shaded area is 1-standard deviation.

## D.2 Zero-shot scores per task

We evaluate zero-shot performance of FBEE<sup>Q</sup> on 15 tasks across 5 domains in DMC every 100k exploration steps. At evaluation time, given a task reward function  $r(s, a)$ , the agents acts with the reward representation  $z_R = \mathbb{E}[r(s, a)B(s, a)]$  for 1000 environment steps. The reward function is bounded to  $[0, 1]$ , hence maximum return per task is of 1000. In practice, we compute the expectation by taking the average over relabeled samples from the current replay buffer. Zero-shot scores across domains for all tasks is shown in Fig. 7. We also show per task performance curves of the variant FBEE<sup>F</sup> in Fig. 8.

## D.3 Other ablations

We additionally implement another ablation of our method, namely FBEE<sup>Q</sup>-POLICY which explicitly learns an exploration policy  $\pi_\theta : \mathcal{S} \rightarrow \mathcal{Z}$  by maximizing the objective in Eq. (8) through gradient descent. Results are shown in Fig. 7. In general we observe that it performs in par with FBEE<sup>Q</sup>-SAMPLING, and we attribute the mismatches in performance to not extensive hyperparameter finetuning.

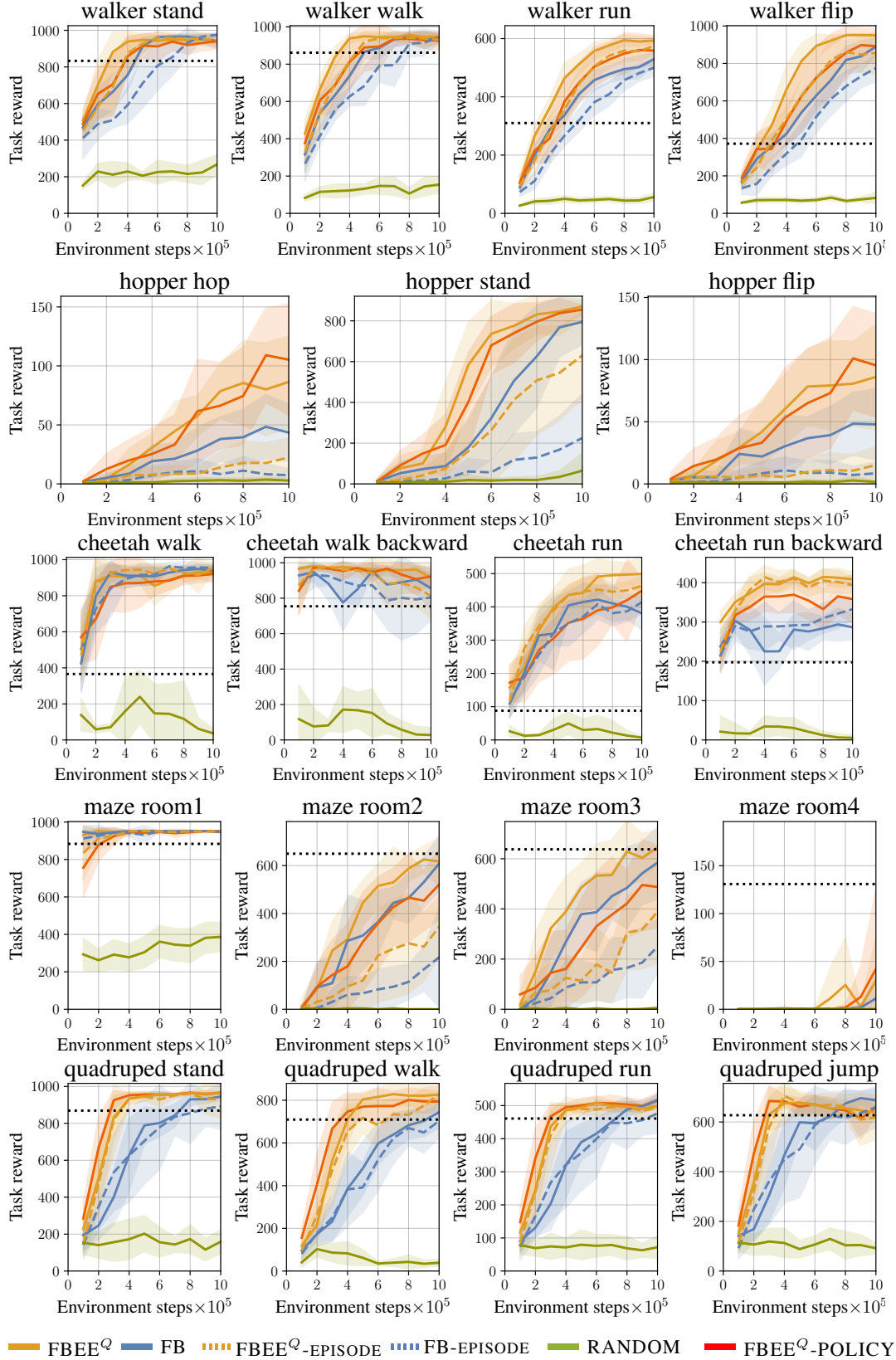


Figure 7: Zero-shot scores for different downstream task as number of environment samples increases. Metrics are averaged over 30 evaluation episodes and 10 independent random seeds. Shaded area is 1-standard deviation. Topline is maximum score of FB-RND (offline method with precollected data). Note: RND buffer for the Hopper task is not available in URLB benchmark (Laskin et al., 2021).

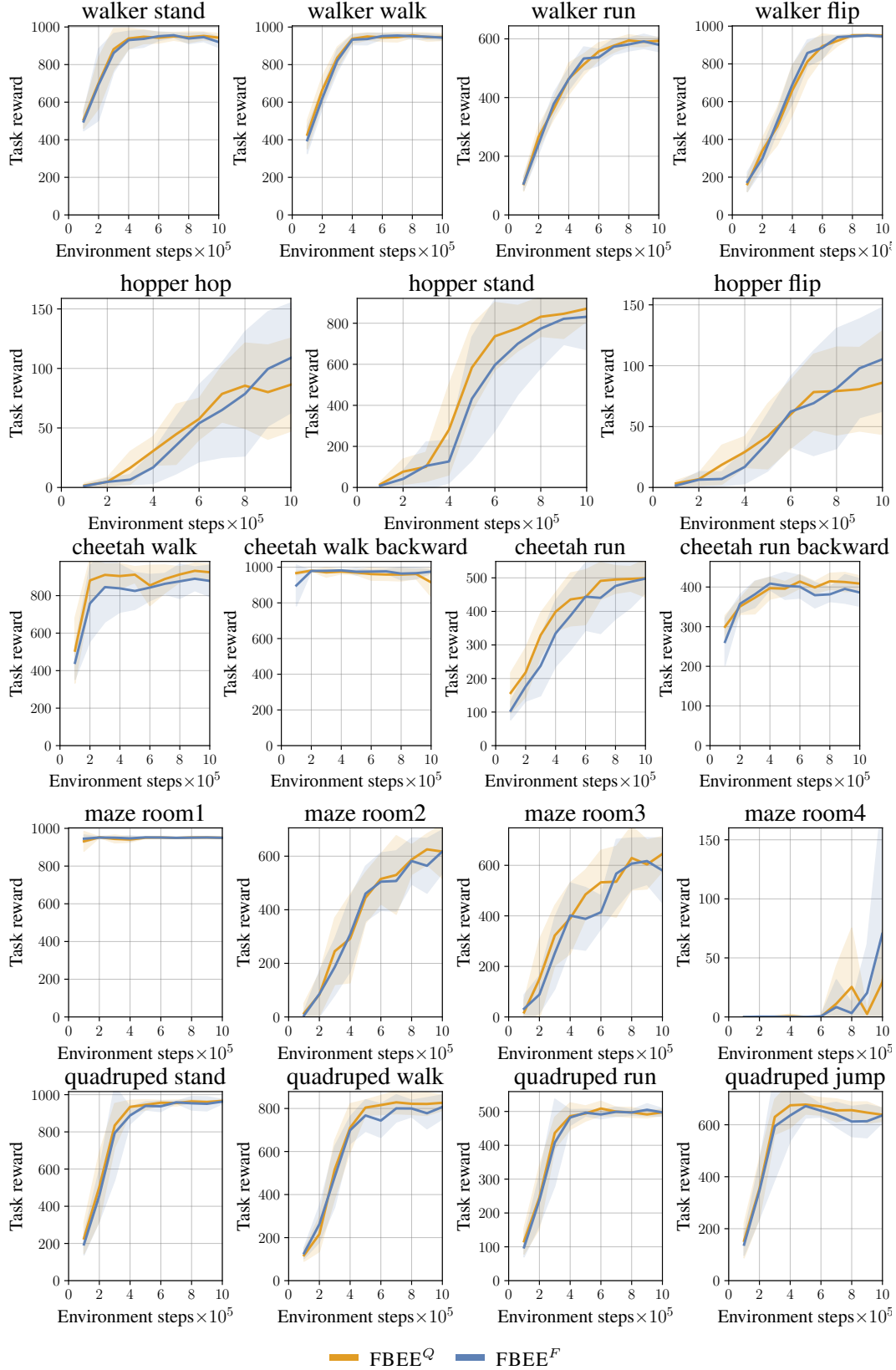


Figure 8: Zero-shot scores comparison when using  $F$ -uncertainty versus  $Q$ -uncertainty exploration for different downstream task as number of environment samples increases. Metrics are averaged over 30 evaluation episodes and 10 independent random seeds. Shaded area is 1-standard deviation.