**Preliminary results of the value function quality study:**

Figure 3 of the draft shows that, when using a value function trained in simulation as an approximate CLF, our fine-tuning method (with a discount factor of 0.0) achieves better sample efficiency for the fine-tuning process than the baseline (that keeps training with the same reward which does not have a CLF term, and keeps the same high discount factor of 0.999). For Figure 3, the value function used as a CLF candidate was obtained with 1,860,000 steps of training data of the initial simulation, which is when the training of the initial simulation using SAC had converged. The reviewer requested us to study the effects that using as CLF an intermediate value function (before convergence of the initial learning process) has on the later fine-tuning process.

In initial results, we have seen that using a poorer value function (obtained with 1,300,000 steps of training data) requires our fine-tuning method to use a larger discount factor (of 0.9 instead of 0.0) and also takes longer to train compared to the results of Figure 3 of the draft. However, even when this poor value function of 1,300,000 training steps is used, our method is able to obtain a similar performance to the baseline which has had access to 1,860,000 steps of training data, as it can be seen in the plots included in this document. These plots are equivalent to the ones of Figure 3, with the exceptions that our method now uses the poorer value function as CLF candidate, and that the trainings have only been run for 3 different random seeds instead of 10, due to the time constraints of the rebuttal process.