

## A HYPERPARAMETERS AND IMPLEMENTATION DETAILS

Hyperparameter	Robomimic	Adroit
Optimizer	Adam	
Batch Size	256	128
Learning Rate	1e-4	
Discount Factor	0.99	
Target Network Update $\tau$	0.01	0.005
MLP Hidden Dim	1024	256
MLP Hidden Layers	3	
History Steps	3	1
$Q$ -Ensemble Size	5	10
UTD Ratio	5	20
Agent Update Interval	2	1

Table 1: **Hyperparameters for RLPD** including both the RLPD baseline and the RLPD backbone of DGN.

Hyperparameters for the base RLPD algorithm (used for both the RLPD baseline and DGN) can be found in Table 1. In all Robomimic tasks, observations are stacked with three steps of history included. Each training run presented is with three seeds.

We present hyperparameters for the DGN learned covariance matrix training and usage in Table 2. The “epochs per update” hyperparameter is the number of epochs for which the DGN learned covariance matrix is trained per DGN update.

For the IBRL baselines, we use the same hyperparameters as in the original IBRL paper (Hu et al., 2023) for the state-based Robomimic tasks. In particular we use dropout of 0.5 for the actor and use the “Soft-IBRL” variant with  $\beta = 10$ . We trained IL policies for IBRL without dropout for ten epochs (for `square`, `can`, and `lift`) and twenty epochs for `tool hang`, choosing the checkpoint with the best evaluation score. The “Underfit BC” policy was trained for 100 steps.

For the IQL baseline, we trained each policy offline for 25,000 steps using the demonstration dataset and then began online fine-tuning. Further hyperparameters for IQL can be found in Table 3. Task configuration parameters are presented in Table 4.

Hyperparameter	Robomimic	Adroit
DGN Update Interval ( $N$ )	1000	2000
Optimizer	AdamW	
MLP Hidden Layers	2	
Dropout	0.5	
Batch Size	128	
MLP Hidden Size	128	256
Weight Decay	3e-2	
Epochs Per Update	2	10
Annealing Timescale ( $\tau$ )	...	30000
Shutoff Success Rate Threshold ( $m$ )	0.5	...
Epochs to Measure Success Rate for Shutoff ( $n$ )	10	...

Table 2: **DGN Hyperparameters.** Hyperparameters for training and using the DGN learned covariance matrix.

We use the Robomimic MH datasets for IBRL multimodal dataset comparison. We use a subset of the MH dataset labeled “worse”, which are successful demonstrations collected by inexperienced operators to incorporate additional diversity. Note that even though it is labeled “worse,” all of the demonstrations are successful.

Hyperparameter	Robomimic	Adroit
Optimizer	Adam	
Batch Size	256	
Learning Rate	1e-4	
Discount Factor	0.99	
Target Network Update ( $\tau$ )	0.01	0.005
History Steps	3	1
MLP Hidden Dim	1024	256
MLP Hidden Layers	3	
$Q$ -Ensemble Size	5	10
UTD Ratio	1	
Agent Update Interval	1	
Expectile	0.8	
Temperature ( $\beta$ )	3	

Table 3: **Hyperparameters for IQL Baseline** for Robomimic and Adroit tasks.

	Robomimic				Adroit		
	Lift	Can	Square	Tool Hang	Pen	Door	Relocate
Time Horizon	200	200	300	900	100	200	200
Num Human Data	100	50	50	200	25	25	25
Warm-Up Episodes	20	40	50	50	0	0	0

Table 4: **Task-Specific Parameters**, including the maximum number of steps in an episode for each task, the number of demonstrations used in each task (except on the ablations over number of demos), and the number of episodes of warm-up before online RL begins. Note for the Adroit tasks, we also use rollouts from a IL policy for the prior data.

## B ADDITIONAL EXPERIMENTS

In this section, we include additional analyses of DGN using different amounts of prior data and different hidden network sizes for training the guidance noise.

**Ablation over number of demonstrations:** We ablate over the number of trajectories in the offline dataset. As shown in Figure 10, the performance of DGN improves as number of demonstrations increases, across both the *Lift* and *Square* tasks. This is intuitive: with more expert trajectories, we can better learn the variance via imitation to provide the implicit imitation signals. Nevertheless, with only 25 demonstrations, DGN already outperforms RLPD by a significant margin in both environments, highlighting the effectiveness of DGN even with limited data.

**Ablation over MLP size:** We present the results for an ablation over the size of the state-dependent covariance MLP in Figure 11. We find that DGN substantially outperforms RLPD for all tested model sizes on the *Lift* and *Square*. However, we find that using too large an MLP can slightly hurt performance, possibly because the largest DGN models are overfitting, leading to slightly worse exploration noise guidance.

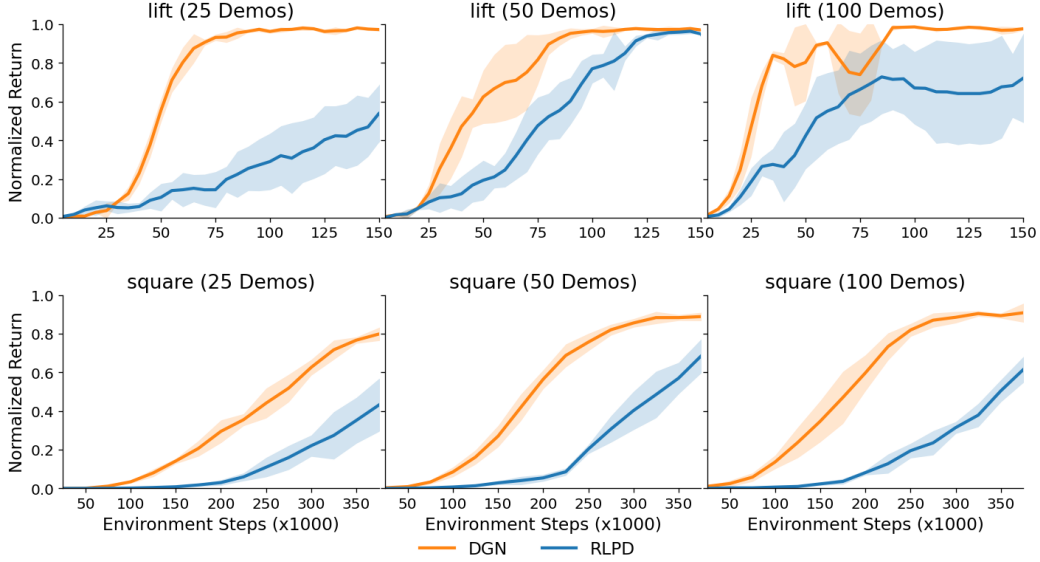


Figure 10: **Ablation over Number of Demos.** We evaluate how changing the number of demonstrations changes the performance of DGN and vanilla RLPD on the `Lift` and `Square` tasks. Adding more demos generally increases the performance of both DGN and RLPD, with DGN matching or outperforming the performance of RLPD in each case.

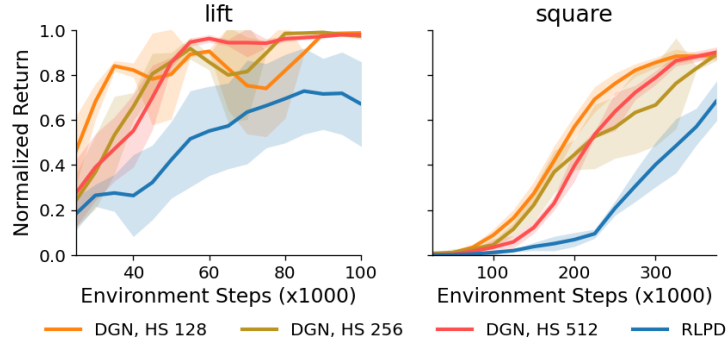


Figure 11: **Ablation over the hidden size** of the MLP of the state-dependent covariance model for the `Lift` and `Square` tasks. The performance on both tasks is strong across all tested model sizes, though slightly worse for larger MLP sizes, possibly indicating overfitting.