

A APPENDIX

A.1 EXPERIMENTS WITH NO DT PRETRAINING

In the following set of experiments, we pretrain only the IDM component of ALPT and not the DT. We show the finetuning performance results for the Narrow set of Atari games in Figure 6

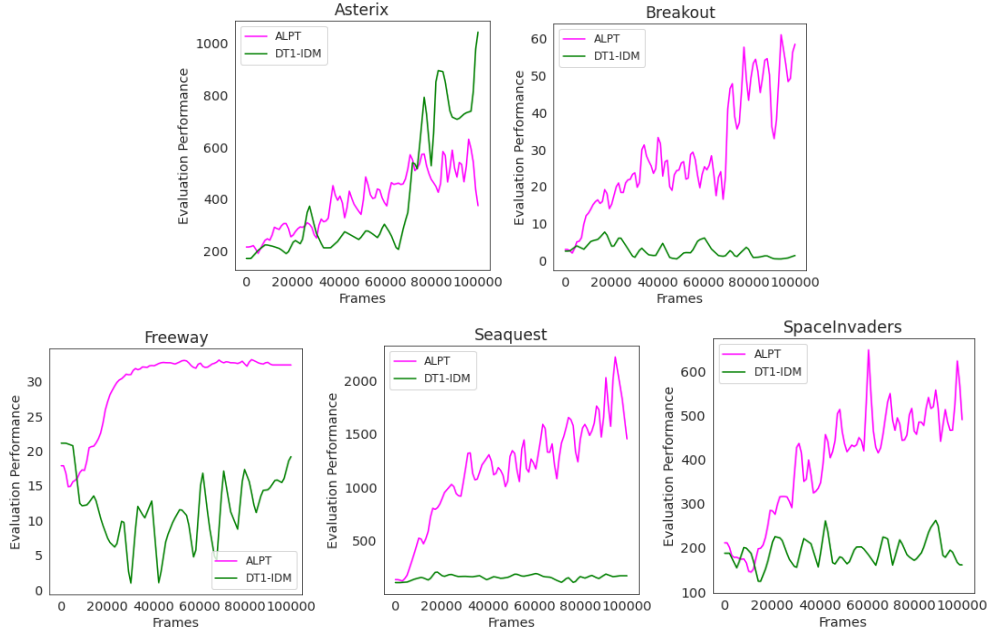


Figure 6: Evaluation game performance during finetuning of ALPT and DT1-IDM. In these experiments we do not pretrain the DT.

A.2 EXPERIMENTS WITH MORE SOURCE AND TARGET GAMES

In this set of experiments, we expand the set of source and target games. As in the previous experiments, each source game provides a fully-labelled dataset of size 100M, while each target game has a dataset of size 100M with only 10k action labels. We evaluate performance using 5 target games ($\{ \text{'Pong'}, \text{'SpaceInvaders'}, \text{'StarGunner'}, \text{'MsPacman'}, \text{'Alien'} \}$) and using either 36 source games (ALPT-36, trained on all 41 game datasets available in Agarwal et al. (2020b) minus the 5 target games) and using 9 source games (ALPT-9, trained on $\{ \text{'Asterix'}, \text{'Breakout'}, \text{'Freeway'}, \text{'Seaquest'}, \text{'Atlantis'}, \text{'DemonAttack'}, \text{'Frostbite'}, \text{'Gopher'}, \text{'TimePilot'} \}$). Note that for each of these ALPT variants, we perform a *single* pretraining phase and then *multiple* finetuning phases (one for each target game), thus showing that a single pretrained model can be transferred to various target games.

We compare pretraining with ALPT to training DT1-IDM on each target game alone. Results are presented in Figure 7 and show that target game performance improves with more source games.

A.3 IMPLEMENTATION DETAILS

In Table 3 we give the implementation details of our IDM and DT transformer architectures.

The IDM model is the same as the DT model, except that it is non-causal. This is enforced by changing the attention mask to a matrix of all 1 values in the IDM.

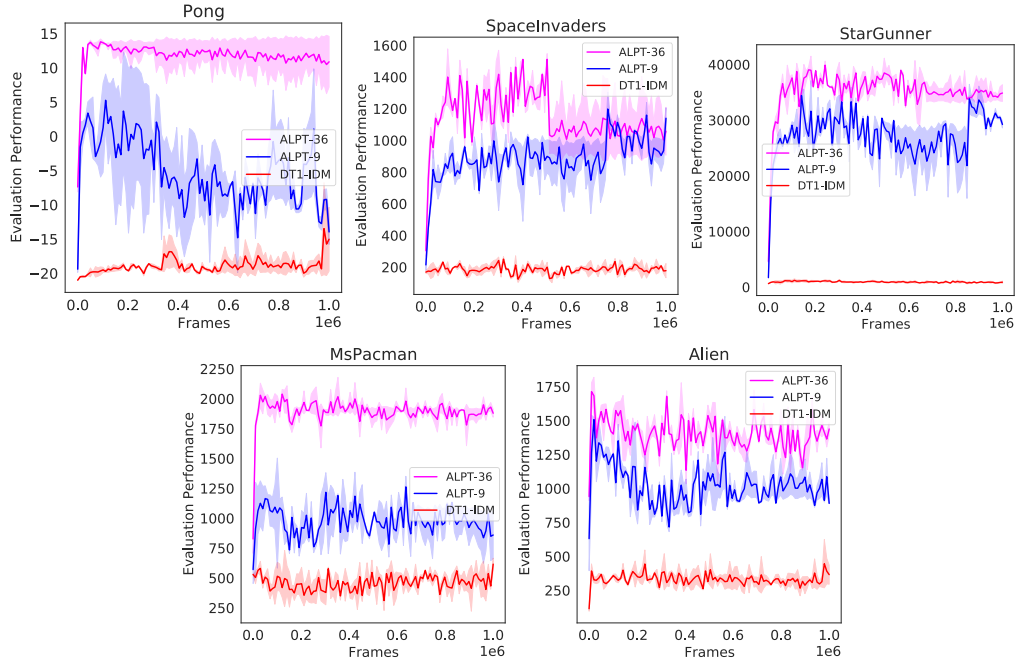


Figure 7: Game performance across ALE environments for the DT1-IDM baseline and ALPT methods. This figure shows the evaluation game performance of our DT policies during finetuning on the limited action target dataset. We show the performance across the different pretraining regimes including ALPT trained on 36 Atari source games (ALPT-36) and ALPT trained on 9 Atari source games (ALPT-9). Higher score is better. The shaded area represents the standard deviation over 3 random seeds. The x -axis shows the number of game frames used during finetuning. We evaluate ALPT on 16 episodes of length 2500 each following (Lee et al., 2022).

Table 3: A summary of the transformer model parameters.

Parameter	Value
Layers	6
Hidden Size	512
Heads	8
Batch Size	256
Weight Decay	5×10^{-5}
Learning Rate	3×10^4
Gradient Clipping	1.0
β_1, β_2	0.9, 0.999
Warm-up Steps	4000
Optimizer	LAMB

Table 4: The final evaluation game performance after training CQL for 100 iterations on a dataset of 10,000 labelled frames from each Atari game.

Game Name	Final Performance
Asterix	227.5
Breakout	12.3
Freeway	10.2
Seaquest	236.0
SpaceInvaders	250.9

A.4 EXPERIMENTS WITH CONSERVATIVE Q-LEARNING (CQL)

In this set of experiments, we examine the performance of Conservative Q-Learning (CQL) (Kumar et al., 2020) trained on a dataset of 10,000 frames, as opposed to 500,000 in the original work (Table 3 of CQL, 1% dataset size), from various Atari games utilized in our experiments. In Table 4 we report the final evaluation performance on the game after training for 100 iterations. All implementation details are consistent with the original implementation in the cited work. We utilize the CQL(\mathcal{H}) method.