Appendix

1 Experimental Details

1.1 Description of Dataset

The DQN Replay Dataset[1] consists of data collected from training DQN agents on 60 Atari 2600 games for a total of 200 million frames per game. The agents were trained while using "sticky actions", i.e., where there is a probability of 25% that the agent will repeat its previous action instead of the current action, making the task more challenging and difficult. For each game, 5 DQN agents were trained with different random initializations, and all of the (state, action, reward, next state) tuples that were encountered during training were recorded in 5 replay datasets, resulting in a total of 300 datasets with approximately 50 million tuples each due to frame skipping. The tuples in the DQN replay dataset are ordered according to the order in which they were experienced by the online DQN during training, so different strategies for collecting data for offline RL benchmarking can be tested by subsampling the replay dataset. For example, the first k million frames of the dataset may represent exploration data with suboptimal returns, while the last k million frames may represent data that is analogous to near-expert data with stochasticity.
Like DT, we subsample the entire dataset randomly to generate smaller offline training sets with varying trajectory quality. Additionally, we focus solely on the replay dataset of the first DQN agent.

1.2 Decision Transformer Baseline

We made several modifications to the baseline DT code:

- Converted the code to PyTorch Lightning framework.
- Identified and addressed state representation differences between the DQN-replay dataset and the Arcade Learning Environment setting.

Our modified version of DT showed improved results in certain games compared to the original DT paper, please refer to Appendix Section 3 for more details. These differences can be attributed to three factors. Firstly, we implemented a more rigorous and extensive evaluation procedure. Specifically, we averaged the agent's performance over 10 different seeds, with each evaluation consisting of 10 rollout episode trials. Secondly, we made adjustments to handle the inconsistencies in state representation. Thirdly, there are differences in the hardware.

1.3 Training Resources

We utilize the NVIDIA RTX2080 GPU for training each model, with an average training time of 4-8 hours. Considering the need to train each environment 10 times with different seeds and different values for the weighting term λ_F, the total training time is proportionally increased. Furthermore, the evaluation process for each model typically requires 1-3 hours to complete all 6 setups.

1.4 More Training Details for FDT experiments

This section summarizes what architecture and other design hyperparameters we have used for FDT:

- Our FDT for Atari games is based on Decision Transformer[2], which is based on minGPT [3].
- As with DT, we use the same DQN encoder for the observations, which is based on [1] with an additional linear layer to project to the embedding dimension. In addition, DT's use of Tanh was retained instead of LayerNorm after embedding each modality, as explained in [2].
- For return-to-go conditioning in our main experiment, FDT is evaluated with various values to assess statistical significance in comparison with DT, such as values within the dataset's possible returns, and values that exceed the dataset's maximum episode returns to evaluate extrapolation.

[1] https://research.google/tools/datasets/dqn-replay/
[2] https://github.com/kzl/decision-transformer
[3] https://github.com/karpathy/minGPT

- All the experiments use the Adam optimizer with an initial learning rate of $6 * 10^{-4}$, and we report it betas as well as the weight decay parameters in Table 1.
- We employ a grid search to determine the best weighting parameters and examine various numbers of feedback as can be seen in Table 3.

1.5 More Evaluating Details for FDT experiments

Using the Atari 2600 Gym environment [3] with pre-processing performed as in [4], yields that our observations are 84×84 grayscale images and each state is stacked with 4 frames. Also, the sticky-actions setting is disabled throughout the evaluation, as in DT, to reduce variability between trials.

2 Hyper-parameters

Tables 1, 2, 3 provide a comprehensive list of hyper-parameters for our proposed FDT method applied to Atari environments. To ensure a fair comparison, we adhere to the default hyper-parameter settings of Decision-Transformer, including the number of Transformer layers, multi-head self-attention heads, embedding dimensions, as well as the learning rate and optimizer configurations, among other relevant parameters. The only exception is Pong, which uses a batch size of 256.
Moreover, FDT has additional hyperparameters that control the selection of states for feedback, loss function, etc. The full list of hyperparameters are presented in Tables 2, 3.

3 Detailed Results

This section contains the complete results of our experiments and analysis, some of which we had to present in a condensed form in our study due to space constraints and to improve readability.

In the main evaluation that compares our proposed approach, FDT and the Decision Transformer over four Atari games, we conducted 10 conditional Return-to-Go (RTG) evaluations. This comprehensive evaluation approach aimed to thoroughly examine and assess the consistency, robustness, and superiority of the performance. The detailed results of our experiments including all RTGs are presented in Tables 6, 7, 8, 9. Moreover, a t-paired statistical hypothesis test was used to determine whether our proposed model has a significant edge over the baseline model; we display the results in tables per setup with asterisks that indicate significant P-values if they fall below 0.05.

Furthermore, in Table 4, we report a comparison of FDT with Decision Transformer, Behavior-Cloning, and CQL on four Atari games. Performance is evaluated with three different seeds, following the evaluation setup from the Decision Transformer paper. In addition, Table 5 presents a comprehensive comparison involving FDT, Decision Transformer and Behavior-Cloning. This analysis is based on our code and performance scores are based on our evaluation protocol as outlined in the paper under Section IV-A. Notably, the agent's performance is averaged over 10 different seeds, with each evaluation consisting of 10 rollout episode trials.

4 More Feedback and Important State Details and Analysis

Table 10 provides a comprehensive overview of the feedback and important state details for each game, considering all the setups used in our experiments. The calculations are based on an average of 10 seeds. For positive feedback samples, we report the total count and the percentage of those samples that were used as the recommended action for negative feedback samples. For negative feedback samples, we report the total count and present the percentage of ignored samples. Additionally, we include the average number of cells and the average number of candidate states within each cell. In general, these values offer insights into the distribution and density of the feedback data.

5 Oracle Details

In this work, we utilize a pre-trained Categorical DQN C51 agent [5] based on [6] as an oracle and operate in discrete action environments to provide high-quality feedback. The model's training runs and performance can be seen in Figure 1.

Table 1: Hyperparameters of FDT for Atari experiments.

Hyperparameter	Value
Number of layers	6
Number of attention heads	8
Embedding dimension	128
Batch size	256 Pong 128 Breakout, Qbert, Seaquest
Context length K	50 Pong 30 Breakout, Qbert, Seaquest
Return-to-go conditioning	[60, 70, 80, 90, 100, 125, 150, 175, 200, 225] Breakout [2500, 5000, 7500, 10000, 12500, 14000, 16500, 19000, 21500, 23000] Qbert [15, 16, 17, 18, 19, 20, 21, 22, 23, 24] Pong [550, 850, 1150, 1450, 1750, 2050, 2350, 2650, 2950, 3250] Seaquest
Nonlinearity	ReLU, encoder GeLU, otherwise
Encoder channels	32, 64, 64
Encoder filter sizes	$8 \times 8, 4 \times 4, 3 \times 3$
Encoder strides	4, 2, 1
Max epochs	5
Dropout	0.1
Learning rate	$6 * 10^{-}4$
Optimizer	Adam
Adam betas	(0.9, 0.95)
Grad norm clip	1.0
Weight decay	0.1
Learning rate decay	Linear warmup and cosine decay (see code for details)
Warmup tokens	$512 * 20$
Final tokens	$2 * 500000 * K$

Table 2: Hyperparameters of the state cell generation and selection for feedback phases.

Hyperparameter	Value
Number of transitions	50M
First-visit	true
Gamma	[0.9, 0.95, 0.99]
Kernel size	7
Stride	7
States in the same cell	Highest and lowest RTG state-action pairs / All pairs in selected cell
Minimum required RTG for cell-action pairs	0
Life loss negative reward	[0, -14] Breakout [0, -5] Pong [0, -500] Qbert [0, -200] Seaquest
Window steps ahead for life loss	10 Breakout 12 Pong 20 Qbert 15 Seaquest
Augment only sparse reward with feedback	true

Table 3: FDT hyperparameters search space.

Hyperparameter	Value
Feedback regularization lambda	[0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001]
Feedback number	[2500, 5000, 10000]
Epochs	2-5

Table 4: A comparison of FDT with Decision Transformer, Behavior-Cloning, and CQL on four Atari games. We present the results of a single return-to-go (the one reported in [2]). We set $\gamma = 0.9$, and do not force a minimum RTG requirement for cell-action pairs. The sampling method is the highest-lowest state-action pair for each cell. Performance is evaluated with three different seeds, following the evaluation setup from the Decision Transformer paper. Mean and standard deviation are reported. Ref. results column indicates sampling with No-Op policy setting, as reported in [2]. Best mean scores are highlighted in bold. Red and green indicate decreased or increased performance, respectively. The results clearly demonstrate that FDT consistently exhibits superior performance compared to DT across all evaluated configurations, with the exception of Seaquest in two setups where No-Op is not used. This finding aligns with our main results. Regarding the Ref. results, the outcomes were mixed. In Breakout, FDT outperformed the baselines by a significant margin. In Qbert, FDT is outperformed by CQL but still achieves better scores than DT. In Seaquest, FDT outperformed all baselines by a large margin and achieved comparable results to DT. Lastly, in Pong, we observed a significant performance gap compared to CQL.

Game	Model	Target RTG	Det.	Det.(+No-Op)	Sampling	Sampling(+No-Op)	Eps.	Eps.(+No-Op)	Ref. [2]
Breakout	FDT	90	**254.67 ± 0.0**	**145.73 ± 28.88**	**100.27 ± 21.28**	**101.0 ± 23.2**	**215.1 ± 26.47**	**152.93 ± 33.25**	**100.27 ± 21.28**
	DT		144.0 ± 0.0	137.1 ± 24.99	89.93 ± 21.99	93.6 ± 22.58	136.63 ± 12.36	131.93 ± 30.62	76.9 ± 27.3
	CQL		-	-	-	-	-	-	61.1 ± 0
	BC		-	-	-	-	-	-	40.9 ± 17.3
	$\Delta_{\%}$		(+77)	(+6)	(+11)	(+8)	(+57)	(+16)	(+30)
Qbert	FDT	14000	**15133.33 ± 0.0**	**13697.73 ± 670.65**	**12588.87 ± 1241.24**	**13232.23 ± 1207.93**	**14714.87 ± 313.24**	**13689.8 ± 788.95**	12588.87 ± 1241.24
	DT		14716.67 ± 0.0	12868.57 ± 1326.86	11874.0 ± 1183.97	13203.57 ± 1053.6	14220.47 ± 474.9	12911.77 ± 1103.93	2215.8 ± 1523.7
	CQL		-	-	-	-	-	-	**14012.0 ± 0**
	BC		-	-	-	-	-	-	2464.1 ± 1948.2
	$\Delta_{\%}$		(+3)	(+6)	(+6)	(0)	(+3)	(+6)	(-11)
Pong	FDT	20	**12.67 ± 0.0**	**15.57 ± 0.86**	**11.2 ± 1.1**	**15.17 ± 1.3**	**11.17 ± 0.78**	**15.7 ± 1.37**	11.2 ± 1.1
	DT		12.33 ± 0.0	12.83 ± 1.62	7.8 ± 1.39	12.07 ± 2.04	9.33 ± 1.33	13.7 ± 2.0	17.1 ± 2.9
	CQL		-	-	-	-	-	-	**19.3 ± 0**
	BC		-	-	-	-	-	-	9.7 ± 7.2
	$\Delta_{\%}$		(+3)	(+21)	(+44)	(+26)	(+20)	(+15)	(-72)
Seaquest	FDT	1150	416.67 ± 0.0	**471.13 ± 94.55**	**1131.33 ± 148.34**	**1078.53 ± 108.89**	405.07 ± 54.42	**454.53 ± 83.06**	**1131.33 ± 148.34**
	DT		**481.33 ± 0.0**	347.6 ± 60.31	984.2 ± 139.18	955.6 ± 116.66	**430.8 ± 68.02**	343.0 ± 60.64	1129.3 ± 189.0
	CQL		-	-	-	-	-	-	779.4 ± 0
	BC		-	-	-	-	-	-	968.6 ± 133.8
	$\Delta_{\%}$		(-16)	(+36)	(+15)	(+13)	(-6)	(+33)	(0)

Table 5: A comparison of FDT with Decision Transformer and Behavior-Cloning, on four Atari games using our modified code and based on our evaluation protocol described in Section **??**. We set $\gamma = 0.9$, and do not force a minimum RTG requirement for cell-action pairs. The sampling method is the highest-lowest state-action pair for each cell. Best mean scores are highlighted in bold. Similar to the findings displayed in Table 4, the results clearly establish that FDT maintains a consistent edge in performance compared to DT and BC across all the evaluated configurations. The only exceptions arise in Seaquest within two setups where No-Op is omitted. Furthermore, when comparing the performance of DT with BC, DT consistently outperforms BC across all the assessed configurations and games, except for Breakout where BC achieves a higher score in 4 out of the 6 configurations.

Game	Model	Target RTG	Det.	Det.(+No-Op)	Sampling	Sampling(+No-Op)	Eps.	Eps.(+No-Op)
Breakout	FDT	90	**174.9 ± 0.0**	**144.58 ± 22.9**	**96.25 ± 21.78**	**98.95 ± 20.3**	**160.88 ± 25.22**	**145.13 ± 28.03**
	BC		144.7 ± 0.0	132.86 ± 25.98	70.27 ± 18.24	70.37 ± 18.6	137.57 ± 22.55	137.69 ± 23.46
	DT		133.9 ± 0.0	132.84 ± 25.14	88.41 ± 18.57	90.62 ± 21.01	133.97 ± 16.98	133.45 ± 22.13
Qbert	FDT	14000	**14920.0 ± 0.0**	**13488.4 ± 721.0**	**12218.5 ± 1241.69**	**13158.66 ± 1082.52**	**14556.98 ± 547.56**	**13247.46 ± 854.17**
	BC		12420.0 ± 0.0	12237.03 ± 1225.63	9642.25 ± 1811.31	10980.39 ± 1360.54	12294.9 ± 699.54	12006.34 ± 1581.62
	DT		13557.5 ± 0.0	12368.71 ± 1151.98	11630.99 ± 1386.17	12674.26 ± 1231.5	13141.31 ± 695.7	12331.05 ± 1213.86
Pong	FDT	20	**9.6 ± 0.0**	**13.32 ± 1.6**	**6.72 ± 1.44**	**12.97 ± 1.96**	**8.66 ± 0.92**	**13.45 ± 2.03**
	BC		−0.4 ± 0.0	6.08 ± 2.6	−6.43 ± 2.26	3.59 ± 3.93	−1.63 ± 1.79	6.0 ± 2.83
	DT		9.4 ± 0.0	12.54 ± 1.67	5.27 ± 1.34	11.62 ± 1.91	7.86 ± 1.17	13.0 ± 2.06
Seaquest	FDT	1150	489.4 ± 0.0	**464.76 ± 68.28**	**1076.78 ± 120.87**	**1073.59 ± 104.14**	461.68 ± 66.18	**458.49 ± 75.54**
	BC		160.2 ± 0.0	216.54 ± 40.68	774.44 ± 137.95	770.0 ± 136.28	202.88 ± 30.52	222.86 ± 37.15
	DT		**516.0 ± 0.0**	402.18 ± 69.28	974.12 ± 126.28	966.16 ± 121.74	**473.7 ± 64.23**	405.64 ± 70.22

Table 6: A comparison of our proposed approach (FDT) and the Decision Transformer in the game of Breakout. The performance evaluation follows our specified procedure, where each score is calculated as an average across 10 seeds, each involving 10 rollout episode trials. Mean and standard deviation are reported. Best mean scores are highlighted in bold. Red and green indicate decreased or increased performance, respectively. Purple indicates FDT achieves nearly the same performance as DT. The results demonstrate that FDT exhibits superior performance compared to DT on all of the evaluated configurations.

Game	Model	Target RTG	*Det.	*Det.(+No-Op)	*Sampling	*Sampling(+No-Op)	*Eps.	*Eps.(+No-Op)
Breakout	FDT	60	**155.0 ± 0.0**	**135.12 ± 22.29**	**85.7 ± 19.75**	**82.47 ± 17.28**	**142.63 ± 25.38**	**131.34 ± 25.23**
	DT		143.6 ± 0.0	126.19 ± 21.29	79.71 ± 18.41	78.49 ± 18.07	121.8 ± 19.02	124.56 ± 23.76
	$\Delta_{\%}$		(+8.0)	(+7.0)	(+8.0)	(+5.0)	(+17.0)	(+5.0)
	FDT	70	**179.1 ± 0.0**	**139.89 ± 22.54**	**87.42 ± 18.33**	**88.62 ± 21.01**	157.29 ± 20.62	**135.47 ± 23.56**
	DT		174.2 ± 0.0	132.54 ± 20.64	85.07 ± 16.73	81.13 ± 15.36	**157.48 ± 22.98**	129.93 ± 23.53
	$\Delta_{\%}$		(+3.0)	(+6.0)	(+3.0)	(+9.0)	(0.0)	(+4.0)
	FDT	80	**167.4 ± 0.0**	**146.33 ± 24.01**	**91.31 ± 20.26**	**91.12 ± 20.26**	**152.44 ± 20.76**	**140.29 ± 23.85**
	DT		130.7 ± 0.0	137.23 ± 22.87	86.36 ± 15.51	88.92 ± 16.99	132.64 ± 18.87	128.92 ± 24.83
	$\Delta_{\%}$		(+28.0)	(+7.0)	(+6.0)	(+2.0)	(+15.0)	(+9.0)
	FDT	90	**174.9 ± 0.0**	**144.58 ± 22.9**	**96.25 ± 21.78**	**98.95 ± 20.3**	**160.88 ± 25.22**	**145.13 ± 28.03**
	DT		133.9 ± 0.0	132.84 ± 25.14	88.41 ± 18.57	90.62 ± 21.01	133.97 ± 16.98	133.45 ± 22.13
	$\Delta_{\%}$		(+31.0)	(+9.0)	(+9.0)	(+9.0)	(+20.0)	(+9.0)
	FDT	100	**176.0 ± 0.0**	**152.32 ± 26.19**	**98.97 ± 18.89**	**96.51 ± 20.11**	**157.13 ± 22.86**	**144.13 ± 23.2**
	DT		134.5 ± 0.0	134.84 ± 25.05	92.08 ± 19.9	96.18 ± 21.96	136.33 ± 17.83	138.1 ± 24.17
	$\Delta_{\%}$		(+31.0)	(+13.0)	(+7.0)	(0.0)	(+15.0)	(+4.0)
	FDT	125	**198.9 ± 0.0**	**162.68 ± 25.2**	**105.36 ± 20.69**	**104.61 ± 21.58**	**174.12 ± 25.94**	**157.63 ± 23.12**
	DT		140.0 ± 0.0	151.81 ± 26.97	95.43 ± 19.97	97.51 ± 19.46	143.9 ± 17.92	150.45 ± 29.76
	$\Delta_{\%}$		(+42.0)	(+7.0)	(+10.0)	(+7.0)	(+21.0)	(+5.0)
	FDT	150	**177.1 ± 0.0**	**169.97 ± 24.88**	**104.58 ± 21.63**	**104.93 ± 23.27**	**162.68 ± 19.73**	**163.71 ± 29.82**
	DT		164.4 ± 0.0	156.78 ± 22.8	101.77 ± 23.3	97.54 ± 18.61	150.52 ± 20.86	150.67 ± 27.35
	$\Delta_{\%}$		(+8.0)	(+8.0)	(+3.0)	(+8.0)	(+8.0)	(+9.0)
	FDT	175	**179.4 ± 0.0**	**171.85 ± 24.42**	**109.09 ± 22.81**	**107.22 ± 24.6**	**176.39 ± 20.07**	**162.79 ± 29.43**
	DT		151.7 ± 0.0	158.51 ± 26.43	101.8 ± 22.2	98.74 ± 23.36	151.54 ± 19.71	157.74 ± 26.29
	$\Delta_{\%}$		(+18.0)	(+8.0)	(+7.0)	(+9.0)	(+16.0)	(+3.0)
	FDT	200	**174.1 ± 0.0**	**170.05 ± 26.08**	**106.63 ± 22.64**	**108.65 ± 21.31**	**167.16 ± 17.18**	**169.6 ± 28.89**
	DT		165.0 ± 0.0	161.3 ± 27.36	102.23 ± 24.37	102.42 ± 22.32	155.05 ± 19.9	156.97 ± 28.25
	$\Delta_{\%}$		(+6.0)	(+5.0)	(+4.0)	(+6.0)	(+8.0)	(+8.0)
	FDT	225	**209.73 ± 0.0**	**170.72 ± 26.88**	**109.42 ± 24.31**	**106.92 ± 22.38**	**185.54 ± 22.28**	**171.34 ± 30.8**
	DT		171.6 ± 0.0	164.06 ± 28.39	99.6 ± 22.09	101.62 ± 24.93	158.0 ± 23.5	157.15 ± 26.91
	$\Delta_{\%}$		(+22.0)	(+4.0)	(+10.0)	(+5.0)	(+17.0)	(+9.0)
	FDT	avg	**179.16**	**156.35**	**99.47**	**99.0**	**163.63**	**152.14**
	DT		150.96	145.61	93.25	93.32	144.12	142.79
	$\Delta_{\%}$		(+19.0)	(+7.0)	(+7.0)	(+6.0)	(+14.0)	(+7.0)

Note: Asterisks denote P-values in a paired t-test that are significant at or less than 0.05.

Table 7: A comparison of our proposed approach (FDT) and the Decision Transformer in the game of Qbert. The performance evaluation follows our specified procedure, where each score is calculated as an average across 10 seeds, each involving 10 rollout episode trials. Mean and standard deviation are reported. Best mean scores are highlighted in bold. Red and green indicate decreased or increased performance, respectively. Purple indicates FDT achieves nearly the same performance as DT. The results clearly indicate that FDT outperforms DT in all of the evaluated configurations.

Game	Model	Target RTG	*Det.	*Det.(+No-Op)	*Sampling	*Sampling(+No-Op)	*Eps.	*Eps.(+No-Op)
Qbert	FDT	2500	**9140.0 ± 0.0**	**9002.2 ± 881.17**	**8044.26 ± 986.13**	**8650.6 ± 1120.29**	**9093.01 ± 508.21**	**8848.75 ± 927.15**
	DT		7757.5 ± 0.0	8092.74 ± 889.4	7615.24 ± 1274.62	8344.66 ± 1072.04	7846.28 ± 499.93	7895.06 ± 940.79
	$\Delta_{\%}$		(+18.0)	(+11.0)	(+6.0)	(+4.0)	(+16.0)	(+12.0)
	FDT	5000	**10602.5 ± 0.0**	**10214.9 ± 705.03**	**9274.22 ± 1024.04**	**9961.56 ± 919.16**	**10150.82 ± 596.29**	**9975.98 ± 907.6**
	DT		9735.0 ± 0.0	9131.78 ± 763.8	8912.75 ± 1179.42	9784.87 ± 878.33	9526.45 ± 569.86	9095.46 ± 951.15
	$\Delta_{\%}$		(+9.0)	(+12.0)	(+4.0)	(+2.0)	(+7.0)	(+10.0)
	FDT	7500	**11360.0 ± 0.0**	**10620.02 ± 627.64**	**9761.87 ± 1005.31**	**10422.9 ± 885.0**	**11069.25 ± 522.38**	**10528.19 ± 876.92**
	DT		10635.0 ± 0.0	9668.62 ± 799.6	9501.55 ± 1194.02	10186.08 ± 913.88	10139.64 ± 575.18	9817.32 ± 829.63
	$\Delta_{\%}$		(+7.0)	(+10.0)	(+3.0)	(+2.0)	(+9.0)	(+7.0)
	FDT	10000	**13165.0 ± 0.0**	**12346.24 ± 634.85**	**11156.56 ± 1126.08**	**11933.35 ± 1100.11**	**13036.47 ± 518.27**	**12098.45 ± 921.4**
	DT		11495.0 ± 0.0	11027.97 ± 956.54	10454.51 ± 1139.2	11449.71 ± 946.33	11461.92 ± 630.59	11105.58 ± 1004.84
	$\Delta_{\%}$		(+15.0)	(+12.0)	(+7.0)	(+4.0)	(+14.0)	(+9.0)
	FDT	12500	**14930.0 ± 0.0**	**13432.48 ± 717.32**	**12104.38 ± 1236.24**	**12922.44 ± 1131.81**	**14505.32 ± 537.33**	**13292.56 ± 1031.26**
	DT		13617.5 ± 0.0	12312.76 ± 1137.7	11599.75 ± 1442.67	12490.98 ± 1126.05	13250.0 ± 705.1	12328.0 ± 1369.1
	$\Delta_{\%}$		(+10.0)	(+9.0)	(+4.0)	(+3.0)	(+9.0)	(+8.0)
	FDT	14000	**14920.0 ± 0.0**	**13488.4 ± 721.0**	**12218.5 ± 1241.69**	**13158.66 ± 1082.52**	**14556.98 ± 547.56**	**13247.46 ± 854.17**
	DT		13557.5 ± 0.0	12368.71 ± 1151.98	11630.99 ± 1386.17	12674.26 ± 1231.5	13141.31 ± 695.7	12331.05 ± 1213.86
	$\Delta_{\%}$		(+10.0)	(+9.0)	(+5.0)	(+4.0)	(+11.0)	(+7.0)
	FDT	16500	**14955.0 ± 0.0**	**13631.26 ± 763.36**	**12233.05 ± 1177.71**	**13061.76 ± 1124.32**	**14420.8 ± 722.19**	**13456.24 ± 943.87**
	DT		13900.0 ± 0.0	12529.56 ± 1209.79	11707.21 ± 1573.52	12806.6 ± 1282.26	13370.65 ± 665.76	12513.9 ± 1177.3
	$\Delta_{\%}$		(+8.0)	(+9.0)	(+4.0)	(+2.0)	(+8.0)	(+8.0)
	FDT	19000	**14985.0 ± 0.0**	**13631.45 ± 764.53**	**12298.49 ± 1264.48**	**13062.26 ± 1210.03**	**14518.81 ± 620.06**	**13457.56 ± 975.58**
	DT		13942.5 ± 0.0	12483.13 ± 1213.39	11840.75 ± 1358.32	12753.79 ± 1336.2	13399.45 ± 664.66	12458.59 ± 1330.46
	$\Delta_{\%}$		(+7.0)	(+9.0)	(+4.0)	(+2.0)	(+8.0)	(+8.0)
	FDT	21500	**14960.0 ± 0.0**	**13631.2 ± 764.12**	**12204.03 ± 1334.73**	**13088.37 ± 1261.61**	**14512.3 ± 655.67**	**13466.94 ± 985.55**
	DT		13942.5 ± 0.0	12481.11 ± 1213.12	11717.97 ± 1426.34	12789.26 ± 1349.96	13407.86 ± 656.79	12456.08 ± 1200.03
	$\Delta_{\%}$		(+7.0)	(+9.0)	(+4.0)	(+2.0)	(+8.0)	(+8.0)
	FDT	23000	**14960.0 ± 0.0**	**13631.2 ± 764.12**	**12261.0 ± 1342.29**	**13068.9 ± 1250.41**	**14524.83 ± 644.62**	**13466.94 ± 985.55**
	DT		13942.5 ± 0.0	12481.11 ± 1213.12	11728.39 ± 1461.72	12802.38 ± 1326.92	13407.86 ± 656.79	12441.83 ± 1206.39
	$\Delta_{\%}$		(+7.0)	(+9.0)	(+5.0)	(+2.0)	(+8.0)	(+8.0)
	FDT	avg	**13397.75**	**12362.93**	**11155.64**	**11933.08**	**13038.86**	**12183.91**
	DT		12252.5	11257.75	10670.91	11608.26	11895.14	11244.29
	$\Delta_{\%}$		(+9.0)	(+10.0)	(+5.0)	(+3.0)	(+10.0)	(+8.0)

Note: Asterisks denote P-values in a paired t-test that are significant at or less than 0.05.

Table 8: A comparison of our proposed approach (FDT) and the Decision Transformer in the game of Pong. The performance evaluation follows our specified procedure, where each score is calculated as an average across 10 seeds, each involving 10 rollout episode trials. Mean and standard deviation are reported. Best mean scores are highlighted in bold. Red and green indicate decreased or increased performance, respectively. Purple indicates FDT achieves nearly the same performance as DT.
The results clearly indicate that FDT outperforms DT in all of the evaluated configurations for all RTGs, except in two specific cases, where we observed a minor degradation in performance. Although this decline is evident, it is important to emphasize that the magnitude of the performance drop is not significant and these instances only appear in just two specific cases out of all the evaluated configurations. Therefore, they do not significantly impact the overall superiority of FDT. The overall trend of FDT outperforming DT in the majority of evaluated configurations remains robust and consistent.

Game	Model	Target RTG	*Det.	*Det.(+No-Op)	*Sampling	*Sampling(+No-Op)	*Eps.	*Eps.(+No-Op)
Pong	FDT	15	**9.8 ± 0.0**	**13.18 ± 1.61**	**7.04 ± 1.31**	**12.37 ± 2.3**	**8.84 ± 1.1**	**13.55 ± 1.75**
	DT		6.3 ± 0.0	11.97 ± 1.78	5.04 ± 1.43	10.9 ± 1.96	6.59 ± 1.02	12.21 ± 1.97
	$\Delta_{\%}$		(+56.0)	(+10.0)	(+40.0)	(+13.0)	(+34.0)	(+11.0)
	FDT	16	**9.1 ± 0.0**	**13.19 ± 1.61**	**7.15 ± 1.23**	**12.55 ± 2.1**	**7.93 ± 1.18**	**13.65 ± 2.14**
	DT		7.1 ± 0.0	12.61 ± 1.81	5.33 ± 1.38	11.09 ± 2.06	6.69 ± 1.1	12.67 ± 2.15
	$\Delta_{\%}$		(+28.0)	(+5.0)	(+34.0)	(+13.0)	(+19.0)	(+8.0)
	FDT	17	**10.0 ± 0.0**	**13.27 ± 1.55**	**7.43 ± 1.29**	**12.81 ± 1.98**	**8.17 ± 1.03**	**13.3 ± 1.97**
	DT		6.5 ± 0.0	12.32 ± 1.57	5.33 ± 1.41	11.27 ± 1.99	5.91 ± 1.01	12.4 ± 2.2
	$\Delta_{\%}$		(+54.0)	(+8.0)	(+39.0)	(+14.0)	(+38.0)	(+7.0)
	FDT	18	**9.1 ± 0.0**	**13.5 ± 1.68**	**7.42 ± 1.33**	**12.84 ± 2.2**	**8.09 ± 1.08**	**13.34 ± 1.86**
	DT		7.5 ± 0.0	12.69 ± 2.01	5.56 ± 1.37	11.45 ± 2.0	6.31 ± 1.05	12.55 ± 1.95
	$\Delta_{\%}$		(+21.0)	(+6.0)	(+33.0)	(+12.0)	(+28.0)	(+6.0)
	FDT	19	**10.3 ± 0.0**	**13.16 ± 1.78**	**7.19 ± 1.35**	**13.03 ± 2.06**	**8.61 ± 1.47**	**13.42 ± 2.35**
	DT		5.4 ± 0.0	12.66 ± 1.83	5.4 ± 1.41	11.43 ± 2.1	6.06 ± 1.14	12.55 ± 2.06
	$\Delta_{\%}$		(+91.0)	(+4.0)	(+33.0)	(+14.0)	(+42.0)	(+7.0)
	FDT	20	**9.6 ± 0.0**	**13.32 ± 1.6**	**6.72 ± 1.44**	**12.97 ± 1.96**	**8.66 ± 0.92**	**13.45 ± 2.03**
	DT		9.4 ± 0.0	12.54 ± 1.67	5.27 ± 1.34	11.62 ± 1.91	7.86 ± 1.17	13.0 ± 2.06
	$\Delta_{\%}$		(+2.0)	(+6.0)	(+28.0)	(+12.0)	(+10.0)	(+3.0)
	FDT	21	**9.1 ± 0.0**	**13.19 ± 1.79**	**6.97 ± 1.32**	**12.58 ± 2.43**	**7.94 ± 1.06**	**13.6 ± 1.93**
	DT		8.0 ± 0.0	12.46 ± 1.83	5.5 ± 1.41	11.59 ± 2.01	6.3 ± 1.25	12.68 ± 2.12
	$\Delta_{\%}$		(+14.0)	(+6.0)	(+27.0)	(+9.0)	(+26.0)	(+7.0)
	FDT	23	7.2 ± 0.0	**12.87 ± 1.63**	**6.49 ± 1.49**	**13.02 ± 2.21**	**7.34 ± 1.07**	**13.34 ± 2.37**
	DT		**9.2 ± 0.0**	11.94 ± 1.95	5.45 ± 1.55	11.46 ± 1.94	7.13 ± 1.31	12.25 ± 2.62
	$\Delta_{\%}$		(-28.0)	(+8.0)	(+19.0)	(+14.0)	(+3.0)	(+9.0)
	FDT	24	**8.4 ± 0.0**	**12.85 ± 1.76**	**6.77 ± 1.54**	**12.53 ± 2.21**	6.95 ± 1.18	**12.96 ± 2.15**
	DT		8.2 ± 0.0	12.4 ± 1.92	5.21 ± 1.49	11.55 ± 2.09	**7.44 ± 1.22**	12.59 ± 2.45
	$\Delta_{\%}$		(+2.0)	(+4.0)	(+30.0)	(+8.0)	(-7.0)	(+3.0)
	FDT	avg	**9.18**	**13.17**	**7.02**	**12.74**	**8.06**	**13.4**
	DT		7.51	12.4	5.34	11.37	6.7	12.54
	$\Delta_{\%}$		(+22.0)	(+6.0)	(+31.0)	(+12.0)	(+20.0)	(+7.0)

Note: Asterisks denote P-values in a paired t-test that are significant at or less than 0.05.

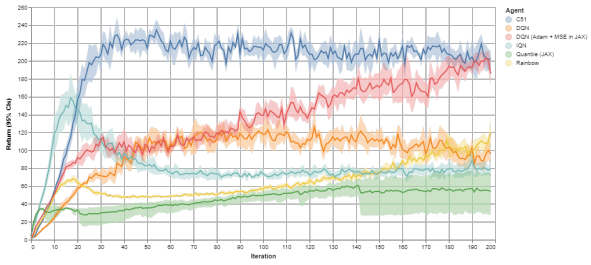
Table 9: A comparison of our proposed approach (FDT) and the Decision Transformer in the game of Seaquest. The performance evaluation follows our specified procedure, where each score is calculated as an average across 10 seeds, each involving 10 rollout episode trials. Mean and standard deviation are reported. Best mean scores are highlighted in bold. Red and green indicate decreased or increased performance, respectively. Purple indicates FDT achieves nearly the same performance as DT. Overall, FDT performed better than DT on all of the evaluated configurations except for two policies across all RTGs, where No-Op is not applied.

Game	Model	Target RTG	Det.	*Det.(+No-Op)	*Sampling	*Sampling(+No-Op)	Eps.	*Eps.(+No-Op)
Seaquest	FDT	550	374.0 ± 0.0	**419.76 ± 65.2**	**864.3 ± 100.39**	**852.16 ± 81.27**	394.74 ± 46.84	**429.0 ± 53.14**
	DT		**421.0 ± 0.0**	363.22 ± 65.9	790.0 ± 98.71	786.66 ± 76.87	**419.74 ± 54.42**	366.28 ± 65.09
	$\Delta_{\%}$		(-13.0)	(+16.0)	(+9.0)	(+8.0)	(-6.0)	(+17.0)
	FDT	850	400.2 ± 0.0	**468.42 ± 75.32**	**988.54 ± 98.74**	**979.74 ± 116.67**	406.84 ± 46.13	**461.1 ± 72.78**
	DT		**461.4 ± 0.0**	391.36 ± 69.1	899.36 ± 96.55	905.28 ± 84.95	**448.78 ± 58.2**	403.22 ± 71.04
	$\Delta_{\%}$		(-15.0)	(+20.0)	(+10.0)	(+8.0)	(-10.0)	(+14.0)
	FDT	1150	489.4 ± 0.0	**464.76 ± 68.28**	**1076.78 ± 120.87**	**1073.59 ± 104.14**	461.68 ± 66.18	**458.49 ± 75.54**
	DT		**516.0 ± 0.0**	402.18 ± 69.28	974.12 ± 126.28	966.16 ± 121.74	**473.7 ± 64.23**	405.64 ± 70.22
	$\Delta_{\%}$		(-5.0)	(+16.0)	(+11.0)	(+11.0)	(-3.0)	(+13.0)
	FDT	1450	394.0 ± 0.0	**475.42 ± 67.34**	**1112.99 ± 159.81**	**1150.54 ± 143.2**	427.54 ± 60.97	**487.95 ± 83.59**
	DT		**522.6 ± 0.0**	390.02 ± 71.5	1014.59 ± 112.57	1047.97 ± 127.23	**479.18 ± 61.68**	395.94 ± 73.54
	$\Delta_{\%}$		(-33.0)	(+22.0)	(+10.0)	(+10.0)	(-12.0)	(+23.0)
	FDT	1750	433.2 ± 0.0	**483.36 ± 71.56**	**1139.62 ± 144.21**	**1145.68 ± 138.47**	455.2 ± 67.47	**485.62 ± 84.86**
	DT		**443.8 ± 0.0**	394.72 ± 59.92	1024.42 ± 131.85	1053.9 ± 141.48	**460.46 ± 56.08**	399.86 ± 76.36
	$\Delta_{\%}$		(-2.0)	(+22.0)	(+11.0)	(+9.0)	(-1.0)	(+21.0)
	FDT	2050	410.4 ± 0.0	**490.28 ± 82.89**	**1181.66 ± 143.18**	**1166.14 ± 145.63**	430.18 ± 59.05	**478.2 ± 86.21**
	DT		**438.2 ± 0.0**	389.14 ± 69.05	1036.28 ± 146.01	1032.94 ± 132.92	**451.7 ± 60.83**	396.92 ± 67.59
	$\Delta_{\%}$		(-7.0)	(+26.0)	(+14.0)	(+13.0)	(-5.0)	(+20.0)
	FDT	2350	**412.2 ± 0.0**	**483.7 ± 69.21**	**1158.26 ± 140.94**	**1160.78 ± 181.49**	427.2 ± 62.66	**477.98 ± 81.88**
	DT		411.4 ± 0.0	384.6 ± 67.98	1068.9 ± 149.31	1087.84 ± 150.03	**436.5 ± 56.15**	393.2 ± 67.13
	$\Delta_{\%}$		(0.0)	(+26.0)	(+8.0)	(+7.0)	(-2.0)	(+22.0)
	FDT	2650	429.4 ± 0.0	**474.9 ± 76.31**	**1152.56 ± 156.13**	**1193.66 ± 179.08**	419.76 ± 44.92	**481.96 ± 86.95**
	DT		**456.4 ± 0.0**	388.28 ± 70.67	1068.74 ± 158.68	1054.52 ± 134.16	**458.96 ± 57.63**	392.5 ± 67.16
	$\Delta_{\%}$		(-6.0)	(+22.0)	(+8.0)	(+13.0)	(-9.0)	(+23.0)
	FDT	2950	436.2 ± 0.0	**468.94 ± 79.27**	**1162.5 ± 164.86**	**1181.28 ± 157.78**	429.82 ± 53.2	**466.74 ± 82.27**
	DT		**451.2 ± 0.0**	401.44 ± 70.57	1065.9 ± 147.0	1051.38 ± 145.0	**450.12 ± 51.51**	401.98 ± 72.25
	$\Delta_{\%}$		(-3.0)	(+17.0)	(+9.0)	(+12.0)	(-5.0)	(+16.0)
	FDT	3250	443.2 ± 0.0	**465.98 ± 78.43**	**1155.9 ± 158.92**	**1175.44 ± 163.55**	431.34 ± 63.12	**468.52 ± 77.92**
	DT		**490.8 ± 0.0**	394.18 ± 72.66	1046.2 ± 154.45	1054.49 ± 151.38	**450.24 ± 62.21**	397.24 ± 67.93
	$\Delta_{\%}$		(-11.0)	(+18.0)	(+10.0)	(+11.0)	(-4.0)	(+18.0)
	FDT	avg	422.22	**469.55**	**1099.31**	**1107.9**	428.43	**469.56**
	DT		**461.28**	389.91	998.85	1004.11	**452.94**	395.28
	$\Delta_{\%}$		(-9.0)	(+20.0)	(+10.0)	(+10.0)	(-6.0)	(+19.0)

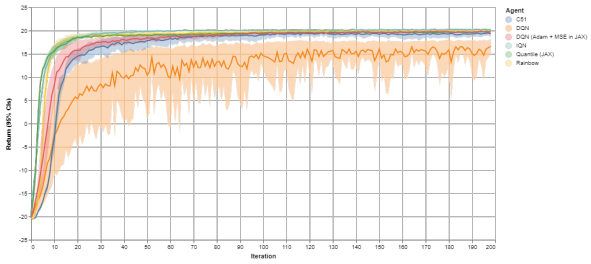
Note: Asterisks denote P-values in a paired t-test that are significant at or less than 0.05.

Table 10: An analysis of the feedback for each of our evaluated datasets and all the setups used in our experiments. The calculations are based on an average of 10 seeds. We count the number of positive feedback samples, and the percentage of them that was used as the recommended action for negative-feedback samples. For negative-feedback, we calculate the percentage of ignored samples. Finally, we calculate the average number of cells, as well as the average number of candidate states in each cell.

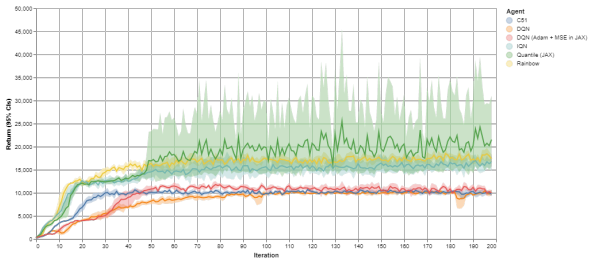
Game	**Settings**			Feedback	Positive Feedback		Negative Feedback		Important States	
	Gamma	Sampling method	Life-loss reward	Budget	Avg #	% rec-meta-action aligned	Avg #	% ignored	Avg #	Avg cells #
Breakout	0.9	default	0	5000	1860	0.63	3140	0.33	28901	199892
	0.9	default	-14	5000	1928	0.61	3072	0.33	28908	
	0.9	all-cell-pairs	-14	5000	2136	0.64	2864	0.3	42717	
	0.95	default	0	5000	1772	0.61	3228	0.33	28909	
	0.99	default	0	5000	1487	0.53	3513	0.36	28930	
	0.9	default	0	2500	944	0.63	1556	0.31	28901	
	0.9	default	-14	2500	1048	0.59	1452	0.28	28930	
	0.9	default	0	10000	3381	0.58	6619	0.33	28901	
	0.9	default	-14	10000	3416	0.57	6584	0.33	28930	
Qbert	0.9	default	0	5000	2193	0.64	2807	0.33	9515	21091
	0.9	default	-500	5000	1981	0.3	3109	0.27	9519	
	0.9	all-cell-pairs	-500	5000	2607	0.25	2393	0.17	18452	
	0.95	default	0	5000	2180	0.6	2820	0.3	9511	
	0.99	default	0	5000	2055	0.51	2945	0.34	9513	
	0.9	default	0	2500	1147	0.64	1353	0.31	9515	
	0.9	default	-500	2500	1038	0.39	1462	0.26	9519	
	0.9	default	0	10000	3510	0.56	6005	0.33	9515	
	0.9	default	-500	10000	3482	0.37	6037	0.3	9519	
Pong	0.9	default	0	5000	1477	0.48	3523	0.35	63886	92245
	0.9	default	-5	5000	1491	0.38	3509	0.32	63877	
	0.9	all-cell-pairs	-5	5000	1843	0.38	3157	0.33	160830	
	0.95	default	0	5000	1451	0.48	3549	0.34	58017	
	0.99	default	0	5000	1154	0.35	3846	0.25	64033	
	0.9	default	0	2500	807	0.55	1693	0.38	63886	
	0.9	default	-5	2500	766	0.46	1734	0.34	63877	
	0.9	default	0	10000	2773	0.42	7227	0.32	63886	
	0.9	default	-5	10000	2341	0.46	7659	0.28	63877	
Seaquest	0.9	default	0	5000	319	0.24	4681	0.26	7444	382148
	0.9	default	-200	5000	315	0.24	4685	0.26	7451	
	0.9	all-cell-pairs	-200	5000	278	0.18	4722	0.19	14461	
	0.95	default	0	5000	335	0.2	4665	0.24	7442	
	0.99	default	0	5000	345	0.2	4655	0.23	7448	
	0.9	default	0	2500	131	0.23	2369	0.26	7444	
	0.9	default	-200	2500	128	0.23	2372	0.26	7451	
	0.9	default	0	10000	521	0.26	6923	0.27	7444	
	0.9	default	-200	10000	521	0.26	6930	0.27	7451	



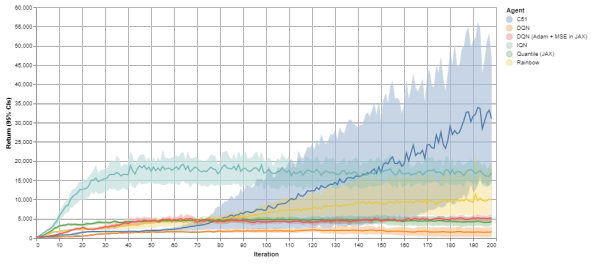
(a) Breakout



(b) Pong



(c) Qbert



(d) Seaquest

Figure 1: C51 model - visualization of the training runs and return value per game.

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[2] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," *Advances in neural information processing systems*, vol. 34, pp. 15 084–15 097, 2021.

[3] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, "The arcade learning environment: An evaluation platform for general agents," *Journal of Artificial Intelligence Research*, vol. 47, pp. 253–279, jun 2013.

[4] R. Agarwal, D. Schuurmans, and M. Norouzi, "An optimistic perspective on offline reinforcement learning," in *International Conference on Machine Learning*, 2020.

[5] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 449–458.

[6] P. S. Castro, S. Moitra, C. Gelada, S. Kumar, and M. G. Bellemare, "Dopamine: A Research Framework for Deep Reinforcement Learning," 2018. [Online]. Available: http://arxiv.org/abs/1812.06110