DECISION TRANSFORMER UNDER RANDOM FRAME DROPPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Controlling agents remotely with deep reinforcement learning (DRL) in the real world is yet to come. One crucial stepping stone is to devise RL algorithms that are robust in the face of dropped information from corrupted communication or malfunctioning sensors. Typical RL methods usually requires considerable online interaction data that are costly and unsafe to collect in the real world. Furthermore, when they are applied to the frame dropping scenarios, they perform unsatisfactorily even with moderate drop rates. To devise a robust and deployable algorithm, we propose Decision Transformer under Random Frame Dropping (DeFog), an offline RL algorithm that enables agents to act robustly in frame dropping scenarios without online interaction. DeFog first randomly masks out data in the offline datasets and explicitly adds the timespan of frame dropping as inputs. After that, a finetuning stage on the same offline dataset with a higher mask rate would further boost the performance. Empirical results show that DeFog outperforms strong baselines under severe frame drop rates like 90%, while maintaining similar returns under non-frame-dropping conditions in the regular MuJoCo control benchmarks and the Atari environments.

1 INTRODUCTION

Imagine you are guiding a drone through a forest. When you look at the images transmitted from the drone, the display delays heavily or even disappears intermittently due to bad communication. Experienced drone pilot would stabilize the drone with the last received frame until the communication recovers to normal.

In this paper, we aim to empower deep reinforcement learning (RL) algorithms with such abilities to control remote agents. In many real world control tasks, the decision makers are separate from the action executor (Saha & Dasgupta, 2018), which introduce the risk of packet loss and delay during network communication. Furthermore, sensors (e.g., camera, IMUs) are sometimes prone to temporary malfunctioning, or limited by hardward restrictions, thus causing the observation to be unavailable at certain timesteps (Dulac-Arnold et al., 2021). These examples lead to the core challenge of devising the desired algorithm: control the agents against frame dropping, i.e., a temporary loss of observations as well as other information.

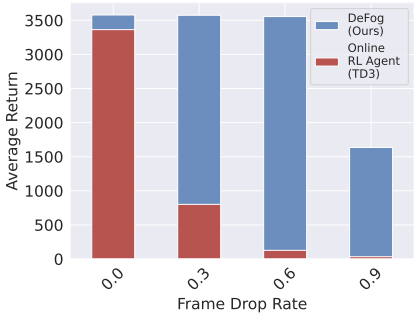

Figure 1: RL performance in the Hopper-v3 environment under different frame drop rates.

Figure 1 illustrates how a regular RL algorithm performs under high frame drop rates. We find that RL agents trained in zero-drop-rate environments encounter a harsh performance degradation when frames drop. This indicates that the frame dropping problem is a severe problem that needs to be resolved if we want to apply RL into remote control tasks.

This problem gradually attracts more attention recently: Nath et al. (2021) adapt vanilla DQN algorithm to a randomly delayed markov decision process; Bouteiller et al. (2020) propose a method that

modifies the classic Soft Actor-Critic algorithm (Haarnoja et al., 2018) to handle observation and action delay scenarios. In contrast to the delayed frames in previous works, we try to solve a more challenging problem where the frames are permanently lost. Moreover, previous methods usually learn in an online frame dropping environment, which can be unsafe and costly.

In this paper, we propose Decision Transformer under Random Frame Dropping (DeFog), an offline reinforcement learner that are robust to frame drops. Specifically, we use Decision Transformer backbones (Chen et al., 2021) to learn from a randomly masked offline datasets and then add the drop-span embeddings to the inputs. Moreover, in continuous control tasks, a finetuning stage with freezed trunk would further increase the robustness of DeFog.

We evaluate our method on continuous and discrete control tasks in MuJoCo and Atari game environments. In these environments, observations will be dropped randomly before they are sent to the agent. Empirical results show that DeFog outperforms various baselines under the frame drop setting, while maintaining performance that are comparable to other offline RL methods in regular no frame dropping environments.

The key contribution of this is that we propose Decision Transformer under Random Frame Dropping (DeFog), a method that is robust to frame drops without any online interaction. The offline nature makes it promising and practical in future real-world remote control tasks. When tested, DeFog achieves superior results in both continuous and discrete control tasks against dropped frames.

2 Related Works

2.1 Control under Frame Dropping and Delay

The observation and control loss or delay is an essential problem in remote control tasks (Balemi & Brunner, 1992; Funda & Paul, 1991). In recent cloud-edge computing systems such as intelligent connected vehicles (Li et al., 2018) and UAV swarm (Bekkouche et al., 2018; Hwangbo et al., 2017), this problem is drawing more attention.

When reinforcement learning is applied to such remote control tasks, a robust RL algorithm is desired. Katsikopoulos & Engelbrecht (2003) first propose the formulation of the Random Delayed Markov Decision Process. Along with the new formulation, a method is proposed to augment the observation space with the past actions. However, previous methods (Walsh et al., 2009; Schuitema et al., 2010) usually stack the delayed observations together, which might require a constant delay duration as a hard threshold.

Hester & Stone (2013) propose to solve this problem by predicting delayed states with a random forest model. Bouteiller et al. (2020) propose to tackle random observation and action delays in a model-free manner by relabelling the past actions with the current policy to mitigate the off-policy problem. Later, Nath et al. (2021) propose to learn a deep Q network (DQN) agent with state augmentation. However, these methods that rely on state augmentation usually assume a maximum delay time. Imai et al. (2021) train a vision-guided quadrupedal robot in the wild against random observation delay by leveraging delay randomization. This work shares the same intuition of the train-time frame masking part, while we are using a decision transformer backbone with additional frame drop interval embedding and a performance-improving finetuning technique.

2.2 Transformers in Reinforcement Learning

Researchers recently formulate the decision making procedure in offline reinforcement learning as a sequence modeling problem using transformer models (Chen et al., 2021; Janner et al., 2021). In contrast to the policy gradient and temporal difference methods, these works advocate the paradigm of treating reinforcement learning as a supervised learning problem (Schmidhuber, 2019), directly predicting actions from the observation sequence and the task specification. The Decision Transformer model (Chen et al., 2021) takes the encoded reward-to-go, state, and action sequence as input to predict the action for the next step, while the Trajectory Transformer (Janner et al., 2021) first discretizes each dimension of the input sequence, maps them to tokens, then predicts the following action's tokens with a beam search algorithm.

The concurrent occurrence of these works attracted much attention in the RL community for further improving upon the transformers. Zheng et al. (2022) increases the model capacity and enables online finetuning of the Decision Transformer by changing the deterministic policy to a stochastic one and adding an entropy term to encourage exploration. Apart from these works, various attempts have been made to improve transformers in multi-agent RL, meta RL, multi-task RL, and many other fields (Meng et al., 2021; Xu et al., 2022; Lee et al., 2022; Reid et al., 2022).

3 METHOD

In this section, we first describe the problem setup and then introduce Decision Transformer under Random Frame Dropping (DeFog), a flexible and powerful method to tackle random frame dropping.

3.1 PROBLEM STATEMENT

In the environment with random frame dropping, the original state transitions of the underlying Markov Decision Process are broken; hence, the observed states and rewards follow a new transition process. We name the new decision process Random Dropping Markov Decision Process (RDMDP). Formally, we define the RDMDP as follows:

Definition 1 (Random Dropping Markov Decision Process (RDMDP))
An RDMDP could be described as $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mu, \mathcal{P}, \mathcal{D}, \mathcal{O}_{\mathcal{S}}, \mathcal{O}_{\mathcal{R}} \rangle$, where $\mathcal{S}, \mathcal{A}$ are the state and action space, $\mathcal{P}\left(s_{t+1} \mid s_t, a_t\right)$ is the state transition possibility, $\mathcal{R}(s_t, a_t)$ is the reward function, $\mu(s_0)$ is the initial state distribution. $\mathcal{D}$ is the Bernoulli Distribution of frame dropping, $\mathcal{O}_{\mathcal{S}}$ is the function that generates the observation, and $\mathcal{O}_{\mathcal{R}}$ the function that generates the cumulative rewards. In the drop frame setting, we assume the cumulative reward $R_t = \sum_{\tau=0}^{t} r_t$ is observed instead of the immediate reward r_t.

At each timestep t, a drop frame indicator $d_t \in \{0, 1\}$ is drawn from the distribution $\mathcal{D}$, with $d = 1$ indicating that the frame is dropped and $d = 0$ the opposite. The observed state and culmulative reward $\hat{s}_t$ and $\hat{R}_t$ of the timestep are updated by

$$\hat{s}_t = \mathcal{O}_{\mathcal{S}}(s_t, \hat{s}_{t-1}, d) = d \cdot \hat{s}_{t-1} + (1 - d) \cdot s_t \tag{1}$$

$$\hat{R}_t = \mathcal{O}_{\mathcal{R}}(R_t, \hat{R}_{t-1}, d) = d \cdot \hat{R}_{t-1} + (1 - d) \cdot R_t \tag{2}$$

The observation and the cumulative reward repeats the last observed one $\hat{s}_{t-1}$ and $\hat{R}_{t-1}$ respectively if the current frame is dropped. If the current frame arrives normally, the observation and culmulative reward is updated. Note that the intermediate rewards obtained during the dropped frame are also added to the culmulative reward since the operation only happens on the remote side. Follow the definition in the Decision Transformer, when a target return R_{target} is set for the environment, we define the real and observed reward-to-go as $g_t = R_{target} - R_t$, and $\hat{g}_t = R_{target} - \hat{R}_t$ respectively.

The optimization goal of our method is to imitate the oracle policy that have access to the true state s_t in a frame-dropping environment through learning from the trajectories generated by the oracle.

3.2 DECISION TRANSFORMER UNDER RANDOM FRAME DROPPING

We choose decision transformer as our backbone because of its expressiveness and flexibility as an offline RL learner. Then, we attack the random frame dropping problem from three perspectives: 1) We modify the offline dataset by randomly masking out observations and reward-to-gos during training and dynamically update the mask ratio; 2) we provide a drop-span embedding that captures the information of dropped frame duration; 3) we find the agent can act robustly against higher dropping rate when finetuning the drop-span emcoder and action predictor after the model is fully converged. A full illustration of our method is shown in Figure 2.

3.2.1 DECISION TRANSFORMER BACKBONE

The Decision Transformer (DT) takes in past K timesteps of reward-to-goes $g_{t-K,t}$, observations $s_{t-K,t}$, and actions $a_{t-K,t}$ as embedded tokens and predicts the next tokens in the same way as the

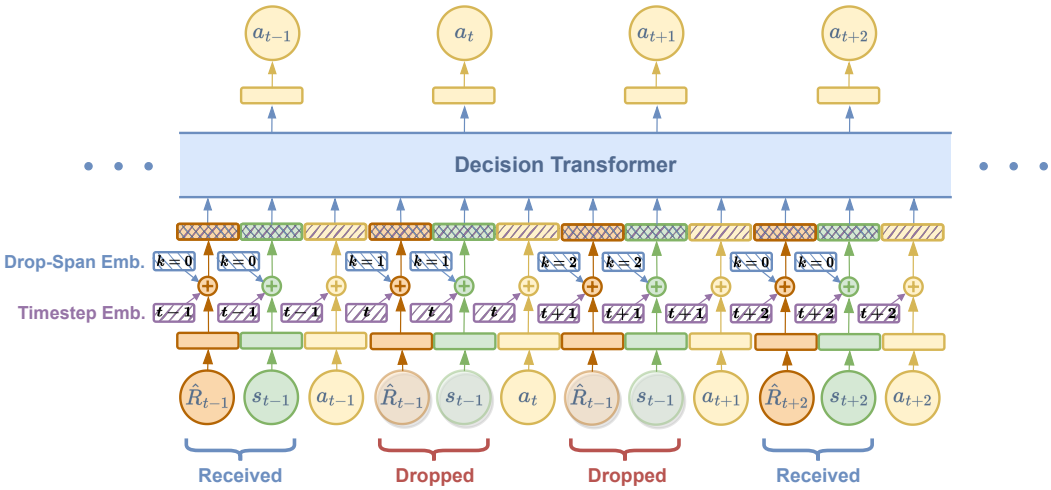

Figure 2: **Decision Transformer under Random Frame Dropping (DeFog).** Reward-to-goes (RTGs) and states are repeated from the previous steps if the current frame is dropped. Timestep and drop-span embeddings, indicating the timestep and number of consecutive frame drops, are added onto the encoded RTGs and states before sent to the Decision Transformer backbone. Since actions are not dropped, only the timestep embeddings are added to the encoded actions. The DT backbone outputs the predicted action embeddings, which is passed through a decoder to obtain the predicted actions.

Generative Pre-Training model (Radford et al., 2018). Let ϕ_g, ϕ_s and ϕ_a denote the reward-to-go, state, and action encoders respectively, the input tokens are obtained by first mapping the inputs to a d-dimensional embedding space, then add a timestep embedding $\omega(t)$ to the tokens.

Let u_{g_t}, u_{s_t} and u_{a_t} denote the input tokens corresponding to the observation, the action and the reward-to-go of the t-th timestep respectively, and v_{g_t}, v_{s_t} and v_{a_t} be their counterparts on the output token side. DT could be formalized as:

$$u_{s_t} = \phi_s(s_t) + \omega(t), \quad u_{g_t} = \phi_g(g_t) + \omega(t), \quad u_{a_t} = \phi_a(a_t) + \omega(t) \tag{3}$$

$$v_{g_{t-K}}, v_{s_{t-K}}, v_{a_{t-K}}, \ldots, g_{s_t}, s_{g_t}, v_{a_t} = DT(u_{g_{t-K}}, u_{s_{t-K}}, u_{a_{t-K}}, \ldots, u_{g_t}, u_{s_t}, u_{a_t}) \tag{4}$$

Online decision transformer (ODT) by Zheng et al. (2022) enables online finetuning of the decision transformer models. We adopt the ODT model architecture because it has larger model capacity. Following this work, we also omit the timestep embedding $\omega(t)$ in the gym environments.

During training time, instead of directly predicting the action a_t, we follow the setting of the ODT to predict a gaussian distribution for action from the output token of state token inputs.

$$\pi_\theta\left(a_t \mid v_{s_t}\right) = \mathcal{N}\left(\mu_\theta\left(v_{s_t}\right), \Sigma_\theta\left(v_{s_t}\right)\right) \tag{5}$$

The covariance matrix Σ_θ is assumed to be diagonal, and the training target is to minimize the negative log-likelihood for the model to produce the real action in the dataset:

$$J(\theta) = \frac{1}{K}\mathbb{E}_{(\mathbf{a},\mathbf{s},\mathbf{g})\sim\mathcal{T}}\left[-\sum_{k=1}^{K}\log \pi_\theta\left(a_k \mid v_{s_k}\right)\right] \tag{6}$$

3.2.2 Train-Time Frame Dropping

To prepare the model for frame dropping, we mask out observations and reward-to-gos mannually from the dataset. During the training stage, we specify some empirical dropping distribution $\hat{\mathcal{D}}$

and sample "drop-mask" periodically from it. A drop-mask is a binary vector the same size as the dataset, It serves as the drop distribution $\mathcal{D}$ of a RDMDP. If a frame in the dataset is marked as "dropped" by the current drop-mask, the observation and the reward-to-go of that frame are then overwritten by the most recent non-dropped frame. We name the timespan between the current frame and the last non-dropped frame of it the drop-span of that frame.

One key consideration for the training scheme is the distribution $\hat{\mathcal{D}}$ of the drop-mask. A natural solution is to assume each frame to have the same possibility p_d to be dropped, and the occurrence of frame dropping is independent. Under this assumption, the random variable sequence of whether each frame is dropped becomes a Bernoulli process. Additionally, we guarantee that the first frame for each trajectory is noy dropped. After a certain training interval, the drop-mask is resampled from $\hat{\mathcal{D}}$ so that those dropped frames of the dataset could be used. We can also change $\hat{\mathcal{D}}$ as the training proceeds, for example, to increase the p_d throughout training. However, we empirically found that a constant p_d is sufficient.

3.2.3 DROP-SPAN EMBEDDING

In a frame dropping scenario, the agent must utilize the information of the past action sequence since the last observation. It is also required for the agent to deal with the missing observation and reward-to-go tokens. Instead of dropping the corresponding tokens in the input sequence, we repeat the last observation or reward-to-go token and explicitly add a unique drop-span embedding to those tokens apart from the original timestep embedding $\omega(t)$.

Let k_t denote the drop-span since the last observation of time t, the drop-span encoder ψ maps integer k_t to a d-dimensional token the same shape as the other observed input tokens $u_{\hat{g}_t}, u_{\hat{s}_t}, u_{a_t}$. The model input with the drop-span embedding becomes:

$$u_{\hat{s}_t} = \phi_s(\hat{s}_t) + \psi(k_t) + \omega(t), \quad u_{\hat{g}_t} = \phi_g(\hat{g}_t) + \psi(k_t) + \omega(t), \quad u_{a_t} = \phi_a(a_t) + \omega(t) \tag{7}$$

Since the actions are decided and executed by the agent itself, they do not face the problem of frame dropping. The drop-span embedding is an analogy to the function of the timestep embedding in the Decision Transformer, but it bears the semantic meaning of how many frames have passed. Compared to the other indirect methods, the explicit use of drop-span embedding achieves better results. Detailed comparison could be found in Section 4.5.

3.2.4 FREEZE TRUNK FINETUNING

The combination of the train-time frame dropping and the drop-span embedding is adequate to make our model rather robust to dropped frames. However, we found that in the continuous control tasks, a fine-tuning procedure could further improve the performance in more challenging scenarios.

Inspired by recent progress on prompt finetuning in both the natural language processing (Liu et al., 2021) and the computer vision (Jia et al., 2022) areas, we propose to finetune the drop-span encoder ψ and the action predictor π_θ after the entire model converges. We term this procedure "freeze trunk finetuning" since the majority of the model parameters are frozen during the finetuning.

The training procedure during this stage is the same as that of the entire model, as we also draw drop-masks from $\hat{\mathcal{D}}$ during this stage to give the drop-span embeddings enough supervision. While the training steps of this stage could be just one-fifth of the main stage, the drop rate p_d can usually be higher. Empirical results prove its ability to improve the model's performance in higher dropping rates among multiple environments. The whole training pipeline of our method could be found at Appendix A.

4 EXPERIMENTAL RESULTS

In this section, we describe our experiment setup and analyze the results. We compare our method with state-of-the-art delay-aware reinforcement learning methods, as well as online and offline reinforcement learning methods in multiple frame dropping setttings. We first evaluate whether DeFog is able to maintain its performance as the drop rate p_d increases. We then explore which key factors

and design choices helped DeFog to achieve its performance. Finally, we provide insights of why DeFog can accomplish control tasks under severe frame dropping conditions.

4.1 EXPERIMENT SETUP

To comprehensively evaluate the performance of DeFog and baselines, we conduct experiments on three continuous control environments with proprioceptive state inputs in the gym MuJoCo environment (Todorov et al., 2012), as well as three discrete control environments with high-dimensional image inputs in the Atari games.

In each of the three MuJoCo environments, we collect offline datasets of three different levels: expert, medium, and medium-replay. While in the three Atari environments, we follow the Decision Transformer to train on an average sampled dataset from a DQN agent's replay buffer. We leave the detailed description of the settings to Appendix B.2.

During evaluation, we test the agents in an environment that has frame drop rates ranging from 0% to 90%. Results are shown by plotting the average return against test-time drop rate for different agents. In the expert datasets, the performance curve of our method is compared against three baselines:

- **Reinforcement learning with random delays (RLRD; Bouteiller et al., 2020).** In order to train delay-robust agents, RLRD adds randomly sampled delay as well as an action buffer to its observation space. For fair comparison, we assume the delay is created by resending the dropped observations, which again has probability p_d of being lost. We correspondingly change the delay-sampling distribution from uniform to geometric. Since RLRD has a maximum delay constraint, we limit the delay value in the augmented observation at its maximum even if frames are still dropped. Since the RLRD performs max-entropy learning of the action, it's not suited for discrete action tasks like Atari, so its performance is only included in the gym MuJoCo environment.
- **Twin-delayed DDPG (TD3; Dankwa & Zheng, 2019).** We train an online expert RL agent under regular non-frame-dropping settings using TD3 for continuous control tasks. We note that TD3 also has the privilege to access the environment.
- **Decision transformer (DT; Chen et al., 2021).** We train the vanilla DT using exactly the same offline dataset without the proposed components in Section 3.2.

We note that the online methods such as RLRD and TD3 are trained directly in the environment. Hence, they do not have any differences in terms of the dataset types. Their performance curves are plotted repeatedly for comparison. We also include a DeFog without finetuning version to evaluate the effectiveness of freeze trunk finetuning. Since TD3 cannot handle frame-dropping scenarios, we only plot TD3 in one row of the figures for better illustration.

4.2 EVALUATION IN THE CONTINUOUS CONTROL TASKS

We first evaluate our performance on three MuJoCo Continuous control environments, namely Hopper-v3, HalfCheetah-v3, and Walker2d-v3, each on the three levels. The results on each dataset are given in Figure 3.

As shown in Figure 3, we find that DeFog is able to maintain performance under severe drop rates. For example, the performance of the finetuned version on Walker2d dataset barely decreases even when the drop rate is as high as 80%. Meanwhile, the performance of the vanilla DT and TD3 agents come close to zero once the drop rate exceeds 67%.

By looking at the starting point of the performance curves, we see that DeFog can achieve the same performance as the vanilla DT agent in non-frame-dropping scenarios. Despite a high train-time drop rate of 80% applied to DeFog none of them are negatively affected when tested with a drop rate of 0%. This hints that DeFog is data efficient and could learn with limited access to the full dataset. As a comparison, the online RLRD failed to achieve the same non-frame-dropping performance as other online baselines.

In the HalfCheetah with expert dataset setting, our method significantly outperforms the RLRD baseline with drop rates lower than 50%; however, in more extreme cases the RLRD takes over. RLRD, with its advantage of interacting with the environment, was able to keep the performance

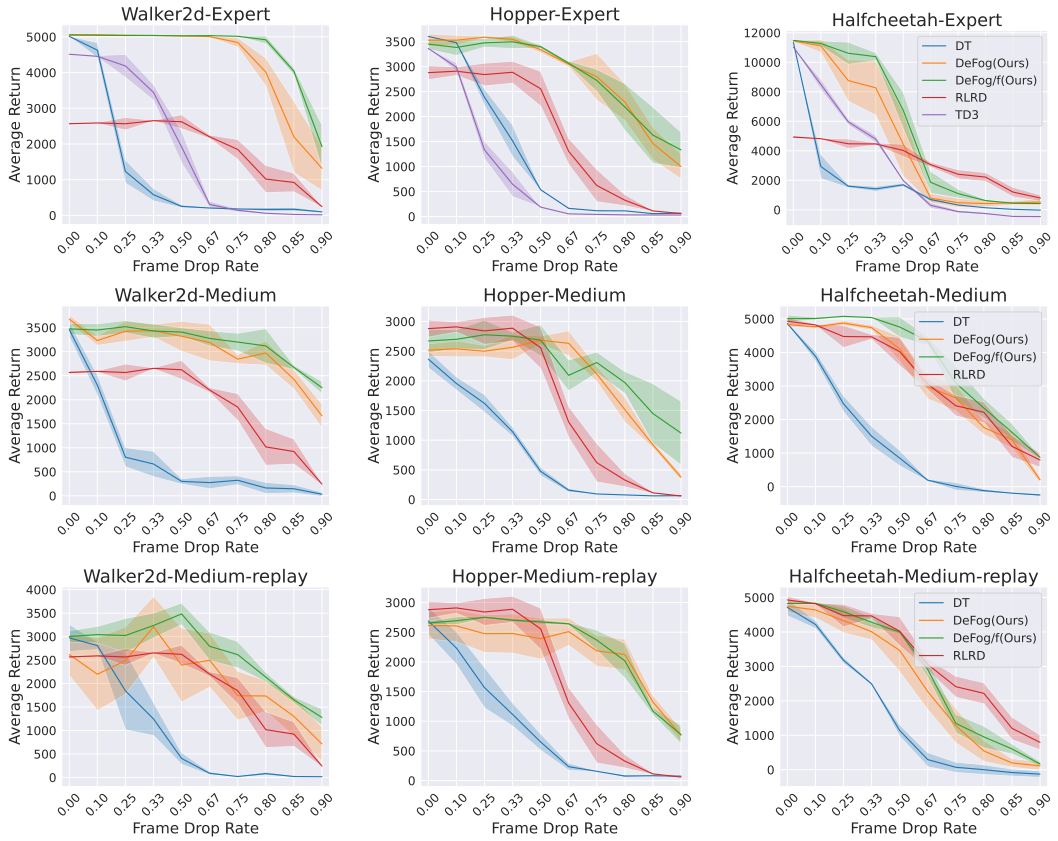

Figure 3: Performance on continous control tasks. There are three baselines: a) TD3: the online TD3 agent. b) DT: the offline Decision Transformer agent trained on the same dataset as DeFog. c): RLRD: the online RLRD agent that is optimized to deal with random frame delay. The full-fledged version of our method is indicated with DeFog/f (Ours), while the result of a non-finetuned version is indicated with DeFog (Ours) for comparison.

better possibly due to HalfCheetah being relatively easy out of these three environments. In the medium and medium-replay datasets, DeFog is limited to data with less expertise, thus obtaining a reduced average return, while RLRD with full access could only reach of maximum level of the same return.

We can also find that freeze trunk finetuning effectively improves DeFog's performance in various settings. In all 9 settings, the finetuned agent obtains better or at least the same results with the non-finetuned ones. The finetuning is especially helpful in high drop rate scenarios as shown in the Hopper-medium and Walker2d-expert settings. We highlight that the finetuning also only needs the offline dataset rather than online interactions.

4.3 Evaluation in the Discrete Control Tasks

In this section, we evaluate our performance on three Atari discrete control environments (Bellemare et al., 2013): Qbert, Breakout, and Seaquest. Following the practice of the Decision Transformer, we use 1% of a fully trained DQN agent's replay buffer to train the agent. The results are shown as in Figure 4. We find that DeFog outperforms the decision transformer. Similarly, we attribute the performance gain to the proposed data masking and drop-span embeddings. We note that in discrete control tasks, finetuning does not provide additional improvement; hence, we omit the result of a finetuned DeFog model.

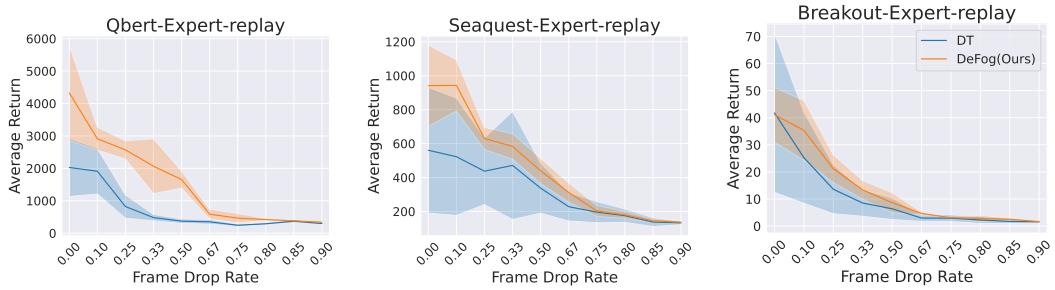

Figure 4: The performance in the discrete Atarti game environments. DT is the offline Decision Transformer agent trained on the same dataset as DeFog. Our method is indicated with DeFog (Ours).

4.4 Visualized Results

To gain a better understanding of DeFog, we visualize the results in the MuJoCo environment.

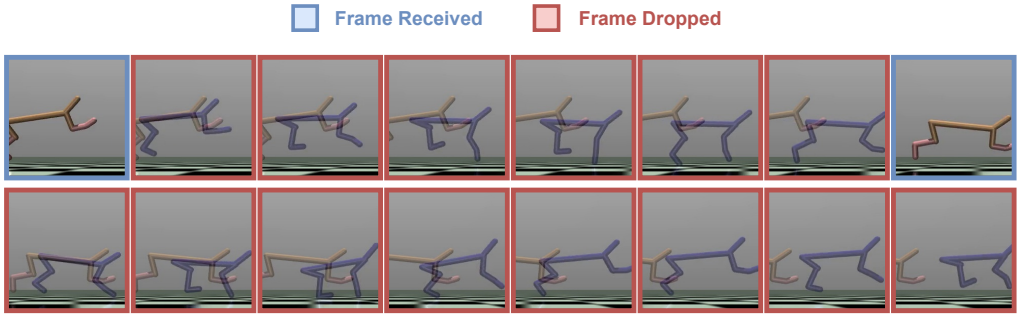

Figure 5: Visualization results of DeFog HalfCheetah agent under 90% frame drop rate. In each frame, the orange cheetah is the observed state $\hat{s}_t$, while the purple cheetah is the actual state s_t.

We first visualize an agent that is trained with DeFogin a frame-dropping environment. We show the frames of a recorded trajectory in Figure 5, under a frame dropping rate of 90%. As shown, the HalfCheetah agent (blue) is able to act correctly even if the observation (semi-transparent yellow) is stuck at 8 steps ago. Once a new observation comes in, the agent immediately adjusts itself to the newest state and continues to perform a series of correct actions.

Building on top of the previous setting, we aim to exploit the capability of DeFog under extreme conditions by increasing the frame drop rate to 100%. In this way, the agent is only able to look at the initial observation and needs to make the rest of the decisions blindly. As shown in Figure 6, the HalfCheetah agent continues to run smoothly for more than 24 frames. This may be due to the Transformer architecture's ability to infer from contextual action history, predicting a sequence of actions even without seeing the next observations. We conjecture that such a phenomenon is analogous to how humans perform a skill such as swing a tennis racket without using the observations.

4.5 Ablation Study

We conduct ablation study on our drop-span embeddings and freeze trunk finetuning parts of DeFog.

For the drop-span embeddings, we implement an alternative method that implicitly embeds the drop span. Concretely, at each time step t, if the current frame is dropped, we change the corresponding timestep embedding $\omega(t)$ to the last received one $t - k_t$. Hence, for a implict embedding method, we have $u_{s_t} = \phi_s(s_t) + \omega(t - k_t), \quad u_{g_t} = \phi_g(g_t) + \omega(t - k_t), \quad u_{a_t} = \phi_a(a_t) + \omega(t)$ .In this way, the agent can similarly have the information of the drop-span. As shown in Figure 7, we see that the proposed drop-span embedding outperforms the implicit embedding, showing the effectiveness and necessity of explicitly providing the drop-span information to DeFog.

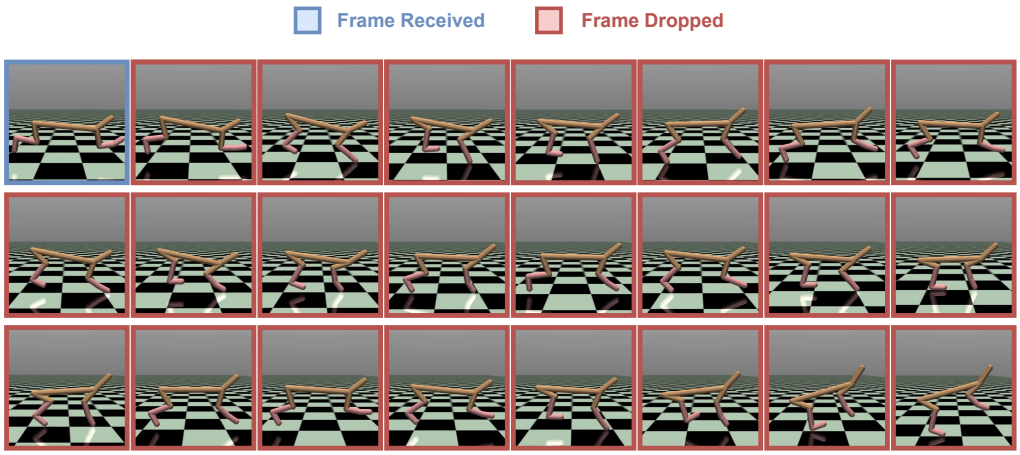

Figure 6: Visualization results of DeFog HalfCheetah agent under 100% frame drop rate. Only the very first observation is received. The scenarios challenges the limit of the DeFog model.

Now we ablate the freeze trunk finetuning method by comparing the the same model with and without the finetuning stage. All comparisons are performed on three MuJoCO medium-replay datasets. As shown in Figure 7, the finetuning largely outperforms the original model on the continuous tasks. We believe that the performance gain in complex continuous control tasks is because the lightweight yet crucial modules (action predictor, drop-span encoder) are easier to be trained for accurate actions. We leave more results in the ablation section in Appendix C.

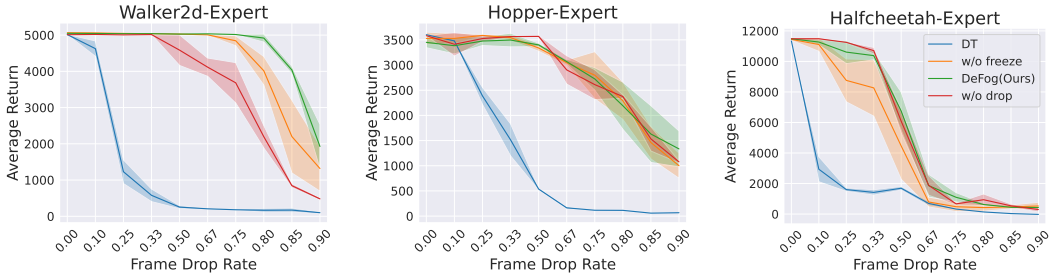

Figure 7: Ablation study of the drop-span embedding and the finetuning stage. "w/o freeze" means without freeze trunk finetuning, and "w/o drop" stands for without drop-span embedding. We find that the proposed methods are effective in the continuous control tasks.

5 Conclusion

In this paper, we propose Decision Transformer under Random Frame Dropping (DeFog), an algorithm based on the Decision Transformer that is robust against a crucial challenge in real-world remote control tasks — frame dropping. DeFog simulates the frame dropping phenomenon with by randomly masking out the observations in offline datasets and embeds frame dropping timespan information explicitly to the model. Furthermore, when finetuned on the drop-span encoder and the action predictor, it becomes more robust to high dropping rates in continuous tasks. Empirical results show that DeFog outperforms strong baselines on continuous and discrete control benchmarks under severe frame dropping settings, where frame drop rates may be as high as 90%. Along with work, we find that a promising future direction is to further handle corrupted observations such as blurred images or inaccurate velocities and to deploy on a real robot.

REFERENCES

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pp. 104–114. PMLR, 2020.

S Balemi and UA Brunner. Supervision of discrete event systems with communication delays. In *1992 American Control Conference*, pp. 2794–2798. IEEE, 1992.

Oussama Bekkouche, Tarik Taleb, and Miloud Bagaa. Uavs traffic control based on multi-access edge computing. In *2018 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6. IEEE, 2018.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279, 2013.

Yann Bouteiller, Simon Ramstedt, Giovanni Beltrame, Christopher Pal, and Jonathan Binas. Reinforcement learning with random delays. In *International conference on learning representations*, 2020.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Stephen Dankwa and Wenfeng Zheng. Twin-delayed ddpg: A deep reinforcement learning technique to model a continuous movement of an intelligent robot agent. In *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*, pp. 1–5, 2019.

Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2020.

J Funda and RP Paul. Efficient control of a robotic system for time-delayed environments. In *Fifth International Conference on Advanced Robotics' Robots in Unstructured Environments*, pp. 219–224. IEEE, 1991.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/haarnoja18b.html`.

Todd Hester and Peter Stone. Texplore: real-time sample-efficient reinforcement learning for robots. *Machine learning*, 90(3):385–429, 2013.

Jemin Hwangbo, Inkyu Sa, Roland Siegwart, and Marco Hutter. Control of a quadrotor with reinforcement learning. *IEEE Robotics and Automation Letters*, 2(4):2096–2103, 2017.

Chieko Sarah Imai, Minghao Zhang, Yuchen Zhang, Marcin Kierebinski, Ruihan Yang, Yuzhe Qin, and Xiaolong Wang. Vision-guided quadrupedal locomotion in the wild with multi-modal delay randomization. *arXiv preprint arXiv:2109.14549*, 2021.

Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems*, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

Konstantinos V Katsikopoulos and Sascha E Engelbrecht. Markov decision processes with delays and asynchronous cost collection. *IEEE transactions on automatic control*, 48(4):568–574, 2003.

Kuang-Huei Lee, Ofir Nachum, Mengjiao Yang, Lisa Lee, Daniel Freeman, Winnie Xu, Sergio Guadarrama, Ian Fischer, Eric Jang, Henryk Michalewski, et al. Multi-game decision transformers. *arXiv preprint arXiv:2205.15241*, 2022.

Yongfu Li, Chuancong Tang, Srinivas Peeta, and Yibing Wang. Nonlinear consensus-based connected vehicle platoon control incorporating car-following interactions and heterogeneous time delays. *IEEE Transactions on Intelligent Transportation Systems*, 20(6):2209–2219, 2018.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

Linghui Meng, Muning Wen, Yaodong Yang, Chenyang Le, Xiyun Li, Weinan Zhang, Ying Wen, Haifeng Zhang, Jun Wang, and Bo Xu. Offline pre-trained multi-agent decision transformer: One big sequence model conquers all starcraftii tasks. *arXiv preprint arXiv:2112.02845*, 2021.

Somjit Nath, Mayank Baranwal, and Harshad Khadilkar. Revisiting state augmentation methods for reinforcement learning with stochastic delays. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 1346–1355, 2021.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Machel Reid, Yutaro Yamada, and Shixiang Shane Gu. Can wikipedia help offline reinforcement learning? *arXiv preprint arXiv:2201.12122*, 2022.

Olimpiya Saha and Prithviraj Dasgupta. A comprehensive survey of recent trends in cloud robotics architectures and applications. *Robotics*, 7(3), 2018. ISSN 2218-6581. doi: 10.3390/robotics7030047. URL https://www.mdpi.com/2218-6581/7/3/47.

Juergen Schmidhuber. Reinforcement learning upside down: Don't predict rewards–just map them to actions. *arXiv preprint arXiv:1912.02875*, 2019.

Erik Schuitema, Lucian Buşoniu, Robert Babuška, and Pieter Jonker. Control delay in reinforcement learning for real-time dynamic systems: A memoryless approach. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3226–3231. IEEE, 2010.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Thomas J Walsh, Ali Nouri, Lihong Li, and Michael L Littman. Learning and planning in environments with delayed feedback. *Autonomous Agents and Multi-Agent Systems*, 18(1):83–105, 2009.

Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua Tenenbaum, and Chuang Gan. Prompting decision transformer for few-shot policy generalization. In *International Conference on Machine Learning*, pp. 24631–24645. PMLR, 2022.

Qinqing Zheng, Amy Zhang, and Aditya Grover. Online decision transformer. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27042–27059. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zheng22c.html.

A ALGORITHM DETAILS

Algorithm 1 Decision Transformer under Random Frame Dropping (DeFog)

The training

Require: n_updates {number of updates}
Require: n_transitions {number of transitions}
Require: update_interval {interval to update train-time frame dropping}
Require: context_len {DT context length}
Require: **sample**(m, n) {sample n numbers uniformly in [0, m]}
Require: **cumcount**(flags) {count cumulative number of 0's before current position (inclusive)}
Require: **Train_DT**(obs, act, rtg, timestep, dropspan) {train GPT model with data}
Require: batch_size, act_buffer, obs_buffer, rtg_buffer

```
N ← n_updates
while N ≠ 0 do
    M ← update_interval
    drop_mask ← sample(1, n_transitions) {whether each frame is dropped in the dataset}
    dropspans ← cumcount(drop_mask) {duration of frames dropped}
    while M ≠ 0 do
        selected_index ← sample(n_transitions, batch_size)
        timestep ← selected_index
        dropspan ← dropspans[selected_index]
        dropped_index ← selected_index − dropspan
        sampled_act ← act_buffer[selected_index: selected_index + context_len]
        sampled_obs ← obs_buffer[dropped_index: dropped_index + context_len]
        sampled_rtg ← rtg_buffer[dropped_index: dropped_index + context_len]
        Train_DT(sampled_obs, sampled_act, sampled_rew, timestep, dropspan)
        M ← M − 1
    end while
    N ← N − 1
end while=0
```

B EXPERIMENT DETAILS

B.1 DATASETS AND SETUP

B.1.1 GYM MUJOCO

We use the D4RL dataset (Fu et al., 2020) that contains data collected by SAC agents. There are three datasets for each environment – expert, medium, and medium-replay. The expert dataset is collected by a fully-trained expert policy, while the medium dataset is collected by an agent about half the performance of the expert. Medium-replay includes the trajectories in a medium agent's buffer, and thus is the most diverse dataset with the lowest average return.

B.1.2 ATARI

For atari game environments, we use the DQN Replay Dataset (Agarwal et al., 2020), which is collected from the replay buffer of a DQN agent during traing of these Atari games. Following the practice of the Decision Transformer, we only use a small portion of the dataset: 1% of the whole dataset, which is 500 thousand of the 50 million transitions observed by an online DQN agent. We define three kinds of datasets for each game as well – epert, medium, and expert-replay. The expert and the medium dataset are collocted from the collected data of a DQN agent's during the final and medium training stages, while a expert-replay dataset is sampled evenly from the whold replay buffer.

B.2 HYPERPARAMETER SETTINGS

B.2.1 GYM MUJOCO

For the gym mujoco environment. we use the same model architechture as the Online Decision Transformer. While the Online Decision Transformer uses different training paramenters for each environment, we keep most of the training parameters the same among different environments

Hyperparameter	Value
Number of layers	4
Number of attention heads	4
Embedding dimension	512
Training context length K	20
Dropout probability	0.1
Activation function	ReLU
Gradient norm clip	0.25

(a) Architechture Parameters

Hyperparameter	Value
Learning rate	1e-4
Weight decay	1e-3
Batch size	256
Total training steps	1e5
Finetune training steps	2e4
Learning rate warmup steps	1e4
Drop-mask update interval	100

(b) Training Parameters

Table 1: Common Parameters for Gym Mujoco

For each dataset, we specify a target reward with it, and report the combination of train time drop-rate and finetuning drop-rate. The environment and dataset related paramenters are as follows:

Environment	Dataset	Target Reward	Training Drop Rate	Finetuning Drop Rate
HalfCheetah	Expert	12000	0.5	0.5
HalfCheetah	Medium	12000	0.8	0.8
HalfCheetah	Medium Replay	12000	0.8	0.8
Hopper	Expert	4000	0.9	0.9
Hopper	Medium	4000	0.5	0.8
Hopper	Medium Replay	4000	0.8	0.8
Walker2d	Expert	5000	0.9	0.9
Walker2d	Medium	5000	0.8	0.8
Walker2d	Medium Replay	5000	0.8	0.8

Table 2: Dataset Specific Parameters for Gym Mujoco

B.2.2 ATARI

For the Atari environments, we use the same model architecture as the Decision Transformer since the Online Decision Transformer doesn't perform experiments on these environments. The hyper-parameters are as follows:

Hyperparameter	Value
Number of layers	6
Number of attention heads	8
Embedding dimension	64
Training context length K	30
Dropout probability	0.1
Activation function	ReLU
Gradient norm clip	1.0

(a) Architechture Parameters

Hyperparameter	Value
Learning rate	6e-4
Weight decay	0.1
Batch size	128
Total training steps	1e5
Finetune training steps	2e4
Learning rate warmup steps	1e4
Drop-mask update interval	1000

(b) Training Parameters

Table 3: Common Parameters for Atari

The environment and dataset related parameters are given in Table 4. Linear increasing frame drop rate means that the drop rate is gradually increased from start to end values.

Environment	Dataset	Target Reward	Training Drop Rate	Finetuning Drop Rate
Qbert	Expert Replay	14000	0.4	0.5
Seaquest	Expert Replay	1150	0–0.8 Linear Increase	0.8
Breakout	Expert Replay	90	0–0.8 Linear Increase	0.8

Table 4: Dataset Specific Parameters for Atari

C Supplementary Results

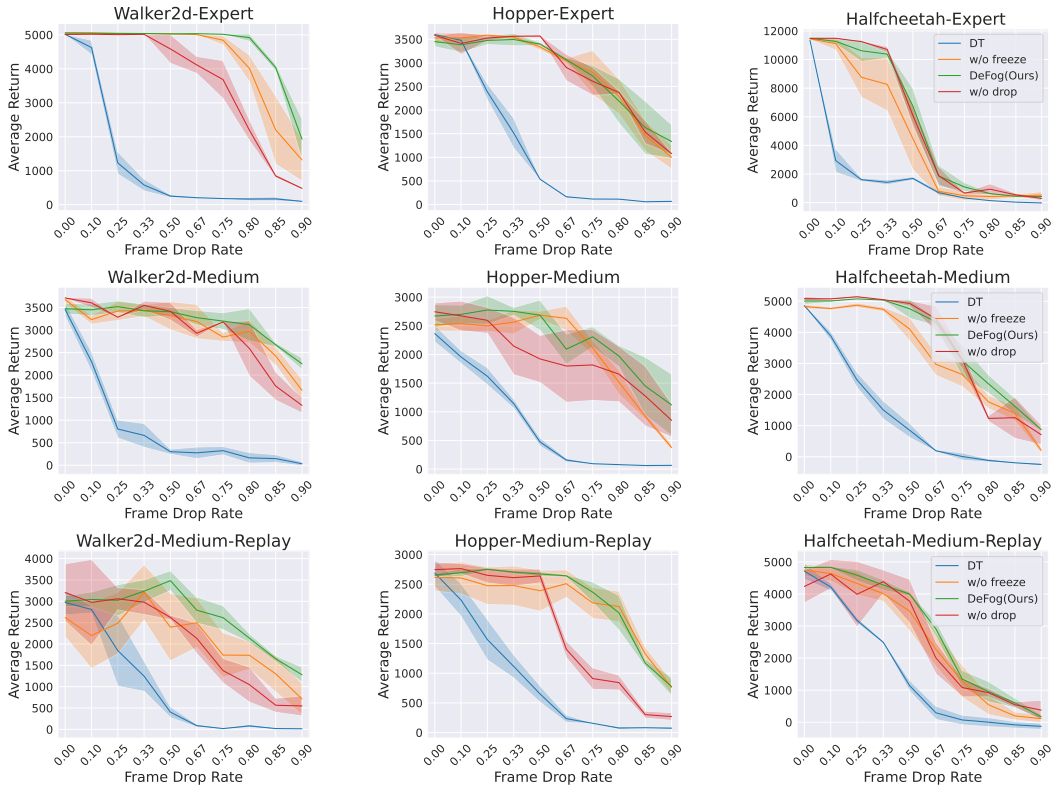

Figure 8: Full ablation results of the gym MuJoCo environments.