1 How did we improve the diffusion model and apply it to RobustRL

Specifically, our method has the following novelties that need to be clarified for you:

- Our Robustlight method first proposes a model-free approach, using the Diffusion model as a policy selector to recover the original state.

- Unlike the native diffusion model, our method adopts a new Beta schedule $\beta_i = 1 - \alpha_i = e^{-\frac{b}{i+a+c}}$ that handles noise information from small to medium noises instead of naive diffusion beta schedule[1][2][3] generating data from pure noise.

- In theory, our objective is to minimize the cross-entropy of the denoised states along the RL trajectory: $\mathcal{L}_{entropy} = \sum_{t=2}^{T} \mathbf{E}_q(s_t)[-\log p_\theta(\tilde{s}_t^0 | a_{t-1}, \tau_{t-1}^{\hat{s}})]$, Following Ho et al.[3], we adopt the variational lower bound (VLB) to optimize the negative log-likelihood: $\mathcal{L}_{VLB} = \sum_{t=2}^{T} \mathbf{E}_{q(s_t)}[-\log p_\theta(\tilde{s}_t^0) + D_{KL}(q(s_t^{1:K} | s_t^0) \| p_\theta(\tilde{s}_{1:K}^t \| a_{t-1}, \tau_{t-1}^s)$. Derived from the equation we get $L_\theta = E_{i \sim \mathcal{U}_K, \epsilon_t \sim \mathcal{N}(0,I), (A_{t-N}, \ldots, A_{t+M-1}) \in D_\nu} \left\| \epsilon_\theta(\tilde{A}_t^i, S_{t-1}, i) - \epsilon_t^i \right\|_2 + \sum_{m=t+1}^{t+M-1} \left\| \epsilon_\theta(\tilde{A}_m^i, \hat{S}_{m-1}, i) - \epsilon_m^i \right\|_2$. Our method better avoids error accumulation by using this non-Markovian loss function, it balances the accuracy of the diffusion model at the current reinforcement learning timestep and the condition shift over a long reinforcement learning time horizon. By optimizing this function, we can reduce the cross-entropy of the recovered states, thereby achieving better data recovery from noise.

- Compared to model-based methods, our model-free Robustlight method interacts better with the environment to train the best diffusion policy without the need for precise environmental modeling, thereby avoiding suboptimal decision-making.

- To enhance policy performance, we employ action gradient methods $A_t = A_t + \mu \nabla_A Q_\phi(\mathbf{S}_t, A_t)$, adjusting the policy based on actions toward better reward performance.

- We design an RL repaint module akin to image repainting. Here, we utilize contextual cues from intact segments to generate missing data segments corrupted by sensor damage. Leveraging the potent generative prowess of the diffusion model, we achieve missing data reconstruction.