We thank reviewers for positive feedback, mentioning DTSIL as an effective novel method (R2,3,4) for a significant problem (R1), extensively evaluated (R1,2,3,4), and systematically discussed (R3). We will incorporate the suggestions.

[**R1**] **Problem statement:** As R1 interpreted, embeddings are high-level state representations which can differentiate meaningfully distinctive states. Instead of directly maximizing expected return, we proposed a novel way to find best demonstrations g^* with (near-)optimal return and train the policy $\pi_\theta(\cdot|g)$ to imitate any trajectory g in the buffer, including g^*. A solution of g^* and θ^* is not necessarily unique. As stated in L69, DTSIL allows for exploiting multiple trajectories with the best rewards found during training. We approximately solve the joint optimization problem $g^*, \theta^* = \arg\max_{g,\theta} \mathbb{E}_{\pi_\theta(\cdot|g)}[\sum_{t=0}^{T} \gamma^t r_t]$ via sampling-based search for g^* over the space of g realizable by the (trajectory-conditioned) policy π_θ and gradient-based local search for θ^*. We will revise and improve Sec. 2.1 to make this clear. **Meaning of $u, \Delta t$ (L120):** For each episode, u denotes the index of state in the given demonstration that is lastly visited by agent. The initial value $u = -1$ (at the beginning of episode) means no state in the demonstration has been visited. r^{im} is imitation reward with a value 0.1. Δt is the number of states in the demonstration to be compared with e_{t+1} to determine reward for each step t (L123). More details were provided in Appendix B.1, especially Fig. 2 for illustrations. We will add these pointers and more descriptions in main text to clarify our algorithm. **Related Work:** We will make the connection between DTSIL and prior works more clear, especially for imitation learning part.

[**R1,R2**] **Embedding clusters:** Pseudocode for organizing clusters was in Appendix A.3. We will add this pointer in L74 and a brief explanation: In the buffer, we keep a representative state embedding for each cluster. If a state embedding e_t in the current episode is close to a representative state embedding $e^{(k)}$, we increase visitation count $n^{(k)}$ of the k-th cluster. If the sub-trajectory $\tau_{\le t}$ of current episode up to step t is better than $\tau^{(k)}$, $e^{(k)}$ is replaced by e_t.

[**R2**] **Supervised learning:** With SL objective, we leverage the *actions* in demonstrations, similarly to behavior cloning, to help RL for imitation of diverse trajectories. DTSIL+EXP without SL performs worse on Montezuma's Revenge (MR) and Pitfall where imitation is difficult due to many obstacles and dangers (Tab. A). **Pseudo-count bonus:** DTSIL discovers novel states mainly by random exploration after the agent finishes imitating the demonstration. The pseudo-count bonus brings improvement over random exploration by explicitly encouraging the agent to visit novel states. Prior works (e.g. CoEX, NGU) commonly use a count-based bonus for exploration (EXP). DTSIL is complementary to EXP; combining both performs better than DTSIL (Tab. A) and PPO+EXP (Tab. 1). We will add the ablative study. **Hyper-parameters:** Assume agent's location in state embeddings is normalized to $[0, 1]$ for each coordinate and the distance metric is l_∞. When clustering embeddings in parametric memory, $\delta = 0.1$ will discretize 2D location space into $\sim$10$\times$10 grid, an intuitively reasonable size. We can remove a hyper-parameter Δt by setting $\Delta t = m$, because the larger $\Delta t \in [1, m]$ leads to better performance (Appendix E.1). DTSIL(+EXP) with $\Delta t = m = 40$, $\delta = 0.1$ achieves scores 8.2 (Apple-Gold), 21365 (MR), 10192 (Pitfall), 1915 (Venture), 7.6 (navigation), 56 (manipulation with discrete actions), comparable with numbers we reported in submission. Thus, DTSIL with a single hyper-parameter setup can perform robustly well and not brittle across various domains. **Off-policy methods:** We listed off-policy methods A2C+SIL and NGU in Tab. 1 in the submission. We additionally run R2D2[1] on Atari (Tab. A) and HER[2] on robotics manipulation with high-dimensional continuous actions, where DTSIL gets a score 20 but HER gets 0. Many off-policy methods tend to discard old experiences with low rewards and hence may prematurely converge to sub-optimal behaviors, but DTSIL using these diverse experiences has a better chance of finding higher rewards in the long term. We will add this comparison and more discussions about off-policy and model-based exploration methods.

[**R3**] We will cite Pathak et al. & Burda et al. as related works and add more discussion: Intrinsic curiosity uses the prediction error as intrinsic reward signals to incentivize visiting novel states, whereas DTSIL instead imitates long trajectories in diverse directions, which can lead to deeper exploration. As R3 suggested, we show additional experiments of ICM[3] and RF[4] for 800M steps (Tab. A).

	DTSIL+EXP	DTSIL+EXP w/o SL	DTSIL	ICM	RF	R2D2
MR	26,314	10,112	5,712	100	10,200	400
Pitfall	11,875	1,966	2,436	0	0	0
Venture	2,135	1,898	1,482	1,813	1,859	1,997

Table A: Comparison with variants of DTSIL and additional baselines, one run for each baseline due to limitation of computational resources. We will report results of more runs in the revision.

[**R4**] **Diversity:** DTSIL's ability to find diverse states does not rely solely on the "imitation error". After visiting the last (non-terminal) state in the demonstration, the agent performs random exploration (because $r_t^{\text{DTSIL}} = 0$) around and beyond the last state until the episode terminates, to push the frontier of exploration. We prevent "the bias led by original demonstrations" by allowing flexibility in following them and replacing them with better trajectories. Different performances under different random seeds are due to huge positive rewards in some states on MR and Pitfall. Once the agent luckily finds these states in some runs, DTSIL can exploit them and perform much better than other runs.

[**R1,R4**] **Presentation:** The important messages about the experiments were summarized as three questions at the start of Sec. 4. Per R1's comments, we will explicitly connect questions and experimental results in the revision. We will also emphasize these take-away messages and point to thorough descriptions in Appendix C, as R4 suggested.

[**R1,R2,R4**] Fig. 3d shows that trajectory-conditioned policy imitates diverse demonstrations well with proper attention weights. Fig. 4 shows DTRA+EXP. We will adjust Tab. 1 & Fig. 3 as suggested and use more legible labels in graphs.

[1] Kapturowski et al. Recurrent experience replay in distributed reinforcement learning. [2] Andrychowicz et al. Hindsight experience replay. [3] Pathak et al. Curiosity-driven exploration by self-supervised prediction. [4] Burda et al. Large-scale study of curiosity-driven learning.