We thank the reviewers for their thoughtful feedback, and note the apparent consensus that our contribution, Sibling Rivalry (SR), is interesting and novel. We would like to emphasize that, as several reviewers have observed, SR is **simple to implement and apply**, **works across a variety of settings, and learns from generic distance-based shaped rewards that do not rely on domain-expertise.** To make SR a valuable resource for the research community, **we will release an open-source implementation that shows SR is light-weight** ($\sim$ 50 lines of added code).

Revisions We will fix all writing issues and revise the title and notation. We will clarify how trajectories are selected for gradient computation and that results with only sparse rewards are omitted because they all fail in our setting.

Intuition (R1, R2) Reward shaping presents significant challenges for on-policy learning of goal-conditioned policies, as such methods often converge to sub-optimal policies [Williams 1992]. Such "local optima" are, e.g., low-entropy policies that fail to reach goal-states and get stuck in regions away from the goal. Manual reward shaping can prevent such outcomes, but is often domain-specific, ad-hoc, and challenging. Instead, SR uses pairs of rollouts to automatically estimate when policies get stuck in local optima and prevents the policy's terminal-state distribution $\rho_g^\pi(s_T)$ from collapsing around the associated regions of state space. In effect, this gives a **dynamic way to encourage exploration away from local optima** while continuing to guide $\rho_g^\pi(s_T)$ towards the goal g. In addition, the SR reward function converges to the sparse reward as the policy learns to reach the task-goal states, preserving the underlying task definition.

Assumptions The key requirements for applying SR are (1) that goal completion can be expressed via a distance metric and (2) that the distance metric and control over episode start state/goal are available during learning. In simulated environments, this availability should exist. We will clarify these requirements and how they are met in our experiments. In real-world environments, it may be impossible to sample sibling rollouts according to their exact definition; we consider sensitivity to noise in the sampling conditions as a direction for future research.

We will include the above and clarify the existing exposition in the paper, with particular attention to improving the presented motivation. Similarly, we will discuss challenges more clearly, such as sensitivity to ϵ in the U-maze (below).

Additional Experiments (R1, R2) We have performed the suggested Corridor and U-shaped maze experiments: the results strengthen the intuition above and reaffirm the effectiveness of SR. We will add the results to the Appendix. **SR solves the Corridor task easily at all lengths tested** (from length 5 to 25) and handles the U-maze task well for the ranges tested (total maze length from 7 to 31). In comparison, **both ICM and HER fail to solve the longer versions of the tasks**. In the most extreme U-maze case, the local optimum is on average ~ 1.5 distance units from the goal but successful navigation requires first moving roughly $10\times$ that distance away from the goal. In this case, the range of inclusion hyperparameter (ϵ) values that produce good results with SR are limited ($0.2 \leq \epsilon \leq 0.8$). Due to the limited rebuttal period, we are unable to report all suggested baseline comparisons here but will include them in the paper.

R1 *...not fully clear...long narrow corridor...U-shape should fail...* Thank you for the insightful suggestion, we hope the above results clarify the robustness of SR. We observed that the two distance terms in the SR reward do not cancel out (see the Corridor results); rather, it is more helpful to think of their distinct effects on the distribution of terminal states $\rho_g^\pi(s_T)$. The distance-to-goal reward draws $\rho_g^\pi(s_T)$ towards g while the distance-to-sibling reward prevents the former reward from collapsing $\rho_g^\pi(s_T)$ around hurtful optima.

...different global optimum...shaped reward baseline weak... Because we only supply shaped rewards at the terminal timestep, the concern about the naive shaped rewards being a weak baseline does not apply. Nevertheless, we will clarify that providing the absolute distance values as reward at each timestep may distort the global optimum.

R2 *...the difference between $V(s, g)$ and $d(s, g)$...how bad a distance can be...* The experiments described above clarify how the quality of the distance function impacts SR. Note that local optima hinder learning even with "good" distance functions, as shown in our bit-flipping and Minecraft experiments.

...any-state-to-any-state problem... We avoid the any-state-to-any-state version of the goal-conditioned RL problem in order to address the realistic scenario where task-relevant goals only occupy a small portion of the state space.

R3 *Sample complexity* We will plot results in terms of sampled timesteps rather than episodes. The number of episodes provides an upper bound on the number of timesteps (`#timesteps` $\leq$ `#episodes * MaxEpisodeLength`). Hence, plotting results in terms of timesteps will shift the learning curves to the left. **We confirmed that SR significantly outperforms baselines also when rewards are plotted vs timesteps.**

Formal analysis Our work contributes an extensive empirical validation of SR, we leave formal analysis for future work.