This reply is for the reviewer 1.

1. About the correctness of Theorem 1 and the connection between RL-based sysnthesis and Theoretical results.
   For an MDP M, the augmented tLDBA $\bar{B}_\varphi$ corresponding to an LTL formula φ, and the product MDP $M^\otimes$ of M and $\bar{B}_\varphi$, we see an optimal policy as an positional policy on the product $M^\otimes$. Then, we define a reward function based on the acceptance condition of $\bar{B}_\varphi$. Therefore, in the counterexample of the reviewer 1, the optimal policy and the reward function should be considered as one on the product $M^\otimes$ and be designed based on the corresponding acceptance condition, respectively.

   An positional policy on $M^\otimes$ corresponds to an finite memoly policy, thus the our counterexample is meaningful as an example that shows we may not obtain any policy satisfying the given LTL formula if we use the corresponding limit-deterministic generalized Büchi automaton without the proper augmentation.

2. The reason we use LDBA.
   The reason we use LDBA is explained in [1]. we know that deterministic Rabin automata (DRA) and non-deterministic Büchi automata (NBA) can recognize all of the ω-regular language. However, there is an example of an MDP M and an LTL formula φ with Rabin index 2, such that, although there is an policy satisfying φ with probability 1, optimal policy obtained from any reward corresponding to the acceptance condition do not satisfy the LTL formula. Further, LDBAs are as expressive as general NBAs. However, in a LDBA (non-generalized), the order of visiting accepting sets is fixed. Therefore, the reward based on the acceptance condition tends to be sparse.

3. Notion of the omega in the exponent etc.
   The omega in the exponent means the infinite connection, namely Σ^ω means $\Sigma\Sigma\ldots$ and $S(\Sigma S)^\omega$ means $S\Sigma S\Sigma S\ldots$. The scalar bounded reward means the reward $r_p \in [0, \infty)$. s_{init} is the initial state.

4. why we employ the notion of a formula being satisfied as the non-zero probability.
   The reason we employ the notion of a formula being satisfied as the non-zero probability is to more generally evaluate an obtained policy. Underlying the notion, the goal is to obtain a policy maximizing the probability of satisfaction efficiently.

5. Contributions in our paper.
   Our contributions are as follows.

   - The sparsity of rewards is relaxed comparing with using LDBA (non-generalized) by introducing the augmentation of LDBAs. Therfore, our proposed algorithm is more sample efficient.

- Recurrent classes of the Markov chain induced by a product MDP $M^{\otimes}$ and a positional policy π on $M^{\otimes}$ are classified as ones that has at least one accepting transition in each accepting set or ones that has no accepting transition in all accepting sets without depending on the policy.

References

[1] E. M. Hahn, M. Perez, S. Schewe, F. Somenzi, A. Triverdi, and D. Wojtczak, “Omega-regular objective in model-free reinforcement learning,” *Lecture Notes in Computer Science*, no. 11427, pp. 395–412, 2019.