We thank the reviewers for their effort and insightful comments during these unprecedented times. In this work, we propose the *first finite-time system identification algorithm for partially observable linear dynamical systems (LDS) in adaptive and closed-loop settings*. Prior estimation methods only work when the actions/controls are iid random noise and do not allow for any exploitation or strategic exploration. This strong limitation significantly restricted the regret minimization and algorithm design in LDS with partial observations [3-6,9,12]. Our proposed estimation algorithm allows the data collection with an adaptive controller and the design of fully adaptive RL methods. We believe this contribution alone has a great interest in both RL and control communities. Ultimately, we deploy this estimation method, propose the first "truly" adaptive control algorithm in partially observable LDS, and obtain the *first polylogarithmic regret* in this challenging setting. Our results provide a clear improvement over the prior works and shed light to further developments in the field.

Relevance of linear systems to RL and machine learning community: (R3). We would like to highlight that the setting of our work is more general than classical LQG since our algorithm can handle time-varying and adversarial cost functions that subsume LQG. Moreover, LQR & LQG settings are MDP & POMDP models with the state, action, reward spaces not constrained to be bounded. They are fundamental models in MDPs and POMDPs:

- LQR & LQG are among few continuous settings where the optimal policies exist (and mainly have closed form) [1].
- For the general model of $x_{t+1} = f(x_t, u_t, w_t)$ with Gaussian w_t, using representation theory results (e.g. Koopman theory or RKHS theory) any such f can be written as linear function of its basis. Thus, MDPs or POMDPs can be written as LDS up to some considerations and a further change of bases. These principles have been intuitively used in deep RL (e.g. Zhang et al. "SOLAR: Deep Structured Representations for Model-Based RL" 2019).

Based on these facts, the study of LDS, with LQG being one of the most challenging ones, is an important problem in RL. Note that prior works in this area, such as [6,12,37,40-46] have been published in recent machine learning conferences (NeurIPS, ICML...). Therefore, we do not see why this paper would be less relevant to our community.

Regret without the persistence of excitation (PE): (R1, R2). In general, PE is standard in control theory since it allows asymptotic convergence of algorithms involving system identification, adaptive prediction, and control (e.g. Boyd & Sastry, *On Parameter Convergence in Adaptive Control*, 1983; Green & Moore, *Persistence of excitation in linear system*, 1986). If PE is absent, we provide two general algorithms stated in Cor. 6.2 and H.1:

1. The agent uses a warm-up period of $O(\sqrt{T})$ after which it commits to a controller yielding a regret of $\sqrt{T}$.
2. This approach is concerned more with adaptive model estimation than regret minimization. The agent adds Gaussian noise to the control input which yields regret of $T^{2/3}$, while adaptively improving the accuracy of model estimates.

R1: We are delighted to hear the kind words of R1 about our novel results.
PE policies: This is a mild condition and most of the well-known controllers (H_2, H_∞) satisfy it. For example, consider a unary DFC $\mathbf{M}$ and the PE condition given in Appendix E.2. For $\mathbf{M}$ to be not PE, we need to have an extremely wide matrix of $p \times O(\bar{H}(n+m))$ dimensions (block row of the matrix that maps past noise to input) to be row rank deficient, where $\bar{H}$ is our choice and a large number. Thus, it is quite hard and pathological to design an LQG with a small neighborhood such that there exists an $\mathbf{M}$ which is not PE for the models in the neighborhood. Moreover, if $\mathbf{M}_\star$ is PE, then there is a neighborhood around $\mathbf{M}_\star$ consisting of all PE controllers. Note that the prior works also rely on PE for consistent estimates. We appreciate R1 for bringing up this discussion. Per R1's suggestion, we added the rigorous version of the above discussion and explanation to the main text. Relaxing the PE requirement is still an important open problem which we will discuss in the conclusion. In light of this, we would like to invite R1 to increase their score.

R2: We thank R2 for their insightful comments about our novel results in the field of online control of LDS.
Regarding stabilizing controller: The majority of the prior works in partially observable LDS consider stable systems [3-5,9,12]. Recently, [6] made a significant effort to generalize aspects of this to stabilizable systems when a stabilizing controller is given. Note that, many partially observed systems cannot be stabilized by a static feedback controller and the assumption of the existence of such controller is somewhat restrictive (see Halevi *Stable LQG controllers* 1994). In the current paper, we provide a general framework of learning and regret analysis in partially obsevable stable LDS, and avoid further complications to convey this core contribution.
At this point, we would appreciate if R1 & R2 could convince R3 of the importance of partially observable LDS in RL.

R3: As we describe in the first paragraph of the rebuttal, we propose the **first finite-time closed-loop learning algorithm of partially observable LDS**. Prior estimation methods only work when actions are random iid noises. We strongly believe that this result alone is a significant contribution to the field. Building on the mentioned estimation method, we use online learning techniques [11] for the final step of controller learning which are also used in [6]. We adapt the controller design of [6] to our formulation of system identification. We derive novel sample complexity requirements to satisfy closed-loop stability and persistence of excitation during the adaptive control period.