We thank the reviewers for their careful reading and useful feedback. Below we briefly address each reviewer's comments; these will be discussed in greater details in the revised version.

Common question on how to choose the problem parameters: As reviewers pointed out, the requirement to know the problem parameters (strong convexity and smoothness) is endemic to many first order methods, including the original Catalyst framework for minimization. It is worth pointing out that for many applications, the strong convexity comes from the regularization term that is usually known. In practice, using *a rough estimate* of the strong convexity or smoothness constant would also work as long as the choice of regularization parameter τ leads to a more balanced auxiliary problem (which is one of the key insights of our framework). Such a rough estimate can be obtained by approximately computing the spectrum of the Hessian or through backtracking. As an alternative, these constants can be simply treated as hyper-parameters and tuned with the stepsize.

To Reviewer 1:

- **Numerical comparison:** Implementations of previous methods such as Minimax-APPA and primal-dual smoothing, *are not available* in the original papers. These algorithms are difficult to implement as they require smoothing and knowing the desired accuracy for each loop, which must be tuned in practice. That is why we pick DIAG as a representative benchmark among these three-loop algorithms for numerical comparison. We will move these comparisons from the appendix to the main paper and provide more numerical evidence.
- **Dependence on $\mathcal{D}_\mathcal{Y}$ and $\mathcal{D}_\mathcal{X}$**: All algorithms in Table 2 have the same dependency on $\mathcal{D}_\mathcal{Y}$. Minimax-APPA assumes $\mathcal{X}$ to be bounded and has logarithmic dependency on $\mathcal{D}_\mathcal{X}$ in the complexity, and DIAG requires $\mathcal{X} = \mathbb{R}^d$. Our algorithm assumes $\mathcal{X}$ is convex and only has logarithmic dependency on $\|x_0 - x^*\|$ in SC-C setting.
- **Negligible cost for checking stopping criterion**: For stochastic methods such as SVRG and SVRE, the cost for checking the stopping criterion is $\mathcal{O}(n)$ in each epoch and does not increase the overall complexity. In addition, the full gradient is already computed in each epoch, so the cost of checking this criterion is still almost negligible.
- **Gain in other setting.** As the reviewer pointed out, there is no gain for the convex-concave setting as EG is already optimal. However, there could be gains in other settings, such as SC-SC and NC-SC settings. In fact, we have recently shown that when extending to (μ_x, μ_y)-SC-SC setting, our framework achieves the complexity of $\mathcal{O}(\ell/\sqrt{\mu_x \mu_y}\log(1/\epsilon))$, which again improves over EG and matches the lower bound up to logarithmic factors.
- **Lower bound**: No matching lower bound is known for the NC-C setting. A valid lower bound from nonconvex minimization is $\mathcal{O}(1/\epsilon^2)$, which may not be tight. For the SC-C setting, we don not know if the current lower bound is missing a log factor or not. We will explicitly discuss the open questions of lower bounds related to our settings.
- **Typos and appendix**: We will fix these issues and clean the appendix. *Thanks for the careful reading!*

To Reviewer 2:

- **Narrow niche**: We should have mentioned that strongly-convex-concave minimax optimization itself has broad applications in game theory, imaging, distributionally robust optimization, etc. The bilinear case of this setting has been studied extensively, but the general case remains unexplored. Besides, the framework we present for SC-C setting can be extended to NC-C setting (as shown in the paper) and potentially other settings such as C-C, SC-SC, and NC-SC settings (see our response to Review 1).
- **Different learning rates:** We agree with the reviewer that some related work uses different learning rates for the two players for NC-C setting, e.g., GDA [Lin et al.,2019] uses $\mathcal{O}(\epsilon^4)$ stepsize for the min player and $\mathcal{O}(1)$ for the max player. Theoretically, they achieve a much worse $\mathcal{O}(\epsilon^{-6})$ complexity than the $\mathcal{O}(\epsilon^{-3})$ complexity achieved by our Catalyst with GDA under *the same constant stepsize* for both players. We will also add numerical comparisons.
- **Related work**: We thank the reviewer for the pointers. We will add these references and discuss them in the paper. Competitive gradient descent [Schäfer & Anandkumar, 2019] and some recent work [Xu et al., 2020] add quadratic regularization terms in the forms of $\|x\|^2$ and $\|y\|^2$. This is different from our framework, as we only add the term $\|y - z_t\|^2$, where z_t comes from extrapolation.

To Reviewer 3:

Pre-defining the number of inner loop or accuracy: This is a good point. In Theorem 4.1, we choose the desired accuracy for inner loop to be $\bar{\epsilon} = \mathcal{O}(\epsilon^2)$. In fact, this can be replaced by an adaptive accuracy, i.e., $\bar{\epsilon}_t = \frac{g(x_0) - g^*}{t+1}$, which will not increase the overall complexity. Hence, the number of inner loops does not necessarily have to be pre-defined.

To Reviewer 4:

We thank the reviewer for the acknowledgement of our contribution. Although Catalyst framework in minimization provide a strong intuition for our work, the analysis for the outer-loop complexity of our framework is an nontrivial extension of their analysis as it requires evaluating solution performance in terms of primal-dual gap (which is much stronger than dual optimality).