We thank all the reviewers and ACs for handling our paper and their constructive comments. Due to space limit, we will focus on responding to the main concerns. Other minor points will be carefully addressed in our next version.

R1: 1) **(9) vs. prox-linear:** We are sorry if our explanation in the submission was unclear. We cannot choose ψ, which comes from (1). But we have the freedom to choose b in (9), which does not depend on problem (1) or (2). If we choose $b(y) := \frac{1}{2}\|y - \dot{y}\|^2$ for any given $\dot{y}$, then (9) becomes $\text{prox}_{\psi/\gamma}(\cdot)$ as explained in line 132 of the submission. In contrast, the prox-linear operator for (2) requires to solve $\min_x\{\phi_0(\tilde{F}_t + \tilde{J}_t(x - x_t)) + \mathcal{R}(x)\}$ (see [28,39]), which does not have closed form solution in general even using proximal operators of ϕ_0 and $\mathcal{R}$. This is due to the composition between ϕ_0 and $\tilde{F}_t + \tilde{J}_t(x - x_t)$. If we use Fenchel conjugate, the dual problem is $\min_z \left\{\mathcal{R}^*(-\tilde{J}_t^T z) + \phi_0^*(z) - \langle \tilde{F}_t - \tilde{J}_t x_t, z\rangle\right\}$, which still does not have closed form solution (see [28,39] for more discussion). Hence, evaluating the prox-linear operator requires solving this complex subproblem (e.g., using primal-dual methods). As a result, the per-iteration complexity of Alg. 1 is better than prox-linear-based methods. Numerical experiments also reveal such a difference.
2) **Why single loop?** As explained, we can choose the quadratic b to have closed form $y^*_{\gamma_t}(\tilde{F}_t) = \text{prox}_{\psi/\gamma_t}(\dot{y} - \gamma_t^{-1} K^T \tilde{F}_t)$. Hence, Alg. 1 is a single loop. Note that variance-reduced algorithms with prox-linear operators, e.g., in [28,39], have double loops regardless of the computation of prox-linear operator. As explained, the prox-linear operator often does not have a closed-form. If we solve it with an additional loop, then these algorithms even have three loops.
3) **Convergence analysis:** Since we exploit the hybrid estimators (14) from [29], Lemma A.2 is indeed adapted from [29]. Lemma B.1 has a similar proof as in [29], but the relation is between Ψ_{γ_t} and $\Psi_{\gamma_{t-1}}$, which requires new result (e) of Lemma A.1. We believe that all other technical proofs are new and do not overlap or recycle from [29]. In fact, Lemmas A.2. and B.1 are not our main results, but Th. B.1 is the key step to prove Th. 3.1 to Th. 3.4. We believe that the proof of Th. B.1. is non-trivial and requires significant technical details and mathematical derivations. Moreover, the adaptive weight β_t is new, and it can remove the initial batches b_0 and $\hat{b}_0$ requirement in Alg. 1 though it sacrifices a $\log$ factor in convergence rate. We will highly appreciate it if the reviewer could take some time to check our proofs.
4) **Title and literature review:** We will adapt the title to make it more precise. Due to the space limit, we did not have much chance to add full literature review. We will add those references and discuss KKT points in our next version.

R3: **Weaknesses concerns:** 1) We believe that our paper has significant novelty compared to previous works. For instance, treating nonstrongly convex ψ with a variance-reduced, single-loop method is new. Analyzing complexity with a single sample, a wide range of mini-batches b, adaptive β_t, and diminishing stepsize is also new.
2) 3) 6) We will certainly implement your suggestions. Due to space limit, we are unable to answer these in detail here. We briefly respond to 2): (a) Hybrid estimator can trade-off variance and bias. Two step-sizes can handle the regularizer $\mathcal{R}$. Obtaining single-loop is due to properties of hybrid estimator (see Lemma A.2) compared to others, like [38].
4) Thank you. Indeed, we were rather selective due to the space limit and certainly missed some references. We were aware of [a], [b], and [d], but since they are deterministic though [d] has small stochastic part, we probably skip them. We will add and discuss them in our next version. The algorithms in [a,b,d] are double or triple loops, more complicated than Alg. 1. The stochastic algorithm in [c] is a single loop but has a worse oracle complexity than ours. These works indeed do not need the PL condition, but they use stronger or different assumptions than ours (see, e.g., (1.2) in [d]).
5) We will elaborate on our comparison in the next version. CIVR uses SARAH (Nguyen et al., 2017) estimator, which has two loops. The details of CIVR in the numerical experiments are given in Supp. Doc. D.
Feedback: 1) We will implement your suggestions and make the paper more readable. 2) Usually, researchers compare stochastic methods via epochs, but we appreciate your suggestion and will add some figures on the loglog scale.

R4: 1) If ψ in (1) is non-strongly convex, then ϕ_0 is nonsmooth, and gradient-based methods are not applicable. Indeed, the smoothing technique is used to make ϕ_0 smooth, but it changes the original problem. This technique is still new when solving complex models (1) or (2). Moreover, we can adaptively update the smoothness parameter γ instead of choosing a tiny value such as $\gamma = \mathcal{O}(\varepsilon)$ as often seen in smoothing techniques (see Th. 3.4).
2) We will adapt the title as suggested.
3) We believe the single loop, constant stepsize, and different batches b are advantages, but we will carefully implement your suggestion. We think that $\mathcal{O}(\varepsilon^{-3})$-complexity is optimal (see [2] for details) for the strongly convex ψ. For nonstrongly convex ψ, our result seems to be the first so far without using prox-linear operator, and achieves the best-known oracle complexity bound. Alg. 1. works with a wide range of batches b, not only one choice as [37,38].

R5: 1) We apologize for missing "Stochastic ... Problems" paper (which is almost concurrent to our work), we will cite it. This paper is indeed similar to [37] but treats a general nonconvex-strongly concave minimax problem. Compared to this paper, Alg. 1 has a single-loop instead of two loops as SREDA in that paper. Alg. 1 still works under a single sample, different mini-batch sizes, and constant and diminishing stepsizes instead of specific mini-batch as in SREDA to achieve such an $\mathcal{O}(\varepsilon^{-3})$ rate. Alg. 1 also tackles the nonstrongly convex ψ, which is new and perhaps harder.
2) We will try to facilitate some parameters as suggested. In theory, we only need b and $\hat{b}_0$, while c_0 and c_1 can be fixed. Other parameters are explicitly defined through b and $\hat{b}_0$ (e.g., (19)). Here, c_1 and c_0 do not depend on any parameter.
3) Indeed, Th.3.1 requires a large initial mini-batch, but it is used only once as opposed to each iteration as in other works, e.g., [37,38,39]. Th. 3.2 does not require a large initial mini-batch (see line 181).
4) Thank you for indicating typos. We will correct them.