We thank all reviewers for their thorough assessment of our paper and suggestions for improvement.

Response to reviewer #1

• ***Experiments, clarity and additional feedback.*** Thank you for your insightful feedback. We agree with you and in the revised paper, we will add an ablation study of RNNs and GRV, implement the combination of benchmarks you suggested, and add more simulation studies to provide a more complete evaluation of our method. We note though that RNNs are not necessarily superior, because the models in other methods can also capture the data generating functions. Finite sample regularization improves the estimation because it adapts all the initial models to solve the nonparametric estimating equation (EE), rather than after the initial models are trained and potentially overfits to the training data. Note that validating causal models is tricky because we do not observe the counterfactuals. We will definitely improve the paper clarity in the revision. As you suggested, we will expand the Appendix to provide a notation table, add several sections introducing the needed background from the various fields such that the paper becomes more accessible and our derivations are easier to follow. Thank you for the excellent suggestions.

• ***Relation to prior work.*** Both GRV and Targeted Regularization (TR) [Shi, et al.] are motivated by solving the EE. However, note that the EIC of DTR in GRV is much more complicated than the EIC of ATE in TR. The EIC of ATE only contains two outcome models and one propensity score model. The EIC of a DTR has many terms as shown in Equation (6-8). This is because the value function estimation involves the DTR, the outcomes, and the treatment assignment mechanism over time. We need to solve the EE by adapting a sequence of outcome and propensity score models over time. There is no direct implementation of TR which can solve the EE across time steps jointly in the DTR learning setting. The GRV regularizer is designed to simultaneously solve the complex EE of DTR learning, which involves models at different time steps; this requires a more technical derivation than TR. In addition, re-estimating nuisance models in DTR learning is more challenging than in ATE estimation. The current related work section gives an overview of the estimator classes and relegates the discussion of the longitudinal TMLE for OPE to the Appendix. In the revision, we will make the connection with TR and TMLE clear in the main body of the paper. Thank you.

Response to reviewer #2

Thank you for the useful suggestions. We agree and we will expand the Appendix to include a discussion of advantages and disadvantages as compared to the other methods, including the running time, stability etc. Thank you.

Response to reviewer #3

• ***Motivations and clarity.*** Thank you for your comments. We agree with you and we will improve the presentation of our method and the adopted architecture in Section 4.3. Due to space limitations, we already relegated some of the discussion about the network architecture to the Appendix. The objective function in Equation (9) has two terms, $\mathcal{L}_t$ and $\mathcal{R}_t$. The term $\mathcal{L}_t$ includes the standard loss function for the outcome regression models and propensity score models in causal inference. In DTR learning, we also need sequential regression to estimate the long-term outcomes; this is why we have the last term in $\mathcal{L}_t$. The regularizer $\mathcal{R}_t$ is motivated by Theorem 1. Minimizing $\mathcal{R}_t$ encourages the fluctuated outcome models to solve the optimal score equation i.e. nonparametric estimating equation, on the R.H.S of Equation (10), so that the resulting estimator is efficient for estimating the value function. In the revised manuscript, we will improve the presentation in Section 4.3 and explain clearly the motivation of each term in our objective function.

• ***Doubly robust (DR) and TMLE.*** The DR estimator achieves efficiency by applying the DR estimating functional on some pre-trained initial outcome and propensity score models. An example of DR estimator is given in Equation (4), where the inverse propensity score (IPS) is multiplied by the cumulated outcome and Q functions. In GRV, the initial estimators are optimized to solve the nonparametric estimating equation jointly with an additional parameter epsilon; this provides additional flexibility to solve the estimating equation. Our estimator in Equation (11) is stable because the IPS is multiplied by a small epsilon. Similarly, it is well known that solving the estimating equation using an additional epsilon in TLME can give a more robust estimator than DR in ATE estimation. The parameter epsilon enables trading-off between finite sample stability and asymptotic efficiency. However, TMLE is not robust when the initial models are not well specified or estimated. This drawback exists even in the simplest setting of ATE estimation and becomes even more severe in DTR learning which involves many outcome and propensity score models. Our method adapts all the initial estimators and the epsilon to solve the estimating equation jointly by regularizing the models in training. By reducing the dependence on the quality of the initial models, our method is thus more robust than TMLE for DTR learning. The connection between our method and TMLE is discussed above Section 4.3 and in the Appendix. We will further improve the discussion and make these points clear in the revision. Thank you.

Response to reviewer #4

Thank you for your useful feedback. In the submitted paper, we have included two simulation studies to evaluate the method performance (in terms of both DTR evaluation and optimization) across various sample sizes. We note that, unfortunately, it is impossible to use offline datasets to benchmark the methods; existing DTR or policy learning papers are also based on simulation studies. We use the MIMIC III dataset for illustration only, to show that the learnt DTR will lead to gradually decreasing the treatment, which corresponds to the clinical literature. The real DTR application can only be demonstrated well through real-world experiments and simulation. In a real-world test-bed, the clinician can check if his/her treatment decision agrees with the DTR before prescribing it. In the revision, we will also explain how such test-beds can be built to evaluate DTR. Thank you.