We would like to thank reviewers for their time and constructive comments. Following are our responses to the raised questions. We start with common questions followed by individual ones.

**Common Q1**: Both **R2** and **R3** question about the claim of the tradeoff between personalization and accuracy. We empirically validated the existence of such tradeoff. For example, by tracking the performance of users in the validation set, we find that the results follow the "80-20 rule", i.e., 80% users whose RMSE results become worse are within top 22.5% over-personalized users. Following reviewers' suggestion, we will include this result into our revised version.

**Common Q2**: Both **R2** and **R3** also question about the applicability of the proposed method. Indeed, we agree with **R2** that the proposed framework is quite generic, and we would like to clarify that the proposed method does not necessarily require a closed form solution such as in matrix factorization. In fact, the proposed framework can be used in many collaborative filtering models as long as the following two theoretical conditions are satisfied. First, the $\bigtriangledown_\Theta L_\Gamma$ term in Eq. (5) can be optimized to (or near) 0 (i.e., the KKT condition). Second, the second partial derivatives of $L_\Gamma$ exist so that we can use the the implicit function existence theorem to obtain Eq. (7). Both conditions can be easily satisfied in many collaborative filtering models including the neural network models mentioned by the reviewers. We plan to extend our idea to such models in our future work.

**Q1 from R1**: **R1** has a concern about overfitting the validation set and the inclusion of a test set. We would like to clarify that this is a misunderstanding and we do have a test set which is separated from both the training set and the validation set. To be specific, as stated in Section 4.1, "we randomly select 80% ratings for training and use the rest 20% ratings for testing". For the training set, we then split it into several folds of debug set and validation set. After the CFDebug finishes training on the training set as illustrated in Fig. 1, we evaluate its performance on the test set. Finally, we repeat the above process five times to do cross-validation. We will make it clearer in the revised version.

**Q2 from R1**: **R1** also concerns about the technical merit of the proposed method. We would like to point out that the key contributions of this work are two fold. First, the data debugging framework that is potentially applicable for a large set of collaborative filtering models (see the response for **Common Q2**). Second, experimental findings and analysis about the tradeoff between personalization and accuracy (e.g., modifying the over-personalized ratings would help improve the overall accuracy).

**Other questions from R1**: Thanks for the advice, we will add more explanations about the C matrix. In Section 4 (D), we did display the movie titles in the case study. We will make it clearer in the revised version.

**Q1 from R2**: **R2** encourages us to experiment on larger datasets. Thank you for the suggestion. We have done experiments on ML-10M. The results also show the effectiveness of the proposed method. For example, with 10% modified ratings, our method gives 2.1% RMSE improvement compared to the original performance (from 0.8336 to 0.8162). We will include these results into our future version.

**Other questions from R2**: As suggested by the reviewer, we will include the top-N studies in the main paper instead of the appendix, and will consider to include the comparison with [7] in our journal version.

**Q1 from R3**: **R3** also asks about the optimal ratio of the results in Table 1 and 2. Empirically, the optimal ratio is around 5%-10% for the studied datasets. We will make this clearer in our revised version.

**Q1 from R4**. **R4** encourages us to analyze the limitation of the proposed method. Thanks for pointing this out. For example, when the ratio of "over-personalized" ratings increases to a certain extent, the overall performance might start to suffer. Much more theoretic work is needed to understand or identify such a theoretic 'transition' point.

**Q2 from R4**: **R4** also suggests using matrix completion view to provide theoretical guarantees. We agree that matrix completion has the potential to provide additional theoretical guarantee.