We thank the reviewers for their time and feedback.

R1: Score 6 (confidence 3)

Q1: "Not novel enough". We disagree. The content of Sec. 3.2 and 3.3 are entirely new. Both R2 and R3 agree.

Q2: "What is the fundamental difference between converting whole network vs only the last layer"? In the beginning of training, weights are changing frequently from task to task, therefore their uncertainty is high. Using only the last layer ignores the uncertainty in the rest of the weights. This could hurt performance a lot in the beginning. We will add more explanation in the paper, and include a small illustrative experiment in the Appendix.

Q3: "What role does the ... regularization term play ... compared with FRCL"? This term ensures that the current DNN outputs (mean) after task t remain close to the DNN after task $t-1$, which is desirable to avoid forgetting. In contrast, FRCL only optimise the 'variance information' to be close, which may not be as effective (see line 44).

Q4: "Is it possible to do task detection?" Yes, it is possible to do this in the same way as FRCL paper.

Q5: "Does sampling according to Λ really makes sense?" Yes, this is closely related to Kernel methods, such as kernel ridge-regression and SVM, where this strategy enables us to pick boundary/high-leverage points. This also leads to an intuitive and computationally cheap method (both R2 and R3 agree).

Q6: "FROMP does not outperform FRORP"... Is random sampling enough? It is not enough, and a careful selection is required when memory size is small. See Split CIFAR results in Figures 3b and 3c. On MNIST, FRORP is similar to FROMP because the task is very simple, and very little is gained from a careful selection of memorable examples. We will add results for 11 tasks on Split CIFAR where we expect to see even more of a difference.

Q7: The complexity of computing Λ. It costs $O(N)$. We will clarify this in the paper.

R2: Score 7 (confidence 3)

Q8: Reliability under distribution shift: This should not be an issue as long as the memory is large enough.

Q9: "How does this compare to simple gradient based sample ranking schemes (e.g. Deep Batch Active learning)". Very good point. Out method turns out to be similar to 'Conf' selection, but we arrived at it with a different approach. We will add this reference and add a discussion in the paper.

Q10: Include experiment with 10s of tasks. Yes, we will add this. Also see our answer to Q14.

R3: Score 8 (confidence 4)

Q11: Add Forward and Backward transfer to all results. Yes, we will add it.

Q12: "Add clarifying assumptions earlier to help categorise this work". Yes, we will add it (around line 73).

Q13: Broader review of recent work in Continual Learning. We agree and will improve it. Due to page limit, our Related Work section is short. For camera-ready, we can devote around half a page to include many other related works.

Q14: Ready to raise score if a challenging and convincing CL result is added. Thanks for the suggestion. Since our main contribution is to propose a new method, we focused on standard benchmarks. We will add the Split CIFAR experiment with 11 tasks (previous papers used only 6 tasks), and the "predictive uncertainty for change-point detection" in the Appendix. Adding many more challenging experiments takes more time and space, and we will explore scalability (e.g., large classes and datasets) and generality (e.g. in RL and Bandits) of the method in a separate future work.

Q15: Behaviour as number of memorable points increases. Good question. The performance saturates as expected. It is not clear how to find the number of examples beforehand, but one could easily design an online, greedy scheme to determine this number.

R4: Score 5 (confidence 4)

Q16: "Limited contributions". We disagree. The content of Sec. 3.2 and 3.3 are entirely new. Both R2 and R3 agree.

Q17: Discuss more related approaches other than regularization ones. Thanks. We will fix this (including adding all the references mentioned, and a comparison with GEM). Also, see Q13.

Q18: "Is optimization constraints violated?" This is a misunderstanding. There are no constraints in our approach.

Q19: Gap between FRORP and FROMP. This is a misunderstanding. Please see our answer to Q6.