We appreciate the reviewers' time and suggestions! We address them all and report new experimental results below.

*Reviewer 1*: • Missing citations:... We will cite the suggested citations and discuss their differences/relations with our method. Although DIH can be helpful to identify noisy data in noisy-label setting (ref.Middle plot in Fig. 1), our curriculum is not specifically designed for noisy labeled data and this paper mainly focuses on clean data setting. DIHCL still achieves 90.34% test-set accuracy under 40% symmetric label noise on CIFAR10 (ref.Top plot in Fig. 1).

• The statement may be revised that "updating instantaneous hardness typically requires extra inference steps of a model over all the samples" ... extra GPU memory... SPADE Alg in [ref2] samples a mini-batch and then selects samples within the mini-batch using the MentorNet. Comparing to DIHCL, SPADE incurs the following extra computational costs: (1) the feature extraction on the sampled mini-batch; (2) training of the MentorNet; (3) repeated training on well-learned and easy samples with correct labels. In addition, DIHCL can be incorporated with [ref2] as a predefined curriculum to train the MentorNet more efficiently. DIHCL requires extra GPU memory linear in the number of samples, which is negligible compared to the GPU memory for network training. We will add a discussion of the memory cost in the next version.

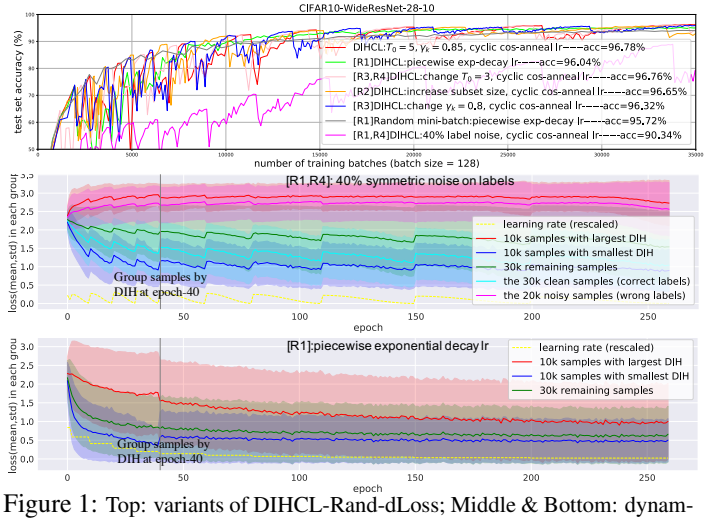

Figure 1: Top: variants of DIHCL-Rand-dLoss; Middle & Bottom: dynamics of DIH-grouped samples under label noise and another lr schedule.

• Is the method specific to cyclic learning rate... DIHCL is applicable to other learning rate schedules. We report the result of DIHCL with a piecewise exponential decay learning rate in Fig. 1. DIHCL improves the test accuracy from 95.72% to 96.04% in this case. We also visualized the dynamics of samples partitioned by DIH at epoch-40 and it shows that the properties in Section 2 also hold for different learning rate schedules.

• Clarity We will simplify the experiment part and use better names for the DIH variants.

*Reviewer 2*: • Compare more baseline methods... Given the limited time for rebuttal, we compared DIHCL with [2]-[4] in Table 1 on CIFAR100 when used to train WideResNet-28-10. We will add a complete comparison in the next version.

• "Increasing the subset size (previous studies)" vs "reducing the subset size (as proposed)" It depends on one's preference for easy v.s. hard samples. If the curriculum always selects the easiest samples (e.g., SPL), the former should be used since only an increasing size can include harder but more informative samples in later stages. We reduce the subset size in DIHCL for two reasons: (1) in early stages, training on sufficient samples yields an accurate estimate of DIH since it is a time-moving average; (2) in later stages, we can reduce unnecessary training costs on easy/memorized samples, as suggested by observations in Section 2. In Fig. 1, we provide a comparison between the two and it shows that the former is slightly worse than the latter on the final test accuracy.

Table 1: [R2]Test accuracy (%) of WideResNet-28-10 on CIFAR100.

| | |
|---|---|
| DIHCL | 82.23 |
| [2] | 75.04 |
| [3] | 71.95 |
| [4] | 76.28 |

• Need variance for in Table 1. Given limited time, we report the mean±variance over 5 trials on two datasets in Table 2. We will add complete variance results in the next version.

Table 2: [R3]Test Acc (mean±variance).

| Curriculum | CIFAR10 | CIFAR100 |
|---|---|---|
| DIHCL-Rand, Loss | $96.74 \pm 0.04$ | $80.80 \pm 0.16$ |
| DIHCL-Rand, dLoss | $96.75 \pm 0.06$ | $80.73 \pm 0.21$ |
| DIHCL-Exp, Loss | $97.07 \pm 0.11$ | $82.31 \pm 0.24$ |
| DIHCL-Exp, dLoss | $96.44 \pm 0.10$ | $81.35 \pm 0.27$ |
| DIHCL-Beta, Flip | $96.48 \pm 0.04$ | $81.13 \pm 0.18$ |

• SOTA results with all techniques. We will try to compete with the baselines in the SOTA setting in the next version.

*Reviewer 3*: • Effects of $T_0$ and $\gamma_k$... $T_0$ is necessary to get an early estimate of DIH and is supported by the theoretical analysis in the Appendix. But a small $T_0 = 3$ suffices to get a stable training because: after the $T_0$ warm-starting epochs, we start from all samples (exploration) and gradually reduce the subset size. In Fig. 1, we compare $T_0 = 5$ and $T_0 = 3$: they produce similar test accuracy. We also compare $\gamma_k = 0.85$ and $\gamma_k = 0.8$: reducing the subset size too fast ($\gamma_k$ is too small) will degrade the accuracy.

*Reviewer 4*: • $T_0$ and hyper-parameter tuning. Please also see our reply to Reviewer 3. DIHCL is not sensitive to the choice of $T_0$ since the earlier epochs after warm starting select almost all samples (exploration), which keeps the DIH's estimation accurate. We do not have computation for a full grid search of all hyperparameters so there could exist better choices. Line 307-312 details how we selected the hyperparameters.

• Method sensitive to noise. The noisy-label experiments in the Appendix use a 100% noise setting, i.e., all the labels are randomized and wrong. The purpose is to show that the pattern of DIH is very different on clean and noisy data, so we can use DIH as an indicator of label noise. In Fig. 1, we report the results under 40% symmetric label noise. It shows (1) DIH contains critical information to identify the noisy-labeled data, and (2) DIHCL is robust to label noise and achieves 90.34% test-set accuracy under 40% symmetric label noise on CIFAR10.