A. Training Details

We run each of our experiment over five trials. Each of these five trials uses the same random seed across all methods so that both the model weights and dataset split are initialized the same way. As methods often use different learning rates, we search over learning rates $\{0.1, 0.05, 0.025\}$, centered around the default of 0.05 in [29], choosing the run that yields the highest mean top-1 accuracy on the validation set across the five trials. The batch size was set to 64 (and 32 for methods that sampled pairs) in training and the temperature parameter in the KD loss was set to 4 as in [29]. We perform learning rate decay three times for each method as with decay rate set to 0.1 as in [29], conditioned on early stopping convergence with patience 50. When continuing training after learning rate decay, we resume from the model with the highest validation accuracy previously. We use the SGD optimizer with a momentum of 0.9 and weight decay of 5×10^{-4} for all experiments. Most baseline methods also include a hyperparameter α that controls the weighting of the original KD loss and their introduced loss term [21, 29]. To be consistent with these prior works, we search over four values of α: $[0, 0.5, 1, 2]$, selecting the best performing baseline on the validation set. We perform this search *only for the benefit of the baselines*; we do not search over any weighting for our method, though future work may find it beneficial to do so.

B. Extending Number of Teacher Calls

Extending the number of teacher calls beyond our few teacher call setting, we can see from Figure 6 that the CKD accuracy converges with that of other baselines at around 16000 teacher calls. Even at larger number of teacher calls closer to the size of the full dataset, CKD still performs quite well among state-of-the-art methods.

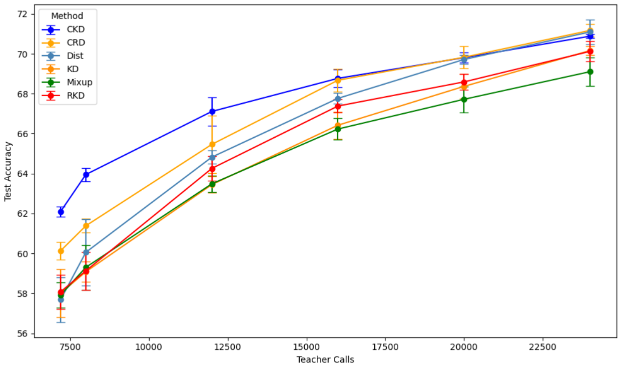

Figure 6. Experimental results on CIFAR-100 when distilling WRN-40-2 to WRN-16-2 using larger numbers of teacher calls. Points and error bars are the mean and standard deviation of runs over five trials.