We sincerely thank all reviewers for their constructive comments and positive support. In particular, we respectively ask R3 to reconsider his/her evaluation since we can address all the comments, as detailed below.

Q1 (R1): Will the estimation error for $T^{\clubsuit}$ be zero if the anchor points are estimated, if so why? And ablation study.

A1: There will be an estimation error for $T^{\clubsuit}$ if there is an estimation error for the anchor points. The same estimation error for the anchor points will also go for the T-estimator. For simplicity, we study the setting where anchor points are given. In this case, the error only comes from estimating $T^{\spadesuit}$. Thus we do not need to do ablation study for the errors of the two matrices $T^{\clubsuit}$ and $T^{\spadesuit}$. Note that even given anchor points, there is an estimation error for the T-estimator because of estimating the noisy class posterior while there is no estimation error for $T^{\clubsuit}$ because the intermediate class posterior is designed to be given. Considering how the estimation error for anchor points affects the two estimators is also very interesting. We leave it as a future work.

Q2 (R1): Will intermediate class labels obtained ... have an estimation error that will affect the error of both matrices?

A2: The intermediate class labels are defined by the ones that maximize the intermediate class posteriors. Since we have given the true intermediate class posterior distribution, i.e., $P(Y'|\boldsymbol{x}) = \hat{P}(\bar{Y}|\boldsymbol{x})$, there will be no estimation error for the intermediate class labels.

Q3 (R1): On the synthetic dataset, dual T-estimator performs better ... Is there an explanation on why this happens?

A3: On the real-world datasets, the estimation error of the dual T-estimator can be larger than that of the T-estimator only with small training sample sizes. It is because the number of images per class is too small to estimate the transition matrix $\hat{T}^{\spadesuit}$ which can be sparse and lead to a large estimation error. We will discuss this in the final version.

Q4 (R3): The justification and explanation of Eq. 3 is not clear. $\hat{P}(\bar{Y}|Y')$ (which is computed using the validation set).

A4: We will add more discussions about Eq.3. It discusses how to make $T^{\spadesuit}$ to be independent of Y because Y is unavailable. Specifically, we have explained a sufficient condition for letting Eq. 3 hold, i.e., let the intermediate class labels be identical to noisy labels. We have also discussed that the condition may be hard to be satisfied, therefore, an estimation error for fitting the noisy labels to intermediate class labels Δ_3 is introduced to our estimation. In Appendix 2, we have also empirically validated the estimation error. In line 193, we have stated that $\hat{P}(\bar{Y}|Y')$ is NOT computed on the validation set. It is computed on the training set. The validation set functions as the same role for both the proposed method and the baselines for model selection. Your concerns on the unfair training/validation split may not stand. However, we agree that to see how the size of validation set influences the performance is interesting. We have redone experiments according to your suggestions to reduce the size of validation set, i.e., using 10% of the training examples as a validation set for the both estimators with five repeated trials (random seeds are from 1 to 5). The estimation errors for transition matrices are illustrated in the following table, which show that the dual T-estimator still significantly outperforms the T-estimator.

	MNIST			*F-MNIST*			*CIFAR-10*			*CIFAR-100*		
	Sym-20%	Sym-50%	Pair-45%	Sym-20%	Sym-50%	Pair-45%	Sym-20%	Sym-50%	Pair-45%	Sym-20%	Sym-50%	Pair-45%
T	0.333±0.002	0.459±0.001	0.644±0.069	0.261±0.012	0.428±0.019	0.687±0.097	0.383±0.004	0.772±0.077	0.861±0.028	0.389±0.001	0.760±0.117	0.894±0.002
Dual T	**0.106**±0.003	**0.271**±0.007	**0.298**±0.021	**0.220**±0.014	**0.315**±0.012	**0.414**±0.020	**0.215**±0.011	**0.267**±0.075	**0.578**±0.026	**0.255**±0.020	**0.516**±0.187	**0.835**±0.014

Q5 (R3): Why use T-Revision models of poorer classification accuracy in the Dual T-estimator paper?

A5: This could be a common concern if we focus on improving the classification accuracy. However, our aim is to reduce the estimation error for the transition matrix. The experiments are set to FAIRLY verify the effectiveness of the Dual T-estimator. As mentioned by the reviewer, the settings of our experiments are different from the original paper, thus the reported accuracy is different. Specifically, to boost the classification performance, different tricks has been employed in the baselines, e.g., transition information is given (e.g., Co-Teaching); choosing different hyper-parameters to estimate anchor points (e.g., using the 97% and 100% largest estimated noisy class posteriors for *CIFAR-10* and *CIFAR-100*, respectively, in Forward); training networks with different epochs to estimate noisy class posteriors (e.g., 40 epochs in Forward while 20 epochs in T-revision on *mnist*); with or without using the validation set to estimate noisy class posteriors (e.g., T-revision uses a validation set while Forward does not). To fairly show the superiority of Dual T, all the baselines in our paper use the same tricks, e.g., using a noisy validation set to choose model for estimating the noisy class posterior and using the largest noisy class posteriors for estimating anchor points. So the classification performance would be worse than that in the original paper. However, the effectiveness of the Dual T-estimator has been successfully illustrated. We will add the discussions in the final version.

Q6 (R4): There is little discussion about the impact of (1) in the paper. e.g. will different types of the classifiers, which have different capabilities to fit the noisy labels, influence the performance of the estimator and how large?

A6: Thanks for the insightful question. We will add more discussions. For example, the estimation error for T-estimator is from estimating the noisy class posterior in (1); the estimator error for Dual T-estimator (or $T^{\spadesuit}$ because $T^{\clubsuit}$ is noise free) is also related to (1) because to push the intermediate class Y' to be close to the noisy class $\bar{Y}$. Studying how different classifiers influence the performance of the estimator via impacting (1) is interesting. In our paper, we assume that Δ_1, which is the estimation error for the noisy class posterior, is bigger than Δ_3, which is the estimation error for pushing the intermediate class Y' to be equal to the noisy class $\bar{Y}$. Note that if $\Delta_1 \geq \Delta_2 + \Delta_3$, the dual T-estimator will outperform T-estimator, where $\Delta_2 = |P(\bar{Y}|Y') - \hat{P}(\bar{Y}|Y')|$ is the error introduced by counting the noisy and intermediate class labels. We discussed when this will hold and empirically validated this assumption in the supplementary material. We agree with the reviewer that it is interesting to study when the assumption is invalid and how different types of classifiers influence it. We would add some empirical study in the supplementary material.