First of all, we would like to thank all the reviewers for their precious comments and we will address all the questions.

To Reviewer #1: Q1: The relationship between $do()$ operation and conditional probability. The backdoor adjustment implementation of $do()$ in Eq.(3) can be considered as the "passive" intervention through observation, instead of a "physical" one. Its detailed derivation is given in [*Causal inference in statistics: A primer, Judea Pearl et al., 2016*], which is essentially a Markov factorization for the graph with broken $D \rightarrow X$. Therefore, their conditional probabilities (or factorization) are not the same. **Q2: The explanations of $f()$ and $g()$.** The numerator $f()$ is the effect of x, *i.e.*, the prediction logits. Since we use a fully connected layer without the bias term as our classifier, it equals to $(w_i)^\top x$. The denominator $g()$ is the propensity score [*An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies, Austin Peter C et al., 2011*], which is a balancing score used to normalize each effects. It can take a variety of forms, like $l2$-norm or capsule-norm. The proposed $\|x\| \cdot \|w_i\| + \gamma\|x\|$ is inspired by the capsule-norm ($\|x\| \cdot \|w_i\| + \|w_i\|$) [42]. We changed the $\|w_i\|$ to $\gamma\|x\|$, because in our causal graph, the effect needs to be normalized by both class-specific and class-agnostic energies of x. **Q3: The range of α.** α is a linear trade-off parameter between the direct and indirect effects. Its range is not limited to $[0, 1]$. **Q4: The scope of Assumption 1.** According to our recent studies in Table 1, the proposed method works well on different imbalance ratios. When the imbalance ratio decreases from 100 to 10, the improvements achieved by TDE start to converge but **not collapse**, because when the dataset is more balanced, the second term of TDE in Eq.(8) is closer to a uniform distribution that affects the prediction less (see Supp Section A).

To Reviewer #3: Q1: Additional experimental comparisons with other methods. The ImageNet-LT and LVIS are the most challenge datasets in long-tailed classification from the perspective of scale and size of vocabulary. Since BBN, LDAM and class-balanced loss didn't reported their results on these datasets, we omitted some comparisons in the original paper. After applying the proposed De-confound-TDE in Long-tailed CIFAR-100/-10, we consistently outperform these previous methods in Table 1. **Q2: The refinement of related works & adjusting some representations in the paper.** Thank you for your advice, we will address these issues in the later revision. **Q3: Reproducibility.** More details can be found in our supplementary codes. The project will be released to Github upon acceptance.

Dataset	Long-tailed CIFAR-100			Long-tailed CIFAR-10		
Imbalance ratio	100	50	10	100	50	10
Focal Loss [24]	38.4	44.3	55.8	70.4	76.7	86.7
Mixup [*Hongyi Zhang et al., ICLR, 2018*]	39.5	45.0	58.0	73.1	77.8	87.1
Class-balanced Loss [11]	39.6	45.2	58.0	74.6	79.3	87.1
LDAM [*Kaidi Cao et al., NeurIPS, 2019*]	42.0	46.6	58.7	77.0	81.0	88.2
BBN [9]	42.6	47.0	59.1	79.8	82.2	88.3
(Ours) De-confound	43.9	48.9	59.5	72.5	78.7	88.1
(Ours) De-confound-TDE	**47.3**	**51.2**	**59.8**	**80.4**	**83.1**	**89.4**

Table 1: **Top-1 accuracy** on long-tailed CIFAR-10 and CIFAR-100 with **different imbalance ratios**. All models are using the same ResNet-32 backbone. Note that we report accuracy rather than error rate (in BBN) for consistency.

To Reviewer #6: Q1: Details about the multi-head strategy and the selection of K. The multi-head strategy means dividing the feature channels into K groups (each has $1/K$ of original dimensions), so it can be considered as sampling multiple independent feature spaces. However, our algorithm is not sensitive to K. As we can see from Supp Table 2 & 3, even when K is set to 1, the proposed De-confound-TDE still outperforms the other methods, which proves that we didn't unfairly take the advantage of multi-head strategy.

To Reviewer #7: Q1: Additional experimental comparisons with other methods. Please refer to our answer of **Reviewer #3 Question 1**. **Q2: A fair comparison without multi-head strategy & the reason of introducing K.** As we discussed in **Reviewer #6 Question 1**, we tested $K = 1$ in Supp Table 3, which shows consistent advantages of the proposed De-confound-TDE over previous methods (in original paper Table 2) even without multi-head strategy (*i.e.*, $K = 1$). Besides, introducing K is also part of our theoretical framework, because sampling multiple feature spaces provides better estimation of the effect. **Q3: The novelty of the proposed de-confounding training and the potential unfair advantage of γ.** Compared with the previous Capsule Norm: $\|x\| \cdot \|w_i\| + \|w_i\|$, we change the additional normalization term $\|w_i\|$ to $\gamma\|x\|$, because the effect needs to be normalized by both class-specific and class-agnostic energies of feature x rather than w in our causal graph. Meanwhile, our algorithm is not sensitive to the selection of γ based on Supp Table 2. More importantly, the same value of $\gamma = 1/32$ can be transferred from ImageNet-LT to LVIS and CIFAR-100/-10-LT, which proves that we didn't use different γ to overfit these datasets. **Q4: The selection of SGD momentum parameter and extending to other optimization methods.** We haven't systematically studied the selection of SGD momentum parameter yet, since it mainly affects the norm of $\bar{x}_T$ while our Assumption 1 only uses its unit direction $\hat{d} = \bar{x}_T / \|\bar{x}_T\|$. As to the other optimization methods, we found that almost all of them contain the similar moving averages of gradient the same as SGD momentum, although they may have different names and symbols, *e.g.*, betas in Adam. Therefore, the proposed causal graph works for them as well. We will follow your suggestion to offer a thorough study on a wider spectrum of optimizers in future.