We thank the reviewers for insightful and constructive comments. We have submitted **code** and **detailed Appdendix** .

Common Question Q1: The covariate shift assumption.

A1: Thanks for reviewers pointing out the covariate shift assumption of this paper. As a fundamental assumption of TransCal, it is inadvertently omitted by us while writing. We will *explicitly state* it in the future version, and discuss the *relevant papers* on covariate shift and (generalized) target/label shift to make the literature review more complete.

Common Question Q2: Will TransCal have a lower accuracy while achieving a better calibration?

A2: As a post-hoc method that softens the overconfident probabilities but *keeps the probability order over classes*, TransCal maintains the **same accuracy** with that before calibration, while achieving a lower ECE (Fig. 1(b)).

R1.1: Whether Eq. (5) can be termed as a bias?

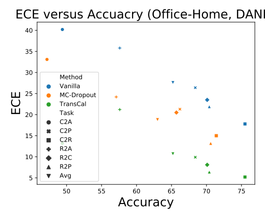
Realizing the gap between the importance weights estimated by LogReg [38, 1, 5] and the (*unknown*) ground-truth ones, we proposed to control the bound M of the weights to reduce the overall estimation error. Further, as reported in Line 235, we ran each experiment 10 times with different sampling data to mitigate the problem of random sampling.

R2.1: The advantage of the adopted post-hoc approaches over the built-in methods, *e.g.* MC-dropout.

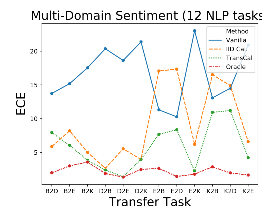
TransCal maintains the **same accuracy** with that before calibration while built-in methods (*e.g.* MC-dropout) may *degrade* prediction accuracy (Fig. 1(b)), and they have to modify the network architecture (*e.g.* adding dropout layers).

Calibration Method	A→C	A→P	A→R	C→A	C→P	C→R	Avg
Before Cal. (Vanilla)	40.2	26.4	17.8	35.8	23.5	21.9	27.6
IID Cal. (MC-dropout)	33.1	21.3	15.0	24.2	20.5	13.2	18.8
IID Cal. (Matrix Scaling)	44.7	28.8	19.7	36.1	25.4	24.1	29.8
IID Cal. (Vector Scaling)	34.7	18.0	11.3	23.4	15.4	11.5	19.4
IID Cal. (Temp. Scaling)	28.3	17.6	10.1	**21.2**	13.2	8.2	16.4
TransCal (ours)	**13.2**	**9.9**	**5.2**	**21.2**	**8.1**	**6.4**	**10.7**

(a) ECE (%) on *Office-Home* for DA method CDAN



(b) ECE vs. Accuracy



(c) *Multi-Domain Sentiment*

R2.2: Why the proposed new Calibration Metric is reasonable?

Among the three typical calibration metrics, BS *conflates* accuracy with calibration and NLL may *over-emphasize* tail probabilities [31], thus we proposed TransCal based on the intuitive and informative one: ECE (Paragraph at Line 149).

R2.3: Why we use the control variate method of [22] instead of the various approaches?

As a *non-intrusive* and *parameter-free* method, control variate is the mainstream, simple and effective variance reduction method. Besides, we further developed *serial* control variate method backed by a theoretical analysis in B.2 of *Appendix*.

R2.4: How will TransCal perform on the source prediction? The calibration result of the source-only model.

TransCal performs well on source prediction and source-only model (ECE decreases $\sim 20\%$ than that before calibration).

R3.1: Experiments on NLP datasets.

TransCal performs well in 12 transfer tasks of a popular NLP dataset: *Amazon Multi-Domain Sentiment* (Fig. 1(c)).

R3.2: The missing experimental analysis on performance of applying the proposed target calibration method.

See common question Q2. We believe there is no need to report the **same accuracy** before and after calibration.

R3.3: There seems to be an error in the derivation of the bias reduction method.

We use LogReg to estimate density ratio from a logistic regression classifier that separates examples from the source and target domains as in Eq. (4). We clarify that $q(x) = 1 - p(x)$ below Line 182 is the output of LogReg, indicating the probability of the target domain that x belongs to. Notations in Line 182 will be updated to avoid such *misunderstanding*.

R3.4: Minor issues on related works (CPCS elaboration), typos, grammar and formally stated algorithm.

Thanks for the valuable suggestions from the reviewer. We will *fully* address these minor issues in the future version.

R4.1: This paper focuses only on the simplest setting of confidence calibration.

As the first transferable calibration work for Domain Adaptation (DA), we adopt the fundamental and mainstream confidence calibration. Thanks for your valuable suggestion, pointing out our future work on more complex settings.

R4.2: The results of calibration methods with vector scaling/matrix scaling.

Both Vector Scaling and Matrix Scaling underperform TransCal and Temp Scaling (Table 1(a)). Matrix Scaling works even worse than the Vanilla model due to overfitting, which was also observed in the results of Guo *et al.*, [16] (Table 2).

R4.3: Will a strong IID calibrator be preferable than a weak transferred calibrator?

Besides the result of IID calibrator Temp Scaling given in Table 2, we add the results of competitive IID calibrators, *e.g.* Vector Scaling, Matrix Scaling and MC-dropout (Table 1(a)). They all underperform TransCal in the Non-IID setup.

R4.4: Will the results be different on another evaluation metric, *e.g.* maximum ECE, accuracy, Brier Score?

See common question Q2 about evaluating on accuracy. The results on Brier Score, NLL and Reliability Diagrams were already given in D.2.5, D.2.4 and D.3 of *Appendix*. They *consistently* demonstrate the efficacy of TransCal.