A Additional Results

A.1 Reliability Diagrams

The *reliability diagram* is an intuitive visualization of the empirical calibration error (ECE) in (9) (DeGroot and Fienberg, 1983; Niculescu-Mizil and Caruana, 2005). In particular, the diagram shows the averaged predicted uncertainty (*i.e.,* $\frac{1}{|\mathcal{T}_b|}\sum_{(x,\mathbf{y})\in\mathcal{T}_b}\hat{f}(x)_{f(x)}$) on the x-axis, and the empirical accuracy (*i.e.,* $\frac{1}{|\mathcal{T}_b|}\sum_{(x,\mathbf{y})\in\mathcal{T}_b}\mathbf{y}_{f(x)}$) on the y-axis. Thus, if bars in the reliability diagram aligns with the diagonal, the ECE of the forecaster in consideration is zero. If the bars are below the diagonal, then the forecaster is over-confident on its uncertainty predictions. Along with the accuracy and confidence plots, each bar is weighted by the fraction of examples in each bin b (*i.e.,* $\frac{|\mathcal{T}_b|}{|\mathcal{T}|}$), which is also reflected in the ECE in (9).

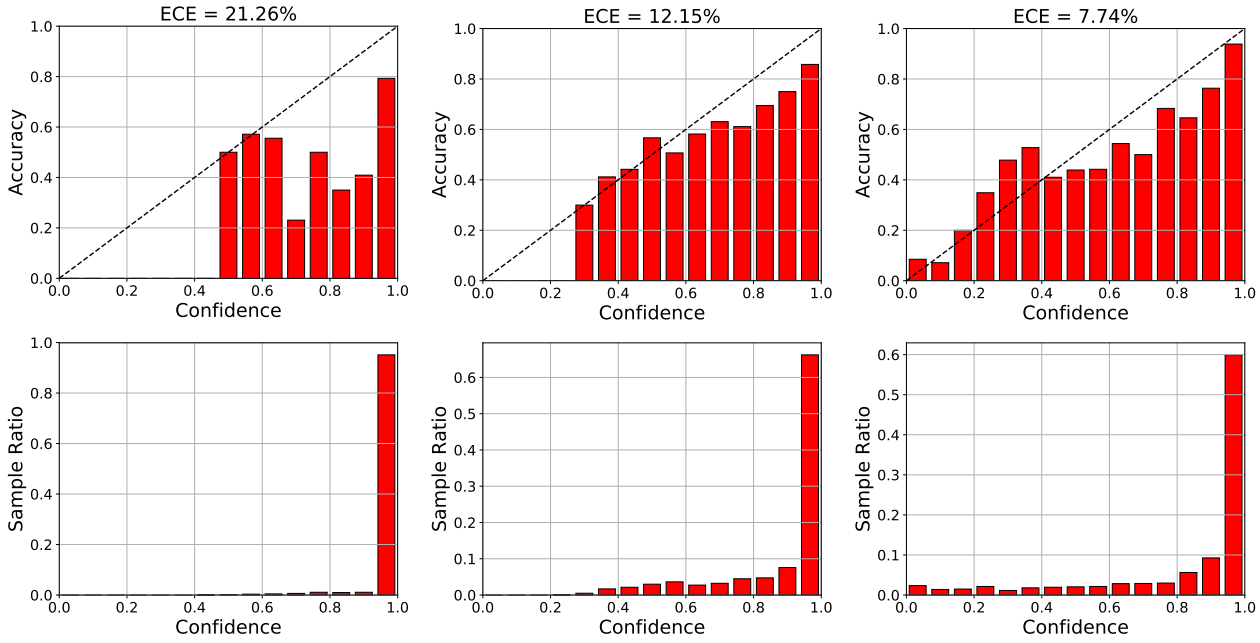
We compare the reliability diagram for the temperature scaling approach and for our approach in Figure 3, Figure 4, Figure 5, Figure 6, Figure 7, Figure 8, and Figure 9.

A.2 Running Time

As with existing approaches to calibrated prediction (Guo et al., 2017), our approach relies on a second phase of training to calibrate the predicted probabilities. We measure the overhead of our calibration step compared to the time used to train the indistinguishable feature map. The results are:

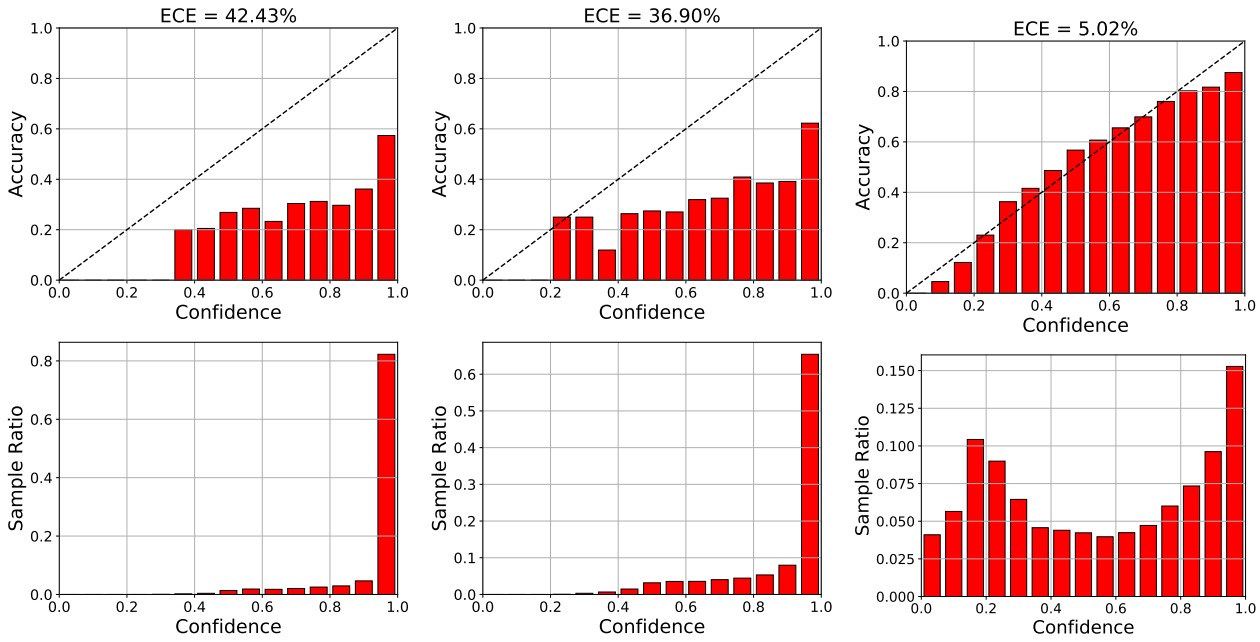
- $\mathcal{M} \to \mathcal{M}$: 7840.8589 sec. (our overhead) vs. 1324.9306 sec. (total)
- $\mathcal{U} \to \mathcal{M}$: 14939.4707 sec. (our overhead) vs. 1001.3495 sec. (total)
- $\mathcal{M} \to \mathcal{U}$: 16900.7378 sec. (our overhead) vs. 1217.413 sec. (total)
- $\mathcal{S} \to \mathcal{M}$: 16437.2544 sec. (our overhead) vs. 581.6673 sec. (total)
- $\mathcal{M} \to \mathcal{S}$: 4437.1197 sec. (our overhead) vs. 1460.2213 sec. (total)
- $\mathcal{A} \to \mathcal{W}$: 16068.9053 sec. (our overhead) vs. 907.5134 sec. (total)
- $\mathcal{D} \to \mathcal{A}$: 9576.0645 sec. (our overhead) vs. 9015.7301 sec. (total)
- $\mathcal{W} \to \mathcal{A}$: 18602.4316 sec. (our overhead) vs. 7217.6859 sec. (total)

In all but one case, the overhead from calibration is less than 1/4 the total time taken (*i.e.,* for both calibration and indistinguishable feature learning). This overhead is reasonable for obtaining calibrated probabilities.



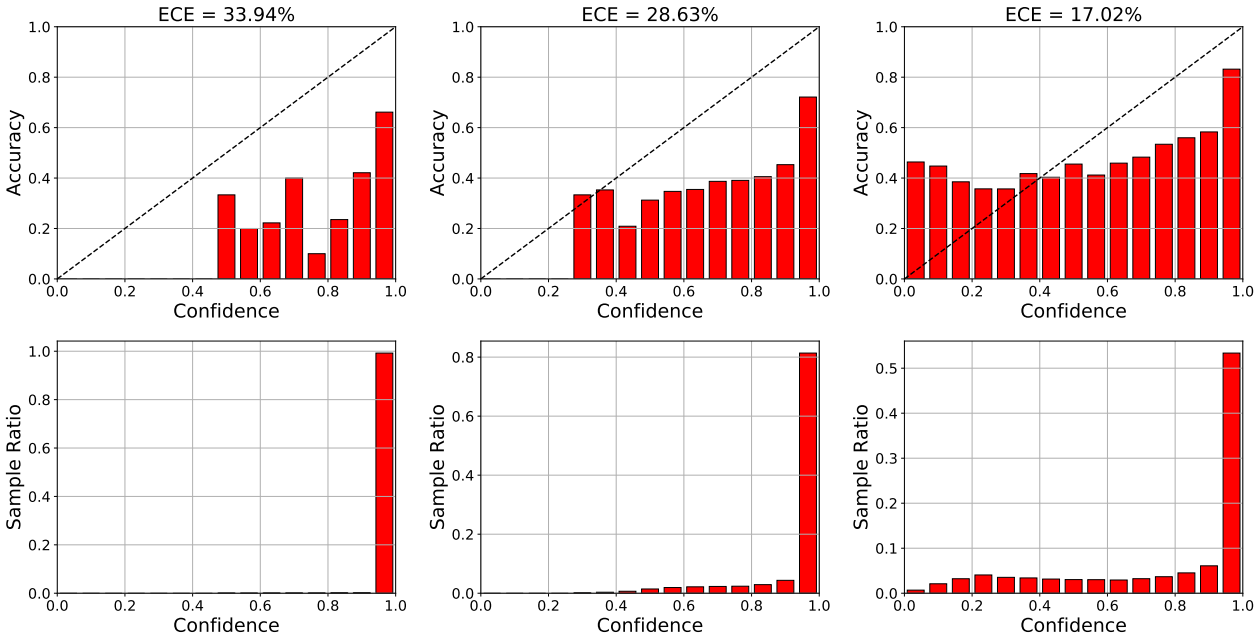
(a) None (b) Temp. scaling (c) Temp.+FL+IW

Figure 3: Reliability diagram of the shift $\mathcal{M} \to \mathcal{U}$ from one experiment among ten.



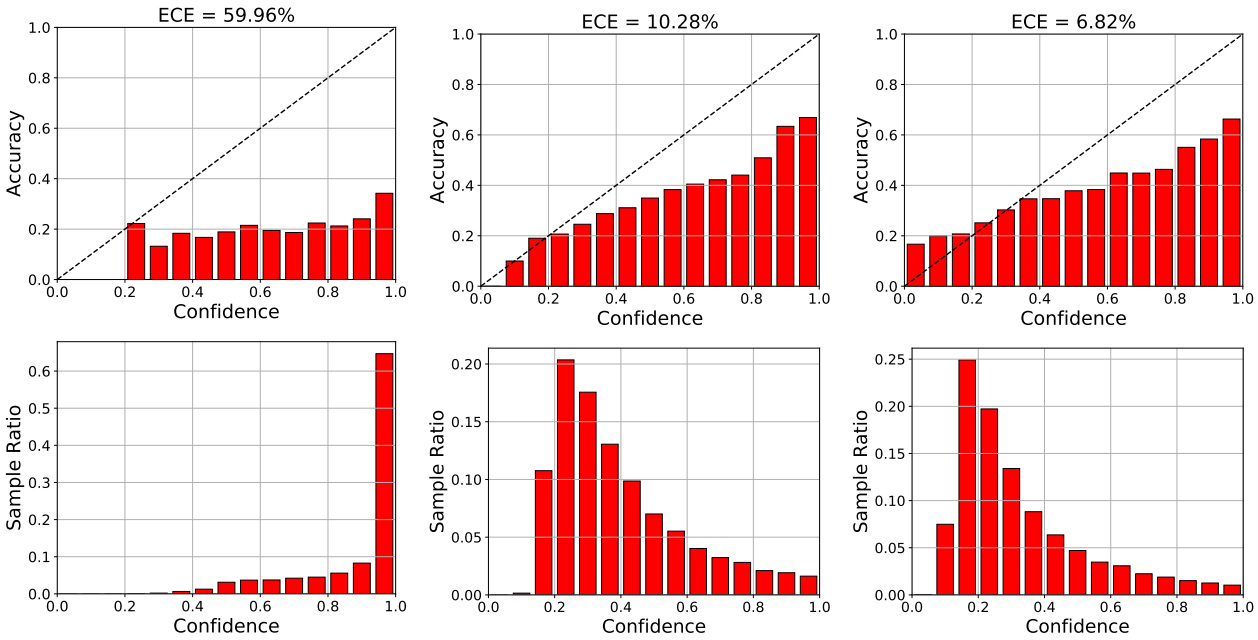
(a) None (b) Temp. scaling (c) Temp.+FL+IW

Figure 4: Reliability diagram of the shift $\mathcal{U} \to \mathcal{M}$ from one experiment among ten.



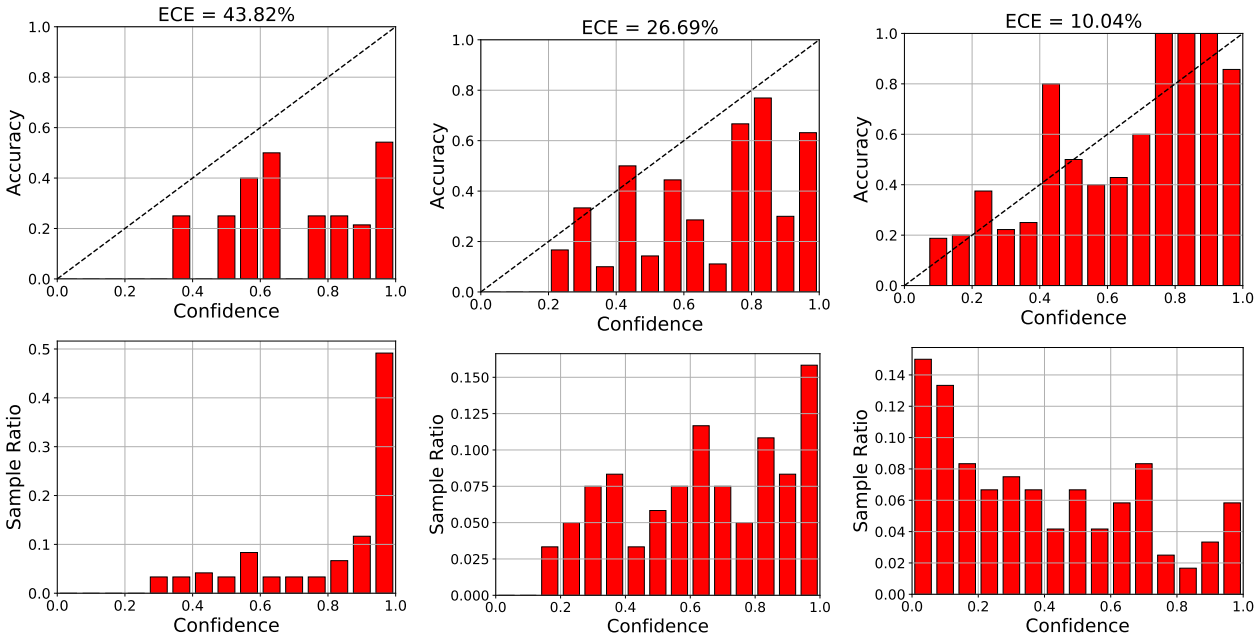
(a) None (b) Temp. scaling (c) Temp.+FL+IW

Figure 5: Reliability diagram of the shift $\mathcal{S} \to \mathcal{M}$ from one experiment among ten.



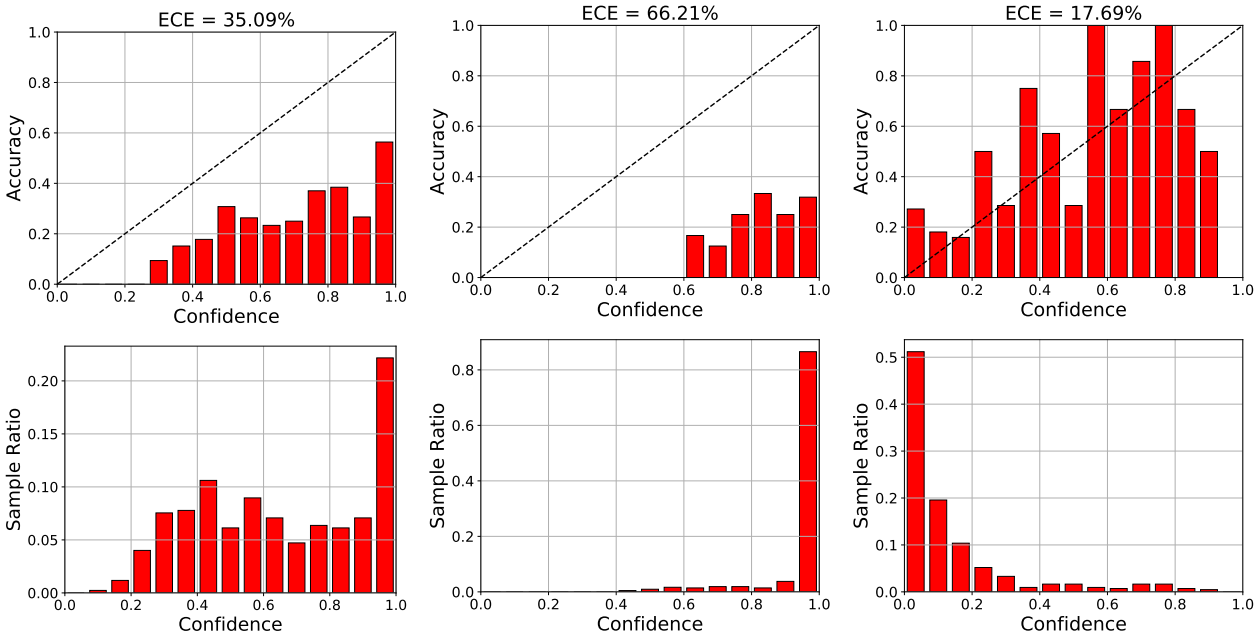
(a) None (b) Temp. scaling (c) Temp.+FL+IW

Figure 6: Reliability diagram of the shift $\mathcal{M} \to \mathcal{S}$ from one experiment among ten.



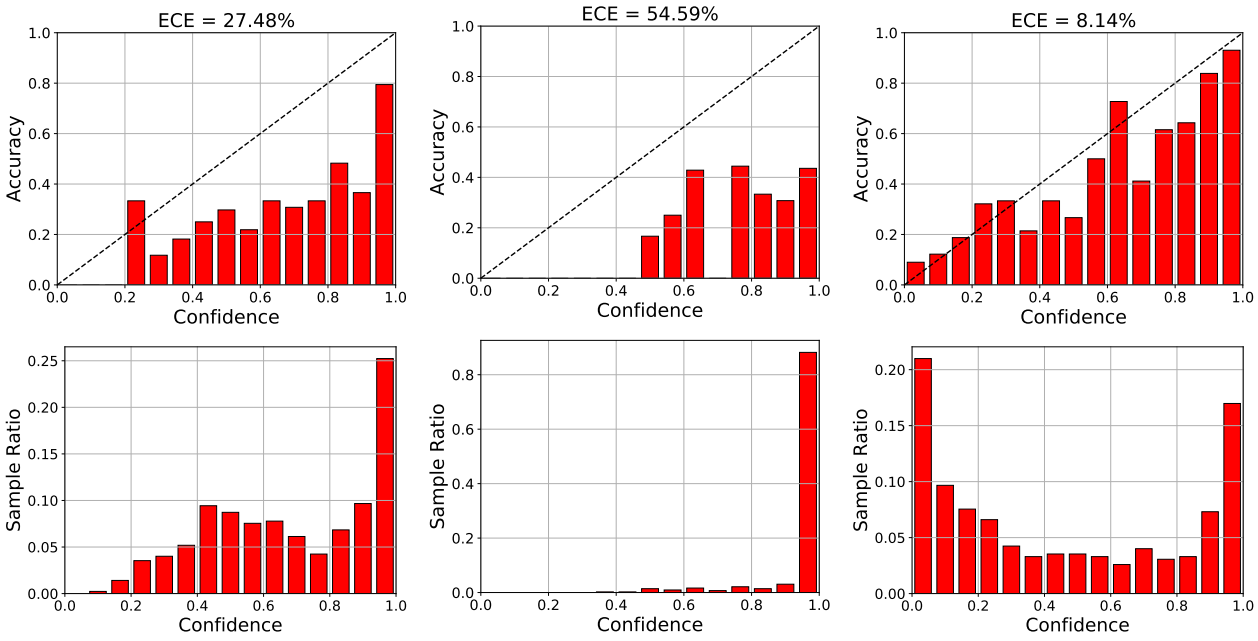
(a) None (b) Temp. scaling (c) Temp.+FL+IW

Figure 7: Reliability diagram of the shift $\mathcal{A} \to \mathcal{W}$ from one experiment among ten.



(a) None (b) Temp. scaling (c) Temp.+FL+IW

Figure 8: Reliability diagram of the shift $\mathcal{D} \to \mathcal{A}$ from one experiment among ten.



(a) None (b) Temp. scaling (c) Temp.+FL+IW

Figure 9: Reliability diagram of the shift $\mathcal{W} \to \mathcal{A}$ from one experiment among ten.