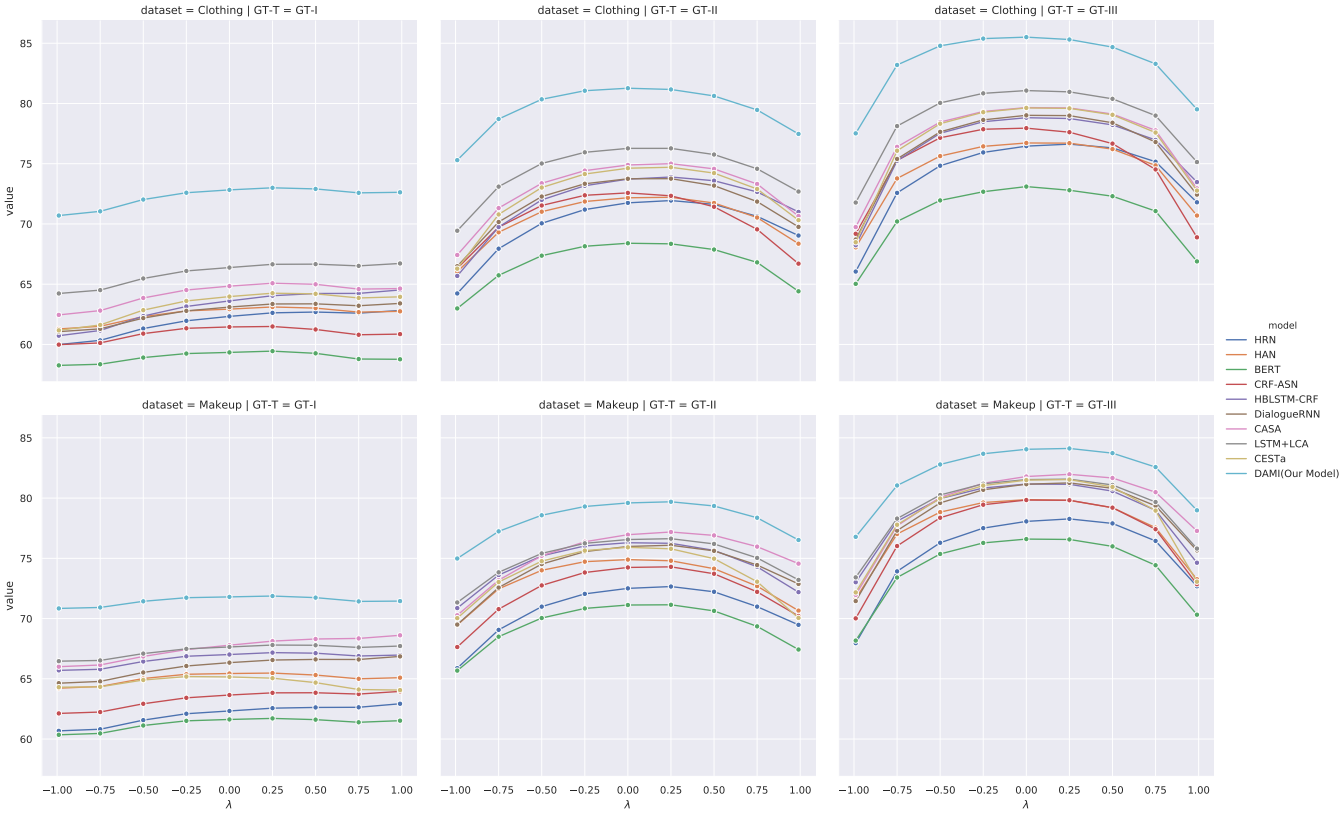

Figure 1: Performances of different λ values on Clothing and Makeup datasets. The performance of baseline models will change in different magnitudes with the change of λ. It demonstrates that the GT-T reflects the performance of different models under different tolerance degrees. We can adjust models and strategies according to the tolerance degrees. Taking GT-II on Makeup dataset as an example, LSTM+LCA performs better than CASA when $\lambda \leq -0.5$, while it is opposite when $\lambda \geq -0.25$. Also, on the Makeup dataset, GT-I score of LSTM+LCA is higher than GT-I score of CASA when $\lambda < -0.25$, while GT-I score of LSTM+LCA is lower than GT-I score of CASA when $\lambda > -0.25$. Overall, our models perform reliably better than baseline models.