Policy-labeled Preference Learning $\left(\log \pi^*(a_t|s_t) + \mathcal{H}^{\boldsymbol{\pi}}(\cdot|s_t) - \mathbb{E}_{\tau \sim \mathbb{P}^{\boldsymbol{\pi}}_{(s_t,a_t)}} \left[\sum_{t} \gamma^l D_{KL}(\boldsymbol{\pi}(\cdot|s_l)||\pi^*(\cdot|s_l)) \right] \right)$ Policy-labeled Likelihood Dataset matched Contrastive KL Regularization