

Dual-Objective Optimization for Refusal

$$\mathcal{L}_{\text{W-DOOR}} = \mathbb{E} \left[\sum_{t=1}^{T} \left(-\beta_t \log \pi_{\theta}(y_t^s \mid x, y_{< t}^s) - \frac{2}{\beta} \log \sigma \left(-\beta \log \frac{\pi_{\theta}(y_t^h \mid x, y_{< t}^h)}{\pi_{\text{ref}}(y_t^h \mid x, y_{< t}^h)} \right) \right]$$