

Contrastive Anchor Sampling

Original Samples

y_1

y_2

$\vdots$

y_{G-1}

Anchor

y_{anchor}

π_{old}

Dual-Stream Advantage Estimation

Style Stream

$$A_i^{style} = \frac{R_i^{style} - \bar{\mu}_t}{\bar{\sigma}_t + \epsilon}$$

Global State
 $\bar{\mu}_t, \bar{\sigma}_t$

A^{style}

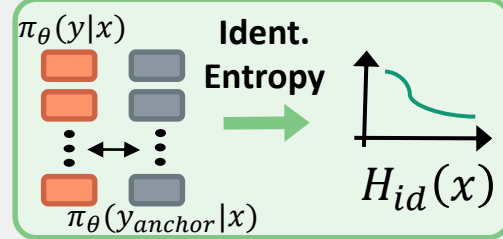
Task Stream

$$A_i^{task} = \frac{R_i^{task} - \mu_{group}}{\sigma_{group} + \epsilon}$$

A^{task}

Entropy-Aware Adaptive Exploitation

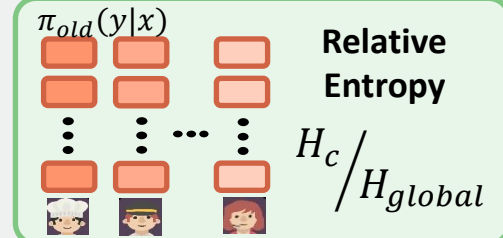
Instance Level Gating



A_i

$\hat{A}_i$

Model Level Regulation



Gradient Update
 $\mathcal{J}_{CRPO}$

β_t^c
Dynamic

$$\mathcal{J}_{CRPO} \approx \mathbb{E}_{\mathcal{G}} \left[\underbrace{\hat{A}_i \cdot \pi_{\theta}}_{\text{Gated Policy}} - \underbrace{\beta_t^c \cdot D_{KL}}_{\text{Adaptive}} \right]$$