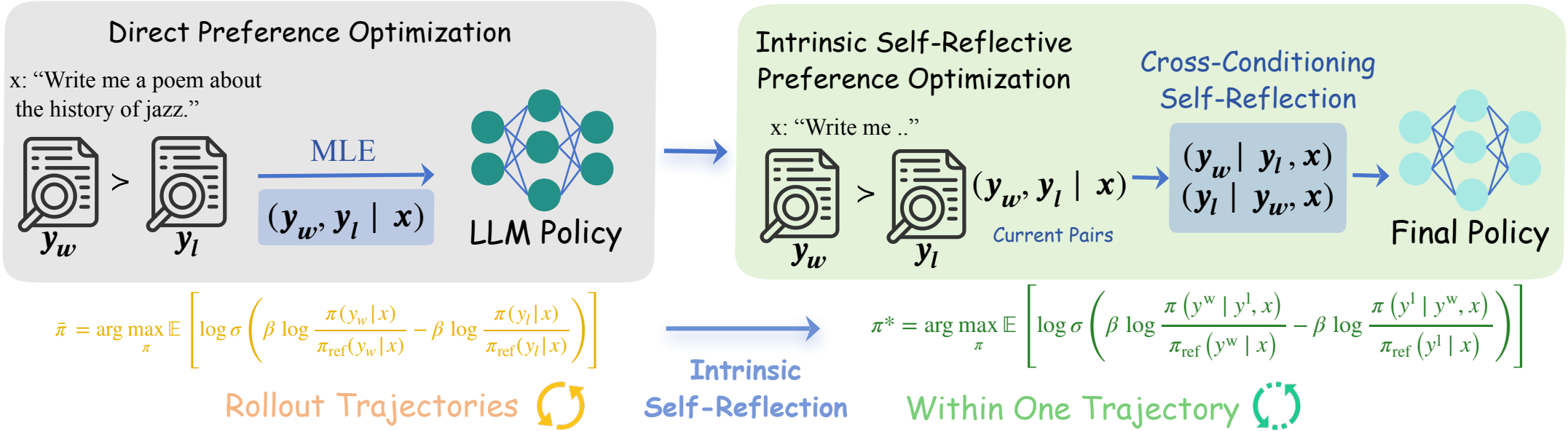
Direct Preference Optimization
x: "Write me a poem about the history of jazz."
y_w
>
y_l
MLE
$(y_w, y_l \mid x)$
LLM Policy
Intrinsic Self-Reflective Preference Optimization
Cross-Conditioning Self-Reflection
x: "Write me .."
y_w
>
y_l
$(y_w, y_l \mid x)$
Current Pairs
$(y_w \mid y_l, x)$
$(y_l \mid y_w, x)$
Final Policy
$\bar{\pi} = \arg\max_{\pi} \mathbb{E}\left[\log \sigma\left(\beta \log \frac{\pi(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$
Rollout Trajectories
Intrinsic Self-Reflection
$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\log \sigma\left(\beta \log \frac{\pi\left(y^w \mid y^l, x\right)}{\pi_{\text{ref}}\left(y^w \mid x\right)} - \beta \log \frac{\pi\left(y^l \mid y^w, x\right)}{\pi_{\text{ref}}\left(y^l \mid x\right)}\right)\right]$
Within One Trajectory