Step 1: IK Computes Position Error

$$\Delta p_{ee} = p_{ee}^{goal} - p_{e}^{current} e$$

(EE goal position - Current EE position)

Step 2: Error Transformed to Reward

$$r_{\text{arm}} = \exp\left(-\frac{\|\Delta \mathbf{p}_{\text{ee}}\|}{2\sigma}\right) \times 0.8$$

- Small error → High reward (≈ 0.8)
- Large error → Low reward (≈ 0)

Step 3: Total Reward Drives Policy

$$r_{\text{total}} = r_{\text{loco}} + r_{\text{arm}}$$

$$\nabla_\theta J = \mathbb{E}\left[\nabla_\theta \log \pi(\mathbf{a}|\mathbf{o}) \cdot r_{\text{total}}\right]$$

(Locomotion + Manipulation reward)

Step 4: Locomotion "Compromises"

Policy discovers:

- Pure walking → Large error → Low r_{total}

Policy learns:

- Adjust legs → Support arm → High r_{total}

Emergent Whole-Body Control

Legs serve global goal (EE tracking) instead of independent locomotion