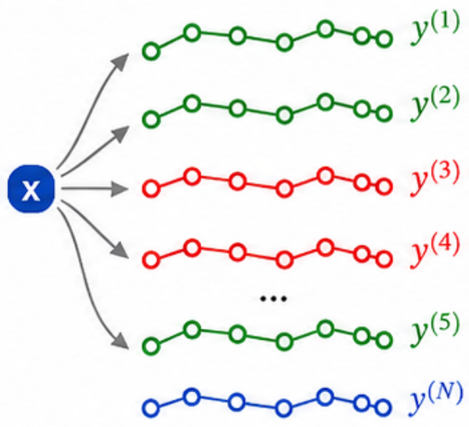
1 Generate on-policy experience

Current policy π_θ samples a rollout group G of complete responses for prompt x.



Diversity comes from sampling: multiple strategies, lengths, and thinking paths.

2 Verify outcomes (sparse but reliable)

The verifier provides a final outcome reward $r_i \in \{0,1\}$. We compute group-relative advantages A_i.

$y^{(1)}$ ✓ $r_1 = 1$

$y^{(2)}$ ✓ $r_2 = 1$

$y^{(3)}$ ✗ $r_3 = 0$

$y^{(4)}$ ✗ $r_4 = 0$

…

$y^{(N)}$ ✓ $r_N = 1$

Group-relative advantage

$$A_i = A_G\left(y^{(i)}\right)$$

(stopped gradient)

Provides outcome alignment across the whole trajectory.

3 Reassess with privileged self-guidance (dense)

A frozen, training-time view of the same policy (privileged context c) scores each token. We compute a trajectory-level self-guidance gain G_i.

Frozen policy π_H (privileged context c)

↓

Token-level log-prob gains

↓

Trajectory gain (clipped, averaged)

$$G_i = G_H\left(y^{(i)} \mid x, c\right)$$

Provides dense reasoning information along the path.

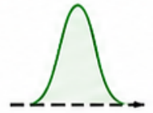
4 Ground self-guidance with verifier outcomes

Combine with sign-gating so self-guidance follows the verifier on success and is reversed on failures.

$$E_i = \eta_A A_i + \beta_G G_i \cdot \operatorname{sgn}(A_i)$$

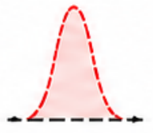
$A_i > 0$ (success) ✓

Use self-guidance (agree with G_i)



$A_i < 0$ (failure) ✗

Reverse self-guidance (disagree with G_i)

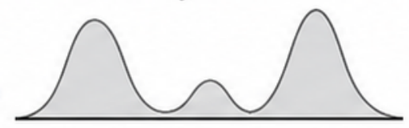


Prevents self-confirming errors and aligns dense guidance with verified outcomes.

5 Internalize by trajectory balance (distributional update)

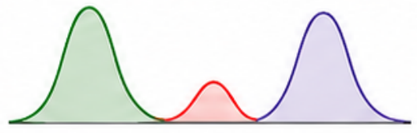
Exponentially reweight the reference π_{ref} with E_i to form a normalized target distribution, then fit it via profiled trajectory balance.

Reference π_{ref}



$$\exp\left(\frac{E_i}{\tau}\right)$$

Target distribution p_G^*



Learns where probability mass should settle over complete responses.