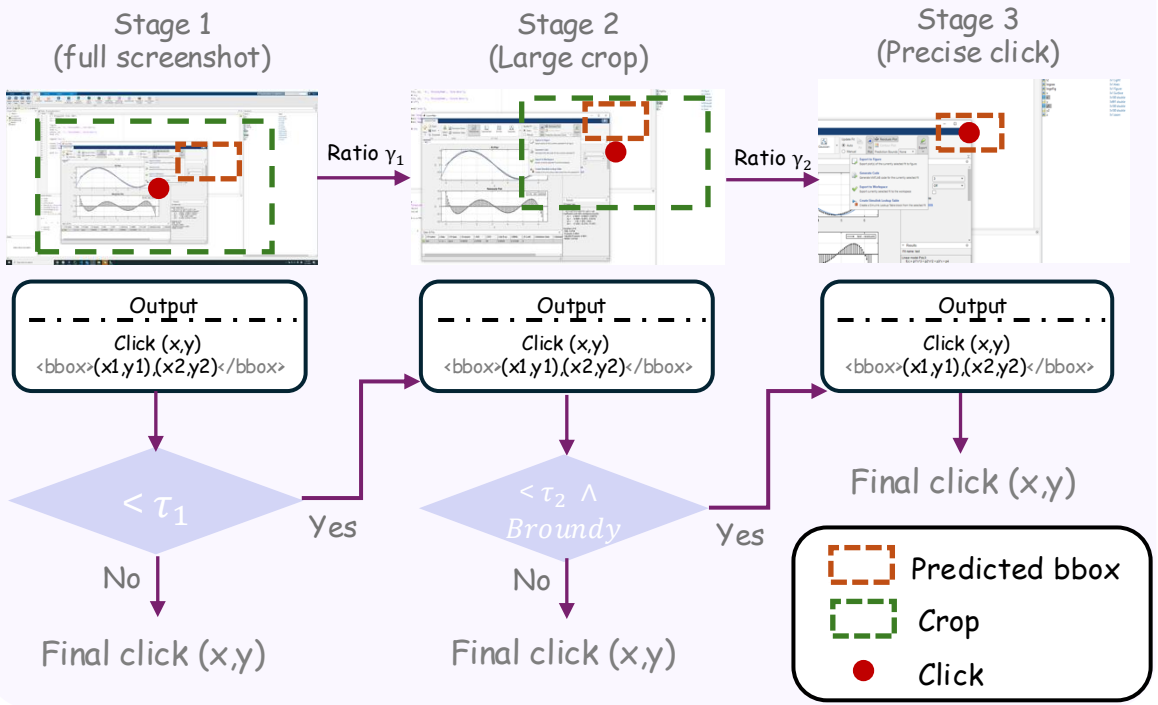
## Coarse to fine policy




## Reward design

$$R(p,B) = \lambda_{fmt}R_{fmt} + \lambda_{click}R_{click} + \lambda_{iou}R_{iou}$$

$$R_{click} = 0.6 \cdot \mathrm{I}[p \in b^*] + 0.4 \cdot \exp(-\beta(\frac{d(p,b^*)}{\sigma(b^*)})^2)$$

$$R_{iou} = IoU(B, b^*)$$

## Difficulty-aware GRPO

Difficulty score

$$L_{GRPO} = -E_{i,g}[\alpha_i \min(r_{i,g}\hat{A}_{i,g}, clip(r_{i,g}, 1-\epsilon, 1+\epsilon)\hat{A}_{i,g})]$$

$Q$

| | | | | |
|---|---|---|---|---|
| $g=1$ | $S_1$ | $S_2$ | $S_3$ | $R_1$ |
| $g=2$ | $S_1$ | $S_2$ | | $R_2$ |
| $g=\cdots$ | $S_1$ | | | $R_{..}$ |
| $g=G$ | $S_1$ | $S_2$ | $S_3$ | $R_G$ |

Relative advantage

$$\hat{A}_{i,g} = R_{i,g} - \frac{1}{G}\sum_{j=1}^{G} R_{i,j}$$

## Auxiliary refinement rewards (per stage)

$$\tilde{R}_t = S_t + \beta^+ \max(S_t - S_{<t}, 0) - \beta^- \max(S_{<t} - S_t, 0) \rightarrow \hat{A}_{i,g}^{(t)}$$

- $S_t$: Click quality
- $\beta^+ \max(S_t - S_{<t}, 0)$: Effective crop
- $\beta^- \max(S_{<t} - S_t, 0)$: Harmful crop

Applied to stage-t response tokens