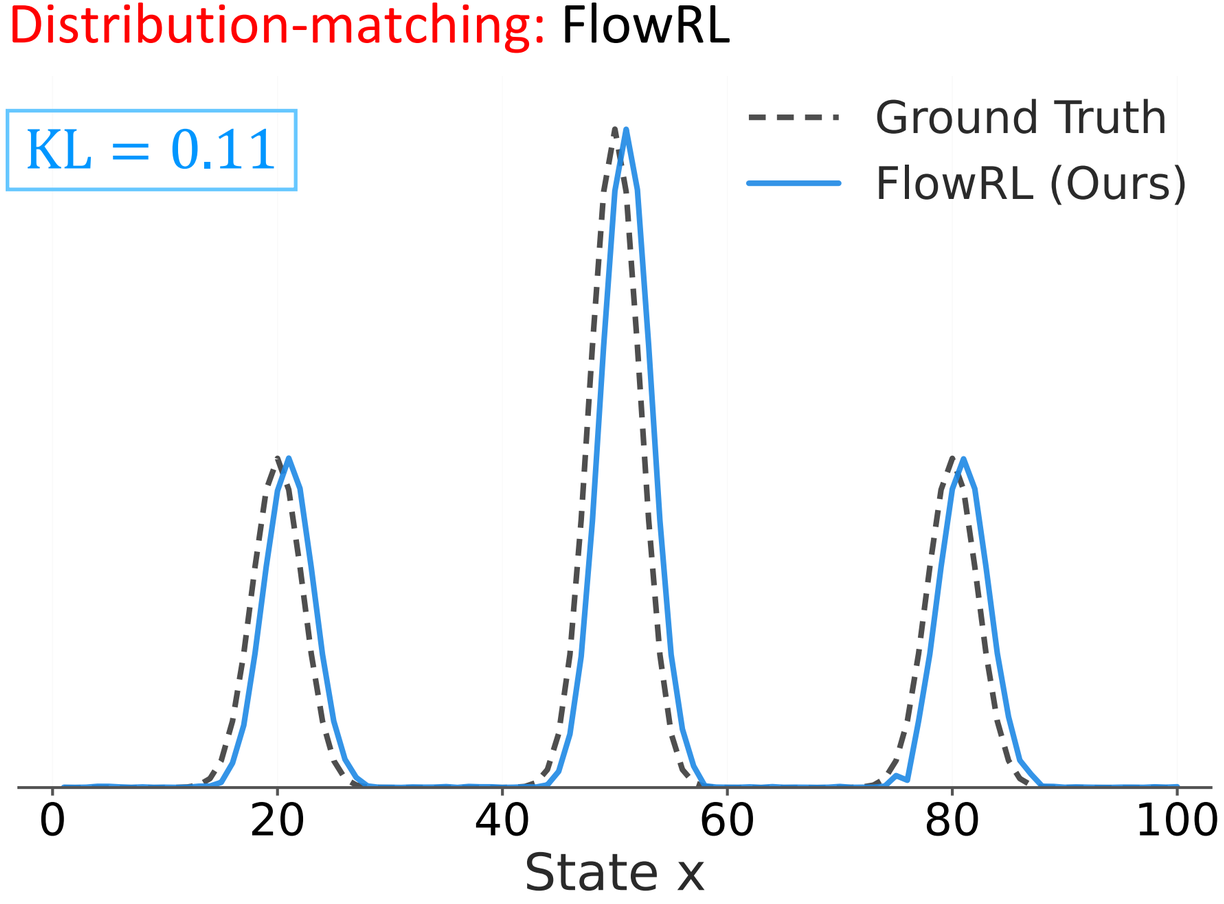
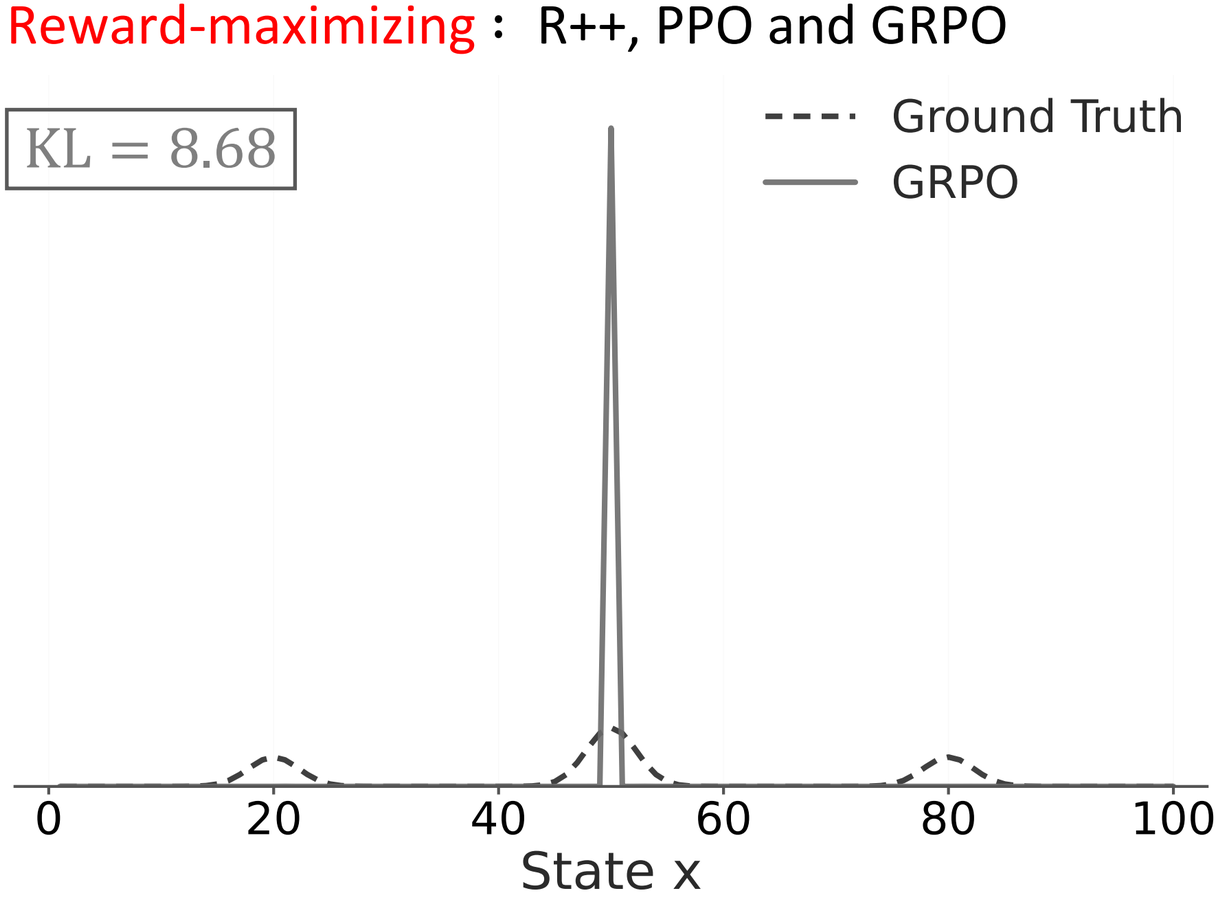

Distribution-matching: FlowRL
KL = 0.11
Ground Truth
FlowRL (Ours)
0
20
40
60
80
100
State x


Reward-maximizing： R++, PPO and GRPO
KL = 8.68
Ground Truth
GRPO
0
20
40
60
80
100
State x