

a. Limitations of existing paradigms

1. Single objective:
max. adversariality

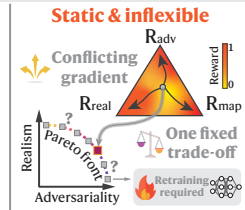
$$\mathbf{R}_{\text{total}} = \mathbf{R}_{\text{adv}} + \mathbf{R}_{\text{real}} + \mathbf{R}_{\text{map}}$$



2. Multi-objective linear scalarization

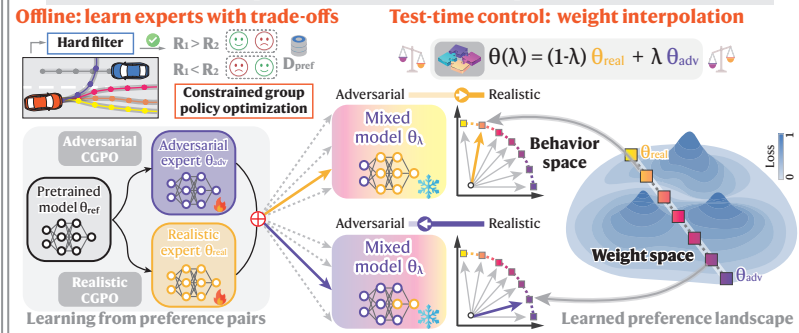
Conflating constraints

$$\mathbf{R}_{\text{total}} = \mathbf{w}_3 \mathbf{R}_{\text{map}} + \mathbf{w}_1 \mathbf{R}_{\text{adv}} + \mathbf{w}_2 \mathbf{R}_{\text{real}}$$



b. SAGE: steerable generation by test-time alignment

From discrete experts to a continuous spectrum of behaviors



c. Closed-loop RL training

Dual curriculum learning

