a. Limitations of existing paradigms
1. Single objective: max. adversariality
$R_{total} = R_{adv} + R_{real} + R_{map}$
Unrealistic
2. Multi-objective linear scalarization
Conflating constraints
$R_{total} = W_3R_{map} +$ Hard
$W_1R_{adv} + W_2R_{real}$ Soft
Static & inflexible
R_{adv}
Conflicting gradient
R_{real}
R_{map}
Reward
1
0
One fixed trade-off
Realism
Pareto front
Adversariality
Retraining required
b. SAGE: steerable generation by test-time alignment
From discrete experts to a continuous spectrum of behaviors
Offline: learn experts with trade-offs
Hard filter
$R_1 > R_2$
$R_1 < R_2$
D_{pref}
Constrained group policy optimization
Test-time control: weight interpolation
$\theta(\lambda) = (1-\lambda)\,\theta_{real} + \lambda\,\theta_{adv}$
Adversarial CGPO
Pretrained model θ_{ref}
Adversarial expert θ_{adv}
Realistic expert θ_{real}
Realistic CGPO
Learning from preference pairs
Adversarial
Realistic
Mixed model θ_λ
Behavior space
Weight space
Loss
1
0
θ_{real}
θ_{adv}
Learned preference landscape
c. Closed-loop RL training
Dual curriculum learning
EGO
ENV
ADV
Co-evolution
Frequency
Intensity