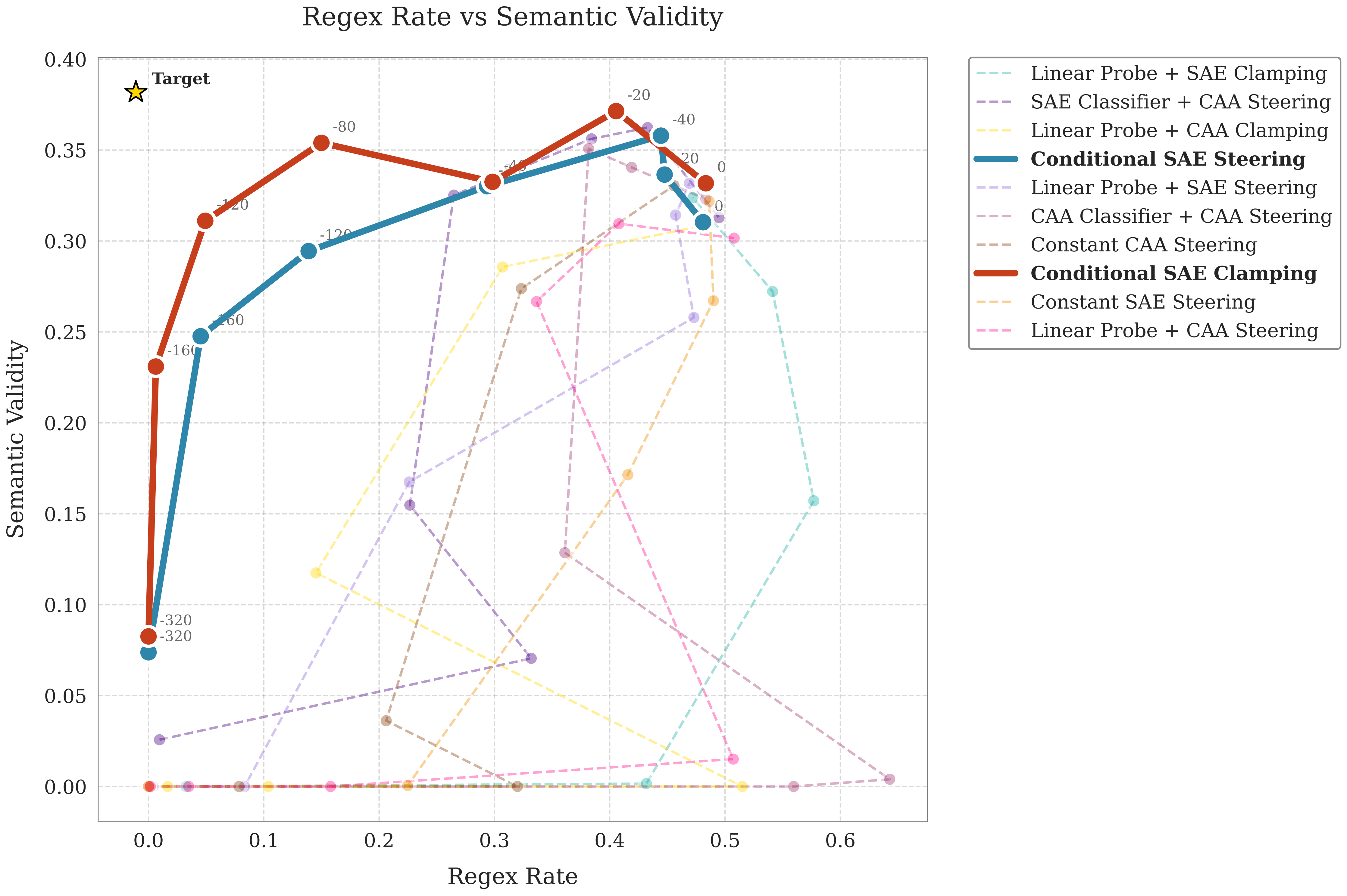
Regex Rate vs Semantic Validity
Target
-20
-40
-80
-40
-120
-120
-160
-160
-320
-320
20
0
0
Semantic Validity
Regex Rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Linear Probe + SAE Clamping
SAE Classifier + CAA Steering
Linear Probe + CAA Clamping
Conditional SAE Steering
Linear Probe + SAE Steering
CAA Classifier + CAA Steering
Constant CAA Steering
Conditional SAE Clamping
Constant SAE Steering
Linear Probe + CAA Steering