Training vs 100-Episode Evaluation Performance

Training vs 100-Episode Evaluation Performance 25.0% 7.7% 10.0% 12 Training Best 25 100-Episode Eval 10 20 18.3% 18.3% 8 Success Rate (%)
0 Success Rate (%) **15.0**% 4.3% 10.0% **7.7**% 4.3% 2 2.7% Spatial Semantic Semantic (Target Block + Target Peg) with (Wher Hot + Target Peg) with Color Spatial Semantic (One Hot + Target Peg) with Color (Target Block + Target Peg) with Color **Conditioning Method** Color Information Impact: Training vs Evaluation 100-Episode Performance Distribution Training Impact **Evaluation Impact** 12 10 Color Impact (% Change) Success Rate (%) -3.3% -5.0% Spatial Spatial Semantic Semantic (Target Block + Target Peg) with Color (Target Block + Target Peg) (Target Block + Target Peg) with Color (Target Block + Target Peg) (Target Bloc Spatial Semantic -5.7% Conditioning Method Individual Run Comparison: Training vs Evaluation Performance Consistency: Training vs Evaluation 12.5 Training Training Std 40 -Evaluation Evaluation Std 10.8 35 10 30 Standard Deviation (%) Success Rate (%) 20 25 25 4.1 4.1 10 2.4 2.4 1.7 1.6 2 -Spatial Spatial Semantic Semantic (Target Block + Target Peg) with (Talget Block + Target Peg) with Color (One Hot + Target Peg) with Color Spatial Spatial Semantic Semantic (Target Block + Target Peg) with (Coher Hot + Target Peg) with Color (Target Block + Target Peg) with Color Conditioning Method