Improving and Accelerating Offline RL in Large Discrete Action Spaces with Structured Policy Initialization

Matthew Landers*, Taylor W. Killian, Thomas Hartvigsen, Afsaneh Doryab | *qwp4pk@virginia.edu

University of Virginia

Institute of Foundation Models
Mohamed bin Zayed University of Artificial Intelligence

PROBLEM

RL in Combinatorial Action Spaces

Requires solving two related problems:

- Searching over an exponential number of joint actions $|\mathcal{A}| = \prod_{d=1}^{A} m_d$
- Ensuring that the chosen sub-actions form coherent combinations

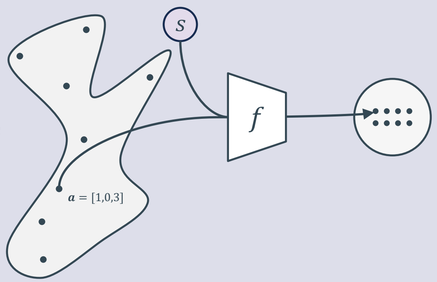
Existing methods either:

- Impose structural priors, which limits representational capacity
- Learn structure and control jointly, which is slow and unstable

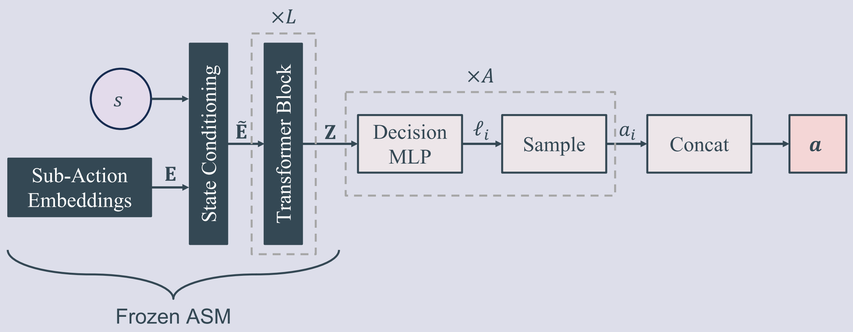
SPIN separates representation learning from control to make RL in combinatorial action spaces tractable and expressive.

APPROACH

SPIN: A Two-Stage Learning Framework



The SPIN Architecture



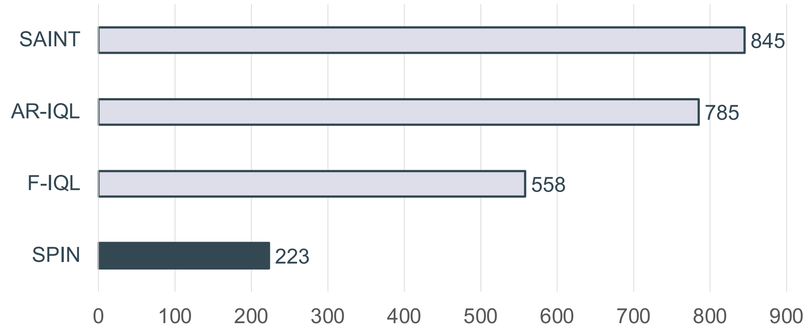
RESULTS

SPIN Achieves SOTA Asymptotic Performance

Dataset	F-IQL	AR-IQL	SAINT	SPIN
Medium	341.8	334.7	343.0	**345.2**
Med-Expert	724.7	717.2	733.3	**753.2**
Rnd-Med-Exp	388.3	395.6	438.9	**499.2**
Expert	778.1	770.3	773.1	**778.7**
Avg. Return	558.2	554.5	572.1	**594.1**

SPIN achieves SOTA performance across discretized DM Control tasks, with the largest gains on heterogeneous datasets.

SPIN Reaches Target Performance Faster Than All Baselines



Total training time in minutes. SPIN reduces time-to-target performance by up to 4x relative to existing methods.

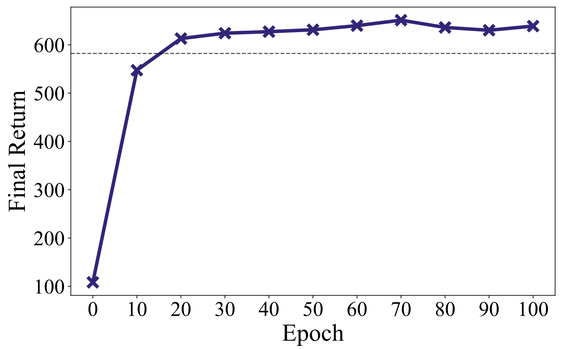
SPIN Scales to Enormous Action Spaces

$\lvert\mathcal{A}\rvert$	F-IQL	AR-IQL	SAINT	SPIN
$\approx 10^{18}$	472.3	526.5	635.1	**647.0**
$\approx 10^{38}$	483.8	457.4	529.1	**629.5**
$\approx 10^{56}$	485.0	557.4	562.5	**703.9**
Avg. Return	480.4	513.8	575.6	**660.1**
Time to Target	545.8	692.6	291.6	**237.0**

SPIN maintains strong performance as action spaces grow exponentially, while baselines degrade or plateau.

ANALYSIS

Representation Quality Drives Policy Performance



Policy return vs. ASM pre-training epochs. Just 20 epochs surpasses fully-trained factored IQL baseline in all tasks.

Representations Capture Meaningful Structure

	Random ASM	Trained ASM
Per-Slot Accuracy	76.6%	**90.0%**
Exact Match Accuracy	0.10%	**4.52%**

Linear probe measures cross-joint coordination: 45× improvement in exact-match accuracy. Exceeds independence baseline ($0.90^{38} \approx 1.83\%$).

ASM Leads to Emergent Rapid Adaptation

Environment	F-IQL	AR-IQL	SAINT	SPIN
Cheetah	30.3%	40.9%	46.1%	**90.6%**
Finger	0.37%	0.70%	0.39%	**68.5%**
Humanoid	2.2%	3.1%	9.3%	**93.4%**
Quadruped	34.1%	38.1%	39.0%	**78.6%**
Average	16.7%	20.7%	23.7%	**82.8%**

SPIN reaches near-SOTA asymptotic performance using just 1% of the training budget.