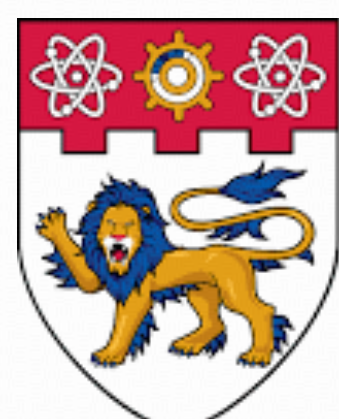




THE UNIVERSITY OF
SYDNEY

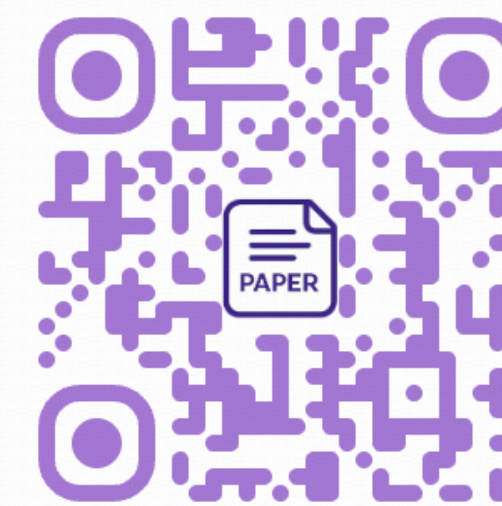


**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

Offline Behavioral Data Selection

Shiye Lei¹ Zhihao Cheng¹ Dacheng Tao²

¹The University of Sydney ²Nanyang Technological University



August 9–13, 2026
KDD2026
Jeju, Korea

Overview

- DOMAIN:** *Data Selection* for Offline Behavioral Cloning (BC).
- GOAL:** *select a compact behavioral subset from* a large offline BC dataset for rapid policy training, a process we term *offline behavioral data selection*.

Millions of Offline BC Data

few-hours policy training 🐻

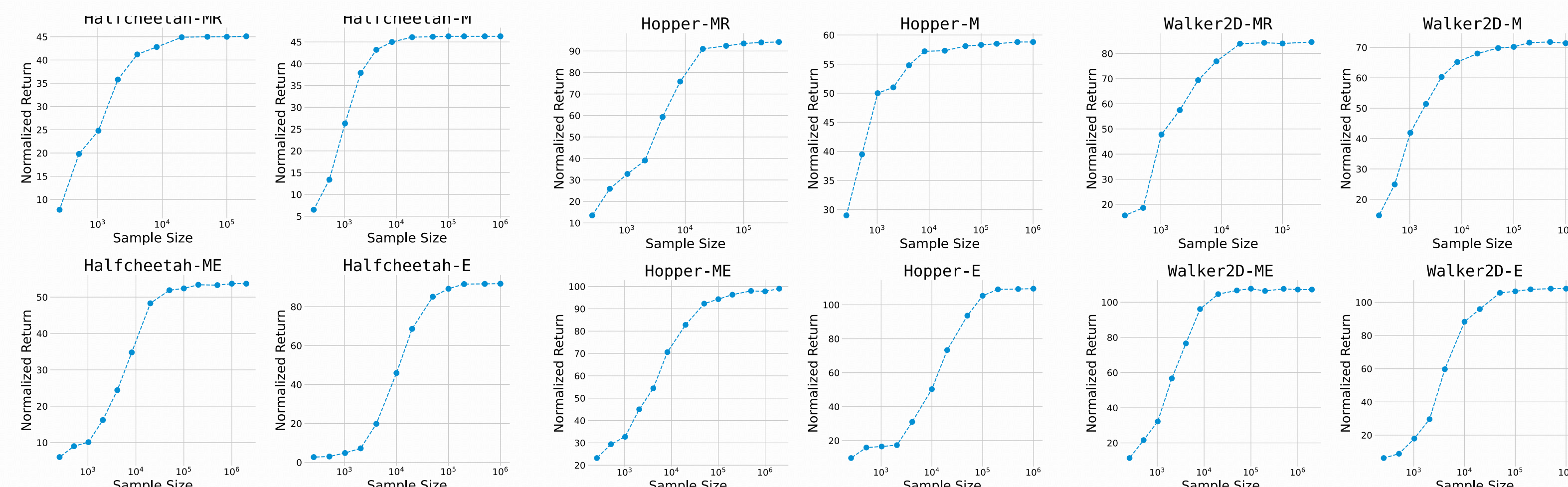
Hundreds of behavioral data

one-second policy training 🐻

- FORMULATION:** $S^* = \arg \min_{\mathcal{S}} |J(\mathcal{S}) - J(\mathcal{D})|$ s.t. $\mathcal{S} \subset \mathcal{D}$, $|\mathcal{S}| = N$

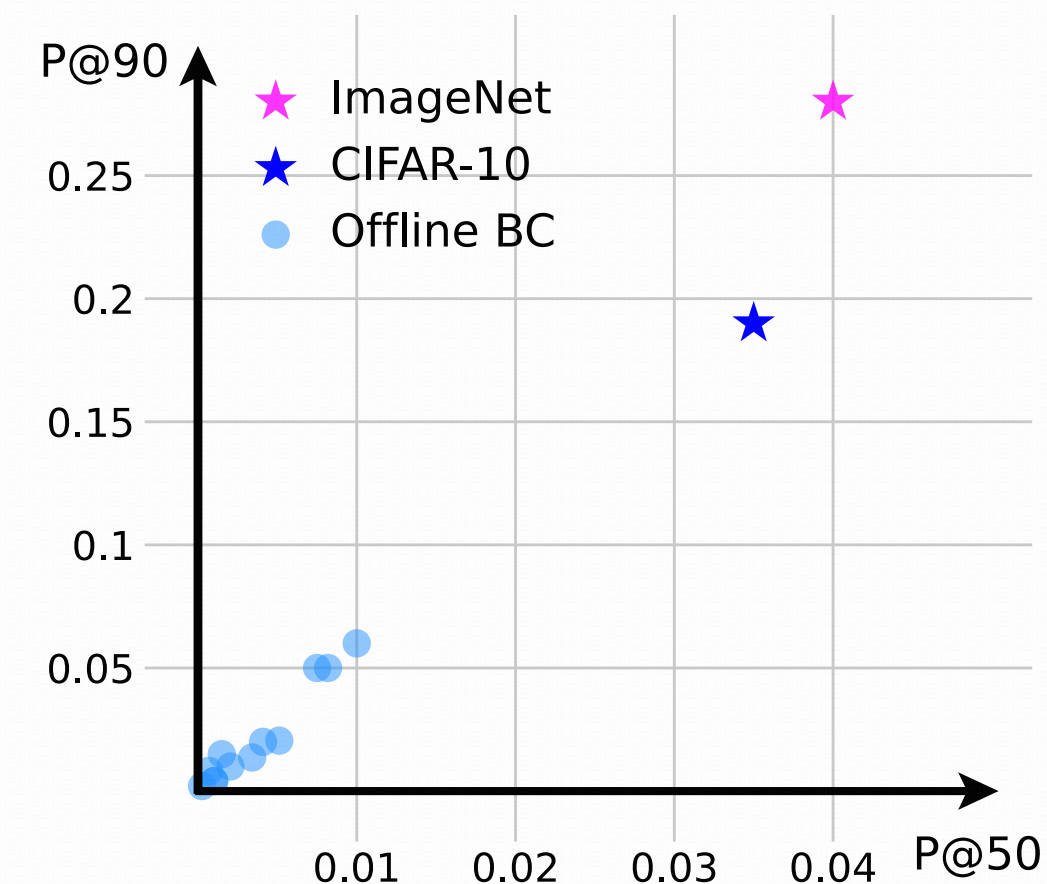
Data Saturation in Offline BC Data

- Offline BC data:** Given offline RL data, we relabel each state with actions predicted by the learned policy, yielding offline BC data.
- Experiment:** We randomly sample subsets of different sizes from the original offline BC dataset and plot the sample size–performance curves below.



- Key finding:** Policy performance gradually plateaus once the sample size exceeds 10,000, even though the full datasets contain hundreds of thousands to millions of samples (the sample-size axis is shown in log scale).

Analysis of Data Saturation

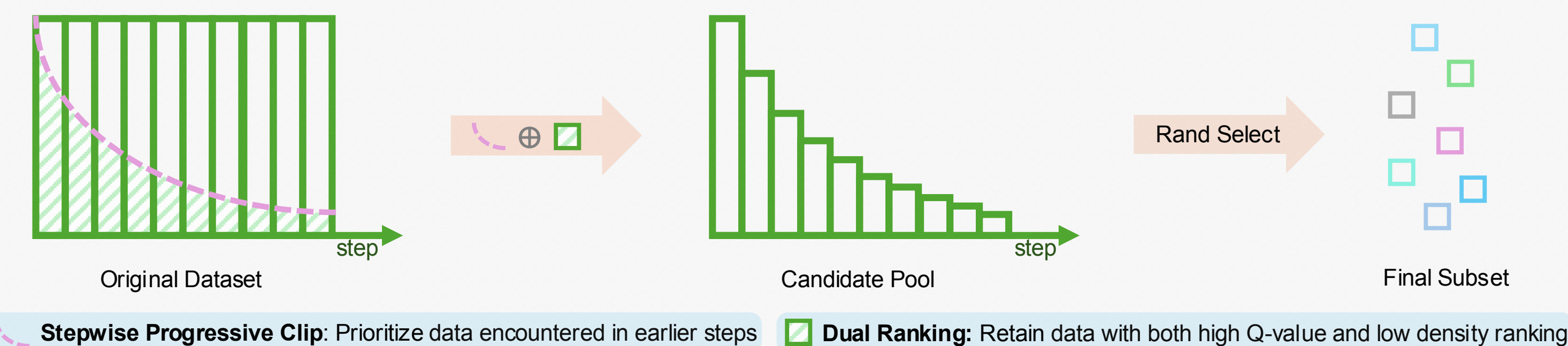


P@50: fraction of data required to reach **50%** of the performance achieved by training on the full dataset.

P@90: fraction of data required to reach **90%** of the performance achieved by training on the full dataset.

Observation: Offline BC data has much smaller P@50 and P@90 than ImageNet and CIFAR-10, indicating much stronger data saturation.

Stepwise Dual Ranking (SDR)



Lemma 1. For two policies π^* and π , let $\epsilon_t = \mathbb{E}_{s \sim d_t^*} [D_{TV}(\pi, \pi^*)]$, we have $|J(\pi) - J(\pi^*)| \leq \sum_{t=0}^{T-1} (T-t)\epsilon_t$.

Stepwise Progressive Clip: Prioritize data from earlier stages during selection

$$\epsilon_t = \mathbb{E}_{s \sim D_t} \left[\frac{d_{\pi^*}^t(s)}{d_{\beta}^t(s)} D_{TV}(\pi, \pi^*) \right]$$

Dual ranking: Prioritize data that rank **both** high in q_{π^*} and low in d_{β}

Stepwise Progressive Clipping + Dual Ranking $\Rightarrow$ Stepwise Dual Ranking (SDR)

Stepwise Progressive Clipping: Define a monotonically decreasing selection ratio

$$\alpha_t = 1 - \lambda \tanh(t/100),$$

where α_t controls the select ratio at step t .

Dual Ranking: At each step t , retain samples that rank **both** above the α_t -quantile in q and below the α_t -quantile in d .

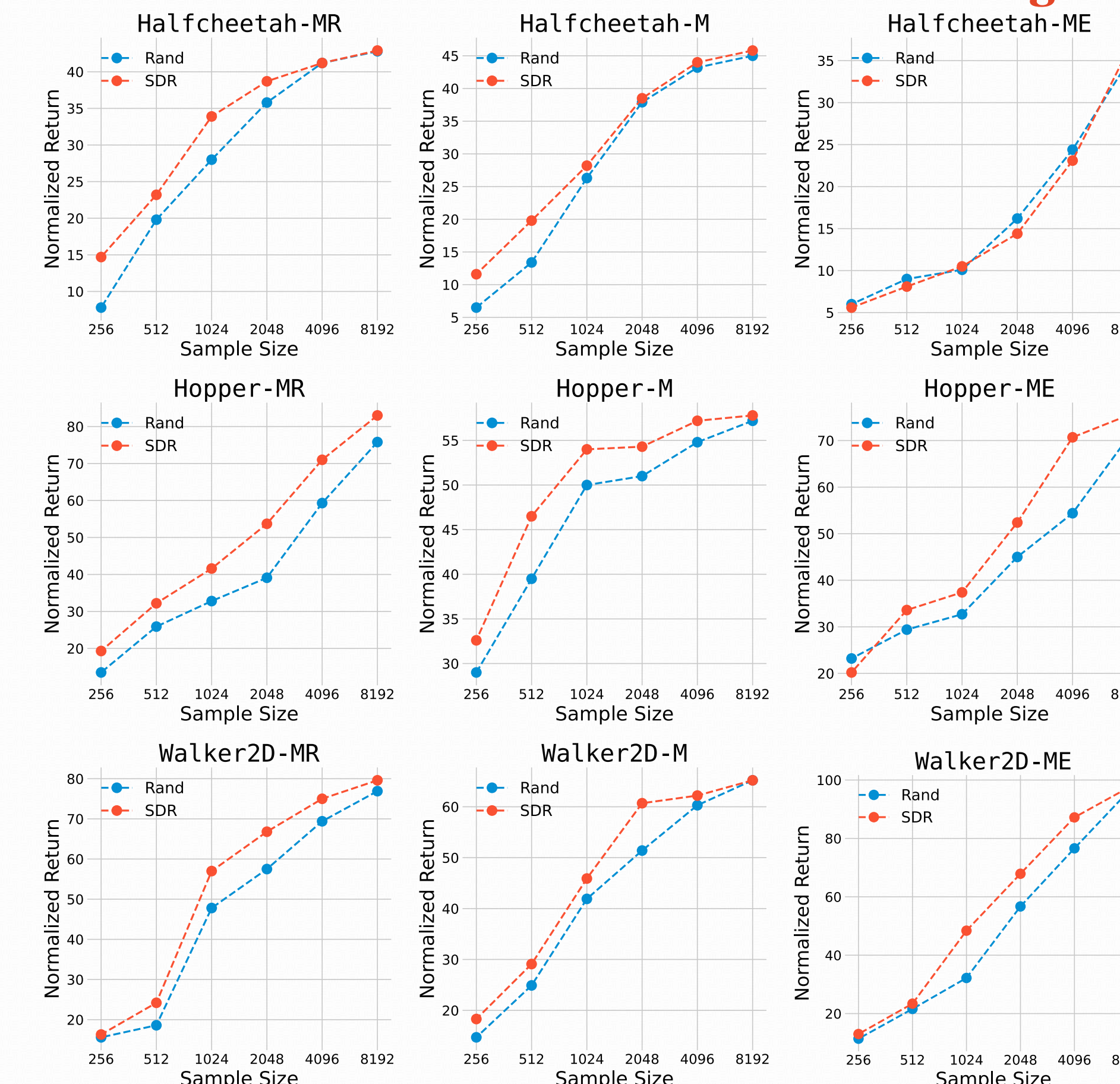
Two-stage Sampling: Apply stepwise dual ranking to construct a larger candidate pool, and then randomly sample a smaller final subset under a fixed budget.

Experiments

Data Selection Performance

Method	Halfcheetach			Hopper			Walker2D			Average
	M-R	M	M-E	M-R	M	M-E	M-R	M	M-E	
Random	28.0	26.3	10.1	32.8	50.0	32.7	47.8	41.9	32.2	27.0
Top Reward	10.6	1.4	2.2	5.9	6.9	3.9	3.6	2.7	7.6	5.0
Top Q-value	4.4	2.3	1.7	6.1	7.7	9.1	3.8	13.1	4.3	5.8
Herding	0.7	1.0	3.3	0.6	4.4	2.9	3.6	2.7	1.1	2.3
GraNd	2.0	2.1	3.1	6.4	14.5	1.3	2.7	5.9	3.3	4.6
GradMatch	3.3	6.8	18.9	29.3	18.3	43.7	23.9	13.6	44.5	22.5
SDR (ours)	33.9	28.2	10.5	41.6	54.0	37.4	57.0	45.9	48.4	40.8

SDR vs. Rand Selection Across Different Budget Sizes



Ablation Study

Method	Halfcheetach			Hopper			Walker2D			Average
	M-R	M	M-E	M-R	M	M-E	M-R	M	M-E	
Random	29.2	28.7	16.8	41.1	46.9	42.6	47.6	43.1	49.1	38.3
DualRank	32.1	28.9	16.1	48.1	48.5	43.2	52.0	45.8	52.1	41.8
StepClip	30.4	29.8	15.2	46.9	48.6	43.5	48.9	45.9	51.6	40.1
SDR	32.4	31.3	16.3	50.1	50.4	48.3	53.2	46.9	56.2	42.8