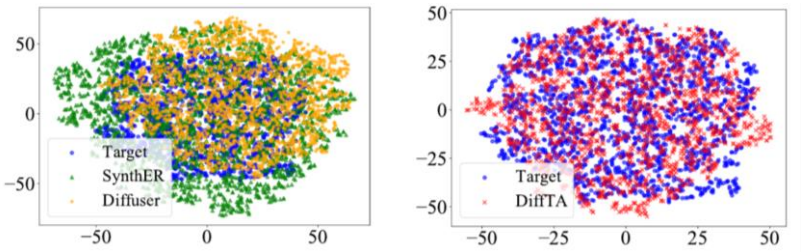
Trajectory Alignment via Diffusion Models in Cross-Domain Offline Reinforcement Learning

Yujia ZHANG, Lin LI, Jianguo WU, Ting GUO, Wei WEI, Jiye LIANG

Frontiers of Computer Science, DOI: 10.1007/s11704-026-52191-9

Problems & Ideas

- Problems of cross-domain offline RL with diffusion models:
  - **Misalignment**: Generated trajectories drift off the target dynamics manifold;
  - **Infeasibility**: Generated actions/transitions are not valid under target dynamics (thus unreliable for policy learning).



t-SNE embeddings show Diffuser and SynthER produce dispersed, misaligned trajectories, while our method DiffTA generate tightly clustered samples aligned with target dynamics, matching its stronger policy performance.

How can we harness DPMs' generative power while ensuring trajectories adhere to the target domain's dynamics?

- Ideas: Make target consistency an explicit **conditioning signal** inside the diffusion denoising process, rather than post-hoc filtering:
  - **Dynamics-aware conditioning (DDS)**: Use a bounded domain similarity score to guide denoising toward the target dynamics manifold;
  - **Utility & feasibility conditioning (VG + PH)**: Steer generation toward transitions with small cross-domain value gap (VG) and penalize action–transition inconsistency via an inverse dynamics check (PH), yielding target-consistent and usable synthetic trajectories.

Main Contributions

- Contributions:
  - We propose DiffTA, a conditional diffusion framework for cross-domain offline RL that enforces target-domain dynamics during generation;
  - We introduce three conditioning signals--Value Guidance, Domain Discrepancy Score, and Policy Harmonization--to improve transferability, dynamics alignment, and feasibility;
  - Experiments under large dynamics shifts show consistent gains over strong baselines, and ablations confirm the contribution of each component.

Table 1 Performance comparison across different baselines under friction-0.5 and gravity-0.5 shifts (Best in Bold)

Shift Type	Environment	IQL	DARA	UTDS	BOSA	SRPO	IGDF	OTDF	DiffTA
friction-0.5	halfcheetah-medium	40.1	43.3	45.1	51.3	47.1	48.7	**60.5**	58.4 ± 2.2
	halfcheetah-expert	63.3	75.4	74.3	80.4	92.4	87.6	88.7	**100.8 ± 3.5**
	hopper-medium	60.4	73.3	76.2	73.3	77.6	64.5	75.3	**83.3 ± 1.9**
	hopper-expert	90.2	87.7	85.4	94.6	86.7	92.3	**98.4**	96.5 ± 2.8
	walker2d-medium	75.4	63.1	65.4	63.7	66.4	76.3	65.3	**84.3 ± 3.1**
	walker2d-expert	90.3	90.4	95.6	93.1	95.3	86.4	91.5	**98.8 ± 2.4**
	ant-medium	55.4	56.3	60.8	58.8	57.7	61.2	49.7	**66.3 ± 3.3**
	ant-expert	72.5	77.3	80.1	71.8	74.4	77.6	75.8	**81.2 ± 5.9**
gravity-0.5	halfcheetah-medium	58.3	56.6	59.3	60.1	67.3	70.3	64.7	**71.8 ± 7.3**
	halfcheetah-expert	60.2	57.7	58.4	65.5	74.4	77.4	78.9	**85.4 ± 2.1**
	hopper-medium	45.4	30.3	55.6	49.3	54.3	60.3	63.5	**74.5 ± 1.2**
	hopper-expert	61.5	63.4	67.8	73.1	70.4	76.3	71.2	**86.7 ± 3.5**
	walker2d-medium	56.7	55.4	56.3	**60.3**	58.7	56.7	58.4	59.4 ± 2.8
	walker2d-expert	60.4	65.3	69.4	68.3	72.4	74.1	68.7	**79.6 ± 3.7**
	ant-medium	33.1	37.5	41.2	43.1	37.6	40.2	37.9	**50.8 ± 5.4**
	ant-expert	34.2	28.4	44.3	44.5	43.5	42.4	47.8	**56.7 ± 4.1**
	Average	59.8	59.2	60.1	65.7	67.3	68.3	68.5	**77.2**
	Friedman test (p-value)				0.00				
	Nemenyi test (p-value)	0.00	0.00	0.01	0.02	0.04	0.11	0.14	–