Appendix

Title of Paper: Trajectory Alignment via Diffusion Models in Cross-Domain Offline Reinforcement Learning

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

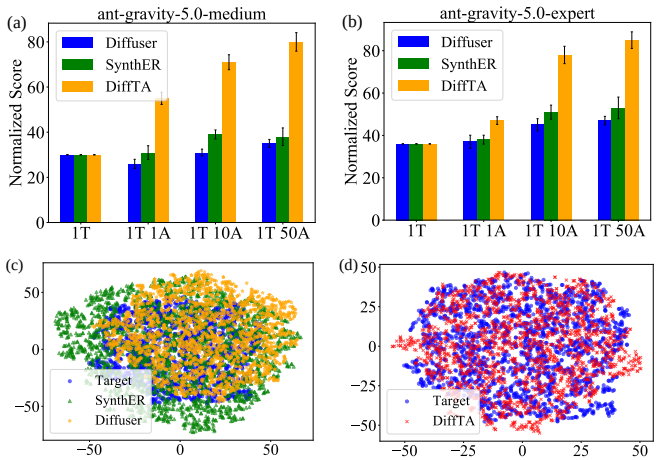

Fig. 1 Comparison of generative methods in cross-domain offline RL. (a) and (b) display performance on the target domain dataset, where "1T" indicates using 10% of the target domain dataset as real samples, and "1A" represents generating an equivalent number of synthetic samples. Here, "ant-gravity-5.0-medium" denotes the ant environment where the target domain's gravity is 5× that of the source, and the dataset type is "medium" as defined in D4RL. (c) and (d) visualize t-SNE embeddings for the ant-medium. Diffuser and SynthER fail to capture target-domain dynamics, yielding dispersed and inconsistent trajectories, whereas DiffTA produces tightly clustered samples well aligned with the target domain, consistent with its superior policy performance

■ 1 Empirical Motivation

As illustrated in Fig. 1, we compute the t-SNE on a shared state-action embedding of both target data and synthetic samples. In this space, DiffTA's generations largely overlap with the target cluster, whereas Diffuser [1] and SynthER [2] produce many samples in low-support regions. This highlights why naively applying such generative planners can fail in cross-domain settings: without explicit guidance, they tend to generate trajectories that fit source dynamics yet violate target physical laws. This dynamics misalignment injects harmful noise into the training buffer, significantly impairing downstream policy learning. This observation motivates our core objective: harnessing the generative capacity of DPMs while enforcing strict consistency with target-domain dynamics.

■ 2 Related Work

2.1 Cross-domain Offline RL

Cross-domain offline RL aims to learn a high-performing target-domain policy from static datasets collected in distinct source environments, where transition dynamics or reward functions differ [3, 4]. Existing approaches generally follow three directions. The first involves domain-weighted reuse, where reward reweighting or discriminator-based filtering [5, 6] identifies source transitions most similar to the target domain. Although effective in small domain shifts, such methods often rely on unstable domain classifiers or over-conservative importance weights, which can discard informative data under large dynamics gaps. The second direction learns dynamics transformations-either forward, inverse, or latent mappings-to convert source trajectories into target-like samples [7, 8]. However, these mappings easily amplify extrapolation errors when domains differ significantly. A third line of work assumes auxiliary simulators [9, 10] or seeks near-dynamics-agnostic policies [11–16], but such assumptions are often impractical in purely offline or safety-critical scenarios.

In contrast, our work introduces a generative perspective to cross-domain offline RL. Instead of filtering or transforming existing data, DiffTA directly synthesizes target-consistent trajectories through conditional diffusion guided by three contextual signals that jointly address transferability, alignment, and feasibility. This design explicitly models how cross-domain information should be selected, aligned, and executed, bridging the gap between data-centric and model-based paradigms. As a result, DiffTA provides a unified and scalable framework for robust cross-domain policy adaptation.

2.2 Diffusion Probabilistic Models

DPMs [17, 18] have emerged as a powerful class of generative models capable of producing diverse and high-fidelity samples across modalities such as images, audio, and video [19, 20]. By learning to reverse a gradual noising process, DPMs achieve stable likelihood-based training and avoid the mode-collapse issues that often affect GANs [21, 22]. Their conditional extensions enable fine-grained control through external guidance signals, such as text prompts, spatial masks, or class embeddings [23, 24], allowing the generative process to be shaped by structured context information. Classifier-free guidance has gained popularity over classifier-guided sampling [25, 26] due to its simplicity and strong performance [27, 28].

Recent works have begun exploring DPMs in RL, primarily for single-domain tasks such as trajectory inpainting, imitation learning, or offline policy optimization [29, 30]. These approaches demonstrate the ability of diffusion models to model complex trajectory distributions but remain limited to environments without domain discrepancies. Unlike prior methods, our work extends DPMs to the fully offline, cross-domain setting, where both dynamics and rewards vary across domains. We equip the diffusion model with three complementary conditioning signals that jointly capture transferability, dynamic alignment, and feasibility. This formulation transforms DPMs from passive trajectory reconstructors into active generators of target-consistent behaviors, enabling robust policy adaptation even under significant cross-domain shifts.

■ 3 Full Experiments

We conduct systematic experiments to assess DiffTA's performance in cross-domain offline RL, focusing on five key questions: (1) Can DiffTA generate high-quality trajectories that match the target domain's

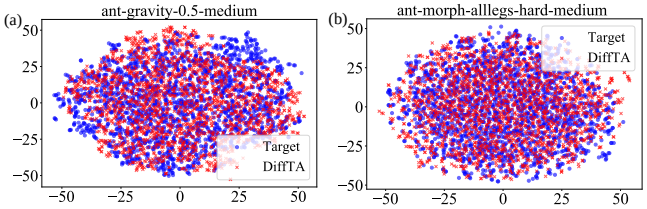

Fig. 2 t-SNE visualization of generated trajectories and target domain trajectories

transition dynamics? (2) How does the method perform when the discrepancy between source and target domain dynamics increases? (3) Is DiffTA robust to variations in target domain data quality and quantity? (4) What is the individual contribution of each core component to the overall performance? (5) Can DiffTA be accelerated without sacrificing performance?

3.1 Datasets and Baselines

Our experiments are conducted on the ODRL benchmark [31], a comprehensive library for evaluating cross-domain offline RL methods. From the benchmark, we select a subset of representative tasks, including ant, hopper, halfcheetah, and walker2d. These tasks feature domain shifts in friction, gravity, kinematics, and morphology to simulate realistic discrepancies. Friction and gravity shifts scale target domain parameters to 0.1, 0.5, 2.0, and 5.0 times the source values, while kinematic and morphology shifts modify joint ranges and limb sizes, respectively. All experiments are conducted with five random seeds. We report both the mean and standard deviation of DiffTA's performance.

DiffTA follows an offline-offline evaluation protocol similar to BOSA [32]. Specifically, source-domain data are sampled from the D4RL benchmark with one million transitions, while target-domain data is collected using an early-stopped SAC [33] policy with parameter perturbations, yielding approximately 0.1 million transitions. This configuration simulates medium-scale offline datasets and reflects realistic domain shift scenarios. For the diffusion model, the diffusion steps are set to 1000 with a cosine noise schedule, U-Net hidden dimensions of (256, 256), a timestep embedding dimension of 128, and a learning rate of 3×10^{-4}. For VG, the dynamics model ensemble size is 5, with network dimensions of (256, 256), a Q-value update frequency of 2, and value clipping in the range $[-10, 10]$. For DDS, the contrastive temperature is 0.1, momentum encoder decay is 0.995, and the batch size is 256. Finally, for PH, the inverse dynamics network architecture is (256, 256) with an action error threshold of 0.1.

We compare DiffTA against offline RL methods (TD3+BC [34] and IQL [35]) and cross-domain approaches (BOSA [32], IGDF [6], DARA [3], SRPO [36], OTDF [16], UTDS [37]). While TD3+BC, IQL, and DARA are directly from ODRL, UTDS, IGDF, SRPO and OTDF are re-implemented based on their original papers and code to align with ODRL. To ensure a fair comparison, all baseline methods are evaluated with the same number of random seeds. Further details are provided in Table 1.

Table 1 Hyperparameter setup for baseline methods

Category	Hyperparameter	Value
Shared Parameters	Actor network architecture	(256, 256)
	Critic network architecture	(256, 256)
	Learning rate	3×10^{-4}
	Optimizer	Adam
	Discount factor (γ)	0.99
	Replay buffer size	10^6
	Activation function	ReLU
	Target update rate	5×10^{-3}
	Batch size (source/target)	128/128
TD3+BC	Normalization coefficient	2.5
IQL	Temperature coefficient	0.2
	Maximum log std	2
	Minimum log std	-20
	Inverse temperature parameter β	3.0
	Expectile parameter τ	0.7
DARA	Temperature coefficient	0.2
	Maximum log standard deviation	2
	Minimum log standard deviation	-20
	Classifier network architecture	(256, 256)
	Reward penalty coefficient (λ)	0.1
BOSA	Temperature coefficient	0.2
	Policy regularization coefficient (λ_{policy})	0.1
	Transition coefficient ($\lambda_{\text{transition}}$)	0.1
	Threshold parameters (ϵ, ϵ')	$\log(0.01)$
	Value weight (ω)	0.1
	CVAE ensemble size	1 (policy), 5 (dynamics)
	Maximum log standard deviation	2
	Minimum log standard deviation	-20
UTDS	In-distribution uncertainty factor (β_1)	0.001
	Out-of-distribution uncertainty factor (β_2)	$3.0 \rightarrow 0.1$
	Exponential decay factor (α)	0.99995
	Batch size	128
	Learning rate	3×10^{-4}
IGDF	Representation dimension (d)	16/64
	InfoGAN learning rate	3×10^{-4}
	Importance coefficient (α)	1.0
	Data selection ratio (ξ)	0.25/0.75
	Update iterations	7000
	Batch size	128
SRPO	Data selection ratio	0.5/0.2
	Delta coefficient (λ)	0.1/0.3
	Batch size	128
	Learning rate	3×10^{-4}
OTDF	CVAE training steps	10000
	CVAE learning rate	0.001
	Number of sampled latent variables M	10
	Standard deviation of Gaussian distribution	$\sqrt{0.1}$
	Cost function	cosine
	Data filtering ratio ϵ	80
	Policy coefficient β	{0.1, 0.5 }

3.2 Performance in Cross-Domain RL

First, we evaluate the fidelity and alignment of the synthesized trajectories. We visualize DiffTA-generated and target-domain trajectories using t-SNE (Fig. 2). The embeddings reveal that DiffTA's trajectories form compact clusters that closely overlap with those from the target domain, demonstrating its capacity to capture underlying dynamics rather than merely reproducing superficial patterns.

Then, we evaluate the effectiveness of DiffTA in cross-domain RL. We employ IQL as the offline RL method to train policies on a combination of DiffTA-generated and target-domain trajectories. To systematically assess the quality of the generated trajectories, we conduct experiments under two distinct dynamic shifts: friction-0.5 and gravity-0.5, and explore scenarios with varying qualities of source domain datasets. This setup reflects real-world conditions where source data may not always be optimal or expert-level. Our evaluation spans four locomotion tasks (ant, hopper, halfcheetah, walker2d), and the results are summarized in Table 2.

Table 2 Performance comparison across different baselines under friction-0.5 and gravity-0.5 shifts (Best in Bold)

Shift Type	Environment	TD3+BC	IQL	DARA	UTDS	BOSA	SRPO	IGDF	OTDF	DiffTA
friction-0.5	halfcheetah-medium	34.7	40.1	43.3	45.1	51.3	47.1	48.7	**60.5**	58.4 ± 2.2
	halfcheetah-expert	70.6	63.3	75.4	74.3	80.4	92.4	87.6	88.7	**100.8 ± 3.5**
	hopper-medium	44.6	60.4	73.3	76.2	73.3	77.6	64.5	75.3	**83.3 ± 1.9**
	hopper-expert	76.4	90.2	87.7	85.4	94.6	86.7	92.3	**98.4**	96.5 ± 2.8
	walker2d-medium	57.3	75.4	63.1	65.4	63.7	66.4	76.3	65.3	**84.3 ± 3.1**
	walker2d-expert	87.7	90.3	90.4	95.6	93.1	95.3	86.4	91.5	**98.8 ± 2.4**
	ant-medium	41.2	55.4	56.3	60.8	58.8	57.7	61.2	49.7	**66.3 ± 3.3**
	ant-expert	34.2	72.5	77.3	80.1	71.8	74.4	77.6	75.8	**81.2 ± 5.9**
gravity-0.5	halfcheetah-medium	55.4	58.3	56.6	59.3	60.1	67.3	70.3	64.7	**71.8 ± 7.3**
	halfcheetah-expert	57.4	60.2	57.7	58.4	65.5	74.4	77.4	78.9	**85.4 ± 2.1**
	hopper-medium	40.3	45.4	30.3	55.6	49.3	54.3	60.3	63.5	**74.5 ± 1.2**
	hopper-expert	55.6	61.5	63.4	67.8	73.1	70.4	76.3	71.2	**86.7 ± 3.5**
	walker2d-medium	53.4	56.7	55.4	56.3	**60.3**	58.7	56.7	58.4	59.4 ± 2.8
	walker2d-expert	64.3	60.4	65.3	69.4	68.3	72.4	74.1	68.7	**79.6 ± 3.7**
	ant-medium	27.5	33.1	37.5	41.2	43.1	37.6	40.2	37.9	**50.8 ± 5.4**
	ant-expert	24.3	34.2	28.4	44.3	44.5	43.5	42.4	47.8	**56.7 ± 4.1**
	Average	51.6	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	0.00	0.01	0.02	0.04	0.11	0.14	–

Table 3 Performance comparison of different methods under various domain shifts on ant-medium (Best in Bold)

Environment	TD3+BC	IQL	DARA	UTDS	BOSA	SRPO	IGDF	OTDF	DiffTA
friction-0.5	41.2	55.4	56.3	60.8	58.8	57.7	61.2	49.7	**66.3 ± 3.3**
friction-5.0	11.2	9.7	12.3	13.2	10.7	15.4	13.4	15.4	**27.3 ± 3.5**
gravity-0.5	27.5	33.1	37.5	41.2	43.1	37.6	40.2	37.9	**50.8** ± 5.4
gravity-5.0	20.4	25.1	30.3	31.2	28.7	34.3	27.6	28.8	**47.6 ± 4.1**
kinematic-anklejoint-easy	98.7	96.3	94.5	101.4	102.3	97.6	98.3	102.5	**108.9 ± 3.2**
kinematic-anklejoint-medium	95.4	99.3	89.7	87.6	93.4	105.6	101.2	100.7	**110.3 ± 2.9**
kinematic-anklejoint-hard	71.3	79.4	82.4	83.7	80.1	85.5	84.3	87.4	**102.3** ± 6.6
morph-alllegs-easy	77.6	82.4	70.3	**85.6**	77.5	80.2	71.3	75.8	83.1 ± 2.3
morph-alllegs-medium	65.4	57.3	47.6	63.5	68.7	70.2	77.3	80.1	**85.4 ± 3.1**
morph-alllegs-hard	13.2	8.7	15.4	11.7	21.2	18.9	20.3	17.6	**35.5 ± 4.7**
Average	52.2	54.7	53.6	58.0	58.5	60.3	59.5	59.6	**71.8**

Three insights emerge: (1) gravity shifts cause greater degradation than friction due to their stronger impact on dynamics; (2) while higher-quality source data tends to improve target performance, it is not always effective under severe domain gaps; (3) standard offline RL methods like TD3+BC and IQL remain competitive in moderate shift settings despite lacking explicit adaptation mechanisms. DiffTA surpasses baselines, highlighting its effectiveness in target-aligned trajectory generation.

To further validate the statistical significance of performance differences among the compared methods, we additionally conduct the Friedman test and the Nemenyi post-hoc test. The Friedman test yields a statistic of 84.7049 with a p-value of 5.50×10^{-15}, strongly rejecting the null hypothesis that all algorithms perform equivalently across tasks. Subsequently, we apply the Nemenyi post-hoc test to assess pairwise differences. Results indicate that DiffTA significantly outperforms the majority of baselines, including TD3+BC, IQL, DARA, UTDS, BOSA, and SRPO (all $p < 0.05$). Although the differences with IGDF and OTDF are not statistically significant ($p = 0.1132$ and $p = 0.1435$, respectively), DiffTA still achieves the highest overall performance, demonstrating consistent robustness and superior generalization under cross-domain shifts.

3.3 Impact of Varying Dynamic Gaps

To evaluate DiffTA's robustness under varying dynamic discrepancies, we conduct experiments on the ant environment by systematically altering friction, gravity, kinematics, and morphology. Both source and target datasets are of medium quality, while the degree of domain shift is progressively increased to assess degradation trends (Table 3).

As the physical gap widens, most baselines show significant performance drops. For example, average returns decline from above 40 at friction-0.5 to around 10 at friction-5.0, highlighting their limited ability to extrapolate beyond training dynamics. In contrast, DiffTA consistently maintains higher returns across all settings. This robustness stems from its dynamics-aware conditioning: VG guides the diffusion process toward transitions with transferable high-value behavior, while DDS adaptively reduces generation probability in regions inconsistent with target dynamics. These mechanisms enable DiffTA to generate synthetic trajectories that bridge dynamic mismatches and preserve task-relevant value structures, allowing effective generalization even under substantial cross-domain shifts.

3.4 Robustness to Target Domain Data

We investigate how the quality and quantity of target-domain data influence cross-domain offline RL, using the walker2d task with a fixed

Table 4 Performance on walker2d-friction-0.5 under varying target-domain data conditions (Best in Bold)

Condition	Target Domain	TD3+BC	IQL	DARA	UTDS	BOSA	SRPO	IGDF	OTDF	DiffTA
Quality	random	2.1	5.7	3.1	5.0	0.7	2.7	4.5	3.7	**11.5** ± 0.6
	medium	57.3	75.4	63.1	65.4	63.7	66.4	76.3	65.3	**84.3** ± 9.1
	medium-replay	50.8	72.1	58.2	62.7	66.2	65.5	71.9	68.7	**80.4** ± **1.0**
	medium-expert	62.9	81.2	59.9	73.1	64.3	72.5	82.1	84.9	**88.6** ± **2.9**
	expert	87.7	90.3	90.4	95.6	93.1	95.3	86.4	**96.4**	96.2 ± 2.4
Quantity	5,000	3.2	6.5	4.4	5.1	2.9	4.7	5.8	6.6	**9.9** ± 2.3
	10,000	9.7	17.4	10.2	12.9	8.1	14.2	13.6	12.4	**19.3** ± **1.3**
	50,000	20.4	27.8	24.3	22.2	19.2	26.1	29.7	25.1	**31.6** ± **1.9**
	100,000	57.3	75.4	63.1	65.4	63.7	66.4	76.3	65.3	**84.3** ± **4.1**
	200,000	75.2	85.1	78.3	88.9	81.5	89.2	84.7	89.7	**95.0** ± **3.8**
	500,000	90.1	93.8	91.7	98.2	94.3	97.5	91.2	94.8	**100.5** ± **2.7**
Average		47.0	57.3	49.7	54.1	50.7	54.6	56.6	55.7	**64.0**

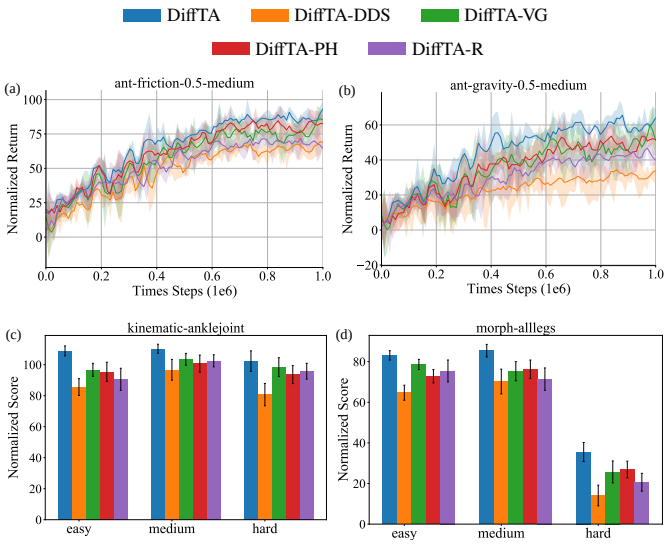

Fig. 3 Normalized Return (mean ± standard deviation) for DiffTA and one-component ablations

medium-quality source dataset (Table 4). Most baselines exhibit substantial degradation as target data become limited or noisy, reflecting their dependence on direct behavioral supervision. In contrast, DiffTA maintains stable performance across different data regimes, owing to its dynamics-aware conditioning that enables the generation of trajectories consistent with target-domain dynamics. This robustness suggests that DiffTA effectively reconstructs missing or low-quality target experiences through generative alignment, highlighting its potential for reliable deployment in practical scenarios where high-quality target data are often scarce.

3.5 Ablation Study on Core Components

To evaluate the contribution of each contextual signal, we perform a component-wise ablation study (Fig. 3). DiffTA-VG, DiffTA-DDS, DiffTA-PH, and DiffTA-R denote variants with Value Guidance, Domain Discrepancy Score, Policy Harmonization, and Reward Context removed, respectively.

Removing DDS causes the most significant performance drop, highlighting its essential role in cross-domain alignment. Without DDS, the diffusion model loses its ability to discern and reconcile discrepancies between source and target dynamics, leading to generated trajectories that drift from target-domain feasibility. Eliminating VG or the reward context also causes a substantial performance decline: VG governs the selection of transferable transitions by constraining generation to value-consistent regions, while the reward context guides the diffusion trajectory toward high-return behaviors. Although PH contributes less to numerical improvement, its absence results in physically inconsistent or out-of-distribution actions, especially when target data are scarce, emphasizing its role in maintaining action-state feasibility during generation. These results demonstrate that DiffTA's contextual signals are interdependent components forming a hierarchical guidance system: VG defines what to transfer, DDS aligns how to transfer, and PH ensures what is transferred remains executable within the target domain. This synergy enables DiffTA to generate trajectories that are transferable, dynamically consistent, and physically realizable, achieving robust cross-domain adaptation.

Table 5 Impact of removing Gaussian dynamics and single-signal ablations on task performance

Task	w/o ensemble	Only VG	Only DDS	Only PH
kinematic-anklejoint	↓ 4.8%	↓ 36.7%	↓ 21.3%	↓ 60.3%
morph-alllegs	↓ 3.6%	↓ 44.2%	↓ 17.4%	↓ 52.3%

While the Gaussian dynamics ensemble in VG may appear complex, it is not redundant. It shares the same state–action–next-state interface as other components and follows a lightweight design, avoiding unnecessary computational overhead. Experimental results show that the Gaussian dynamics ensemble is essential in VG, particularly for cross-domain generation and model uncertainty control. As shown in Table 5, when the entire Gaussian dynamics ensemble is ablated, performance drops by 4.8% and 3.6% on the kinematic-anklejoint and morph-alllegs tasks, respectively.

We further evaluated the interchangeability of the guidance signals by conditioning on each in isolation. Results reveal that no single signal suffices: DDS-only conditioning leads to performance drops of 21.3% and 17.4%; VG-only conditioning suffers substantial declines of 36.7% and 44.2%; and PH-only conditioning collapses by 60.3% and 52.3%, reverting to the performance levels of generic baselines like Diffuser. These findings demonstrate that VG, DDS, and PH provide distinct, non-redundant constraints targeting value transferability,

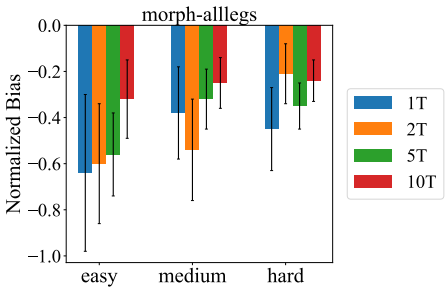

Fig. 4 Value estimation bias of the target critic under dynamics-ensemble rollouts across target-data regimes

Table 6 One-step transition distance on ant-medium-morph-alllegs across three difficulty settings

Setting	DiffTA	DiffTA w/o DDS	SynthER
Easy	**1.29**	2.34	2.69
Medium	**2.07**	2.67	3.56
Hard	**1.43**	2.78	2.96

dynamics alignment, and physical executability, respectively, making them collectively indispensable for effective adaptation.

We further investigate the sensitivity of VG to target data scarcity, where dynamics model errors could propagate to value estimation. We conduct a regime-wise analysis on the ant-medium-morph-alllegs task by varying target dataset sizes across {1T, 2T, 5T, 10T}, where T represents 5% of the source data size. For each regime, we train the dynamics ensemble for 10k steps and quantify the estimation bias. Specifically, on a fixed evaluation set ($N = 256$), we compare the ensemble-predicted value against the ground-truth Monte Carlo return. As shown in Fig. 4, the estimation bias decreases as target data availability increases. Notably, the bias remains consistently negative due to the twin-critic minimization operator. This indicates a mild underestimation rather than overestimation, which serves as a safe conservative bound for policy learning. Despite limited supervision, DiffTA does not exhibit significant performance degradation (consistent with Section 3.1), demonstrating that the framework remains effective and robust even under severe target data scarcity.

Additionally, to validate that DDS captures transition-level dynamics rather than mere distributional similarity, we performed a one-step transition consistency check. Using the ant-medium-morph-alllegs task across three difficulty settings, we sampled a fixed set of 256 state-action pairs (s, a) from the target dataset. For each pair, we computed the mean vector distance between the ground-truth next state s' and the prediction $\hat{s}'$ generated by DiffTA, DiffTA w/o DDS, and SynthER. As shown in Table 6, DiffTA achieves the lowest transition error in all settings.

Lastly, we report a sensitivity check for diffusion steps and the VG ensemble size. As shown in Table 7 and Table 8, performance is sensitive to using too few diffusion steps but improves only marginally once the step count is sufficiently large, indicating diminishing returns. In contrast, results are relatively stable across different VG ensemble sizes: a small ensemble already works well, and increasing the ensemble further does not yield consistent gains.

Table 7 Sensitivity to diffusion steps

Steps	500	1000	2000	5000	10000
Performance	↓ 8.4%	0.0%	↑ 4.2%	↑ 5.5%	↑ 1.6%

Table 8 Sensitivity to ensemble size

Ensemble size	2	10	20	50
Performance	↓ 1.1%	0.0%	↑ 1.5%	↓ 0.3%

Table 9 Training time (normalized) and task performance under friction-0.5 for IGDF, DiffTA, DiffTA+EDP, and DiffTA+EDP+SiTo

Method	Training Time	halfcheetah	walker2d
IGDF	1.0×	48.7 ± 3.8	76.3 ± 3.5
DiffTA (original)	5.1×	58.4 ± 2.2	84.3 ± 9.1
DiffTA + EDP	2.4×	54.5 ± 3.1	79.6 ± 5.4
DiffTA + EDP + SiTo	1.7×	54.2 ± 4.1	78.3 ± 2.5

3.6 Training Efficiency and Acceleration

To investigate whether DiffTA can be accelerated without compromising its generative fidelity or policy performance, we analyze its computational profile and introduce a lightweight variant inspired by Efficient Diffusion Policy (EDP) [38]. We adopt EDP's deterministic sampling approximation that replaces stochastic diffusion with a closed-form mean trajectory update, preserving conditional guidance while reducing the number of denoising steps by an order of magnitude. Building on EDP, we further apply a similarity-based token pruning method (SiTo [39]).

As shown in Table 9, the EDP variant achieves a 21× speedup in trajectory generation, indicating that dynamics alignment relies primarily on contextual conditioning rather than stochastic sampling. When combined with SiTo (DiffTA+EDP+SiTo), the framework further reduces training overhead from 5.1× to 1.7× relative to IGDF. Despite these significant accelerations, the performance drop is minimal (≈ 7% in halfcheetah and walker2d), confirming that our guidance signals (VG, DDS, PH) robustly preserve generation quality even under approximate deterministic updates.

References

[1] Janner M, Du Y, Tenenbaum J B, Levine S. Planning with diffusion for flexible behavior synthesis. arXiv preprint arXiv:2205.09991, 2022

[2] Lu C, Ball P, Teh Y W, Parker-Holder J. Synthetic experience replay. In: Proceedings of the Advances in Neural Information Processing Systems. 2024, 46323–46344

[3] Liu J, Zhang H, Wang D. Dara: Dynamics-aware reward augmentation in offline reinforcement learning. arXiv preprint arXiv:2203.06662, 2022

[4] Eysenbach B, Asawa S, Chaudhari S, Levine S, Salakhutdinov R. Off-dynamics reinforcement learning: Training for transfer with domain classifiers. arXiv preprint arXiv:2006.13916, 2020

[5] Xu K, Bai C, Ma X, Wang D, Zhao B, Wang Z, Li X, Li W. Cross-domain policy adaptation via value-guided data filtering. In: Proceedings of the Advances in Neural Information Processing Systems. 2023, 73395–73421

[6] Wen X, Bai C, Xu K, Yu X, Zhang Y, Li X, Wang Z. Contrastive representation for data filtering in cross-domain offline reinforcement learning. arXiv preprint arXiv:2405.06192, 2024

[7] Desai S, Durugkar I, Karnan H, Warnell G, Hanna J, Stone P. An imitation from observation approach to transfer learning with dynamics mismatch. In: Proceedings of the Advances in Neural Information Processing Systems. 2020, 3917–3929

[8] Zhang G, Zhong L, Lee Y, Lim J J. Policy transfer across visual and dynamics domain gaps via iterative grounding. arXiv preprint arXiv:2107.00339, 2021

[9] Hanna J, Stone P. Grounded action transformation for robot learning in simulation. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2017, 3834–3840

[10] Karnan H, Desai S, Hanna J P, Warnell G, Stone P. Reinforced grounded action transformation for sim-to-real transfer. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. 2020, 4397–4402

[11] Niu H, Ji T, Liu B, Zhao H, Zhu X, Zheng J, Huang P, Zhou G, Hu J, Zhan X. H2o+: an improved framework for hybrid offline-and-online rl with dynamics gaps. arXiv preprint arXiv:2309.12716, 2023

[12] Fickinger A, Cohen S, Russell S, Amos B. Cross-domain imitation learning via optimal transport. arXiv preprint arXiv:2110.03684, 2021

[13] Hejna D, Pinto L, Abbeel P. Hierarchically decoupled imitation for morphological transfer. In: Proceedings of the International Conference on Machine Learning. 2020, 4159–4171

[14] Raychaudhuri D S, Paul S, Vanbaar J, Roy-Chowdhury A K. Cross-domain imitation from observations. In: Proceedings of the International Conference on Machine Learning. 2021, 8902–8912

[15] Lyu J, Bai C, Yang J W, Lu Z, Li X. Cross-domain policy adaptation by capturing representation mismatch. In: Proceedings of the International Conference on Machine Learning. 2024, 33638–33663

[16] Lyu J, Yan M, Qiao Z, Liu R, Ma X, Ye D, Yang J W, Lu Z, Li X. Cross-domain offline policy adaptation with optimal transport and dataset constraint. In: Proceedings of the International Conference on Learning Representations. 2025

[17] Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. In: Proceedings of the International Conference on Machine Learning. 2015, 2256–2265

[18] Nichol A Q, Dhariwal P. Improved denoising diffusion probabilistic models. In: Proceedings of the International Conference on Machine Learning. 2021, 8162–8171

[19] Yang L, Zhang Z, Song Y, Hong S, Xu R, Zhao Y, Zhang W, Cui B, Yang M H. Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys, 2023, 56(4): 1–39

[20] Croitoru F A, Hondru V, Ionescu R T, Shah M. Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10850–10869

[21] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A C. Improved training of wasserstein gans. In: Proceedings of the Advances in Neural Information Processing Systems. 2017

[22] Jiangzhou D, Songli W, Jianmei Y, Lianghao J, Yong W. Dgrm: Diffusion-gan recommendation model to alleviate the mode collapse problem in sparse environments. Pattern Recognition, 2024, 110692

[23] Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E L, Ghasemipour K, Gontijo Lopes R, Karagol Ayan B, Salimans T, others . Photorealistic text-to-image diffusion models with deep language understanding. In: Proceedings of the Advances in Neural Information Processing Systems. 2022, 36479–36494

[24] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023, 3836–3847

[25] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis. In: Proceedings of the Advances in Neural Information Processing Systems. 2021, 8780–8794

[26] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, 10684–10695

[27] Ng W Z T, Chen J, Zhang T. Off-dynamics conditional diffusion planners. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. 2024, 7461–7468

[28] Ajay A, Du Y, Gupta A, Tenenbaum J B, Jaakkola T S, Agrawal P. Is conditional generative modeling all you need for decision making? In: Proceedings of the International Conference on Learning Representations. 2023

[29] Van L L P, Nguyen M H, Kieu D, Le H, Tran H T, Gupta S. Dmc: Nearest neighbor guidance diffusion model for offline cross-domain reinforcement learning. arXiv preprint arXiv:2507.20499, 2025

[30] Ada S E, Oztop E, Ugur E. Diffusion policies for out-of-distribution generalization in offline reinforcement learning. IEEE Robotics and Automation Letters, 2024, 9(4): 3116–3123

[31] Lyu J, Xu K, Xu J, Yang J W, Zhang Z, Bai C, Lu Z, Li X, others . Odrl: A benchmark for off-dynamics reinforcement learning. In: Proceedings of the Advances in Neural Information Processing Systems. 2024, 59859–59911

[32] Liu J, Zhang Z, Wei Z, Zhuang Z, Kang Y, Gai S, Wang D. Beyond ood state actions: Supported cross-domain offline reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 13945–13953

[33] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of the International Conference on Machine Learning. 2018, 1861–1870

[34] Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration. In: Proceedings of the International Conference on Machine Learning. 2019, 2052–2062

[35] Kostrikov I, Nair A, Levine S. Offline reinforcement learning with implicit q-learning. arXiv preprint arXiv:2110.06169, 2021

[36] Xue Z, Cai Q, Liu S, Zheng D, Jiang P, Gai K, An B. State regularized policy optimization on data with dynamics shift. In: Pro-

ceedings of the Advances in Neural Information Processing Systems. 2024, 32926–32937

[37] Bai C, Wang L, Hao J, Yang Z, Zhao B, Wang Z, Li X. Pessimistic value iteration for multi-task data sharing in offline reinforcement learning. Artificial Intelligence, 2024, 326: 104048

[38] Kang B, Ma X, Du C, Pang T, Yan S. Efficient diffusion policies for offline reinforcement learning. In: Proceedings of the Advances in Neural Information Processing Systems. 2023, 67195–67212

[39] Zhang E, Tang J, Ning X, Zhang L. Training-free and hardware-friendly acceleration for diffusion models via similarity-based token pruning. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 9878–9886