Explore 3D Dance Generation via Reward Model from Automatically-Ranked Demonstrations

Zilin Wang[1*], Haolin Zhuang[1*], Lu Li[1*], Yinmin Zhang[2], Junjie Zhong[3], Jun Chen[1], Yu Yang[1], Boshi Tang[1], Zhiyong Wu[1†]
[1] Shenzhen International Graduate School, Tsinghua University
[2] University of Sydney, [3] Waseda University

1. Introduction

1.1 Task

- Music-conditioned 3D dance generation
  - ✓ Input: condition music & initial movement
  - ✓ Output: dance movements aligned with give music
  - ✓ Making more people aware of and enjoy the art of dance

1.2 Motivation

- Shortcomings in supervised learning approaches
  - ✓ Weak generalization for unseen music
  - ✓ Fragility of auto-regressive models
  - ✓ Misalignment between generated dances and human preferences

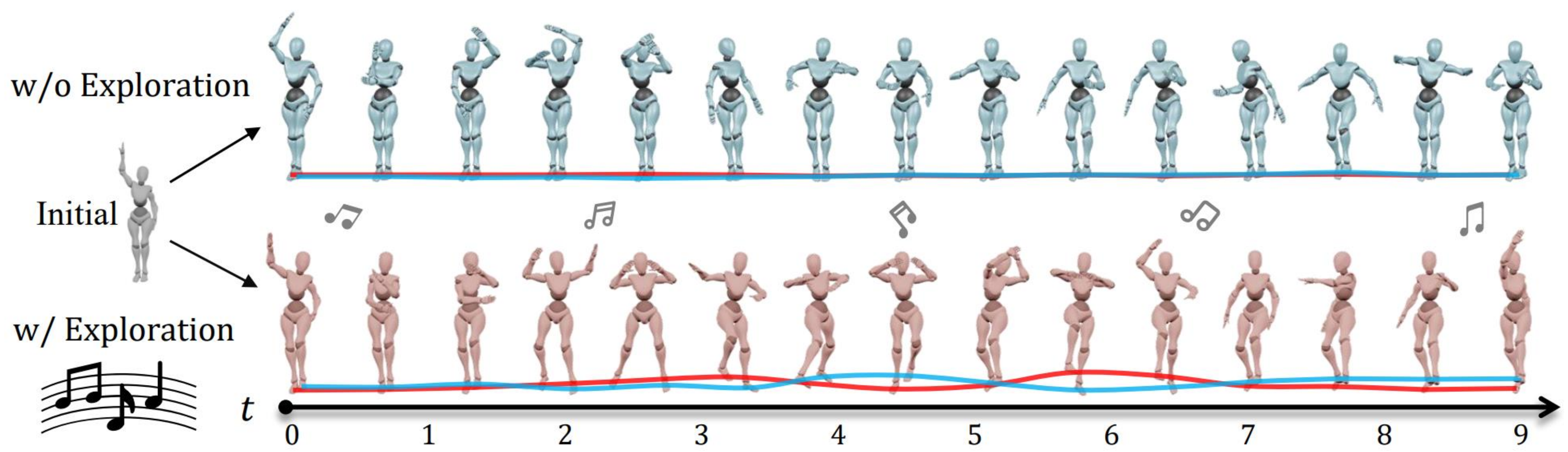

Figure 1. Visualizations. Red and blue lines represent right and left leg movements, respectively. *Top:* Dance examples generated by the policy lack exploration, exhibiting limited leg movements' diversity and quality. *Bottom:* Dance examples generated by the policy reinforced via exploration align with human preferences, showcasing increased leg movements' diversity and quality.

2. Methodology

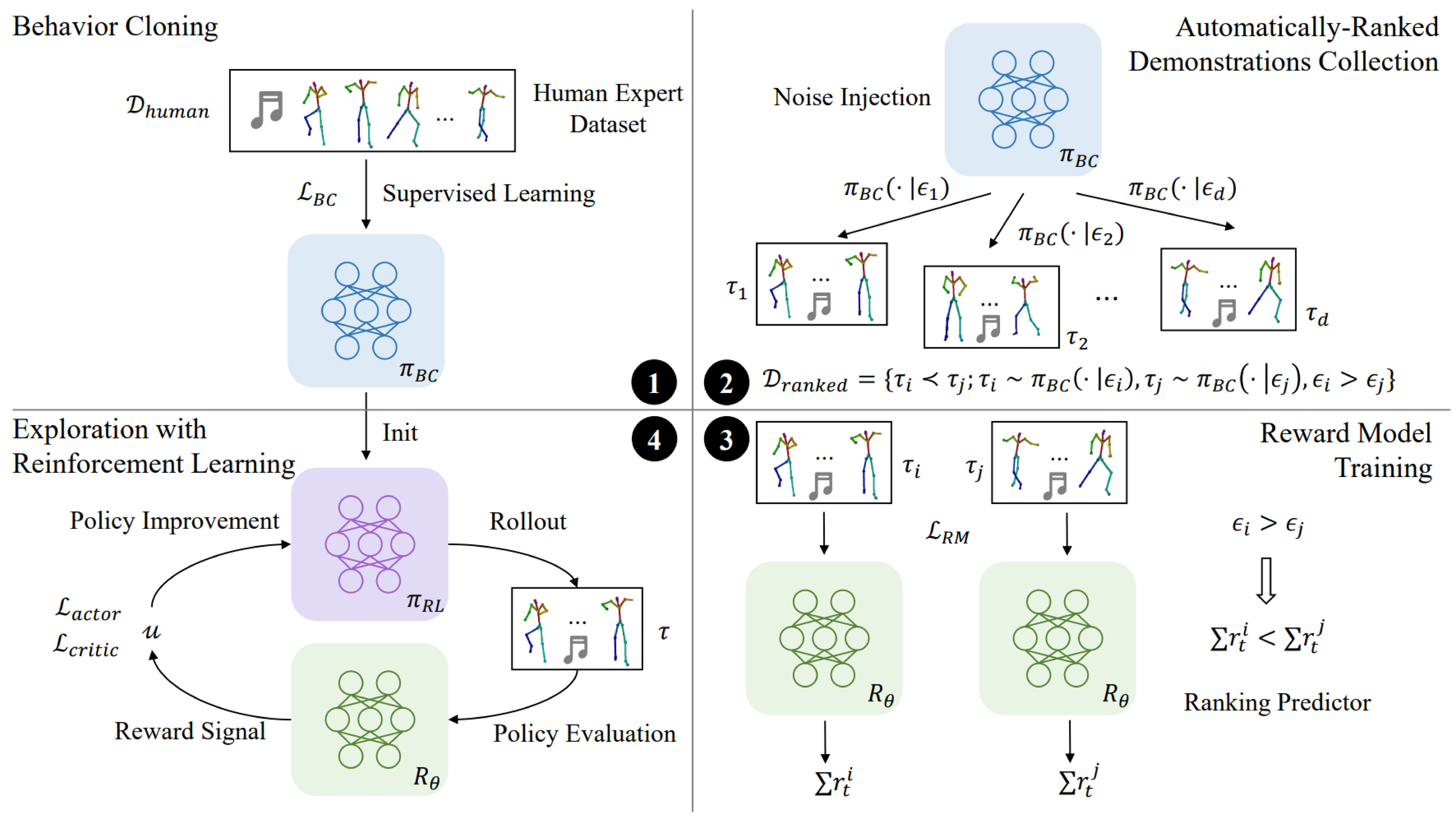

Figure 2. Diagram of our E3D2: (1) An initial policy π_{BC} is distilled from the human expert dataset through behavior cloning. (2) Automatically ranked dance demonstrations are collected by π_{BC} with different levels of noise. (3) A reward model R_θ is trained from these automatically ranked demonstrations to rank the quality of dance trajectories. (4) A reinforcement learning policy π_{RL} is initialized with π_{BC} and optimized to obtain the optimal dance policy, guided by the reward model R_θ.

3. Experiments

3.1 Comparisons with State-Of-The-Arts

Table 1. Evaluation results on test set of different dance generation frameworks. To ensure a fair comparison with baselines, we report the results of Bailando without RL fine-tuning on the test set.

	Motion Quality		Motion Diversity		
	$FID_k \downarrow$	$FID_g \downarrow$	$DIV_k \uparrow$	$DIV_g \uparrow$	$BAS \uparrow$
Ground-Truth	17.10	10.60	8.19	7.45	0.2484
FACT (Li et al. 2021)	37.31	34.87	5.75	5.47	0.2175
Bailando (Siyao et al. 2022)	28.62	9.95	6.27	6.22	0.2220
E3D2 (Ours)	**26.25**	**8.94**	**7.96**	**6.49**	**0.2232**

3.2 Does exploration provide more alignment?

Table 2. Human-based evaluation results. We conduct a human evaluation to ask annotators to select the preferred dances through pairwise comparison.

	Win	Fail	No Preference
Ours vs. FACT	94.4%	4.2%	1.4%
Ours vs. Bailando	66.7%	28.7%	4.6%

3.3 Is a learned reward function more effective than a hand-designed one?

Table 3. Performance of hand-designed reward. 'Steps' is the interaction numbers between the agent and the environment. The hand-designed reward only considers BAS and orientation, leading to decreasing performance on other metrics during the optimization.

Steps	$FID_k \downarrow$	$FID_g \downarrow$	$DIV_k \uparrow$	$DIV_g \uparrow$	$BAS \uparrow$
0M	28.62	9.95	6.27	6.22	0.2220
1M	45.39	15.41	4.17	3.49	0.2338
2M	46.25	17.20	4.63	3.46	0.2374
3M	43.10	18.59	4.82	2.90	0.2283
4M	47.80	22.15	4.97	2.47	0.2388
5M	56.30	24.58	5.52	3.56	0.2442

3.4 Higher level noise leads to the worse demonstrations?

Table 4. Ablation on the impact of noise in the training set. The performance of the BC policy gradually decreases as the noise level increases. $\overline{u}$ represents the average total reward across all trajectories in the training set.

ϵ	$FID_k \downarrow$	$FID_g \downarrow$	$DIV_k \uparrow$	$DIV_g \uparrow$	$BAS \uparrow$	$\overline{u}$
0.02	13.94	2.71	8.01	6.20	0.2782	206.31
0.25	40.45	22.39	4.41	2.40	0.2501	127.68
0.50	48.59	29.80	3.72	1.61	0.2547	52.09
0.75	53.79	33.35	3.31	1.32	0.2451	-20.24
1.00	57.18	35.67	3.04	1.17	0.2427	-91.53

3.5 What is the performance of the reward model?

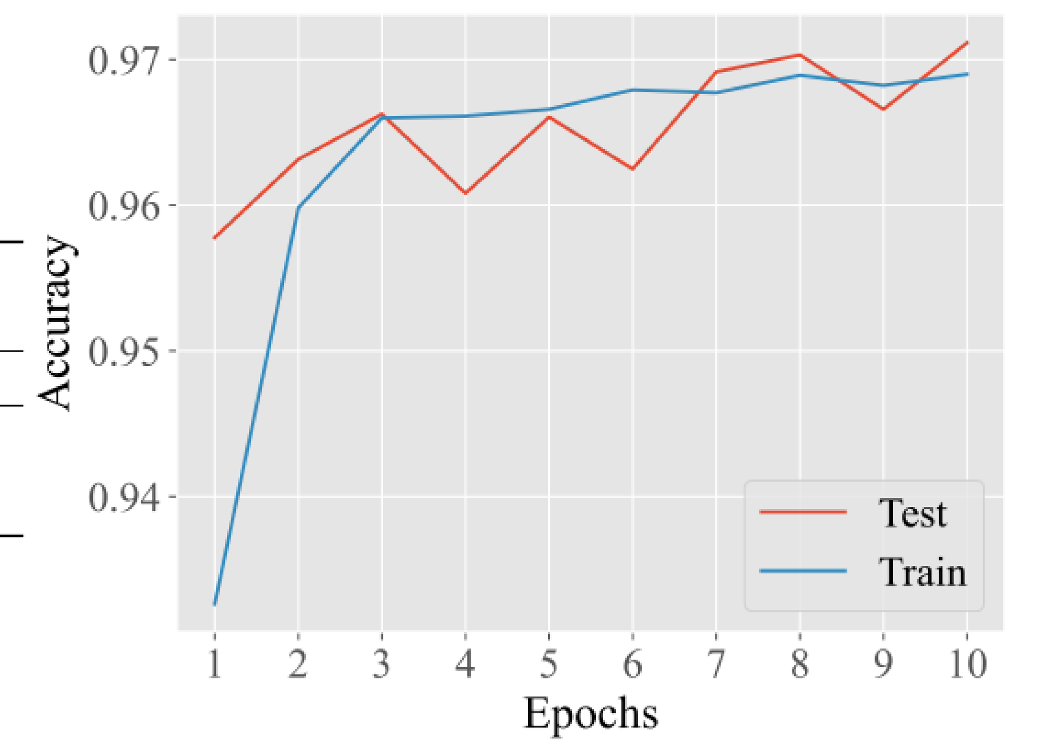

Figure 3. Reward model accuracy: The classification accuracy of the reward model on dances generated by policies with varying levels of noise during training. The reward model exhibits excellent generalization on the test set.

Table 5. Pose prediction accuracy. We evaluate the behavior cloning policy on both seen and unseen music. 'Complete Pose': both the codes of upper and lower half bodies are correct; 'Partial Pose': at least one code is correct. These results demonstrate the limited generalization capabilities of supervised learning approaches.

Dataset	Complete Pose	Partial Pose
Music Seen	54.69%	73.44%
Music Unseen	2.32%	7.52%

Table 6. Performance of behavior cloning policy on seen and unseen music. The significant gap indicates the limited generalization of supervised learning approaches.

Dataset	$FID_k \downarrow$	$FID_g \downarrow$	$DIV_k \uparrow$	$DIV_g \uparrow$	$BAS \uparrow$
Music Seen	8.48	1.88	8.28	6.86	0.2854
Music Unseen	28.62	9.95	6.27	6.22	0.2220

The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Paper ID: 6233

Visual comparisons

Contact: wangzl21@mails.tsinghua.edu.cn