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]The University of Sydney
[3]Waseda University
{wangzl21, zhuanghl21, lilu21}@mails.tsinghua.edu.cn, yinmin.zhang@sydney.edu.au, junjiezhong@ruri.waseda.jp
{y-chen21, yy20, tbs22}@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn

Abstract

This paper presents an Exploratory 3D Dance generation framework, E3D2, designed to address the exploration capability deficiency in existing music-conditioned 3D dance generation models. Current models often generate monotonous and simplistic dance sequences that misalign with human preferences because they lack exploration capabilities. The E3D2 framework involves a reward model trained from automatically-ranked dance demonstrations, which then guides the reinforcement learning process. This approach encourages the agent to explore and generate high quality and diverse dance movement sequences. The soundness of the reward model is both theoretically and experimentally validated. Empirical experiments demonstrate the effectiveness of E3D2 on the AIST++ dataset.

Introduction

Music-conditioned 3D dance generation is an emerging field that combines the art of dance and the science of machine learning, fostering a novel and creative fusion. By utilizing music as a guiding condition, dance generative models create dance poses synchronized with the melody and rhythm of the music. Several studies (Huang et al. 2021, 2022; Li et al. 2021; Siyao et al. 2022) utilize generative networks to auto-regressively generate dance sequences in supervised learning, with music as the condition and human dance poses as the supervisory signal. These approaches are capable of producing complete dance movements, as significant advancements in the field of dance generation.

Nevertheless, we observe that supervised learning approaches often exhibit the following three shortcomings: (1) Weak generalization for unseen music, which affects diversity and quality, (2) Fragility of auto-regressive models, which are prone to severe compounding rollout errors, particularly when data is scarce, leading to the potential collapse of the dance sequence, and (3) Misalignment between generated dances and human preferences, which stems from the excessive focus on mimicking human movements without considering human preferences (*e.g.*, movement difficulty and aesthetic appeal). Inspired by the learning process of human dancers, novice dancers not only require the mechanical imitation of movements from dance experts but also continuous practice and exploration to develop their skills. Furthermore, receiving feedback from experts plays a crucial role in reinforcing their movements, ultimately helping them become proficient dancers.

In this work, we argue that the aforementioned three issues arise from the lack of **exploration** capacity in current dance generation models. We expect trained dance agents to explore various movements within the dance space while receiving accurate signals indicating which movements are desirable, thereby increasing the probability of generating such movements. Based on this assumption, we propose the **Exploratory 3D Dance generation framework, E3D2,** to address the issue of exploration. To achieve this, we model the music-conditioned dance generation task as a Markov Decision Process (MDP) and employ Reinforcement Learning (RL) to endow the dance agent with the ability to explore. For the reward signal, we utilize Inverse Reinforcement Learning (IRL) to train a Reward Model (RM) from automatically ranked dance demonstrations, which guides the exploration and exploitation of the dance agent. As shown in Figure 2, we firstly use Behavior Cloning (Michie, Bain, and Hayes-Miches 1990) to train an initial dance generation policy, allowing the agent to learn basic dance movements. Then, we inject increasing levels of noise into multiple cloned initial dance generation policies to acquire multiple policies with decreasing performance, generating different quality dance demonstrations. Next, we train a reward model with these automatically ranked dance demonstrations. Finally, under the guidance of the learned reward model, we encourage the dance agent to explore using reinforcement learning, ultimately obtaining the optimal dance generation policy.

Our design enables the dance policy to address the above issues through exploration: (1) To tackle the limited diversity and quality of generated dances, reinforcement learning encourages the dance agent to efficiently explore a broader range of state-action pairs, where new movements emerge

[*]These authors contributed equally.
[†]Corresponding author.

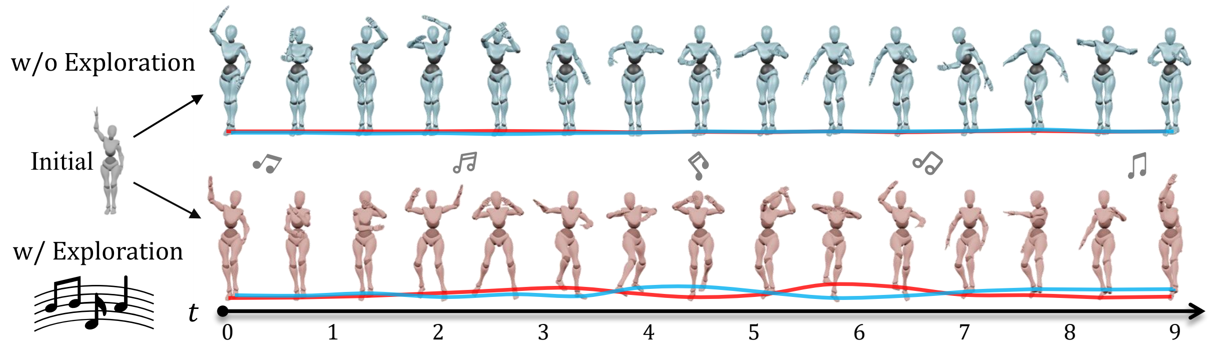

Figure 1: Visualizations. Red and blue lines represent right and left leg movements, respectively. *Top*: Dance examples generated by the policy lack exploration, exhibiting limited leg movements' diversity and quality. *Bottom*: Dance examples generated by the policy reinforced via exploration align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

naturally, either by combining the sub-actions from various human dances or by creating entirely novel dances. This results in dance movements with increased diversity and quality, as shown in Figure 1. Moreover, the consistent distribution of music and dance movement in the environment ensures the stability of dynamics, such as transition probability, allowing the learned reward to generalize. This theoretically guarantees the generalization of the learned policies, guided by the learned reward, for both seen and unseen environments. (2) Regarding the fragility of auto-regressive models under supervised learning, which suffers from severe compounding rollout errors of single-step decisions with respect to long planning horizons, our proposed method is optimized by sequence-based reward with trial and error. Through sequence-based exploration and exploitation, our proposed method focuses on the generated dance trajectories rather than its single-step error, avoiding the compounding errors (Asadi et al. 2019; Janner, Li, and Levine 2021; Janner et al. 2022). (3) To address the misalignment between the policy and human preferences, our proposed reward model is able to distinguish the differences between attractive and ordinary dances due to the assumption that the dance generation with a higher noise level aligns less with human preferences. During the exploration and exploitation, human preference is incorporated into the dance policy through the guidance of the reward model.

Empirical experiments on the AIST++ dataset (Li et al. 2021) demonstrate that the proposed E3D2 outperforms the behavior cloning (pure supervised) method across multiple metrics. Moreover, we perform an in-depth analysis and provide a theoretical proof (in Appendices section[1]) of the reward model. The contributions of this article are three-fold:

- We illuminated three issues, weak generalization, fragility, and misalignment, in existing supervised dance generation methods attributable to a lack of exploration capability.
- To address the deficiency of exploration, we propose an Exploratory 3D Dance generation framework, E3D2, which encourages dance agents to explore by introducing the inverse reinforcement learning method with a learned reward model that reflects human preference.
- Empirical experiments demonstrate the effectiveness and generalization performance of our reward model and E3D2 over supervised models.

[1]https://arxiv.org/abs/2312.11442

Related Works

Music-Conditioned Dance Generation

Music-conditioned dance generation is a cross-modal task involving auditory and visual integration. Existing methods for music-conditioned dance generation can be broadly classified into two categories: retrieval-based methods and direct generation methods. Retrieval-based methods (Fukayama and Goto 2015; Ye et al. 2020; Chen et al. 2021a; Au et al. 2022) divide dances into fixed length units and choreograph by concatenating these units according to the melody of the music. Unfortunately, the fixed length and Beats Per Minute (BPM) of the segmented dance units imposed significant restrictions on the rhythm of the music used to drive the dance. To tackle these issues, direct generation methods (Ahn et al. 2020; Huang et al. 2021, 2022; Zhuang et al. 2022; Valle-Pérez et al. 2021; Wang et al. 2022; Gao et al. 2022; Li et al. 2022, 2020; Tseng, Castellon, and Liu 2023) have been proposed which generate dance motion from scratch. These methods are trained in a supervised learning fashion, with music as the conditioning input and real human dance as the supervisory signal. In this work, we focus on exploratory capabilities during training to improve the quality and diversity of the generated dance sequences.

Preference-Based Inverse Reinforcement Learning

The goal of Preference-based Inverse Reinforcement Learning (PIRL) (Sugiyama, Meguro, and Minami 2012; Christiano et al. 2017) is to learn a reward function from expert preferences. Compared with learning the reward model directly from expert behaviors through conventional IRL methods (Russell 1998; Ng and Russell 2000; Abbeel and Ng 2004), *e.g.*, Adversarial Inverse Reinforcement Learning (AIRL) (Fu, Luo, and Levine 2018), PIRL have been effectively applied in many high-dimensional state spaces

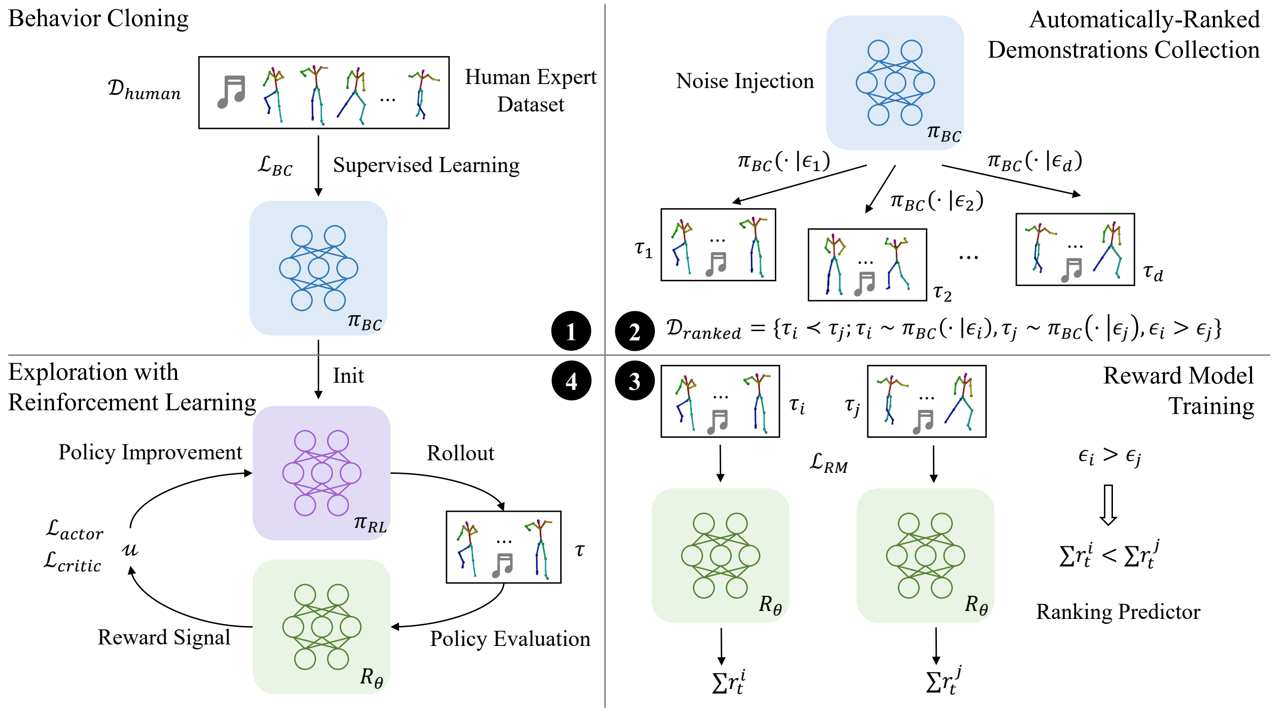

Figure 2: Diagram of our E3D2: (1) An initial policy π_{BC} is distilled from the human expert dataset through behavior cloning. (2) Automatically ranked dance demonstrations are collected by π_{BC} with different levels of noise. (3) A reward model R_θ is trained from these automatically ranked demonstrations 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.

(Brown, Goo, and Niekum 2020; Ibarz et al. 2018) where AIRL may not work effectively (Tucker, Gleave, and Russell 2018). Besides, PIRL could also serve as a way to introduce the human feedback (RLHF) (Christiano et al. 2017; Warnell et al. 2018; MacGlashan et al. 2017), make the model better aligned with human preferences (Stiennon et al. 2020; Wu et al. 2021; Nakano et al. 2021; Ganguli et al. 2022; Glaese et al. 2022). To address the issue of sub-optimal demonstrations, T-REX (Brown et al. 2019) trained a reward model conditioned on states with expert-provided ranking information and then trained an agent that surpasses the sub-optimal demonstrator using the reward model. Based on T-REX, D-REX (Brown, Goo, and Niekum 2020) proposed a generation method of automatically ranked demonstrations by injecting different levels of noise into the behavior cloning policy. D-REX is highly relevant to the demonstration collection of E3D2. However, our main focus is not so much that we proposed a novel PIRL algorithms, or our successful adoption of D-REX in dance generation, but rather our methods solve exploration capability deficiency plaguing existing music-conditioned 3D dance generation models, that were previously unaddressed and holds significant importance.

Preliminary

Given a music-driven dance dataset $\mathcal{D} = \{(\boldsymbol{m}^i, \boldsymbol{p}^i)\}_{i=1}^N$ consisting of N sequence pairs, where $\boldsymbol{m}^i \in \mathcal{M}$ is a music feature sequence, and $\boldsymbol{p}^i \in \mathcal{P}$ is the corresponding dance sequence, $\mathcal{M}$ and $\mathcal{P}$ represent the music feature space and the dance motion space, respectively. We treat music-conditioned dance generation as a sequential decision problem (Sutton and Barto 2018) and model it as a Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, R, P, \gamma, T)$, where $\mathcal{S}, \mathcal{A}$ represent state and action spaces, T and R represent the termination of the episode and reward function, and $\gamma \in (0, 1)$ represents the discount factor. We identify two entities, the *environment* and the *agent*, where the environment is determined by MDP and the agent is determined by the policy π. And m_t and $\hat{p}_t$ represent the music feature and dance pose at timestep t, respectively. To sufficiently consider the consistency of the dance generation sequence, we instantiate the MDP by extending the state with history information. At the beginning of each episode, $t = 0$, the dance *agent* receives the initial state $s_0 = \{m_{init}, p_{init}, m_0\} \in \mathcal{S}$, which is randomly sampled from the dataset by the *environment*, where $\mathcal{S}$ is the state space with $s_t \in \mathcal{P}^{t+1} \times \mathcal{M}^{t+2}$ and m_{init} and p_{init} are the initial music feature and dance pose, respectively. Then, the agent generates an action $a_0 \sim \pi(\cdot|s_0) \in \mathcal{A}$ according to the policy π, where the action space $\mathcal{A} = \mathcal{P}$ and thus $a_t = \hat{p}_t$. Following this, the environment receives the action $\hat{p}_0$ and obtains the next state $s_1 = \{m_{init}, p_{init}, m_0, \hat{p}_0, m_1\}$ using the deterministic state transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S}) = s_t.\text{extend}(\{\hat{p}_t, m_{t+1}\})$. After that, the reward r_t of taking action a_t at state s_t is obtained from the reward function $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. The agent then continues to make decisions based on s_1 to determine the next action a_1. This process continues until the termination of the episode T, where the termination state $s_{T-1} = \{m_{init}, p_{init}, m_0, \hat{p}_0, \cdots, m_{T-1}, \hat{p}_{T-1}\}$ is reached (there is no subsequent music feature m_T in the termination state) and target dance sequence $\{p_{init}, \hat{p}_0, \cdots, \hat{p}_{T-1}\}$ is obtained. The objective of our learning algorithm is to train a dance agent with optimal policy $\pi^*(a|s)$ to maximize the expected discounted return $J(\pi^*) = E_{\tau \sim \pi^*}[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t)]$, where $\tau = \{m_{init}, p_{init}, m_0, \hat{p}_0, m_1, \hat{p}_1, \cdots, m_{T-1}, \hat{p}_{T-1}\}$ and $r(s_t, a_t) = R(s_t, a_t)$ means the reward value, abbreviated as r_t. To simplify the notation, in the following parts, unless otherwise specified, $p_t = \hat{p}_t$.

Methodology

Our framework comprises four main steps, as illustrated in Figure 2. The subsequent sections provide a comprehensive depiction of each component.

Behavior Cloning

To supply demonstration data for reward model training and establish the initial skill required for efficient exploration, we learn an initial policy, π_{BC}, from the human expert dataset $\mathcal{D}_{human}$ in a supervised learning manner. Specifically, we follow (Siyao et al. 2022) including network architecture (*i.e.*, Transformer), objective (*i.e.*, cross-entropy loss $\mathcal{L}_{BC}$), and action space discretization (*i.e.*, VQ-VAE).

Automatically-Ranked Demonstrations Collection

In this section, we will describe how to use π_{BC} to collect automatically ranked demonstrations. Specifically, we obtain policies with performance between the behavior cloning policy π_{BC} and a completely random policy by injecting noise of different levels into the pretrained behavior cloning policy, similar to D-REX (Brown, Goo, and Niekum 2020). Empirically (in Discussion section), we show that given a noise schedule $\mathcal{E} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_d)$ where the $\epsilon_i, i \in \{1.\cdots, d\}$ means the noise range in $[0, 1]$ and are ordered as $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_d$. Intuitively, the dance agent's performance $J(\cdot)$ is likely to have the following ranking: $J(\pi_{BC}(\cdot|\epsilon_1)) < J(\pi_{BC}(\cdot|\epsilon_2)) < \cdots < J(\pi_{BC}(\cdot|\epsilon_d))$.

In practice, we collect demonstrations by using the ϵ-greedy strategy to inject noise into the policy. That is, at each decision-making step, the agent has a probability of ϵ to uniformly sample an action a from the action space $\mathcal{A}$, and a probability of $1 - \epsilon$ to decide on the action a based on its learned policy π_{BC}. For each noise ϵ_i, K dance trajectories are generated to construct the dataset for training the reward model. Finally, the dataset contains $d \times K$ trajectories with the following ranking relationship:

$$D_{ranked} = \{\tau_i \prec \tau_j; \tau_i \sim \pi_{BC}(\cdot|\epsilon_i), \tau_j \sim \pi_{BC}(\cdot|\epsilon_j), \epsilon_i > \epsilon_j\}, \tag{1}$$

where $\tau_i \prec \tau_j$ means τ_i is worse than τ_j.

Reward Model

In this section, given the automatically-ranked demonstrations dataset D_{ranked}, we will discuss the network architecture and training method of the reward model. Intuitively, as a discriminative model, the reward model has greater potential for extrapolation tasks compared to auto-regressive models.

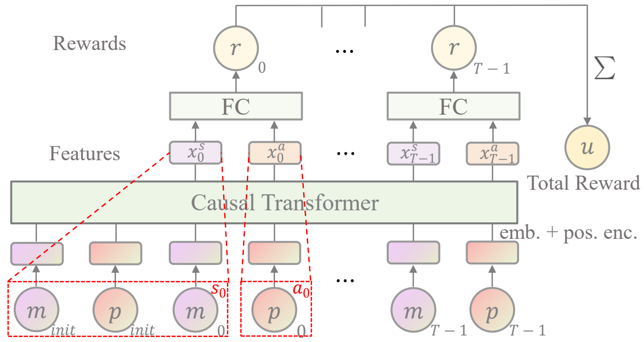

Figure 3: Overview of our reward model. Tokens are generated by combining music and pose embeddings with position encoding, followed by a Causal Transformer to extract features. The extracted features are then fed into a fully connected layer to predict the total reward.

Network Architecture An overview of the reward model is shown in Figure 3. Given a dance trajectory $\tau = \{m_{init}, p_{init}, m_0, p_0, m_1, p_1, \cdots, m_{T-1}, p_{T-1}\}$ of T timesteps generated by the interaction between the agent and the environment, which contains two modalities, music and dance poses, with a total length of $2(T + 1)$. We interleave the two modalities in the trajectory to ensure compatibility with the standard causal attention mechanism and feed them into the reward model R_θ. Then, we apply a linear layer for each modality to map the raw inputs into an embedding space, added by a learned timestep embedding, which is shared by different modalities embedding similar to Decision Transformer (Chen et al. 2021b). Subsequently, these tokens will be fed into a Causal Transformer with multiple layers of masked multi-head attention to produce output features with equal length $\{x_t^s, x_t^a\}_{t=0}^{T-1}$, where x_t^s represents the feature of state s_t, and x_t^a represents the feature of action a_t. Then, the corresponding state and action features will be fed into a fully connected layer to generate the reward r_t that can be obtained by taking action a_t under the current state s_t. By applying the reward function $R_\theta(s_t, a_t)$, we obtain the total reward $u = \sum_t r_t$ for the entire sequence.

Training the Reward Model For training, we first sample a pair of trajectories τ_i, τ_j of different quality from the automatically ranked demonstration dataset $D_{ranked} = \{\tau_1, \cdots, \tau_m\}$, where $\tau_i \sim \pi_{BC}(\cdot|\epsilon_i), \tau_j \sim \pi_{BC}(\cdot|\epsilon_j)$ and $\epsilon_i \neq \epsilon_j$. Next, we obtain the quantitative metric of each trajectory through the reward model, *i.e.*, the total reward $u_l = \sum_{s_t, a_t \in \tau_l} R_\theta(s_t, a_t), l \in \{i, j\}$. Here we use the total reward u instead of individual reward r_t as the ranking criterion because the performance of the policy are decided by the entire sequence rather than each state-action pair. Then, we define a ranking predictor (Bradley and Terry

1952) based on the reward function R_θ:

$$\mathrm{R}(\tau_i \prec \tau_j;\theta) = \Big(\exp \sum_{s_t,a_t\in\tau_j} R_\theta(s_t,a_t)\Big)/ \Big(\exp \sum_{s_t,a_t\in\tau_i} R_\theta(s_t,a_t) + \exp \sum_{s_t,a_t\in\tau_j} R_\theta(s_t,a_t)\Big). \quad (2)$$

Then, we optimize the network using cross-entropy loss:

$$\mathcal{L}_{RM} = -\mathbb{E}_{(\tau_i,\tau_j,y)\in\mathcal{D}}\Big[(1-y)\log \mathrm{R}(\tau_i \prec \tau_j;\theta) + y \log \mathrm{R}(\tau_j \prec \tau_i;\theta)\Big], \quad (3)$$

where $y = \mathrm{int}(\epsilon_i < \epsilon_j)$. Intuitively, the cross-entropy loss trains a classifier to predict the quality of two trajectories correctly. This ranking loss (Cao et al. 2007) is based on the classic Bradley-Terry (Bradley and Terry 1952) and Luce-Shephard models of preferences (Luce 2012), and its effectiveness has been demonstrated in previous works (Kim et al. 2023; Park et al. 2022; Hejna III and Sadigh 2023).

Inference During Reinforcement Learning Consistent with the ranking criterion in the training process, in subsequent reinforcement learning, the learned reward model R_θ provides a sparse reward $\hat{r}_t$ for each state-action pair in dance sequence:

$$\hat{r}_t = \begin{cases} \sum_t r_t, \text{if} \quad t = T-1 \\ 0, \text{else} \end{cases} \quad (4)$$

We provide a total reward at the end of the dance sequence.

Exploration with Reinforcement Learning

The policy network π_{RL} (or π_ϕ) is parameterized by ϕ and seeks to maximize the expected return of the trajectory τ:

$$\phi = \arg\max_\phi \mathbb{E}_{\tau\sim\pi_\phi}[R(\tau)] = \arg\max_\phi \sum_\tau p_{\tau\sim\pi_\phi}(\tau)R(\tau), \quad (5)$$

where $p_{\tau\sim\pi_\phi}(\tau)$ represents the probability of generating trajectory τ given policy π_ϕ. $R(\tau) = \sum_{t=0}^{T-1}\gamma^t R(s_t,a_t)$ represents the discounted return of trajectory τ. Combined with the Markov Decision Process (MDP) defined in the Preliminary section, for $p_{\tau\sim\pi_\phi}(\tau)$, we have:

$$\begin{aligned} p_{\tau\sim\pi_\phi}(\tau) &= p(s_0)\prod_{t=1}^{K-1}\pi_\phi(a_t|s_t)p(s_{t+1}|s_t,a_t) \\ &= p(m_{init},p_{init},m_0)\prod_{t=1}^{K-1}\pi_\phi(p_t|m_{init},\cdots,m_t). \end{aligned} \quad (6)$$

The music-conditioned generation is a deterministic environment, where the state and action are given and the next state is deterministic. Therefore, the $p_{\tau\sim\pi_\phi}(\tau)$ is only relative to π_ϕ and initial state. Additionally, since the probability of the initial state $p(s_0) = p(m_{init},p_{init},m_0)$ is determined solely by the environment and not affected by the policy parameters ϕ, we have:

$$\begin{aligned} \phi &= \arg\max_\phi \sum_\tau p_{\tau\sim\pi_\phi}(\tau)R(\tau) \\ &= \arg\max_\phi \sum_\tau \prod_{t=1}^{T-1}\pi_\phi(p_t|m_{init},p_{init},m_0,\cdots,m_t)R(\tau), \end{aligned} \quad (7)$$

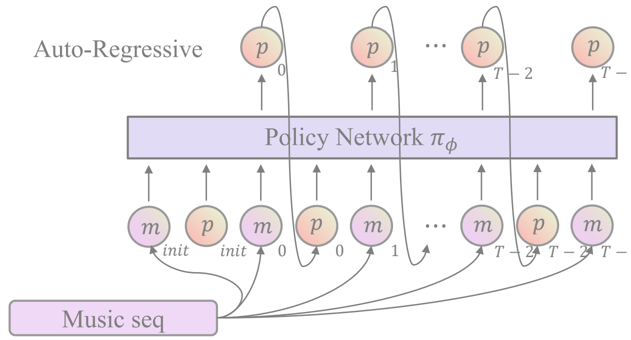

Figure 4: Overview of our policy network. Given a music sequence and an initial pose, the policy network π_ϕ auto-regressively predicts the subsequent poses.

where the second term $R(\tau) = \sum_t \gamma^t \hat{r}_t$ is determined by the learned reward model R_θ. According to the first term, we can directly apply an auto-regressive model for dance pose generation, *e.g.*, Transformer, similar to (Chen et al. 2021b; Janner, Li, and Levine 2021; Zheng, Zhang, and Grover 2022; Xu et al. 2022). As shown in Figure 4, when π_{RL}, initialized with π_{BC}, collects samples, the environment provides fixed music sequence $\{m_{init}, m_0, \cdots, m_{T-1}\}$ and an initial action p_{init}. The policy π_{RL} generates the entire dance pose sequence $\{p_{init}, p_0, \cdots, p_{T-1}\}$ in an auto-regressive manner.

The training of the policy adopts the Proximal Policy Optimization algorithm (PPO) (Schulman et al. 2017).

Experiements

Experiments Setup

Dataset We conduct the training and experiments on the AIST++ dataset (Li et al. 2021), which is the largest public available dataset for aligned 3D dance motions and music. AIST++ dataset includes 992 60-Frame Per Second (FPS) 3D dance motion sequences in SMPL (Loper et al. 2015) format. In line with (Li et al. 2021; Siyao et al. 2022), we split these data into 952 sequences for training and 40 sequences for subsequent experiments.

Implementation Details For audio preprocess, we employ Librosa to extract music features. Specifically, we extract the following features: Mel Frequency Cepstral Coefficients (MFCC), MFCC delta, constant-Q chromagram, tempogram, and onset strength, yielding a 438-dimensional musical feature vector. More details and hyper-parameter settings can be found in Appendices.

Comparisons with State-Of-The-Arts

We compare E3D2 to state-of-the-art including FACT (Li et al. 2021) and Bailando (Siyao et al. 2022), which is also our behavior cloning policy. Following (Siyao et al. 2022), we generate 40 dance clips for each method in the AIST++ test set and cut the generated dances into the length of 20 seconds for further experiments.

	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**

Table 1: Evaluation results on test set of different dance generation frameworks. To ensure a fair comparison with baselines, we report the results of (Siyao et al. 2022) without RL fine-tuning on the test set.

We conduct objective evaluations following (Siyao et al. 2022), including the quality and diversity of generated dances, and the alignment score between the dance and music beats. Specifically, for the quality of generated dances, we calculate the Fréchet Inception Distance (FID) (Heusel et al. 2017) between the generated dance and all dance sequences of AIST++ dataset on the kinetic feature (FID_k) (Onuma, Faloutsos, and Hodgins 2008) and the geometric feature (FID_g) (Müller, Röder, and Clausen 2005). For the diversity, we calculate the average Euclidean distance (DIV) (Li et al. 2021) of the generated dances on the kinetic feature (DIV_k) and the geometric feature (DIV_g). For the alignment between the dance and music beats, we calculate the Beat Align Score (BAS) (Liu et al. 2022b; Siyao et al. 2022).

Table 1 report the comparison with state-of-the-art methods. According to the comparison, the proposed E3D2 outperforms baseline frameworks in all aspects, demonstrating the effectiveness of the exploration. Specifically, with exploration, E3D2 improves 8.28% and 10.15% than the Behavior Cloning (BC) policy Bailando on FID_k and FID_g, respectively. This indicates that the reward model prefers movements that are more similar to those of humans and high-quality. And for the motion diversity, exploration helps the policy improve 21.23% and 4.16% on DIV_k and DIV_g, respectively. The results on BAS also indicate the improvement of our method. More comparisons and visualizations in wild musics are available in demo page[2].

Discussion

This section provides a comprehensive analysis of the reward model, including the effectiveness, advantages over the hand-designed reward, the empirical soundness of the training process, as well as the accuracy and generalization.

Does Exploration Provide More Alignment?

To further present the alignment with human preferences brought by exploration and reward model, we conduct a subjective test with 24 subjects, who are asked to select which dance segment they preferred through pairwise comparisons given a certain piece of music. The results are shown in Table 2. Human evaluation shows the superior performance of our approach compared to frameworks without exploration, which is consistent with objective metrics diversity, quality and beat align score.

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

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

Is a Learned Reward Function More Effective Than a Hand-Designed One?

To explore the differences between learned and hand-designed reward, we compare a hand-designed reward proposed in (Siyao et al. 2022):

$$r = r_b + \gamma_c r_c, \tag{8}$$

where r_b and r_c represent the beat alignment reward and orientation reward, respectively. The former aligns dance movements with the rhythm of music, while the latter constrains the consistency of the upper and lower half body. γ_c is the balance weight.

Table 3 presents various evaluation results of dance generation from the agent trained with the hand-designed reward. As the interaction steps increase, the dance gradually deviates from human movement patterns and its diversity is not as good as the behavior cloning policy. This is because the hand-designed reward only considers the beat alignment and consistency of upper and lower body movements, which results in a lack of diversity and similarity to human movements being overlooked. Besides, the design of reward requires a lot of task-specific prior knowledge (Wirth et al. 2017; Liu et al. 2022a). In contrast, our reward is learned from automatically ranked demonstrations without any domain knowledge. This reward is expected to implicitly learn various aspects of the dance, allowing a comprehensive evaluation of the dance and a correct optimization of the dance policy.

Higher Level Noise Leads to the Worse Demonstrations?

The learned reward model is based on the assumption that the behavior cloning policy significantly outperforms a completely random policy and that increasing levels of noise lead

[2]https://sites.google.com/view/e3d2

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

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.

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

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.

to an increasingly worse policy. To validate this assumption, we select different noisy levels $[0.02, 0.25, 0.50, 0.75, 1.00]$ and evaluate the performance of behavior cloning policies injected with these noises on the training set music. As shown in Table 4, as the noise level increases, the quality of the generated dances generally decreases.

What Is the Performance of the Reward Model?

Table 4 demonstrates that the reward model can provide appropriate evaluations, *i.e.*, total reward, for the dance sequences. For generalization, Figure 5 illustrates the classification accuracy of the reward model on dances with different rankings conditioned on training and test set musics during the training process. At the final epoch, the reward model achieves an accuracy of around 97% on both the training and test sets. In contrast, we evaluate the pose code prediction accuracy of the BC policy on training and test sets, with the results shown in Table 5. 'Complete Pose' refers to the correct prediction of both the upper and lower half body codes, while 'Partial Pose' indicates that at least one code is correct. The BC policy suffers from severe overfitting to the training set and struggles to generalize to unseen music, achieving only 2.32% complete pose accuracy and 7.52% partial pose accuracy. Moreover, we assess the performance of the BC policy in generating dance sequences under the conditions of seen and unseen music, as shown in Table 6. The significant gap indicates the weak generalization of the supervised learning approaches. In comparison, the reward model exhibits excellent generalization performance. A plausible explanation for these findings is that the reward model, as a discriminative model, easily achieves better generalization than the generative behavior cloning model, enabling reinforcement learning policy to demonstrate superior generalization compared to behavior cloning policy.

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

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.

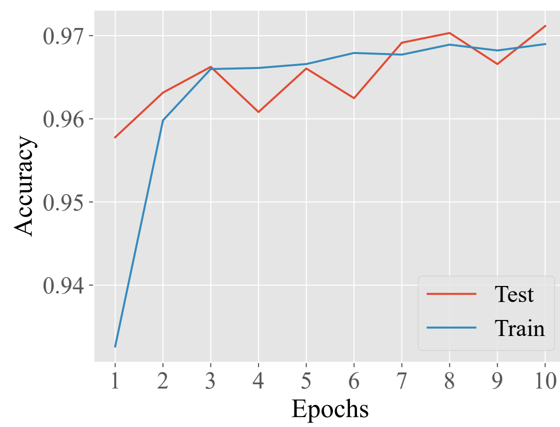

Figure 5: Reward model accuracy: The classification accuracy of the reward model on dances generated by policies with varying levels of noise during training. The reward model exhibits excellent generalization on the test set.

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

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

Conclusion

In this paper, to address the problem of the lack of exploration ability in current music-driven dance models, we propose a novel dance generation framework, E3D2. We first train a reward model on automatically ranked dance demonstrations, and then, we train the dance policy using reinforcement learning with the learned reward model, resulting in more diverse and human-aligned dances. Extensive experiments demonstrate the effectiveness of E3D2.

Acknowledgements

This work is supported by National Natural Science Foundation of China (62076144), Shenzhen Science and Technology Program (WDZC20220816140515001, JCYJ20220818101014030) and Shenzhen Key Laboratory of next generation interactive media innovative technology (ZDSYS20210623092001004).

References

Abbeel, P.; and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, 1.

Ahn, H.; Kim, J.; Kim, K.; and Oh, S. 2020. Generative Autoregressive Networks for 3D Dancing Move Synthesis From Music. *IEEE Robotics and Automation Letters*, 5(2): 3501–3508.

Asadi, K.; Misra, D.; Kim, S.; and Littman, M. L. 2019. Combating the compounding-error problem with a multi-step model. *arXiv preprint arXiv:1905.13320*.

Au, H. Y.; Chen, J.; Jiang, J.; and Guo, Y. 2022. Choreo-Graph: Music-conditioned Automatic Dance Choreography over a Style and Tempo Consistent Dynamic Graph. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3917–3925.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345.

Brown, D.; Goo, W.; Nagarajan, P.; and Niekum, S. 2019. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, 783–792. PMLR.

Brown, D. S.; Goo, W.; and Niekum, S. 2020. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Conference on robot learning*, 330–359. PMLR.

Cao, Z.; Qin, T.; Liu, T.-Y.; Tsai, M.-F.; and Li, H. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, 129–136.

Chen, K.; Tan, Z.; Lei, J.; Zhang, S.-H.; Guo, Y.-C.; Zhang, W.; and Hu, S.-M. 2021a. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)*, 40(4): 1–13.

Chen, L.; Lu, K.; Rajeswaran, A.; Lee, K.; Grover, A.; Laskin, M.; Abbeel, P.; Srinivas, A.; and Mordatch, I. 2021b. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34: 15084–15097.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Fu, J.; Luo, K.; and Levine, S. 2018. Learning Robust Rewards with Adverserial Inverse Reinforcement Learning. In *International Conference on Learning Representations*.

Fukayama, S.; and Goto, M. 2015. Music content driven automated choreography with beat-wise motion connectivity constraints. *Proceedings of SMC*, 177–183.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Gao, J.; Pu, J.; Zhang, H.; Shan, Y.; and Zheng, W.-S. 2022. PC-Dance: Posture-controllable Music-driven Dance Synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1261–1269.

Glaese, A.; McAleese, N.; Trębacz, M.; Aslanides, J.; Firoiu, V.; Ewalds, T.; Rauh, M.; Weidinger, L.; Chadwick, M.; Thacker, P.; et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Hejna III, D. J.; and Sadigh, D. 2023. Few-shot preference learning for human-in-the-loop RL. In *Conference on Robot Learning*, 2014–2025. PMLR.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Huang, R.; Hu, H.; Wu, W.; Sawada, K.; Zhang, M.; and Jiang, D. 2021. Dance Revolution: Long-Term Dance Generation with Music via Curriculum Learning. In *International Conference on Learning Representations*.

Huang, Y.; Zhang, J.; Liu, S.; Bao, Q.; Zeng, D.; Chen, Z.; and Liu, W. 2022. Genre-Conditioned Long-Term 3D Dance Generation Driven by Music. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4858–4862. IEEE.

Ibarz, B.; Leike, J.; Pohlen, T.; Irving, G.; Legg, S.; and Amodei, D. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31.

Janner, M.; Du, Y.; Tenenbaum, J.; and Levine, S. 2022. Planning with Diffusion for Flexible Behavior Synthesis. In *International Conference on Machine Learning*, 9902–9915. PMLR.

Janner, M.; Li, Q.; and Levine, S. 2021. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34: 1273–1286.

Kim, C.; Park, J.; Shin, J.; Lee, H.; Abbeel, P.; and Lee, K. 2023. Preference Transformer: Modeling Human Preferences using Transformers for RL. In *The Eleventh International Conference on Learning Representations*.

Li, B.; Zhao, Y.; Zhelun, S.; and Sheng, L. 2022. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1272–1279.

Li, J.; Yin, Y.; Chu, H.; Zhou, Y.; Wang, T.; Fidler, S.; and Li, H. 2020. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13401–13412.

Liu, R.; Bai, F.; Du, Y.; and Yang, Y. 2022a. Meta-Reward-Net: Implicitly Differentiable Reward Learning for Preference-based Reinforcement Learning. In *Advances in Neural Information Processing Systems*.

Liu, X.; Wu, Q.; Zhou, H.; Xu, Y.; Qian, R.; Lin, X.; Zhou, X.; Wu, W.; Dai, B.; and Zhou, B. 2022b. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10462–10472.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16.

Luce, R. D. 2012. *Individual choice behavior: A theoretical analysis*. Courier Corporation.

MacGlashan, J.; Ho, M. K.; Loftin, R.; Peng, B.; Wang, G.; Roberts, D. L.; Taylor, M. E.; and Littman, M. L. 2017. Interactive learning from policy-dependent human feedback. In *International Conference on Machine Learning*, 2285–2294. PMLR.

Michie, D.; Bain, M.; and Hayes-Miches, J. 1990. Cognitive models from subcognitive skills. *IEE control engineering series*, 44: 71–99.

Müller, M.; Röder, T.; and Clausen, M. 2005. Efficient content-based retrieval of motion capture data. In *ACM SIGGRAPH 2005 Papers*, 677–685.

Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Ng, A. Y.; and Russell, S. J. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 663–670.

Onuma, K.; Faloutsos, C.; and Hodgins, J. K. 2008. FMDistance: A Fast and Effective Distance Function for Motion Capture Data. In *Eurographics (Short Papers)*, 83–86.

Park, J.; Seo, Y.; Shin, J.; Lee, H.; Abbeel, P.; and Lee, K. 2022. SURF: Semi-supervised Reward Learning with Data Augmentation for Feedback-efficient Preference-based Reinforcement Learning. In *International Conference on Learning Representations*.

Russell, S. 1998. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Siyao, L.; Yu, W.; Gu, T.; Lin, C.; Wang, Q.; Qian, C.; Loy, C. C.; and Liu, Z. 2022. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11050–11059.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Sugiyama, H.; Meguro, T.; and Minami, Y. 2012. Preference-learning based inverse reinforcement learning for dialog control. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*.

Tseng, J.; Castellon, R.; and Liu, K. 2023. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 448–458.

Tucker, A.; Gleave, A.; and Russell, S. 2018. Inverse reinforcement learning for video games. *arXiv preprint arXiv:1810.10593*.

Valle-Pérez, G.; Henter, G. E.; Beskow, J.; Holzapfel, A.; Oudeyer, P.-Y.; and Alexanderson, S. 2021. Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics (TOG)*, 40(6): 1–14.

Wang, Z.; Jia, J.; Wu, H.; Xing, J.; Cai, J.; Meng, F.; Chen, G.; and Wang, Y. 2022. GroupDancer: Music to Multi-People Dance Synthesis with Style Collaboration. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1138–1146.

Warnell, G.; Waytowich, N.; Lawhern, V.; and Stone, P. 2018. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Wirth, C.; Akrour, R.; Neumann, G.; Fürnkranz, J.; et al. 2017. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136): 1–46.

Wu, J.; Ouyang, L.; Ziegler, D. M.; Stiennon, N.; Lowe, R.; Leike, J.; and Christiano, P. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Xu, M.; Shen, Y.; Zhang, S.; Lu, Y.; Zhao, D.; Tenenbaum, J.; and Gan, C. 2022. Prompting decision transformer for few-shot policy generalization. In *International Conference on Machine Learning*, 24631–24645. PMLR.

Ye, Z.; Wu, H.; Jia, J.; Bu, Y.; Chen, W.; Meng, F.; and Wang, Y. 2020. Choreonet: Towards music to dance synthesis with choreographic action unit. In *Proceedings of the 28th ACM International Conference on Multimedia*, 744–752.

Zheng, Q.; Zhang, A.; and Grover, A. 2022. Online decision transformer. In *International Conference on Machine Learning*, 27042–27059. PMLR.

Zhuang, W.; Wang, C.; Chai, J.; Wang, Y.; Shao, M.; and Xia, S. 2022. Music2Dance: DanceNet for Music-Driven Dance Generation. *ACM Trans. Multimedia Comput. Commun. Appl.*