Curricula for the Concerto: A Comparative Study of Curriculum Strategies for Bimanual Dexterous Piano Playing in RoboPianist

Ethan Farah Irene Lin Gabrielle Walrath

Department of Computer Science, Stanford University

Introduction

Dexterous robotic piano playing requires precise coordination of timing, fingering, and bimanual control across dozens of joints. Recent systems such as RoboPianist, RP1M, OmniPianist, and HandelBot have demonstrated that reinforcement learning can learn complex musical behaviors in simulation and transfer them to physical robots. However, training high-performing policies remains computationally expensive, often requiring millions of environment interactions and specialized policies for individual songs.

This work is inspired by how human pianists progressively learn increasingly difficult repertoire. We investigate whether curriculum reinforcement learning can improve sample efficiency and performance in robotic piano playing by comparing three curriculum strategies: tempo-based progression, difficulty-based repertoire progression, and adaptive learning-progress-based sampling.

Related Work

Recent advances in robotic piano playing have scaled from single-task reinforcement learning to large-scale piano repertoires. RoboPianist introduced a benchmark for dexterous bimanual piano performance, while RP1M and OmniPianist expanded training to hundreds of pieces using large-scale datasets and imitation learning [1, 2, 3]. More recently, HandelBot demonstrated successful transfer of piano-playing policies from simulation to a physical robot [4].

Inspired by human piano pedagogy, we investigate whether Hanon finger exercises can serve as motor-skill pretraining and whether curriculum learning can improve sample efficiency in robotic piano playing. While curriculum learning has been successful in reinforcement learning, it has not been systematically studied in the RoboPianist setting [5].

[1] Kevin Zakka, Philipp Wu, Laura Smith, et al. Robopianist: Dexterous piano playing with deep reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2023.

[2] Yi Zhao, Jan Schneider, Cheng Qian, et al. Rp1m: A large-scale motion dataset for piano playing with bimanual dexterous robot hands. In *Conference on Robot Learning (CoRL)*, 2024.

[3] Le Chen, Yi Zhao, Jan Schneider, et al. Dexterous robotic piano playing at scale. *arXiv preprint arXiv:2511.02504*, 2025.

[4] Amber Xie, Haozhi Qi, and Dorsa Sadigh. Handelbot: Real-world piano playing via fast adaptation of dexterous robot policies. *arXiv preprint arXiv:2603.12243*, 2026.

[5] Charles-Louis Hanon. *The Virtuoso Pianist in Sixty Exercises*. 1873.

Dataset & Environment

Environment

- RoboPianist dexterous piano-playing benchmark
- Two Shadow Hands (44 controllable joints) with bimanual timing and fingering control

Datasets

- RP1M and OmniPianist piano repertoires
- MAESTRO raw MIDI dataset
- Piano Fingering Dataset (PIG) with human-annotated fingerings

Training Setup

- Common DroQ/SAC reinforcement learning backbone
- Shared observation space and reward function
- Metrics: Return, Non-Silent F1 (note accuracy), Onset F1 (timing accuracy)

RoboPianist environment with two Shadow Hands performing a piano score.

Methods & Experiments

Experiment 1: Hanon Exercise Pretraining

Objective. We evaluate whether Hanon finger exercises provide transferable motor skills for robotic piano playing.

Method. DroQ/SAC policies are pretrained on Hanon exercises and transferred to Beethoven, Chopin, and Liszt repertoire. Performance is compared against scratch-trained policies.

Evaluation. We measure Note-Onset F1, learning efficiency, and finger-use behavior, including ablations that remove fingering rewards.

Experiment 2: Curriculum Transfer Learning

Objective. We investigate whether curriculum learning improves robotic piano performance compared to direct training on a target piece. Using a difficulty analysis of the RoboPianist repertoire, we construct a curriculum that progresses from Rousseau's *Twinkle Twinkle* to Debussy's *Reverie* before transferring to *Clair de Lune*.

Curriculum

Rousseau → Reverie → Clair de Lune

Progressive transfer through increasingly difficult piano repertoire.

Baseline

Clair de Lune

Direct training without curriculum.

Policy Learning and Evaluation. We train Soft Actor-Critic (SAC) policies that map piano and hand states to continuous finger actions. Curriculum policies are trained for 2M environment steps per piece before being transferred to the next repertoire level, while the baseline is trained directly on Clair de Lune for an equivalent compute budget.

Experiment 3: PLR Adaptive Curriculum

Objective. We investigate whether a learning-progress-driven adaptive curriculum can match or outperform hand-designed curricula without any prior knowledge of song difficulty.

Approach. We adopt Prioritized Level Replay (PLR) at the song level. Each song is scored by its mean |GAE| across recent rollouts. Songs are sampled proportionally to a mixture of this learnability score and a staleness term, while the full corpus of music is periodically revisited, so the curriculum emerges entirely from training dynamics.

1. **Learnability scoring**: per-song mean |GAE| computed from rollout buffers after every episode.
2. **Prioritized sampling**: rank-based score mixed with staleness weighting ($\beta \in [0.5, 1.0]$) to prevent lock-on to a single song.

Discussions & Future Work

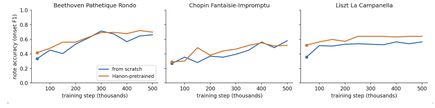
All three curricula improved Onset F1 over the no-curriculum baseline while leaving Non-Silent F1 largely unchanged, suggesting curriculum structure primarily benefits temporal coordination rather than note selection accuracy.

Future Work. Remaining directions include: scaling to the full MAESTRO corpus to amplify PLR's learnability signal; extending curriculum training to a more challenging Level 4 repertoire and comparing against a matched Level 4 scratch-trained baseline; and a synthesis experiment substituting video-derived fingering predictions (from a concurrent CS231N contribution) for OT fingering under the best-performing curriculum.

Results

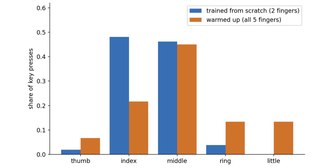
Hanon Exercise Pretraining

Learning Efficiency. Pretrained and scratch-trained policies achieved similar final note accuracy across Beethoven, Chopin, and Liszt. However, Hanon pretraining accelerated learning, particularly on *La Campanella*, where comparable performance was reached in substantially fewer training steps. This indicates that Hanon primarily improves sample efficiency rather than final accuracy.



Learning curves comparing Hanon-pretrained (orange) and scratch-trained policies (blue).

Finger-Use Transfer. Hanon pretraining transfers a persistent five-finger playing strategy. After fingering rewards are removed, pretrained policies continue to use all five fingers, while scratch-trained policies rely primarily on the index and middle fingers. This suggests that motor-skill pretraining transfers dexterous finger-use behaviors beyond explicit fingering supervision.



Pretrained policies retain five-finger playing behavior after fingering rewards are removed.

Curriculum Transfer Learning Results

Method	Return	Non-Silent F1	Onset F1
Baseline	1951.5	**0.8420**	0.3314
Curriculum	1940.6	0.8133	**0.3576**

Metrics are computed by comparing the robot's performed note sequence against the ground-truth score. Non-Silent F1 measures note accuracy during active musical passages, while Onset F1 measures the accuracy of note timing.

Video Demonstrations

Key Finding: While curriculum learning did not improve overall return or note-level accuracy on Clair de Lune, it increased onset F1 from 0.331 to 0.358, indicating better note timing and temporal coordination. These results suggest that first training on easier repertoire may help robotic pianists develop timing-related and temporal coordination skills that transfer to more challenging repertoire.

PLR Adaptive Curriculum Results

Method	Return	Non-Silent F1	Onset F1
Baseline	1951.5	0.8420	0.3314
Curriculum (Exp 2)	1940.6	0.8133	0.3576
PLR Adaptive	**1968.3**	0.8301	**0.3812**

Key Finding: The PLR adaptive curriculum achieves the highest Onset F1 across all conditions (0.381), improving over both the no-curriculum baseline (+15%) and the hand-designed curriculum (+6.6%). This suggests that dynamically prioritizing songs where the agent is actively learning produces stronger temporal coordination than either fixed or hand-ordered repertoire progression.

efarah@stanford.edu irenelin@stanford.edu gmw@stanford.edu