

# 비디오 생성 모델 기반 레퍼런스 합성을 통한 4족 로봇의 스케이트보딩 동작 학습

변경언<sup>0</sup>, 권태수, 이윤상<sup>†</sup>  
한양대학교 컴퓨터소프트웨어학과  
{byeon9980, taesoo, yoonsanglee}@hanyang.ac.kr

## Learning Quadruped Skateboarding from Synthesized References via Video Generation Models

Kyung-Eon Byeon<sup>0</sup>, Tae-soo Kwon, Yoon-Sang Lee<sup>†</sup>  
Department of Computer Science, Hanyang University

### 요약

본 연구는 충분한 양의 모션 캡처 데이터나 촬영 영상을 확보하기 어려운 개의 스케이트보딩 동작을 대상으로, 비디오 생성 모델을 이용해 합성 레퍼런스 영상을 생성하고 이를 활용하여 4족 로봇의 스케이트보딩 스킬을 학습하는 방법을 제안한다. 이를 위해 비디오 생성 모델 기반 영상 합성, 3D 포즈 추정 및 리타겟팅, 그리고 모방 학습으로 파이프라인을 구성했다.

### 1. 서론

최근 심층 강화 학습을 통해 물리 시뮬레이션 환경의 캐릭터나 로봇이 모션 캡처 데이터를 모방하도록 제어 정책을 학습하는 연구가 활발히 이루어지고 있다 [1,2]. 그러나 이는 전용 장비와 통제된 환경이 필요해 진입 장벽이 높고, 동물이나 위험한 동작의 경우 피사체 확보가 어렵다.

한편, 레퍼런스 없이 보상만으로 학습하는 강화 학습 [3]은 동작 또는 구간 별 정교한 보상 설계가 요구되며, 특히 접촉이 많은 경우에는 더욱 그러하다. 반면 모방 학습은 레퍼런스 교체만으로 다양한 동작을 학습할 수 있으며, 여러 동작이 결합된 복합 동작에서도 별도의 보상 설계가 필요하지 않다. 이에 모방 학습이 효율적인 방법으로 자리 잡았으며, 모션 캡처의 한계를 극복하기 위해 영상에서 직접 포즈를 추정해 활용하는 다양한 방법이 제안되었다 [4,5]. 하지만 실제 촬영 영상에서는 occlusion이나 motion blur로 포즈 추정이 불안정해진다.

우리는 비디오 생성 모델을 활용해 레퍼런스 영상을 합성하는 접근을 제안한다. 비디오 생성 모델은 카메라 시점과 재생속도 등을 프롬프트로 제어할 수 있어 포즈 추정에 유리한 영상을 제작할 수 있고, 촬영이 까다로운 동작 역시 제약 없이 확보할 수 있다.

본 연구는 (1) 비디오 생성 모델로 스케이트보딩 레퍼런스 영상 합성, (2) 3D 포즈 추정 및 리타겟팅, (3) 목표 속도를 추종하는 모방 학습으로 구성된 파이프라인을 통해 4족 로봇의 스케이트보딩 동작을 학습한다. 그 결과, 모션 캡처 없이도 공통된 정책 학습 구조로 네 가지 스킬을 모두 학습했으며, 다양한 동작 영역으로의 확장 가능성을 보였다.

### 2. 방법

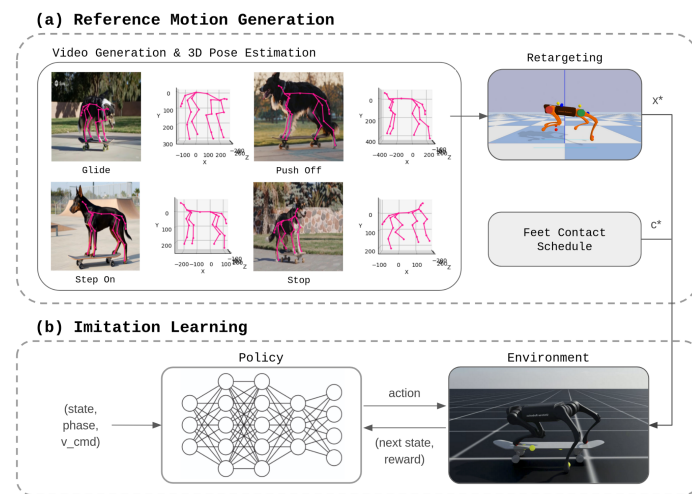


그림 1: 전체 프레임워크 개요

#### 2.1. 레퍼런스 모션 생성

**영상 합성.** 비디오 생성 모델(Sora, Grok)을 활용하여 4개의 레퍼런스 영상을 합성한다: glide(보드 위 활주), push off(발로 지면을 밀어 보드 가속), stop(발을 지면에 내려

\* 구두발표논문, 요약논문 (Extended Abstract)

\* 본 논문은 요약논문 (Extended Abstract)으로서, 본 논문의 원본 논문은 현재 타 학술대회에 제출 준비중임.

\* 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2024년도 문화체육관광 연구개발사업 (RS-2024-00399136) 및 과학기술 정보통신부 및 정보통신 기획평가원의 가상융합지능화핵심기술 개발사업 (RS-2026-25508404)의 지원을 받아 수행되었음.

† 교신저자

보드 정지), step on(정지 상태에서 발을 보드로 복귀). 프롬프트는 네 다리와 발의 움직임이 잘 드러나는 측면 시점을 유지하고 선명한 영상을 생성하도록 작성했다.

**3D 포즈 추정.** 합성된 영상에서 4족 동물의 키포인트는 Digidogs[6]를 사용해 추정한다. Digidogs는 합성 데이터로 학습된 모델로, 본 연구의 합성 영상과 도메인이 유사해 fine-tuning 없이 안정적인 포즈 추정이 가능했다.

**리타겟팅.** 추정된 키포인트는 힙과 발 위치를 기준으로 [2]의 공개 코드를 활용해 Unitree A1 로봇의 관절 공간으로 리타겟팅했다.

## 2.2. 모방 학습

**Phase 기반 진행도 표현.** 각 스킬의 길이가 다르므로, 레퍼런스 모션의 진행도를  $[0, 1]$  범위의 phase 변수로 표현해 정책의 관측에 포함시켰다.

**접촉 대상.** 각 발의 접촉 대상(없음, 지면, 보드)을 원-핫 벡터로 인코딩하여 보상 항에 포함시켰다. 레퍼런스 모션의 접촉 대상 정보는 프레임 단위로 직접 라벨링했다.

**보상.** 레퍼런스 모션은 정책 관측이 아닌 보상 항에만 활용하여 레퍼런스에 대한 의존도를 낮췄다. 보상은 공통 항(레퍼런스 추종 및 안정적인 스케이트보딩)과 task-specific 항(스킬별 강조 요소)으로 구성된다. 스킬별 강조 요소는, glide는 데크의 수평 유지, push off는 보드 가속, stop은 보드 감속, step on은 4발이 모두 보드에 위치하는 것을 다룬다. 정책 학습은 IsaacLab에서 PPO 알고리즘으로 진행했다.

## 3. 구현 세부 사항

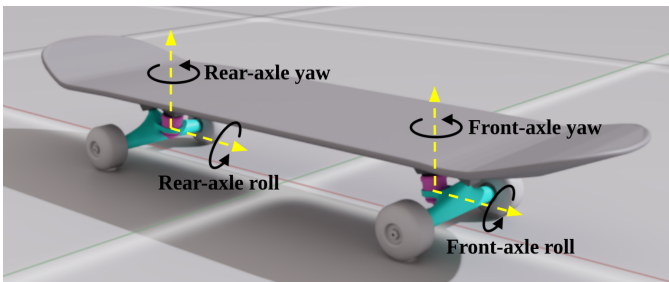


그림 2: 스케이트보드 모델 (분홍: 부상, 민트: 트럭)

**스케이트보드 모델링.** 스케이트보드의 데크-부싱은 fixed joint, 부싱-트럭은 2-DoF rotational joint, 트럭-바퀴는 revolute joint로 연결된다 (그림 2). 이러한 구조는 데크 roll을 트럭 yaw 회전으로 변환하는 실제 보드의 부싱 메커니즘을 반영하며, 앞뒤 부싱의 독립성으로 인해 데크 roll뿐 아니라 앞뒤 하중 분포에 따른 데크 roll의 편차까지 표현할 수 있다.

## 4. 실험 결과 및 결론

그림 3과 같이, 학습된 정책은 레퍼런스 영상과 유사한 접촉 패턴을 보이며 목표 속도를 안정적으로 추종했다. 모든 스킬은 동일한 학습 구조에서 task-specific 보상 항만 달리하여 안정적으로 수렴했다. 또한 접촉 대상 정보를 보상 항에서 제거할 경우 phase가 끝까지 증가하지 못하고 학습이 실패했다.

본 연구는 합성 레퍼런스 영상을 사용하여 모션 캡처 없이 4족 로봇의 스케이트보딩 동작 학습이 가능함을 보였다. 다만 합성 영상의 품질이 프롬프트에 따라 편차가 있어, 품질 평가 자동화와 이를 활용한 프롬프트 최적화 프로세스가 있으면 더 효율적으로 영상을 확보할 수 있을 것이다. 향후에는 turn, manual 등 보다 복잡한 보드 트릭으로의 확장 and 실제 로봇으로의 sim-to-real 적용을 목표로 한다.

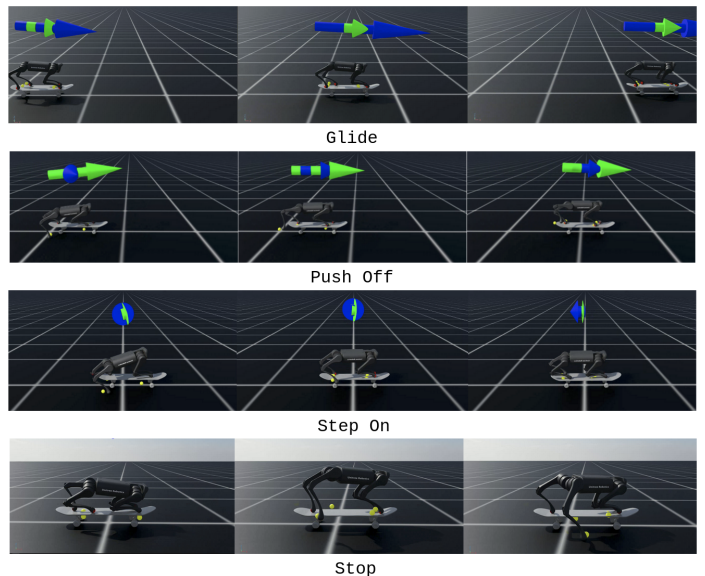


그림 3: 학습 결과

## 참고문헌

- [1] X. B. Peng et al., "DeepMimic: Example-Guided Deep Reinforcement Learning of Physics-Based Character Skills," ACM Trans. Graph., 37(4), 2018.
- [2] X. B. Peng et al., "Learning Agile Robotic Locomotion Skills by Imitating Animals," RSS, 2020.
- [3] H. Liu et al., "Discrete-Time Hybrid Automata Learning: Legged Locomotion Meets Skateboarding," RSS, 2025.
- [4] A. Allshire et al., "Visual Imitation Enables Contextual Humanoid Control," CoRL, 2025.
- [5] Q. Feng et al., "PhysHMR: Learning Humanoid Control Policies from Vision for Physically Plausible Human Motion Reconstruction," SIGGRAPH Asia, 2025.
- [6] M. Shooter, C. Malleson, and A. Hilton, "DigiDogs: Single-View 3D Pose Estimation of Dogs Using Synthetic Training Data," WACVW, pp. 80-89, 2024.