한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 31, No. 3, P. 21~34

월드모델을 통한 Latent Diffusion Model 예측 고도화

나예현° Richard.Y.Park 서상영 권태수*

한양대학교

{2024141977@hanyang.ac.kr, 2024222199@hanyang.ac.kr , syz6478@hanyang.ac.kr, taesoo@hanyang.ac.kr}

Enhancing Prediction Robustness of Latent Diffusion Models through World Models

Yehyeon NA° Richard Y. Park Sangyeong Seo Taesoo Kwon*

Hanyang University

요 약

인간 동작 예측은 자율 시스템부터 몰입형 가상 환경에 이르기까지 다양한 응용 분야에서 필수적인 요소이다. 그러나 기존 대부분의 모델은 일상적인 동작 예측에 초점을 맞추고 있어, 스포츠, 가상현실(VR), 보조 로보틱스 등과 같은 역동적이고 표현력 있는 상황에서는 효과가 제한적이다.
이러한 한계를 해결하기 위해, 본 연구는 클래스 내 변동이 크고 고에너지 동작을 포함하는 LaFAN 데이터셋을 대상으로 한다. 우리는 다음의 세 가지 핵심 아이디어를 결합한 새로운 프레임워크를 제안한다: (1) 시간적 일관성을 향상시키기 위한 LDM, (2) 안정적인 장기 예측을 위한 강화학습 기반 접근법, (3) 노이즈 최적화를 위한 double sampling 연구 결과, 제안하는 모델은 사실적이고 일관된 동작 시퀀스를 생성하였으며, 복잡하고 표현력 있는 동작 예측을 위한 높은 가능성을 입증하였다.

Abstract

Human motion prediction is essential for applications ranging from autonomous systems to immersive virtual environments. However, most existing models focus on everyday motion, limiting their effectiveness in dynamic, expressive contexts like sports, VR, and assistive robotics. To address this, we target the LaFAN dataset, which captures high-energy movements with significant intra-class variation. We propose a novel framework combining three ideas: (1) a latent diffusion model for improved temporal consistency, (2) a reinforcement learning approach for stable long-term prediction, and (3) a robust noise optimization strategy to counter diffusion sampling noise. Experiments show our model produces realistic, coherent motion sequences, highlighting its potential for complex and expressive motion prediction.

키워드: 동작 예측, 강화학습, 생성모델
Keywords: MotionPrediction,Reinforcement Learning,Generative Model

1 Introduction

인간 동작 예측은 다양한 응용 분야에서 핵심적인 요소로 작용해 왔다. 기존의 동작 예측 연구는 크게 두 가지 주요 범주로 분류할 수 있다. 가장 널리 연구된 분야는 로봇이나 차량이 안전하게 행동할 수 있도록 인간의 움직임을 예측하는 응용이다. 이러한 응용은 자율주행 및 인간-로봇 상호작용에 관한 관심이 높아지면서 본격적으로 주목을 받기 시작하였다. 예를 들어, 도로를 건너려는 보행자의 움직임을 예측함으로써 차량이 미리 속도를 줄이거나 멈출 수 있게 된다 [1]. 유사하게, 로봇공학 분야에서는 인간과 밀접하게 협업하는 로봇이 주변에서 안전하게 움직이기 위해 동작 예측이 요구된다. 이러한 안전 중심 응용에서는 일상적이고 일반적인 동작을 견고하게 예측하는 것이 핵심이다 [2]. 두 번째 범주는 스포츠 훈련, 보조 로보틱스, VR 엔터테인먼트, 위험 감시 등에서처럼 역동적이거나 특수한 동작을 정확히 예측

*corresponding author: Taesoo Kwon / Hanyang University (taesoo@hanyang.ac.kr)

Received : 2025.06.13./ Review completed : 1st 2025.06.30. 2nd 2025.07.10./ Accepted : 2025.07.14.
DOI : 10.15701/kcgs.2025.31.3.21
ISSN : 1975-7883(Print)/2383-529X(Online)

해야 하는 응용이다. 이러한 분야에서는 좁은 동작 범위 내에서 역동적이고 개별적인 동작 패턴을 정밀하게 예측하는 능력이 요구된다. 예컨대 스포츠 훈련에서는 선수의 자세나 움직임을 예측함으로써 부상 위험을 줄이고 수행 능력을 향상할 수 있다 [3]. 보조 로보틱스 분야에서는 사용자의 움직임을 실시간으로 예측함으로써 로봇이 적절한 저항이나 지지력을 제공할 수 있어, 운동 기능 향상과 재활 효과를 높일 수 있다 [4]. VR 엔터테인먼트에서는 적은 수의 센서로부터 사용자의 의도된 동작을 예측함으로써, 반응성 있고 현실적인 아바타 행동을 생성할 수 있다. 위험 감시 분야에서는 비정상적이거나 위험한 행동을 사전에 예측하여 즉각적인 개입을 가능하게 한다. 이러한 응용의 공통점은 일반적이고 일상적인 활동보다는, 제한된 수의 고속이거나 특수한 동작 패턴을 정밀하게 예측해야 한다는 점이다.

현재까지의 대부분의 연구는 첫 번째 범주, 즉 다양한 일상 동작을 예측하는 응용을 중심으로 설계되었으며, AMASS, Human3.6M, CMU 동작 캡처 데이터셋 등 대규모 데이터셋을 기반으로 평가되어 왔다 [5, 6]. 이들 데이터 셋은 보통 제한된 공간에서 수집된 짧은 길이의 동작들로 구성되어 있어, 스포츠나 엔터테인먼트 환경에서 요구되는 고속 동작 예측에는 부적합하다. 현재 공개된 데이터셋 중에서 이러한 동적 움직임을 가장 잘 포착한 것은 LaFAN 데이터셋으로, 게임 개발을 목적으로 제작되었으며, 전문 배우가 넓은 공간에서 수행한 고속 동작이 포함되어 있다. LaFAN은 AMASS나 Human3.6M과는 근본적으로 다른 동작 프로파일을 제공하며, 빠르고 표현력 있는 고에너지 동작이 요구되는 시나리오에서 모델 성능을 평가하기에 특히 적합하다. 그러나 이 데이터셋은 동작 예측 모델에게 독특한 도전 과제를 제시한다. 특히 의미론적으로 유사한 동작 간에도 수치적인 거리 차이가 매우 크다는 특성이 있다. 예를 들어, 느린 걷기 동작과 서 있는 동작 간의 수치적 거리는 매우 작지만, 서로 다른 속도로 수행된 두 개의 빠른 달리기 동작 간의 거리는 훨씬 크다. 이러한 차이는 모델이 동일 동작 유형 내의 변화를 일반화하기 어렵게 만들며, 결과적으로 모델은 유사한 동적 동작 간을 보간하지 못하고, 오히려 학습 중에 본 특정 동작 패턴에 과적합 (overfit)되는 경향을 보인다. 이에 따라 예측 결과는 일관 성과 안정성이 떨어지는 문제가 발생한다. 예를 들어, 표현력을 강조한 확산 기반 (diffusion-based) 예측 모델의 경우, 개별 프레임은 자연스럽게 보일 수 있지만, 전체 예측 동작 시퀀스는 시간적으로 일관성이 떨어지고 각 프레임이 급격하게 변하는 문제가 있다. 이러한 현상은 과적합의 결과이거나, 맥락 (context) 또는 동작 정체성 (identity)을 보존하지 못한 것으로 해석될 수 있다. 일관된 맥락 보존형 동작 생성을 목표로 하는 두 번째 응용 범주에서는 이러한 가변성은 단점으로 작용한다. 반대로, 시간적 유사성을 잘 반영하는 연구들은 장기 예측 시 예측이 평균 동작으로 수렴하는 문제가 빈번하게 나타난다. 이러한 도전 과제는 보조 로보틱스나 가상 아바타와 같은 실제 응용에서 동작 예측 모델의 실용성을 크게 저하시킨다.

본 연구는 이러한 한계를 극복하고 더 정확하며 시간적으로 일관된 인간 동작 예측을 달성하기 위해 세 가지 핵심 기술적 혁신을 제안하였다:

Transformer 기반 U-Net을 활용한 LDM: Transformer와 Latent Diffusion Model의 결합은 최근 동작 예측 과제에서 점점 더 주목받고 있다. 우리의 하이브리드 모델은 다양한 아키텍처의 강점을 통합하여 활용한다: U-Net의 계층적 특징 추출 (hierarchical feature extraction), Transformer의 장거리 의존성 모델링 (long-range dependency modeling), 그리고 잘 설계된 임베딩 공간에서 작동할 때 특히 뛰어난 성능을 보이는 확산 기반 접근법 (diffusion approach)의 장점을 결합하였다.

월드 모델을 이용한 강화학습: 우리는 기존에 보유한 프레임 데이터를 활용하여 Neural ODE를 학습시켰으며, 이를 통해 데이터 공간상에서 월드 모델을 구축하였다. 이렇게 생성된 가상의 특징 환경 (feature environment)은 데이터셋만을 기반으로 구성된 시뮬레이션 환경으로, 강화학습에 활용되었다. Neural ODE를 통해 외부 시뮬레이터 없이도 원본 데이터셋만으로 일관되고 그럴듯한 가상 환경을 구축하여 강화학습을 진행하였다 [7, 8].

Dual Sampling: 우리는 기존의 diffusion 모델에서 p-sampling을 통해서 동작을 예측하지만 이는 상한은 존재하지만 하한은 존재하지 않는 문제가 있다.이를 해결하기 위해서 p-sampling의 하한을 위한 pd-sampling을 생성하여서 노이즈의 최적화 연구를 진행하였다 [9].

2 Related Work

이 장에서는 본 연구와 관련된 세 가지 핵심 분야를 중심으로 기존 연구들을 소개한다.

2.1 Latent Diffusion Model

Latent Diffusion Model (LDM)은 입력 특징 공간이 아닌 학습된 잠재 공간 (latent space)에서 확산 과정을 수행함으로써, 생성 능력을 유지하면서도 차원수를 효과적으로 감소시킨다 [10, 11]. LDM는 높은 품질과 효율성 덕분에 조건부 생성 (conditional generation)에서 널리 사용되고 있다. 기존 Diffusion Model의 생성을 향상시키기 위해 다양한 아이디어가 제안되었으며, 예를 들어, 변분 오토인코더 (Variational AutoEncoder, VAE)를 활용한 잠재 공간의 압축 [12], 점진적인 노이즈 제거 과정을 포함한 확산 과정 [13], GAN 프레임워크와의 통합 [14] 등이 있다.

LDM는 텍스트-이미지 생성 [15], 비디오 복원 [16] 등 다양한 작업에 성공적으로 적용되어 왔으며,

최근에는 인간 동작 예측 [17] 및 동작 계획 (motion planning) [18, 19] 분야에서도 활용되며, 좌표 공간이 아닌 잠재 공간 내에서 더 현실적이고 다양한 동작 생성을 가능하게 하고 있다.

본 연구와 가장 유사한 선행 연구는 BeLFusion을 통해서 LDM을 사용하게 되었다. 이 연구는 LDM을 동작 예측에 도입하며, 자세 표현과 동작 역학을 분리한 행동 기반 잠재 공간에서 확산 과정을 수행하였다 [17].

기존의 방법들이 좌표 수준에서의 다양성에 초점을 맞춘 것과 달리, BeLFusion은 본 연구와 마찬가지로 잠재 임베딩을 활용하여 행동 수준의 일관성을 유지하는 데 중점을 두고있다.우리의 연구도 Belfusion에서 사용하는 LDM의 매커니즘을 사용하여서 동작을 예측하였다.

2.2 동작 예측을 위한 강화학습

강화학습은 동적인 환경에서의 순차적 의사결정 과정을 모델링하는 강력한 프레임워크로 부상하였다. 특히, 로보틱스 및 물리 기반 캐릭터 제어 분야에서 복잡하고 고차원의 공간 내에서 안정적이고 목표 지향적인 동작을 학습하는 데 있어 강화학습은 탁월한 성과를 보여주고 있다 [20, 21].

동작 예측 분야에서 강화학습은 사전에 정의된 동작 궤적을 넘어서 일반화 가능한 적응형 정책 (adaptable policies)을 학습할 수 있도록 한다. 이는 과거에서 미래 동작으로의 직접적인 매핑에 의존하는 전통적인 지도학습 방식과 달리, 환경과의 상호작용을 통해 행동을 최적화함으로써 복잡하고 불확실한 상황에서 더 뛰어난 일반화 성능을 제공한다 [22, 23].

또한 일부 연구에서는 LSTM이나 GRU와 같은 순환 신경망 (RNN)을 활용하여시간적 이력을 압축된 잠재 상태로 인코딩하여 월드 모델을 구현하는 방식으로 문제를 정식화하였다 [24, 25]. 하지만 이러한 방법들은 동작 예측 (task prediction)에 있어 큰 난제 중 하나는 부분 관측성 문제 (partial observability)를 포함하고 있다. 센서 데이터는 흔히 잡음이 많거나 희소한 특성을 가지며, 움직임 이면의 실제 의도를 명확하게 파악하는 것은 본질적으로 불가능하다. 이러한 이유로 전체 상태 정보를 정확하게 관측하는 데 어려움이 따른다.

본 연구에서는 부분 관측성 문제를 해결하기 위해 Neural ODE를 도입하여, 현실 세계의 복잡한 동역학을 효과적으로 근사하였다. Neural ODE는 연속적인 시간 변화를 모델링하는 능력이 뛰어나, 관측이 불완전한 환경에서도 상태의 추정과 예측 성능을 향상시키는 데 크게 기여한다.

이렇게 학습된 Neural ODE를 기반으로 현실 세계의 복잡한 환경을 모사하는 월드 모델 (world model)을 구축하였다. 월드 모델 내부에서 에이전트는 직접 계획하거나 학습할 수 있는 환경을 제공받게 되며, 이를 통해 데이터 효율성을 크게 향상시킬 수 있었다. 그리고 이렇게 학습된 월드 모델을 통해서 우리는 부분관측만조건인 POMDP (Partial Observe Markov Decision Process)를 통해서 강화학습을 진행하여서 부분관측 문제를 해결하고 노이즈 견고성을 강화하였다.

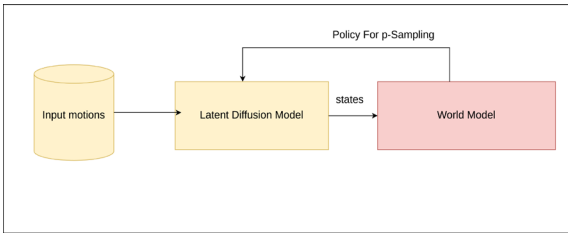


Figure 1: Whole Model.

2.3 Dual sampling

Diffusion 모델은 최근 생성 모델 분야에서 우수한 성능을 입증하며 활발하게 연구되고 있다. 특히, 확률 밀도 기반의 Sampling 방법인 p-sampling은 노이즈를 점진적으로 제거하여 현실적인 데이터를 생성하는 데 널리 사용되어 왔다. 그러나 이러한 방법은 노이즈 수준에 상한 (upper bound)만 존재하고 하한 (lower bound)이 설정되지 않아, 노이즈가 작아지는 경우에 대한 견고성을 보장하지 않아서, Sampling 과정의 안정성과 견고성 (robustness)이 떨어지는 단점이 있다 [26] . 이러한 문제를 해결하기 위해 최근의 연구들에서는 기존의 p-sampling 방식에 대응하는 dual sampling을 도입하여 노이즈에 대한 하한을 설정함으로써 모델의 견고성을 높이려는 접근법이 제안되고 있다 . 확률적 Sampling (DDPM)과 결정론적 Sampling (DDIM)의 dual 구조를 활용하여, 노이즈의 최소 수준을 수학적으로 보장함으로써 노이즈가 과도하게 감소하는 현상을 방지하는 방법을 제안하였다. 또한, Diffusion Model의 확률 미분 방정식 (SDE) 프레임워크를 기반으로 리스크 민감도 (risk-sensitivity)를 최적화하여, 노이즈의 최소값과 최대값을 함께 고려함으로써 Sampling 과정의 안정성을 향상시키고자 하였다 [27].본 연구 역시 이러한 흐름을 따라 p-sampling과 dual sampling 인 pd-sampling을 결합하여 Sampling 과정에서 노이즈 수준에 대한 하한과 상한을 동시에 제어함으로써 보다 안정적이고 견고한 생성 성능을 달성하고자 한다.

3 Overview

Figure 1에 나타난 바와 같이, 본 시스템은 두 가지 모듈과 한가지 최적화 과정으로 구성된다:

1. LDM.
2. 월드 모델.
3. Dual Sampling.

LDM은 과거 10프레임을 입력으로 받아, 미래 10프레임 이후의 단일 동작을 예측한다. 즉, 1프레임부터 10프레임까지의 미래 동작들을 예측하기 위해서는, 과거 20프레임을 슬라이딩 윈도우 방식으로 입력으로 사용하여 각 미래 프레임에 대해 총 10번 반복하여 예측을 수행한다.

월드 모델은 LDM으로부터 예측된 10개의 동작을 입력으로 받아 이를 개선한다. 이를 손실 함수 기반의 학습을 통해 LDM 예측에 대한 노이즈 견고성을 향상시킨다.구체적으로, 월드 모델은 LDM으로부터 예측된 프레임들과 q-Sampling된 노이즈를 환경 상태로 간주하고, 강화학습을 통해 장기적이고 일관된 예측을 위한 정책을 최적화한다.

마지막으로 전체적인 모델은 아니지만 노이즈 최적화를 위한 Dual sampling을 실행해 노이즈에 대한 경계를 잡아 주었다.다음 절에서는 각 모듈에 구조와 동작원리를 순차적으로 설명한다.

4 Latent Diffusion Model

Figure 2에 나타난 바와 같이, 본 연구에서는 고차원 동작 데이터를 보다 효율적으로 처리하기 위해 이를 압축된 잠재 공간으로 사영하는 LDM 프레임워크를 채택하였다. 이러한 전략은 연산 및 메모리 효율을 크게 향상시켜, 고해상도 입력 데이터에 대해서도 효율적인 학습을 가능하게 한다. 해당 잠재 공간에서는 제어된 양의 노이즈가 체계적으로 주입된다. LDM은 이 노이즈를 점진적으로 제거함으로써, 원래의 데이터를 복원한다. LDM의 학습 과정은 다음과 같은 목적 함수로 표현된다:

$$L = E_{\mathbf{z}_t, c_{\text{manifold}}, \epsilon \sim \mathcal{N}(0,I), t} \left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, c_{\text{manifold}}, t)\|_2^2 \right], \quad (1)$$

여기서 $\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \epsilon, \quad \alpha_t, \sigma_t \in R.$ 는 노이즈 확산 과정에서의 하이퍼파라미터이며, c_{manifold}는 동작 데이터의 임베딩 다양체 구조에 대한 조건을 나타낸다.

우리는 본 LDM 프레임워크 내에 Transformer가 결합된 U-Net 구조 (Transformer-Supported U-Net)를 제안한다. 이 모델은 U-Net의 계층적 특징 추출 능력과 Transformer의 장거리 시공간 의존성 모델링 능력을 결합하여, 예측의 품질과 시간적 일관성을 크게 향상시킨다. 이 인코더-디코더 구조는 다음과 같은 주요 구성 요소들로 이루어진다. 첫째, Manifold-Aware Pose Encoder는 현재 동작과 노이즈 벡터를 사전에 정의된 다양체 표현을 활용하여 인코딩한다. 생성된 임베딩은 U-Net 내에서 스킵 연결을 통해 디코딩 시 의미론적 구조를 보존하도록 전달된다. 둘째, Transformer의 attention 모듈은 동작, 노이즈, 다양체 특징을 텐서 곱 방식으로 통합한 후, Transformer를 통해 전역적인 시공간 의존성을 학습한다. 이러한 구조는 고정밀 동작 복원을 가능하게 할 뿐만 아니라, 강화학습 파이프라인과의 높은 호환성도 유지한다. 특히 Transformer-Supported U-Net은 매우 역동적인 동작 입력으로부터 부드럽고 시간적으로 일관된 예측을 생성하는 데 매우 효과적이다. Transformer 기반 Attention 모듈의 출력은 두 가지 구성 요소로 분기된다. 첫 번째 구성 요소는 denoised motion sequence를 예측하며, 이는 단기 동작 예측에 사용된다. 두 번째 출력은 fake noise generator로 지정되며, 월드 모델과의 통합을 위해 설계된 노이즈를 생성한다. 이 노이즈는 다양한 시나리오 생성을 가능하게 하여, 환경 시뮬레이션의 견고성을 높이는 데 기여한다. 이러한 이중 출력 구조는 GAN에서의 노이즈 생성 메커니즘과 유사하게 작동한다. 네트워크 구조에 대한 보다 자세한 내용은 Section A을 참조하기 바란다.

5 World Model

Figure 3에 나타난 바와 같이, 본 연구에서는 강화학습 기반의 장기 동작 예측을 위해 월드 모델을 사용한다. 이 월드 모델은 자기지도학습에 기반한 모듈로, 실제 인간 동작 및 그 변화를 학습하도록 설계되었다. 우리의 월드 모델은 다음과 같은 명제로부터 출발한다:

명제 (Proposition): 인간 동작 예측에서 입력 동작 표현의 특성상 예측 공간은 컴팩트 (compact)하다. 이를 모델링하기 위해, 우리는 시간과 자세 (프레임)의 곱공간을 월드 모델 다양체 (world model manifold)로 사상하는 연산자 (operator)를 정의하며, 이는 POMDP 프레임워크 내에서 최적화된다.

$$K : F^k \to W, \quad (2)$$

여기서 k는 예측 윈도우 (10 프레임)를 나타내고, $F = R^{35}$는 프레임 공간을 나타내며, 이는 Neural ODE에 의해 예측된 동작 공간으로, 각 벡터는 인체의 루트 구성과 관절 각도를 인코딩한다.

W는 월드 모델 다양체 공간을 나타내며, 10프레임으로 구성된 동작 시퀀스를 구조적으로 임베딩한 공간이다. 집합 W는 다음과 같은 함수를 포함한다:

$$W = \{f \mid f : W \to W_{\text{best}} \text{ using POMDP}\},$$

이는 불확실성 하에서 최적 상태 전이를 보장하는 함수 집합이다.

연산자 K는 닫힌 (closed) 컴팩트 연산자로서, 동작 데이터로부터 월드 모델 다양체로의 사상에서 안정성과 유계성 (boundedness)을 보장한다. 또한 K는 Neural ODE를 사용하여 구성되며, 시간에 따른 인간 동작의 연속적이고 미분 가능한 표현을 제공한다. 따라서, K를 통한 임베딩은 인간 동작을 구조화된 방식으로 학습할 수 있도록 하며, POMDP 기반 최적화를 활용하여 최적의 정책을 학습할 수 있도록 한다.

5.1 Loss Functions

모든 모델은 입력 프레임으로부터 목표 동작을 예측하는 방식으로, 직접 정책 최적화와 유사한 방식으로 동시에 학습된다. 전체 손실 함수는 다음과 같이 네 개의 독립적인 항목으로 구성된다:

$$L_{total} = L_{output} + L_{noise_{fake}} + L_{world}. \quad (3)$$

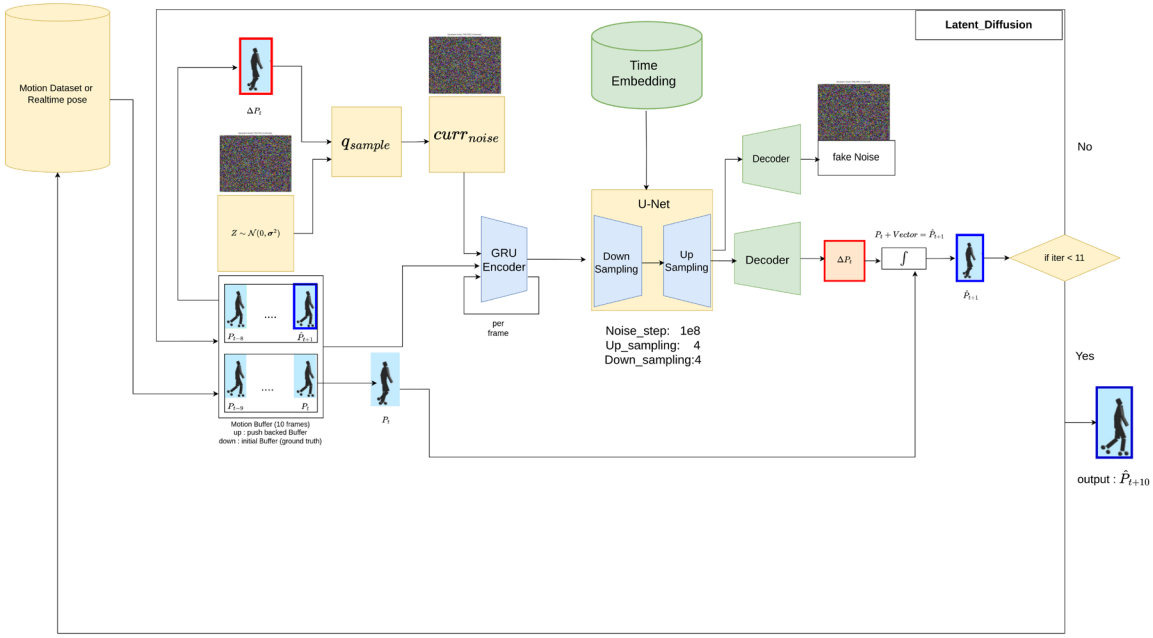


Figure 2: Latent Diffusion Model.

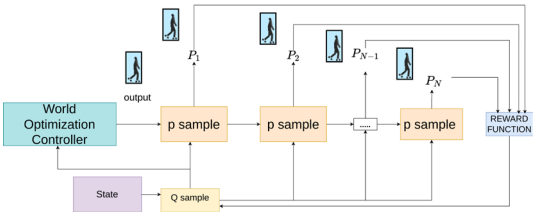


Figure 3: World Optimization Controller.

출력 손실 L_{output}은 LDM의 출력 동작과 해당하는 정답 동작 (ground truth motion) 간의 L^2 차이를 직접적으로 측정한다:

$$L_{output} = \sum_{k \in N} \|diffusionPredict(s_k) - \bar{P}_k\|_2^2. \quad (4)$$

$L_{noise_{\text{fake}}}$ 이는 LDM로부터 생성된 노이즈 출력을 기반으로 계산된 GAN 유사 손실(GAN-like loss)을 나타낸다:

$$L_{noise_{\text{fake}}} = \|noise - noise_{\text{fake}}\|_2^2. \quad (5)$$

$L_{noise_{\text{fake}}}$는 LDM의 q-Sampling 과정에서 사용된 원래 노이즈와 LDM이 예측한 (또는 "가짜" 로 생성한) 노이즈를 비교하는 손실 함수이다. 이 손실 항은 월드 모델의 출력이 LDM이 기대하는 노이즈 분포와 호환되도록 보장하여, LDM이 적절한 노이즈 경계 내에서 유효한 예측을 생성할 수 있도록 한다. 월드 모델의 강화학습 손실인 L_{world}는 사전 학습된 노이즈 경계 조건을 활용하여, 월드 모델 프레임워크 내에서 센서 입력 프레임 공간에서의 POMDP를 해결한다. 이 손실 함수는 부분 관측 하의 Neural ODE를 활용하여 강화학습을 수행하며, 할인 누적 보상 (discounted cumulative reward)을 통해 학습을 진행한다. 여기서 r은 할인 계수 (discount factor)를 의미한다:

$$L_{world} = \sum_{k \in Z} r^k \cdot \left\|\bar{P}_k - worldModelPredict(s_k)\right\|_2^2. \quad (6)$$

이러한 손실 함수들을 통합함으로써, 전체 프레임워크는 동작 예측의 정확도와 강화학습의 효율성을 동시에 최적화하고자 한다. 전통적인 지도 학습을 넘어, 우리는 월드 모델을 활용하여 미래 프레임을 생성하고 최적화를 수행한다. 또한, 본 접근법은 프레임 시퀀스 상에서 강화학습에 적합한 공간을 구성하기 위해 Neural ODE 기반의 수식화를 통합하였다. 학습 과정에 대한 보다 자세한 내용은 Section C을 참조하기 바란다.

6 Dual sampling

본 연구에서는 기존의 p-sampling을 넘어서 pd-sampling을 추가하여서 단순 sampling이 아닌 노이즈 에대한 경계값을 설정하는 sampling을 통해서 노이즈에 대한 최적화를 진행하였다. 이에 대한 sampling 의 공식은 아래와 같다:

$$\begin{aligned} \text{normal sample} &= \frac{p \text{ sample} + pd \text{ sample}}{2} \\ &= \frac{(\mu + \sigma \cdot \epsilon_1) + (\mu - \sigma \cdot \epsilon_2)}{2} \\ &= \mu + \epsilon_3 \end{aligned} \quad (7)$$

여기서 μ는 생성된 평균 (mean)을 나타내며, ϵ은 각각의 Sampling 과정에서 사용된 노이즈를 나타낸다. 우리는 normal sam-

pling을 통해서 기존의 LDM의 생성에 대한 장점을 유지하면서도 노이즈 에 대해서 견고한 동작을 생성할 수 있었다.

7 Experimental results

구체적으로, 동작 유형별로 분류했을 때, 학습 데이터셋에는 전체 데이터셋 중 Aiming의 40%, Dance의 42%, Fight의 33%, Ground의 50%, Jump의 33%, Multi-Action의 75%, Obstacle의 44%, Push의 33%, Run의 25%, Splinter의 50%, Walk의 16%가 포함된다. 이 동작들은 주로 Subject 1과 Subject 2에 의해 수행되었다. 나머지 동작들은 평가용 데이터로 사용되었다.

학습 데이터에는 스타일적으로 유사한 동작들이 포함되어 있으나, 평가 데이터와 동일한 동작 시퀀스는 포함되지 않아 일반화 성능을 공정하게 평가할 수 있도록 구성하였다.

첫 번째 연구에서는 제안하는 동작 예측 프레임워크를 세 가지 기준 모델과 비교하여 평가하였다. 첫 번째 기준 모델인 BeLFusion은 본 연구와 유사하게 LDM을 활용한다 [17]. 마지막 두 기준 모델인 TransFusion은 Transformer 기반의 Diffusion Model을 채택하여 동작 예측을 수행한다 [28, 29]. 모든 모델은 RTX 3070 Ti GPU와 AMD Ryzen 5 5600X 6코어 CPU가 장착된 머신에서 학습되었다. 모델별 학습 시간은 다음과 같이 상이하였다: BeLFusion과 TransFusion은 약 2일정도의 학습 시간이 소요되었고, 한편, 제안하는 우리 모델은 아키텍처의 복잡성과 추가적인 월드 모델에 의한 강화학습 과정으로 인해 약 3일정도의 학습 시간이 소요되었다.

우리의 방법은 BeLFusion 및 TransFusion과 같은 확산 기반 (diffusion-based) 모델들과 동일한 타임스텝 수 (time steps)를 사용하여 학습되었다. 예측 정확도를 평가하기 위해, 예측된 동작과 정답 동작(ground-truth)을 다음과 같은 지표를 사용하여 비교하였다: **MPJPE** (Mean Per Joint Position Error) – 관절 위치 오차의 평균, 단위는 미터 (m), **MPJRE** (Mean Per Joint Rotation Error) – 관절 회전 오차의 평균, 단위는 도 (degree). Table 1는 각

Table 1: Model-wise Comparison of MPJRE and MPJPE Performance.

Model	Performance		Aim	Ground	Fight	Push	Multiple	Walk
Ours	MPJPE (m)	max	**0.039**	**0.079**	**0.068**	**0.097**	**0.063**	**0.124**
		avg	**0.014**	**0.021**	**0.017**	**0.024**	**0.022**	**0.052**
	MPJRE (deg)	max	**4.388**	**4.389**	**12.534**	**10.501**	**7.810**	**10.412**
		avg	**1.660**	**1.788**	**2.831**	**3.509**	**3.468**	**2.886**
Belfusion	MPJPE (m)	max	0.357	0.393	0.425	0.409	0.275	0.279
		avg	0.032	0.051	0.066	0.039	0.040	0.033
	MPJRE (deg)	max	8.125	12.457	19.470	14.911	17.241	15.521
		avg	1.995	2.410	4.932	3.519	3.649	3.112
Transfusion	MPJPE (m)	max	0.347	0.416	0.398	0.435	0.335	0.291
		avg	0.046	0.070	0.051	0.047	0.050	0.042
	MPJRE (deg)	max	8.928	11.658	15.190	14.813	11.888	16.003
		avg	2.962	3.293	4.612	4.255	4.327	3.885

동작 범주별로 오류를 측정한 후, 전체 범주에 대한 최대 오차 및 평균 오차를 제시한다. 우리의 방법은 모든 연구에서 가장 우수한 결과를 달성하였다.

더 중요한 점은, 다른 모든 접근 방식이 매우 역동적인 동작에 대해 어려움을 겪었으며, 그 결과 예측이 전반적으로 노이즈가 많은 경향을 보였다는 것이다. 반면, 월드 모델을 결합한 우리 모델은 더 높은 품질의 동작 예측을 생성하였으며, 이는 첨부된 비디오에서도 각각의 모델의 출력을 비교하면 확인할 수 있다.

8 Ablation Study

제안하는 모델의 전체 성능에 있어 각 구성 요소인 잠재 동작 U-Net, 그리고 월드 모델의 개별적인 기여도를 명확히 평가하기 위해 Ablation Study를 수행하였다.

Table 2에 요약된 결과는 각 구성 요소가 성능 향상에 의미 있게 기여하고 있음을 보여준다. 각각 LDM 모델에서의 Latent 파트의 제거, 그리고 그 후 LDM에서 월드 모델 의 제거 단계로 ablation study를 진행하였다. 이는 단순 확산인 Standard Diffusion을 통해서 기존의 Diffusion을 실험 하였고,그 후 우리모델에서 강화학습 부분인 월드 모델을 제거한 Standard LDM을 실험하였다.마지막으로 월드 모델까지 포함한 우리의 모델로의 발전을 연구로 진행하였다. 실험 결과를 보면,월드 모델을 제외했을 경우 성능이 크게 저하되었으며, 이는 구성 요소가 모델의 성능에 있어 매우 중요한 역할을 하고 있음을 뒷받침한다.

Table 2: MPJRE and MPJPE Performance Comparison (Ablation Study).

Model	Performance		Aiming	Dance	Fall	Splinter	Walk
Ours	**MPJPE (m)**	min	0.011	0.015	**0.005**	**0.001**	**0.007**
		max	**0.077**	**0.174**	**0.079**	**0.078**	**0.169**
		average	**0.014**	**0.056**	**0.025**	**0.016**	**0.052**
	MPJRE (deg)	min	0.630	1.615	**0.717**	**0.039**	**0.664**
		max	**4.388**	**12.781**	**12.770**	**12.505**	**10.412**
		average	**1.660**	5.952	**3.196**	**2.979**	**2.886**
Standard DM	MPJPE (m)	min	0.013	0.012	0.011	0.007	0.016
		max	0.491	0.176	0.330	0.234	0.114
		average	0.251	0.052	0.089	0.076	0.049
	MPJRE (deg)	min	1.767	1.814	1.703	1.481	1.857
		max	16.548	24.111	31.337	29.396	15.104
		average	6.451	7.168	10.191	10.590	6.915
Standard LDM	MPJPE (m)	min	0.004	0.011	0.009	0.022	0.011
		max	0.245	0.187	0.352	0.170	0.144
		average	0.075	0.067	0.102	0.062	0.052
	MPJRE (deg)	min	0.494	1.415	0.995	1.042	0.984
		max	5.962	12.887	13.358	14.648	11.412
		average	2.095	5.590	5.255	4.667	3.887

그리고 우리는 Dual sampling을 진행한 것과 아닌 것을 비교하는 연구를 추가로 진행하였다. 이를 통해서 우리는 Dual Sampling을 통한 노이즈의 최적화가 적용 가능함을 알 수 있었다.

Table 3: MPJRE and MPJPE Performance Comparison (No Sampling vs. Ours).

Model	Performance	Dance	Run	Walk
Ours	MPJRE (deg)	8.141	6.304	6.606
	MPJPE (m)	0.0201	0.0664	0.077
No Sampling	MPJRE (deg)	9.925	8.4255	9.24377
	MPJPE (m)	0.1307	0.1470	0.1348

9 Conclusion and Discussion

본 연구는 기존 확산 기반 (Diffusion) 모델이 갖는 노이즈 민감성 문제를 극복하고, 보다 정확하고 시간적으로 일관된 인간 동작 예측을 달성하기 위해 세 가지 핵심 기술을 제안한다.

첫째, 트랜스포머 기반의 U-Net 구조를 활용한 LDM을 도입하였다. 최근 동작 예측 과제에서는 트랜스포머와 LDM의 결합이 점차 주목받고 있으며, 본 연구에서는 트랜스포머의 시계열 정보에 대한 장기 의존성 처리 능력을 활용하여, 노이즈 환경 하에서의 경계를 효과적으로 형성함으로써 노이즈에 대한 견고성을 확보하였다.제안하는 하이브리드 모델은 다양한 아키텍처의 장점을 통합한다. U-Net의 계층적 특징 추출 능력, 트랜스포머의 장거리 의존성 모델링, 그리고 구조화된 임베딩 공간에서 효과적으로 작동하는 확산 기반 접근법(diffusion approach)의 이점을 조합하였다.이는 기존의 Diffusion Model에서는 역문제(inverse problem) 해결 시 데이터 정보가 부족하여, 연속적인 동작을 생성할 때 노이즈가 잔존하고 결과적으로 동작의 불안정한 현상이 발생하는 문제를 해결하고제노이즈에 대한 견고성을 향상시켰다.

두번째로, 우리의 모델은 기존의 강화학습이 가지는 부분 관측성의 문제를 Neural ODE를 통한 월드 모델의 구현을 통해서 해결하였다.기존의 강화학습에서는 환경에 대한 전체 정보를 통해서 학습이 진행되는데,이는 현실적으로 전체 환경을 예측하는 어려움이 실제 세계에 있고,전체 정보를 전부 집어 넣어서 학습을 시키면,시간적으로 큰 손실이 있으며 과적합의 문제 또한 발생할 수 있다.우리의 모델은 월드 모델의 장점이 특징 공간 (feature space) 만을 사용하여서 과적합에 대해서 견고성을 유지함과 동시에 Neural ODE를 통해서 부분 정보를 통해서 실세계에서의 부분 관측성 문제를 해결 하였다.이는 우리의 월드 모델을 통한 강화학습이 노이즈에 대한 견고서을 높이는데 있어서 실세계와 바로 연결된다는 장점을 시사하고 있다.

마지막으로, Dual Sampling 전략을 통한 노이즈에 대한 최적화 전략으로 Dual Sampling을 통해서 우리의 연구는 기존의 Diffusion Model 이 가지고 있는 p-sampling 에서의 단점을 해결하고 단순한 동작의 생성을 넘어서서 고품질의 동영상을 실시간적으로 예측하는 결과를 얻을수 있었다 [7, 8].

하지만 우리의 연구는 아직 실시간 동작 생성에서 13frame 이상에서는 결과가 좋지 않다는 단점이 있다.앞으로의 연구에서는 실시간 고품질 동작 생성에 더해서 장기간 동작 예측에 대한 추가적인 연구가 필요해 보인다.

감사의 글

본 연구는 G-deep XR(2024.04) - RS-2024-00399136 [중앙행정기관 - 문화체육관광부/전문기관 - 한국콘텐츠진흥원] 과 중견 연구(2024.05) - RS-2024-00354549 [중앙행정기관 - 과학기술정보통신부/전문기관 - 한국연구재단]을 통해 진해되었다.

References

[1] G. Aydemir *et al.*, "Adapt: Efficient multi-agent trajectory prediction with adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[2] C. Zhong *et al.*, "Spatio-temporal gating-adjacency gcn for human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[3] W. H. Tai *et al.*, "Cutting-edge research in sports biomechanics: From basic science to applied technology," *Bioengineering (Basel)*, vol. 10, no. 6, p. 668, 2023.

[4] M. H. Abidi, "Multimodal data-based human motion intention prediction using adaptive hybrid deep learning network for movement challenged person," *Scientific Reports*, vol. 14, p. 30633, 2024.

[5] N. Mu *et al.*, "Most: Multi-modality scene tokenization for motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[6] Y. Gao *et al.*, "Multi-transmotion: Pre-trained model for human motion prediction," *8th Annual Conference on Robot Learning (CoRL)*, 2024.

[7] D. Ha *et al.*, "World models," *arXiv preprint arXiv:1803.10122*, 2018.

[8] Zhao *et al.*, "Ode-based recurrent model-free reinforcement learning for pomdps," *Advances in Neural Information Processing Systems*, vol. 36, pp. 65 801–65 817, 2023.

[9] I. Bae *et al.*, "Non-probability sampling network for stochastic human trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[10] A. van den Oord *et al.*, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] R. Rombach *et al.*, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[12] D. P. Kingma *et al.*, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[13] J. Ho *et al.*, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851.

[14] Z. Wang *et al.*, "Diffusion-gan: Training gans with diffusion," *arXiv preprint arXiv:2206.02262*, 2022.

[15] W. Peebles *et al.*, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[16] G. Zheng *et al.*, "Cami2v: Camera-controlled image-to-video diffusion model," *arXiv preprint arXiv:2410.15957*, 2024.

[17] G. Barquero *et al.*, "Belfusion: Latent diffusion for behavior-driven human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[18] T. V. Wouwe *et al.*, "Diffusionposer: Real-time human motion reconstruction from arbitrary sparse sensors using autoregressive diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[19] A. Serifi *et al.*, "Robot motion diffusion model: Motion generation for robotic characters," in *SIGGRAPH Asia 2024 Conference Papers*, 2024.

[20] Haarnoja *et al.*, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1861–1870. [Online]. Available: https://proceedings.mlr.press/v80/haarnoja18b.html

[21] X. B. Peng *et al.*, "Deeploco: Developing locomotion skills using hierarchical deep reinforcement learning," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2017.

[22] S. Wu *et al.*, "E-motion: Future motion simulation via event sequence diffusion," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 105 552–105 582.

[23] Z. Zhang *et al.*, "Predicting long-term human behaviors in discrete representations via physics-guided diffusion," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024.

[24] J. Martinez *et al.*, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[25] Y. Tang *et al.*, "Long-term human motion prediction by modeling motion context and enhancing motion dynamic," *arXiv preprint arXiv:1805.02513*, 2018.

[26] G. Chen *et al.*, "Unsupervised sampling promoting for stochastic human trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[27] S. Xu *et al.*, "Stochastic multi-person 3d motion forecasting," in *The Eleventh International Conference on Learning Representations*, 2022.

[28] Zhou *et al.*, "Transfusion: Predict the next token and diffuse images with one multi-modal model," *arXiv preprint arXiv:2408.11039*, 2024.

[29] X. Sun *et al.*, "Defeenet: Consecutive 3d human motion prediction with deviation feedback," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[30] Y. Qu, T.-T. Wong, and P.-A. Heng, "Manga colorization," *ACM Transactions on Graphics*, pp. 1214–1220, 2006.

[31] P. Alliez, D. Cohen-Steiner, O. Devillers, B. Levy, and M. Desbrun, "Anisotropic polygonal remeshing," *ACM Transactions on Graphics*, pp. 485–493, 2003.

[32] L. Fussell *et al.*, "Supertrack: Motion tracking for physically simulated characters using supervised learning," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 6, pp. 1–13, 2021.

[33] Yao *et al.*, "Controlvae: Model-based learning of generative controllers for physics-based characters," *ACM Transactions on Graphics*, vol. 41, no. 6, 2022. [Online]. Available: http://dx.doi.org/10.1145/3550454.3555434

[34] Chen *et al.*, "Mgf: Mixed gaussian flow for diverse trajectory prediction," *arXiv preprint arXiv:2402.12238*, 2024.

[35] Z. Zhang *et al.*, "Real-time motion prediction via heterogeneous polyline transformer with relative pose encoding," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 57 481–57 499.

[36] S. Xu *et al.*, "Diverse human motion prediction guided by multi-level spatial-temporal anchors," in *European Conference on Computer Vision*. Springer Nature Switzerland, 2022.

[37] W. Mao *et al.*, "Contact-aware human motion forecasting," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 7356–7367.

[38] M. Meng *et al.*, "Forecasting human trajectory from scene history," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 24 920–24 933.

[39] J. Sun *et al.*, "Stimulus verification is a universal and effective sampler in multi-modal human trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[40] D. Zhu *et al.*, "Ipcc-tp: Utilizing incremental pearson correlation coefficient for joint multi-agent trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[41] C. Xu *et al.*, "Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[42] F. G. Harvey *et al.*, "Robust motion in-betweening," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 60–1, 2020.

[43] Y.-H. Park *et al.*, "Unsupervised discovery of semantic latent directions in diffusion models," *arXiv preprint arXiv:2302.12469*, 2023.

[44] S. Shi *et al.*, "Motion transformer with global intention localization and local movement refinement," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 6531–6543.

[45] M. Li *et al.*, "Skeleton-parted graph scattering networks for 3d human motion prediction," in *European Conference on Computer Vision*. Springer Nature Switzerland, 2022.

[46] C. Xu *et al.*, "Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[47] D. Guo *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[48] A. Liu *et al.*, "Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model," *arXiv preprint arXiv:2405.04434*, 2024.

[49] P. Xu *et al.*, "Context-aware timewise vaes for real-time vehicle trajectory prediction," *IEEE Robotics and Automation Letters*, vol. 8, no. 9, pp. 5440–5447, 2023.

[50] W. Wu *et al.*, "Smart: Scalable multi-agent real-time motion generation via next-token prediction," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 114 048–114 071.

[51] C. Feng *et al.*, "Macformer: Map-agent coupled transformer for real-time and robust trajectory prediction," *IEEE Robotics and Automation Letters*, vol. 8, no. 10, pp. 6795–6802, 2023.

[52] G. Xu *et al.*, "Learning semantic latent directions for accurate and controllable human motion prediction," in *European Conference on Computer Vision*. Springer Nature Switzerland, 2024.

[53] X. Lin *et al.*, "Progressive pretext task learning for human trajectory prediction," in *European Conference on Computer Vision*. Springer Nature Switzerland, 2024.

[54] C. Xing *et al.*, "Scene-aware human motion forecasting via mutual distance prediction," in *European Conference on Computer Vision*. Springer Nature Switzerland, 2024.

[55] J. Yue *et al.*, "Human motion prediction under unexpected perturbation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[56] W. Mao *et al.*, "Weakly-supervised action transition learning for stochastic human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[57] T. Ma *et al.*, "Progressively generating better initial guesses towards next stages for high-quality human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[58] T. Salzmann *et al.*, "Motron: Multimodal probabilistic human motion forecasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[59] C. Diller *et al.*, "Forecasting characteristic 3d poses of human actions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[60] X. Sun *et al.*, "Overlooked poses actually make sense: Distilling privileged knowledge for human motion prediction," in *European Conference on Computer Vision*. Springer Nature Switzerland, 2022.

[61] X. Gao *et al.*, "Decompose more and aggregate better: Two closer looks at frequency representation learning for human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[62] S. Xu *et al.*, "Interdiff: Generating 3d human-object interactions with physics-informed diffusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[63] L.-H. Chen *et al.*, "Humanmac: Masked motion completion for human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[64] J. Jeong *et al.*, "Multi-agent long-term 3d human pose forecasting via interaction-aware trajectory conditioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[65] S. Park *et al.*, "Nodeik: Solving inverse kinematics with neural ordinary differential equations for path planning," in *2022 22nd International Conference on Control, Automation and Systems (ICCAS)*. IEEE, 2022.

[66] D. Agrawal *et al.*, "Pose and skeleton-aware neural ik for pose and motion editing," in *SIGGRAPH Asia 2023 Conference Papers*, 2023.

[67] D. Holden *et al.*, "Learning motion manifolds with convolutional autoencoders," in *SIGGRAPH Asia 2015 Technical Briefs*, 2015, pp. 1–4.

[68] W. Hong *et al.*, "Generative adversarial exploration for reinforcement learning," in *Proceedings of the First International Conference on Distributed Artificial Intelligence*, 2019.

[69] I. Goodfellow *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[70] B. Zhang *et al.*, "Real-time action recognition with enhanced motion vector cnns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[71] J. Zhang *et al.*, "Video watermark technique in motion vector," in *Proceedings XIV Brazilian Symposium on Computer Graphics and Image Processing*. IEEE, 2001.

[72] H. Wang *et al.*, "Spatio-temporal manifold learning for human motions via long-horizon modeling," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 1, pp. 216–227, 2019.

[73] E. Corona *et al.*, "Context-aware human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[74] Z. Cao *et al.*, "Long-term human motion prediction with scene context," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer International Publishing, 2020.

[75] Z. Liu *et al.*, "Towards natural and accurate future motion prediction of humans and animals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[76] Q. Cui *et al.*, "Learning dynamic relationships for 3d human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[77] A. Hernandez *et al.*, "Human motion prediction via spatio-temporal inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[78] E. Aksan *et al.*, "A spatio-temporal transformer for 3d human motion prediction," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021.

[79] R. Sekar *et al.*, "Planning to explore via self-supervised world models," in *International Conference on Machine Learning*. PMLR, 2020.

[80] S. Lin *et al.*, "Common diffusion noise schedules and sample steps are flawed," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[81] A. Abuduweili *et al.*, "Enhancing sample generation of diffusion models using noise level correction," *arXiv preprint arXiv:2412.05488*, 2024.

[82] W. Zhu *et al.*, "Social motion prediction with cognitive hierarchies," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 77 675–77 690.

[83] R. T. Q. Chen *et al.*, "Neural ordinary differential equations," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[84] S. Andrist *et al.*, "Look like me: Matching robot personality via gaze to increase motivation," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 3603–3612.

[85] D. Helbing *et al.*, "Self-organizing pedestrian movement," *Environment and Planning B: Planning and Design*, vol. 28, no. 3, pp. 361–383, 2001.

[86] B. J. Kim *et al.*, "The disappearance of timestep embedding in modern time-dependent neural networks," *arXiv preprint arXiv:2405.14126*, 2024.

[87] F. Liu *et al.*, "Timestep embedding tells: It's time to cache for video diffusion model," *arXiv preprint arXiv:2411.19108*, 2024.

[88] S. Rouchier *et al.*, "Calibration of simplified building energy models for parameter estimation and forecasting: Stochastic versus deterministic modelling," *Building and Environment*, vol. 134, pp. 181–190, 2018.

[89] Tall *et al.*, "Flow-box theorem and beyond," 2011.

[90] V. D. Oord *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[91] Aron *et al.*, "Weak-star continuous analytic functions," *Canadian Journal of Mathematics*, vol. 47, no. 4, pp. 673–683, 1995.

[92] Chen *et al.*, "Neural ordinary differential equations," *Advances in neural information processing systems*, vol. 31, 2018.

[93] Ashburner *et al.*, "Diffeomorphic registration using geodesic shooting and gauss–newton optimisation," *Neuroimage*, vol. 55, no. 3, pp. 954–967, 2011.

[94] Kovachki *et al.*, "Neural operator: Learning maps between function spaces with applications to pdes," *Journal of Machine Learning Research*, vol. 24, no. 89, pp. 1–97, 2023.

A Latent Diffusion

우리는 [11]에서 LDM 의 구조를 사용하여 입력 (input), 분석 (analysis), 출력 (output)의 세 부분으로 구성하였다.

A.1 Input

Figure 2에 나타난 바와 같이, 우리는 두 가지 인코더 입력을 사용한다. 첫 번째 입력은 동작간의 차이 $P_{t+1} - P_t$를 인코더 입력으로 사용한다.

A.2 Analysis

분석 단계에서는 확산 과정을 통해 잠재 공간을 예측한다. 이 단계에서는 확산 단계 수를 제어하기 위해 노이즈 스케일(noise scale)을 약 10^8로 설정한다. 정확도와 해석 가능성 (interpretability)을 더욱 향상시키기 위해, U-Net 아키텍처 내에 동작 다양체 (motion manifold) 지원을 통합하였다.

우리 모델은 Transformer 기반의 동작 다양체 학습을 지원하도록 확장된 U-Net 구조를 활용한다. 구체적으로, 인코더와 Transformer, 그리고 스킵 연결 (skip connection)을 함께 사용하는 포즈 인코더 (pose encoder)를 통해 다양체 정보를 임베딩한다.

우리는 다음과 같은 변환 파이프라인 (transformation pipeline)을 제안한다:

$$\text{Current Pose} + \text{Noise} + \text{Manifold} \longrightarrow \text{General Manifold}, \quad (8)$$

변환 과정은 Transformer 모듈의 어텐션 메커니즘 (attention mechanism)에 의해 조정된다. 일반 다양체 (General Manifold)는 다음과 같은 과정을 통해 노이즈에 견고한 동작 예측을 향상시킨다. 포즈 인코더는 입력 포즈로부터 동작 벡터를 생성하도록 설계되어 있으며, 이는 노이즈와 의미 있는 동작 패턴을 분리하는 데 기여한다.

이를 통해 U-Net 모델은 디노이징 맥락 내에서 보다 효과적으로 작동할 수 있게 되며, 동작 예측 작업에서의 재구성 및 생성 성능 모두를 향상시킨다.

A.3 Output

출력 모듈은 두 개의 디코더로 구성된다. 첫 번째 디코더는 R^{35} 공간에서 동작 포즈(motion poses)를 생성하며, $\mathbf{D}_{t+1}, \ldots, \mathbf{D}_{t+10}$ 순으로 프레임을 생성하고, 이들은 월드 모델 처리 (world-model processing)를 위해 동작 버퍼에 저장된다. 여기서 $\mathbf{D}_t$는 다음을 나타내는 표기이다: *diffusionPredict*$(\mathbf{s}_t)$. 두 번째 디코더는 GAN과 유사한 방식으로 합성된 노이즈 (synthesized noise)를 출력하며, 원래의 노이즈를 월드 모델과 정렬되는 분포로 변환한다. 이를 통해 월드 모델은 원래 노이즈가 연속적인 경계를 정의하는 노이즈 공간 내에서 동작할 수 있게 된다.

B World Model

여기서는 월드 모델의 구조에 대해 설명하고자 한다. 월드 모델은 강화학습을 통한 문제 해결의 핵심 구성 요소로 작용한다. 우리는 훈련된 Neural ODE와 LDM을 활용하여 고동적 (high-dynamic) 동작 예측 과제를 해결하기 위한 강화학습 환경을 구성하였다 [79]. 이 환경은 입력 센서 프레임을 효율적으로 압축함으로써, 복잡한 동작 역학 (motion dynamics)에 대한 보다 효과적이고 확장 가능한 학습을 가능하게 한다.

B.1 Input

LDM의 입력으로 사용함으로써, 생성 과정을 기하학적으로 정렬된 공간 내에서 유도하는 방식이다. 직접적인 영향은 Neural ODE을 월드 모델 내에 직접 통합함으로써 발생하며, 이를 통해 연속 시간 역학 (continuous-time dynamics)을 보다 자연스럽게 모델링할 수 있게 한다.

LDM은 환경 상태 (environment states)를 두 가지 구성 요소로 정의한다: 예측된 포즈(predicted poses)와 합성된 (또는 가짜) 노이즈 (synthesized or fake noise).

- **상태(states):**

  - **예측된 포즈(Predicted poses):** LDM은 10개의 포즈 시퀀스를 예측하며, 이는 강화학습 에이전트의 학습에 사용된다.
  - **가짜 노이즈(Fake noise):** 이 구성 요소는 생성적 적대 신경망 (GAN)과 유사하게 동작하며, 월드 모델 내에서의 변분 데이터 Sampling을 유도하기 위해 가짜 노이즈를 생성한다.

예측된 포즈와 합성된 노이즈를 함께 활용함으로써, 월드 모델은 샘플 효율성 (sample efficiency)을 향상시키고, 보다 자연스러운 강화학습을 가능하게 한다.

B.2 Learning

월드 모델의 학습 과정은 Sampling, 동작 예측, 그리고 보상 기반 최적화 (reward-based optimization)를 중심으로 이루어진다. 모델은 포즈 공간에서 향후 프레임을 자연스럽게 예측하는 방법을 학습하며, 보상 기반 손실 함수 (reward-guided loss function)를 통해 예측 정확도를 향상시킨다 [66]. 이 학습 단계는 강화학습에 있어 핵심적이며, 월드 모델이 다중 sampled 궤적 (sampled trajectories)을 통해 동작 예측을 정교화하고, 현실 세계의 동역학 (real-world dynamics)에 부합하는 보다 사실적인 미래 상태를 생성할 수 있도록 한다.이러한 구조화된 학습 메커니즘을 활용함으로써, 월드 모델은 동작 패턴을 효과적으로 시뮬레이션하고 예측할 수 있으며, 강화학습 환경 내에서의 성능을 더욱 향상시킬 수 있다. 우리는 Neural Ordinary Differential Equations (Neural ODEs)를 활용하여, POMDP의 전 상태 공간에 걸친 동역학을 모델링할 수 있는 연속 연산자 (continuous operator)를 학습한다 [94]. 다음과 같은 사상을 정의한다:

$$V : X \to S$$

여기서 X는 시스템의 잠재 공간 또는 입력 공간 (input space)과 같은 컴팩트 한 공간이며, S는 전체 환경 또는 동작 다양체를 나타내는 무한 차원 상태 공간 (infinite-dimensional state space)일 수 있다.

이 정의에서 V는 잠재 공간작 예측 작업에서의 재구성 및 생성 성능 모두를 향상시킨다.

B.3 Output

출력 모듈은 두 개의 디코더로 구성된다. 첫 번째 디코더는 R^{35} 공간에서 동작 포즈(motion poses)를 생성하며, $\mathbf{D}_{t+1}, \ldots, \mathbf{D}_{t+10}$ 순으로 프레임을 생성하고, 이들은 월드 모델 처리 (world-model processing)를 위해 동작 버퍼에 저장된다. 여기서 $\mathbf{D}_t$는 다음을 나타내는 표기이다: *diffusionPredict*$(\mathbf{s}_t)$. 두 번째 디코더는 GAN과 유사한 방식으로 합성된 노이즈 (synthesized noise)를 출력하며, 원래의 노이즈를 월드 모델과 정렬되는 분포로 변환한다. 이를 통해 월드 모델은 원래 노이즈가 연속적인 경계를 정의하는 노이즈 공간 내에서 동작할 수 있게 된다.

C World Model

여기서는 월드 모델의 구조에 대해 설명하고자 한다. 월드 모델은 강화학습을 통한 문제 해결의 핵심 구성 요소로 작용한다. 우리는 훈련된 Neural ODE와 LDM을 활용하여 고동적 (high-dynamic) 동작 예측 과제를 해결하기 위한 강화학습 환경을 구성하였다 [79]. 이 환경은 입력 센서 프레임을 효율적으로 압축함으로써, 복잡한 동작 역학 (motion dynamics)에 대한 보다 효과적이고 확장 가능한 학습을 가능하게 한다.

C.1 Input

LDM의 입력으로 사용함으로써, 생성 과정을 기하학적으로 정렬된 공간 내에서 유도하는 방식이다. 직접적인 영향은 Neural ODE을 월드 모델 내에 직접 통합함으로써 발생하며, 이를 통해 연속 시간 역학 (continuous-time dynamics)을 보다 자연스럽게 모델링할 수 있게 한다. LDM은 환경 상태 (environment states)를 두 가지 구성 요소로 정의한다: 예측된 포즈(predicted poses)와 합성된 (또는 가짜) 노이즈 (synthesized or fake noise).

- **상태(states):**
  - **예측된 포즈(Predicted poses):** LDM은 10개의 포즈 시퀀스를 예측하며, 이는 강화학습 에이전트의 학습에 사용된다.
  - **가짜 노이즈(Fake noise):** 이 구성 요소는 생성적 적대 신경망 (GAN)과 유사하게 동작하며, 월드 모델 내에서의 변분 데이터 Sampling을 유도하기 위해 가짜 노이즈를 생성한다.

예측된 포즈와 합성된 노이즈를 함께 활용함으로써, 월드 모델은 샘플 효율성 (sample efficiency)을 향상시키고, 보다 자연스러운 강화학습을 가능하게 한다.

C.2 Learning

월드 모델의 학습 과정은 Sampling, 동작 예측, 그리고 보상 기반 최적화 (reward-based optimization)를 중심으로 이루어진다. 모델은 포즈 공간에서 향후 프레임을 자연스럽게 예측하는 방법을 학습하며, 보상 기반 손실 함수 (reward-guided loss function)를 통해 예측 정확도를 향상시킨다 [66]. 이 학습 단계는 강화학습에 있어 핵심적이며, 월드 모델이 다중 sampled 궤적 (sampled trajectories)을 통해 동작 예측을 정교화하고, 현실 세계의 동역학 (real-world dynamics)에 부합하는 보다 사실적인 미래 상태를 생성할 수 있도록 한다.이러한 구조화된 학습 메커니즘을 활용함으로써, 월드 모델은 동작 패턴을 효과적으로 시뮬레이션하

고 예측할 수 있으며, 강화학습 환경 내에서의 성능을 더욱 향상시킬 수 있다. 우리는 Neural Ordinary Differential Equations (Neural ODEs)를 활용하여, POMDP의 전 상태 공간에 걸친 동역학을 모델링할 수 있는 연속 연산자 (continuous operator)를 학습한다 [94]. 다음과 같은 사상을 정의한다:

$$V : X \to S$$

여기서 X는 시스템의 잠재 공간 또는 입력 공간 (input space)과 같은 컴팩트 한 공간이며, S는 전체 환경 또는 동작 다양체를 나타내는 무한 차원 상태 공간 (infinite-dimensional state space)일 수 있다. 이 정의에서 V는 잠재 공간 내의 동역학에서 연속적이고 유계인 (bounded) 연산자로 작동하며, 신경망 (neural networks)을 통해 근사하기에 적합한 특성을 갖는다.

Proposition. X를 컴팩트 위상 공간 (compact topological space)이라 하고, M이 약한 위상 (weak topology)을 갖는 공간이라고 하자 (예: 일반적인 동작 다양체). 이때 연산자 $V : X \to S$가 연속이라면, 이는 S에서 약-* 컴팩트(weak-*)한 상(이미지)을 유도한다 [91].

즉, 연산자 V는 약-* 수렴(weak-* convergence)의 관점에서 학습될 수 있으며, 이는 Neural ODE 아키텍처 내에서 일반화 가능한 신경 연산자 (neural operator)를 학습하는 데 충분하다.

Proof 이는 바나흐–알라우글루 정리 (Banach–Alaoglu Theorem)에 따른 결과로, 노름 공간의 쌍대 공간 (dual space)에서 닫힌 단위 공 (ball)은 약-* 위상에서 컴팩트를 보장한다. 따라서 X가 컴팩트하고 V가 연속일 경우, $V(X)$는 S의 약-* 위상에서 여전히 컴팩트한 상 (image)을 유지한다.

결과적으로, 연산자 V를 신경 연산자(neural operator)로 학습하는 것은 약 수렴(weak convergence) 가정 하에서도 가능하며 안정적이다.

Interpretation. 이 프레임워크는 입력 공간의 컴팩트성 과 상 (image)의 약-* 컴팩트성 (weak-* compactness)에 기반하여, 고차원 (high-dimensional) 또는 부분 관측 시스템 (partially observed systems)에서도 세계 동역학 (world dynamics)을 일반화 가능한 방식으로 모델링할 수 있도록 한다.

이를 통해 Neural ODE는 학습 가능한 유계 연산자 (bounded operator)를 통해 일반화된 동작 (motion)과 동역학 (dynamics)을 표현할 수 있게 된다. 이러한 관점에서, 본 연구는 정확한 장기 동작 예측을 지원하는 컴팩트하면서도 일반적인 다양체 (manifold)를 구성한다.

Internal Structure of the World Model. Figure 4에 나타난 바와 같이, 월드 모델 내부에서는 최종적인 월드 포즈 데이터를 생성하기 위해 보조 신경망 (auxiliary neural networks)이 사용된다.

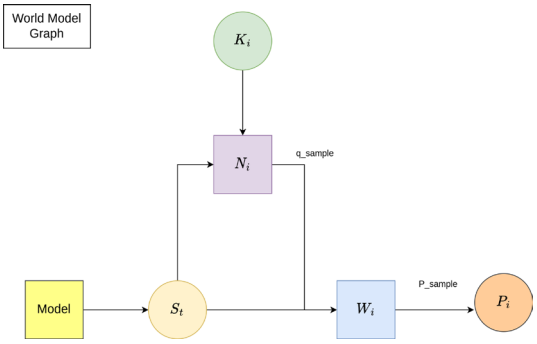


Figure 4: World Optimization Controller - A More Detailed Version.

이러한 신경망들은 다양한 동작 패턴에 대한 전문가 지식 (expert knowledge)을 통합하고 있으며, 이를 통해 보다 정확하고 견고한 동작 예측이 가능하다 [62].

특징 맵 (feature map)의 표현력을 향상시키기 위해, VAE의 구조를 도입하였다. VAE는 월드 모델 내에서 특징 맵을 구성하고, 이를 월드 데이터와 결합하여 다음과 같은 동작 프레임 시퀀스 $\mathbf{P}_t, \mathbf{P}_{t+1}, \cdots, \mathbf{P}_{t+10}$를 예측한다.

이 예측 결과는 확산 프레임워크 내의 p-sampling 정책을 통해 생성된다.

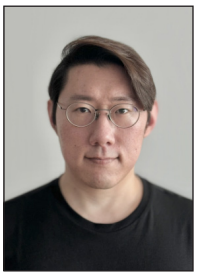
C.3 Output

이후 월드 모델은 손실 함수에 의해 유도되는 p-sampling 과정을 위한 정책을 출력한다. 환경에 의해 구성된 상태를 입력으로 받으면, 월드 모델은 최종 프레임들을 예측하게 된다.

〈저자소개〉

나예현

- 2020-2024: 한양대학교 컴퓨터소프트웨어학과 학사
- 2024-현재:한양대학교 컴퓨터소프트웨어학과 석사
- https://orcid.org/0009-0007-8089-8656

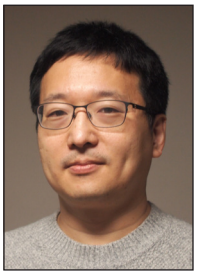


Richard Young Park

- 2024-현재 한양대학교 컴퓨터소프트웨어학과 박사
- https://orcid.org/0009-0001-2681-8244

서상영

- 2019-2025 강원대학교 컴퓨터정보통신공학전공 학사
- 2025-현재 한양대학교 컴퓨터소프트웨어학과 석사
- https://orcid.org/0009-0005-0269-7651



권태수

- 1996-2000:서울대학교 전기 컴퓨터 공학부 학사
- 2000-2002:서울대학교 전기컴퓨터공학부 석사
- 2002-2007:한국과학기술원 전산학전공 박사
- https://orcid.org/0000-0002-9253-2156