한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 31, No. 3, P. 109~117

전문가 혼합 구조를 활용한 텍스트 기반 3D 모션 생성 성능 향상 연구

선재영[1,O] 홍성은[2] 김경민[2] 우승우[4] 황효석[3,*] 강형엽[2,*]

경희대학교 인공지능학과[1], 고려대학교 컴퓨터학과[2], 경희대학교 소프트웨어융합학과[3],
국립과학수사연구원[4]

cogongnam@khu.ac.kr, {seong_eun, kgm031189}@korea.ac.kr, hyoseok@khu.ac.kr, siamiz_hkang@korea.ac.kr, seungwoo82@korea.kr

Text-based 3D Motion Generation with a Mixture of Experts for Enhancing Performance

JaeYoung Seon[1,O] Seong-Eun Hong[2] KyeongMin Kim[2] SeungWoo Woo[4]
Hyoseok Hwang[3,*] HyeongYeop Kang[2,*]

Department of Artificial Intelligence, Kyung Hee University[1], Department of Computer Science and Engineering, Korea University[2], Department of Software Convergence, Kyung Hee University[3], National Forensic Service[4]

요 약

자연어로부터 의미에 부합하는 모션 시퀀스를 생성하는 텍스트 기반 3D 인간 모션 생성은 컴퓨터 그래픽스와 인공지능 분야에서 중요한 연구 주제로 부상하고 있다. 그러나 기존 텍스트 기반 모션 생성 모델들은 전체 모션을 통합적으로 처리하여 신체 부위별 동작 특성을 충분히 반영하지 못하는 한계가 있었다. 이를 해결하기 위해 본 연구에서는 Mixture of Experts (MoE) 구조를 활용하여 신체 부위별로 특화된 전문가 모델을 학습하는 방법을 제안한다. 제안된 모델은 Transformer가 출력한 모션 특징 벡터를 바탕으로 Gating Network를 통해 각 신체 부위에 특화된 전문가를 선택하고, 이를 통해 정교한 부분 동작을 생성한다. 실험을 통해 MoE 기반 구조가 기존의 모션 생성 방식과 비교하여 모델의 표현력과 효율성을 동시에 향상시키며, 다양한 텍스트 묘사에 대해 보다 자연스럽고 일관된 모션 생성이 가능함을 검증하였고, 모션 품질과 텍스트-모션 일치도 측면에서 기존 방법 대비 성능이 향상되었음을 확인하였다.

Abstract

Text-based 3D human motion generation, which aims to produce motion sequences that align with the semantics of natural language, has emerged as a significant research topic in the fields of computer graphics and artificial intelligence. But existing text-to-motion models often treat the entire motion sequence holistically, limiting their ability to capture the distinct characteristics of individual body parts. To address this limitation, we propose a method that employs a Mixture of Experts (MoE) architecture, where each expert specializes in motions of specific body parts. The proposed model uses a Gating Network to selectively activate relevant experts based on motion features extracted by a Transformer, enabling the generation of fine-grained, part-specific motions. Through experiments, we demonstrate that the MoE-based architecture improves both the expressiveness and efficiency of the model compared to existing approaches, enabling more natural and coherent motion synthesis for a wide range of textual descriptions. Furthermore, our method shows improved performance in terms of motion quality and text-motion alignment.

*corresponding author: Hyoseok Hwang / Kyung Hee University (hyoseok@khu.ac.kr), HyeongYeop Kang / Korea University (siamiz_hkang@korea.ac.kr)

Received : 2025.06.13./ Review completed : 1st 2025.06.30. 2nd 2025.07.10./ Accepted : 2025.07.14.
DOI : 10.15701/kcgs.2025.31.3.109
ISSN : 1975-7883(Print)/2383-529X(Online)

키워드: 텍스트 기반 모션 생성, Mixture of Experts, 3D 모션 생성, Transformer, VQ-VAE
Keywords: Text based motion generation, Mixture of Experts, 3D motion generation, Transformer, VQ-VAE

1. 서론

자연어 설명으로부터 자연스러운 3D 인간 모션을 생성하는 기술은 컴퓨터 그래픽스, 게임, 가상현실 등 다양한 분야에서 중요한 응용 가능성을 가지고 있어 최근 활발한 연구가 진행되고 있으며, 특히 메타버스 기술의 발전과 함께 자연스럽고 표현력 있는 인간 모션 생성의 중요성이 더욱 부각되고 있다. 이러한 텍스트 기반 모션 생성 기술은 자연어의 복잡한 의미 구조를 이해하고, 이를 시공간적으로 일관성 있는 3D 모션으로 변환해야 하는 도전적인 문제이다.

자연어는 본질적으로 모호성과 복잡성을 내포하고 있으며, 동일한 텍스트 설명도 다양하게 해석될 수 있다. 예를 들어, “천천히 걷기”와 같은 간단한 설명도 개인의 신체적 특성, 감정 상태, 주변 환경 등에 따라 매우 다른 모션으로 표현될 수 있다. 또한, “손을 흔들며 뛰기”와 같은 복합적인 동작은 상체와 하체가 서로 다른 움직임을 동시에 수행해야 하므로, 각 신체 부위의 특성을 정확히 이해하고 조합하는 것이 필수적이다.

기존의 텍스트 기반 모션 생성 모델들은 주로 전체 모션 시퀀스를 하나의 통합된 표현으로 처리하는 방식을 채택하고 있다. 그러나 인간의 모션은 본질적으로 상체, 하체, 팔, 다리 등 각 신체 부위가 서로 다른 특성과 움직임 패턴을 가지고 있으며, 서로 다른 모션 특성을 요구한다. 따라서 이러한 신체 부위별 특성을 고려하지 않고 전체 모션을 일률적으로 처리하는 것은 생성되는 모션의 품질과 다양성을 제한할 수 있다.

신체 부위별 특화된 모션 생성을 위해 일부 연구들은 신체 부위별로 독립적인 생성 모듈을 구성하는 방법을 제안하였다. 하지만 이러한 접근법은 모델 크기의 급격한 증가를 초래하며, 이에 따라 더 많은 학습 데이터와 계산 자원이 필요하다는 문제가 존재한다.

이러한 문제를 해결하기 위해 본 논문에서는 Mixture of Experts (MoE) 구조를 텍스트 기반 모션 생성에 적용하는 새로운 접근법을 제안한다. MoE 는 여러 개의 전문가 모델을 두고 입력에 따라 적절한 전문가를 활성화하여 활용하는 기법으로, 복잡한 문제를 여러 하위 문제로 분할하여 각각을 전문적으로 처리할 수 있다는 장점이 있다. 본 연구에서는 이를 활용하여 각 전문가가 특정 신체 부위나 모션 패턴에 특화되도록 학습함으로써, 보다 정교하고 자연스러운 모션 생성을 목표로 한다.

본 연구는 MoE 구조를 통해 텍스트의 의미 단위와 모션 구성 요소 간의 정렬을 보다 명시적이고 구조화된 방식으로 달성하고자 한다. 텍스트 입력의 각 토큰이 개별 전문가를 통해 처리됨으로써, 텍스트와 모션 사이의 의미적 정합성이 강화된다. 또한 희소 활성화 (sparse activation)를 기반으로 한 전문가 선택 메커니즘을 도입하여, 모델 전체의 파라미터 수는 확장 가능하되, 실제 연산 비용은 유지함으로써 계산 효율성과 확장성을 동시에 확보할 수 있다.

본 연구의 핵심 기여는 다음과 같이 요약할 수 있다.

- 텍스트 기반 모션 생성 과정에서 텍스트의 의미 구조와 모션 구성 요소 간의 정렬을 구체화할 수 있는 MoE 기반 구조를 제안한다.
- 모든 전문가가 입력을 처리한 후 선택적으로 특정 전문가의 출력을 활용하는 희소 활성화 구조를 설계함으로써, 계산 효율성을 확보하고 모델의 확장 가능성을 높인다.
- 제안하는 MoE 기반 구조가 기존의 통합형 또는 부위별 독립 구조에 비해 신체 부위 간 표현 분리 및 의미 정렬에 더 효과적임을 다양한 실험을 통해 입증한다.

2. 관련 연구

2.1 텍스트 기반 모션 생성

텍스트 기반 모션 생성은 자연어로 기술된 문장을 입력으로 받아 의미에 부합하는 모션을 생성하는 기술로, 최근 다양한 딥러닝 모델들이 이 분야에 적용되고 있다.

초기 연구들은 Variational AutoEncoder (VAE) [1]를 통해 연속 잠재 공간에서 텍스트 의미를 반영한 샘플을 생성한 뒤 이를 복원하는 구조를 갖는다. TEMOS [2]는 텍스트와 모션을 잠재 공간에 공동 임베딩하고, cross-attention 기반 디코더를 통해 고품질의 모션 시퀀스를 생성하였다. 이외에도 T2M [3]은 VAE의 잠재 변수 분포를 활용하여 동일한 문장에서 다양한 스타일의 모션을 생성할 수 있음을 보였으며, 표현 다양성 측면에서 큰 진전을 이루었다. 그러나 이러한 방식은 연속 공간 표현의 한계와 장기 시퀀스에서의 정합성 부족 문제를 일부 내포하고 있다.

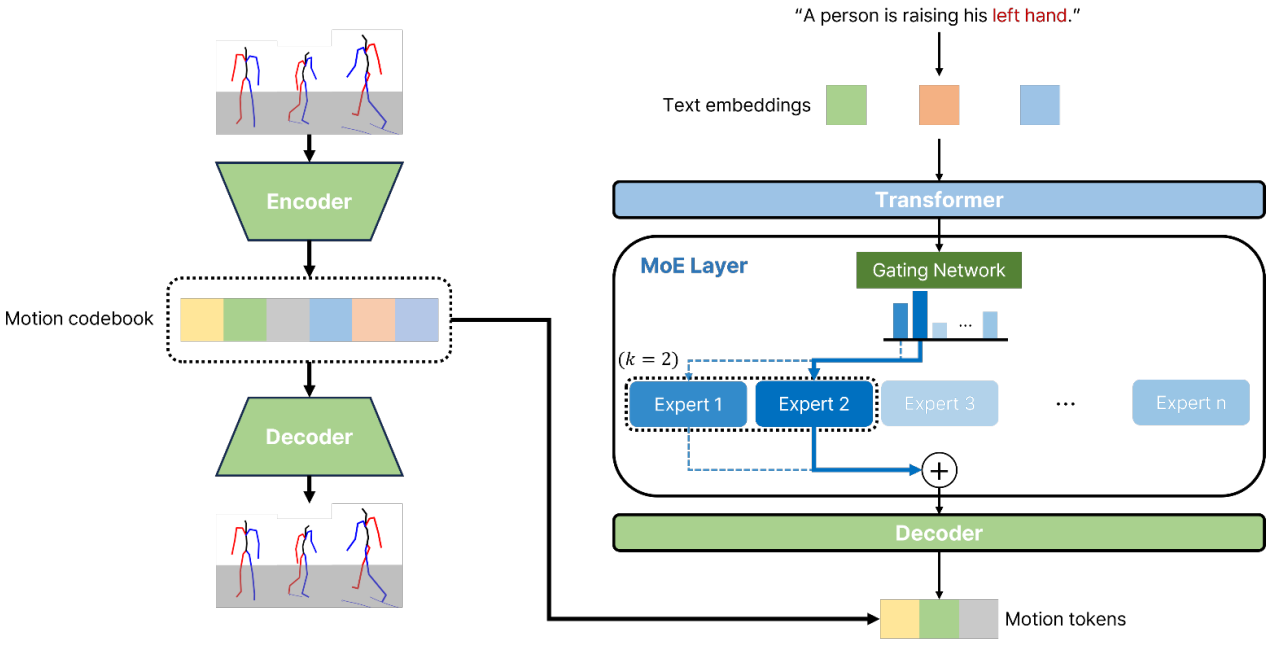

Figure 1: Overall architecture of proposed method. (a) Motion VQ-VAE encodes motion into discrete tokens; (b) MoE-based transformer generates motion by semantically decomposing text via expert routing.

이후 제안된 diffusion 모델 [4] 기반 접근은 확률적 노이즈 제거 과정을 통해 점진적으로 모션을 복원하는 방식이다. MDM [5]은 텍스트 조건부 모션 생성을 위해 diffusion 과정을 Transformer [6]와 결합하여 구현하였으며, 시간적 일관성과 정밀한 움직임 표현에서 우수한 성능을 입증하였다. 또한 MLD [7]는 잠재 공간에서의 diffusion을 수행함으로써 더 효율적인 샘플링과 고품질 모션 생성을 동시에 달성하였다. 이러한 계열의 방법들은 복잡하고 다양한 모션 표현에서 정량적, 정성적으로 뛰어난 결과를 보였다.

최근 연구들에서는 Vector-Quantised VAE (VQ-VAE) [8]를 활용하여 모션 시퀀스를 이산 토큰 시퀀스로 변환한 후, 이를 Transformer로 모델링하는 방식을 채택한다. 이산 표현을 통해 장기 구조 학습이 용이하며, 토큰 단위의 예측을 통해 보다 안정적인 생성 성능을 달성하였다. MoMask [9]는 모션 마스킹 및 복원을 기반으로 텍스트 조건 하에서 정교한 모션 토큰 생성을 달성하였으며, 이는 학습 안정성과 샘플 다양성 측면에서 강점을 보였다. 해당 계열의 모델들은 VQ-VAE의 강점인 표현 정수성과 Transformer의 순차적 구조 모델링 능력을 결합하여 자연스러운 모션 시퀀스를 효과적으로 생성할 수 있음을 보여준다.

2.2 신체 부위별 모션 모델링

인간의 동작은 본질적으로 다양한 신체 부위의 조합으로 구성되며, 각 부위는 고유한 움직임 특성을 지닌다. 이러한 특성을 반영하기 위해, 일부 연구에서는 상체와 하체를 별도로 분리하여 모션을 생성하는 방식이 제안되었다 [10]. 이 구조는 각 부위의 표현력을 높이고 복잡한 동작의 조합을 효과적으로 처리할 수 있는 장점을 갖는다.

또 다른 접근으로는 신체 부위를 5개 영역 (양팔, 양 다리, 몸통)으로 나누고, 지역-전역 어텐션을 통해 부위별 특징을 계층적으로 추출하는 방법이 있으며 [11], 해당 방식은 텍스트의 세부 내용에 따라 다양한 부위의 표현을 정교하게 반영하는 데 효과적이다.

더 나아가, 일부 연구에서는 각 신체 부위에 특화된 모델을 별도로 설계하고, 문장에서 추출한 의미적 단위를 각 모델에 할당하여 모션을 생성하는 구조가 제안되었다 [12]. 이 방식은 문장의 구조를 기반으로 상체와 하체 등 신체 부위별 동작을 명시적으로 분리하여 모델링함으로써, 복잡한 문장에 대한 세분화된 동작 생성을 가능하게 한다.

이처럼 부위별 모션 모델링은 동작의 세밀한 제어와 다양성 확보 측면에서 유용한 방향으로 인식되고 있다.

2.3 Mixture of Experts

MoE [13]는 입력에 따라 서로 다른 전문가 모델을 선택적으로 활성화하여 복잡한 문제를 분할 정복 방식으로 처리하는 구조이다. 초기에는 간단한 회귀 문제를 해결하기 위한 방식으로 제안되었으며, 최근에는 대규모 모델을 효율적으로 운용하기 위한 핵심 기술로 재조명되고 있다.

대규모 언어 모델에서는 MoE 구조가 이미 중요한 확장 전략으로 자리잡고 있으며, 각 입력 토큰에 대해 최적의 전문가를 선택하여 처리하는 방식이 성공적으로 활용되고 있다. 예를 들어, Switch Transformer 와 같은 모델은 수 조 개의 파라미터를 효율적으로 운용하면서도 높은 성능을 유지할 수 있었으며, 이 과정에서 MoE 는 연산 자원의 효율성과 표현력의 균형을 동시에 달성하는 데 핵심적인 역할을 하였다 [14].

3. 방법론

3.1 전체 구조 개요

본 논문에서 제안하는 모델은 T2M-GPT [15]의 아키텍처를 베이스라인으로 삼되, Transformer 내부에 MoE 구조를 통합하여 확장한 형태이다.

기본 프레임워크는 두 단계로 구성되며, 먼저 VQ-VAE 를 통해 연속적인 3D 모션 데이터를 이산적인 토큰 시퀀스로 양자화하고, 이를 복원하는 디코더를 학습한다. 이후 텍스트 설명으로부터 모션 시퀀스를 예측하는 Transformer 기반 생성기를 학습하는 방식이다.

기존 모델에서 Transformer 는 텍스트를 조건으로 전체 모션을 단일 특징 벡터 흐름으로 생성하였다. 본 연구에서는 해당 구조를 확장하여 Transformer 내부에 일부 MoE 레이어를 삽입하고, 각 시점의 모션 특징 벡터에 대해 전문가 선택 기반의 부분 처리 방식을 도입하였다. 이를 통해 모션의 전체적인 일관성은 유지하면서도, 세밀한 동작 표현을 각 전문가가 나누어 학습할 수 있는 구조를 갖는다.

전체 모델은 다음과 같이 구성된다: (1) 텍스트 인코딩은 Transformer 인코더로 구성된 CLIP [16] 기반 특징 벡터 추출기를 활용하고, (2) 모션 디코딩은 기존의 VQ-VAE 구조를 유지하며, (3) 모션 토큰 시퀀스를 생성하는 Transformer 내부에 MoE 레이어를 삽입하여, 입력 특징 벡터의 의미에 따라 전문가 조합을 동적으로 선택하고 활성화하는 방식으로 처리한다.

3.2 MoE 레이어

MoE 레이어는 Transformer 블록의 피드포워드 단계 일부를 대체하며, 입력된 특징 벡터에 대해 다음과 같은 과정을 수행한다. 먼저 디코더 이전 단계에서 얻은 특징 벡터 $h \in \mathbb{R}^d$에 대해, 선형층 $W_g \in \mathbb{R}^{d \times n}$ 을 적용하여 n 개의 전문가에 대한 gating logit ($g_1, g_2, \dots, g_i, \dots, g_n$)을 얻는다 여기서 d 는 특징 벡터의 차원 수이고 n은 전문가의 개수이다. 이 값들은 아래 식과 같이 softmax 함수를 통해 확률 분포로 정규화된다.

$$G(h) = \text{softmax}(W_g h) = (g_1, g_2, \dots, g_n)$$

이후 g_i 값이 큰 순서대로 상위 k 개의 전문가 인덱스 $i_1, \dots, i_k$ 를 선택한다. 여기서 각 i_j 는 게이팅 분포에서 상위 확률을 갖는 전문가의 인덱스를 의미하며, 해당 인덱스에 대응되는 전문가 E_{i_j} 가 실제로 활성화되어 사용된다. 본 모델에서는 모든 전문가가 동일한 특징 벡터 h를 처리한 후, 선택된 전문가의 출력만을 가중합으로 조합하는 구조를 따른다.

$$h_{\text{out}} = \sum_{j=1}^{k} \tilde{g}_{i_j} E_{i_j}(h)$$

이때 $\tilde{g}_{i_j}$은 선택된 전문가들에 대한 정규화된 가중치이며, 다음과 같이 정의한다.

$$\tilde{g}_{i_j} = \frac{g_{i_j}}{\sum_{m=1}^{k} g_{i_m}}$$

이와 같은 구조는 의미 기반 feature routing 을 실현함과 동시에, 모델 계산량을 줄이고 확장성을 높이는 데 기여한다.

3.3 전문가의 역할 학습 및 신체 부위별 분화

본 모델의 MoE 구조는 명시적인 신체 부위 라벨 없이도, 전문가들이 입력 특징 벡터의 의미에 따라 자율적으로 서로 다른 역할을 암묵적으로 학습하도록 설계되었다. 이러한 구조는 전문가에게 명확한 기능을 사전에 할당하지 않고도 기능적 분화가 자연스럽게 발생하는 것을 목표로 한다.

전문가의 역할 학습은 텍스트 기반 모션 생성에서 중요한 요소이다. 자연어는 다양한 의미 단위를 포함하며, 이를 반영하는 모션 역시 시공간적으로 복잡한 구조를 가진다. 따라서 입력 특징 벡터의 의미적 구성에 따라 특정 전문가들이 반복적으로 선택되는 구조는, 모델이 다양한 의미 구성 요소를 모션의 부분적 표현으로 효과적으로 분산시키는 데 도움이 된다. 이러한 선택은 앞서 기술했듯이 softmax 기반 게이팅과 top-k 전문가 선택 메커니즘을 통해 이루어지며, 학습이 진행됨에 따라 전문가들이 서로 다른 유형의 모션 표현에 기여하게 된다.

이와 같이 전문가 간 역할이 명시적으로 정해지지 않았지만, 학습 과정에서 입력에 대한 전문가 선택이 점차 분화되는 현상은 모델이 의미 기반 feature routing 을 암묵적으로 수행하고 있음을 보여준다. 이로 인해 모델은 전체 모션 시퀀스 생성 시 텍스트의 의미 구조를 보다 정밀하고 세분화된 방식으로 반영할 수 있으며, 이는 결과적으로 더 자연스럽고 일관성 있는 모션 생성으로 이어진다.

4. 실험 및 결과

4.1 실험 세팅 및 평가 지표

본 논문의 실험은 텍스트 기반 3D 모션 생성 성능을 평가하기

Table 1: Comparison with other text to motion generation methods on HumanML3D test set.

Methods	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity →	MModality ↑
	Top-1	Top-2	Top-3				
Real Motion	0.511	0.703	0.797	0.002	2.974	9.503	-
T2M-GPT [15]	0.491	0.680	0.775	0.116	3.118	9.761	1.856
TM2T [19]	0.424	0.618	0.729	1.501	3.467	8.589	2.424
FineMoGen [20]	0.504	0.690	0.784	0.151	2.998	9.263	2.696
MDM [5]	0.320	0.498	0.611	0.544	5.566	**9.559**	**2.799**
ParCo [12]	**0.515**	**0.706**	**0.801**	0.109	**2.927**	9.576	1.382
Ours	0.476	0.669	0.767	**0.106**	3.151	9.570	1.829

위해 HumanML3D 데이터셋 [3]을 사용하였다. 해당 데이터셋은 14,616개의 문장-모션 쌍을 포함하며, 평균 13초 분량의 인간 동작 클립과 이에 상응하는 자연어 설명으로 구성되어 있다. 실험은 9:1:1 비율로 나눈 학습/검증/테스트 셋을 기준으로 기존 연구들과 동일하게 진행하였다.

정량 평가는 다음 다섯 가지 지표를 기반으로 수행하였다.

- Fréchet Inception Distance (FID): 생성된 모션이 실제 모션과 얼마나 유사한지를 평가하는 지표로, Recurrent Neural Ntwork (RNN) 기반 모션 특징 추출기에서 추출한 특징 벡터를 기반으로 계산된다. 생성된 모션의 품질을 측정하는 지표이며, 값이 낮을수록 생성된 모션의 분포가 실제 모션의 분포와 가깝다는 것을 의미한다.
- R-Precision: 입력 텍스트와 생성된 모션 간의 의미 일치도를 정량화한 지표로, 텍스트-모션 쌍의 정답 순위를 기준으로 측정한다. 값이 높을수록 의미적으로 정확한 모션 생성이 가능함을 의미한다.
- Multimodal Distance: 다양한 텍스트에 대해 생성된 모션들이 동일하거나 유사한 양상을 보일 경우 낮은 점수를 기록하며, 모션 다양성보다는 의미 일치에 더 민감한 특성을 가진다.
- Diversity: 동일한 텍스트 조건에서 생성된 여러 모션 샘플 간의 평균 거리를 측정하여, 생성된 모션의 다양성을 평가한다. 값이 높을수록 다양한 모션 표현이 가능함을 나타낸다.
- Multimodality: 하나의 텍스트 설명에 대해 다양한 모션 스타일을 생성할 수 있는 능력을 정량화한 지표로, 문장 조건에 대한 표현력과 생성의 비결정성을 반영한다.

4.2 세부 구현 정보

입력 모션 데이터는 HumanML3D 데이터셋 [3]의 표현 방식을 따르며, SMPL [17] 기반 22개 관절에 대한 다양한 신호로 구성된다.

하나의 모션 시퀀스 M는 총 F개의 프레임과 $J = 22$개의 관절 정보를 포함하며, 구체적으로 다음과 같은 요소들로 구성된다: 루트 관절의 y축 각속도 $\dot{r} \in \mathbb{R}^{F\times 1}$, xz 평면 상의 선속도 $v_{\text{root}} \in \mathbb{R}^{F\times 2}$, 루트 관절의 높이 $h \in \mathbb{R}^{F\times 1}$, 루트를 제외한 관절의 로컬 위치 $p \in \mathbb{R}^{F\times(J-1)\times 3}$, 6D 회전 표현 $r \in \mathbb{R}^{F\times(J-1)\times 6}$, 관절별 속도 $v \in \mathbb{R}^{F\times J\times 3}$, 그리고 양 발의 접지 여부를 나타내는 접지 신호 $c \in \mathbb{R}^{F\times 4}$ 등이다. [3]이와 같이 구성된 모션 표현은 학습의 안정성을 높이고, 네트워크가 다양한 모션 세부 특성을 학습하는 데 도움을 준다.

텍스트-모션 Transformer는 앞서 제시한 MoE 레이어를 포함한 아키텍처를 기반으로 하며, 학습은 teacher forcing 방식으로 수행된다. 최적화는 AdamW 옵티마이저 ($\beta_1 = 0.5, \beta_2 = 0.9$) [18]를 사용하였으며, 초기 학습률은 2e-4로 설정 후 처음 20만 step 동안 고정한 뒤 1e-5까지 선형적으로 감소시켰다.

모델 학습은 NVIDIA RTX A5000 환경에서 진행되었으며, 전체 학습 시간은 배치 사이즈 128 기준으로 약 54시간이 소요되었다. 이는 기존 모델 [15] 대비 약 30% 감소된 시간이다.

4.3 정량 평가 결과

Table 1에서 확인할 수 있듯이 본 모델은 기존 텍스트 기반 모션 생성 모델과 비교하여 HumanML3D 데이터셋 벤치마크에서 우수한 생성 성능을 기록하였다.

특히, 제안된 모델은 MoE 구조를 통해 표현력 있는 모션 생성을 가능하게 하였으며, 전체적인 FID 수치는 기존 방법들보다 현저히 낮은 값을 나타냈다. 이는 제안된 구조가 복잡한 자연어의 의미를 보다 세밀하게 반영하고, 각 시점의 모션 특징을 다양한 전문가를 통해 적절히 분산하여 학습할 수 있도록 설계된 점에서 기인한다.

또한, Diversity와 Multimodality 측면에서도 우수한 균형을 달성하였다. 이는 제안된 모델이 다양한 텍스트 입력에 대해 서로 다른 모션 스타일을 일관되게 생성할 수 있음을 시사한다.

이와 같은 결과는 전문가의 암묵적 분화를 유도하는 sparse routing 구조가, 텍스트 의미와 모션 표현 간의 정렬을 효과적으로 수행함을 정량적으로 뒷받침한다.

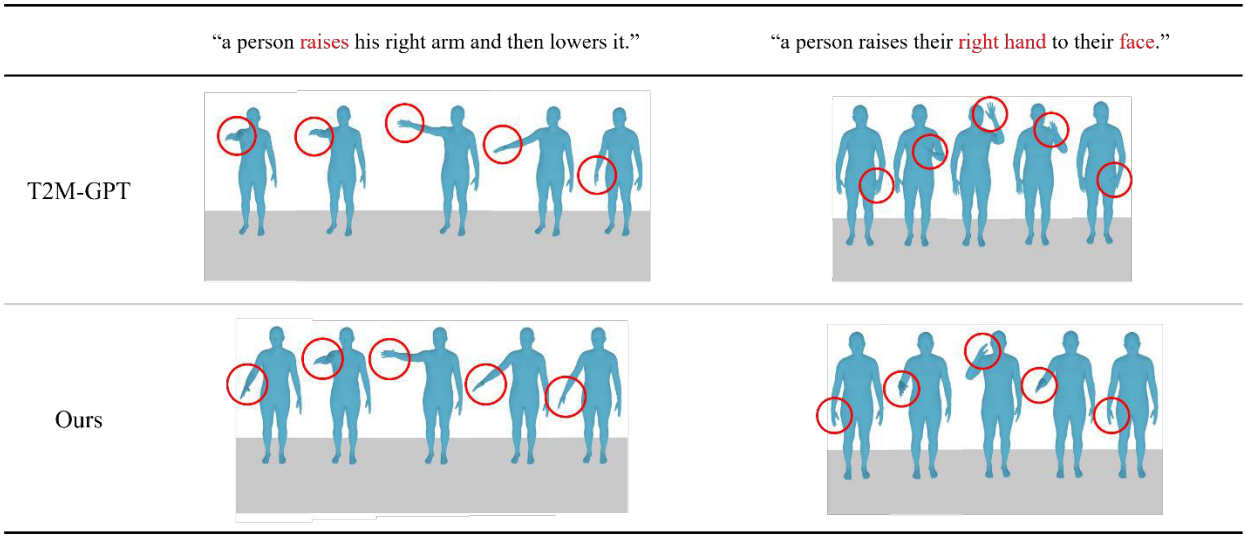

Figure 2: Qualitative evaluation results.

4.4 정성 평가 결과

본 절에서는 정성적 분석을 위해 다양한 자연어 설명을 입력으로 사용하여 생성된 모션 시퀀스를 시각화하고, 기존 모델들과 비교한다.

Figure 2에서 확인할 수 있듯이, 본 연구에서 제안한 MoE 기반 구조는 복합적인 동작에서도 각 신체 부위의 움직임이 보다 정교하게 조화를 이루는 양상을 보였다. 이러한 결과는 기존의 단일 모델 기반 구조와 비교했을 때, 제안된 MoE 기반 구조가 입력 텍스트로부터 유도되는 복합적 동작 구성 요소를 보다 효과적으로 분리 및 통합하여 시공간적으로 정교한 모션 표현을 가능하게 함을 시사한다. 이는 각 전문가가 선택적으로 활성화되어 모션 생성을 분담함으로써, 다양한 의미적 요소가 반영된 세분화된 표현을 가능하게 만들기 때문이다.

4.5 제거 연구

본 절에서는 제안한 MoE 구조의 설계 요소가 성능에 미치는 영향을 정량적으로 분석하기 위해 두 가지 관점에서 수행한 제거 연구 결과를 분석한다.

Table 2: Ablation results of different MoE methods.

MoE methods	Top-3 ↑	FID ↓	MM-Dist ↓
Ours (Dense)	**0.770**	0.129	3.168
Ours (Sparse)	0.767	**0.106**	**3.151**

- Dense MoE 와 sparse MoE 비교: Dense MoE 구조는 모든 전문가의 출력값 평균을 사용하는 방식이며, Sparse MoE 는 제안한 방식처럼 상위 k 개의 전문가만을 선택하여 출력을 혼합한다. Table 2 에서 확인할 수 있듯이 sparse MoE 구조는 dense MoE 에 비해 FID 및 Multimodal Distance 측면에서 더 우수한 수치를 기록하였다. 이는 전문가 선택의 희소성이 표현력을 유지하면서도 불필요한 계산을 줄이고, 각 전문가의 특화된 학습을 촉진한다는 점을 나타낸다.

Table 3: Ablation results of different number of experts setting.

Number of experts (n)	Top-3 ↑	FID ↓	MM-Dist ↓
$n = 2$	0.778	0.130	3.105
$n = 6$	0.767	**0.106**	3.151
$n = 10$	**0.787**	0.234	**3.067**

- 전문가 개수 (n)에 따른 성능 차이: MoE 구조에서 사용되는 전문가의 개수 (n)을 변화시키며 성능을 비교하였다. 본 실험에서는 $n = 2, 6, 10$ 의 세 가지 설정을 대상으로 동일한 top-k 조건 $(k = 2)$ 하에 비교를 수행하였다. Table 3에서 확인할 수 있듯이, 전문가 수가 적은 $n = 2$에서는 모션의 다양성과 표현력이 상대적으로 저하되었고, 전문가 수가 많은 $n = 10$에서는 생성된 모션의 품질이 크게 하락했다. 반면, $n = 6$ 의 설정에서는 균형 잡힌 성능을 보였다. 이는 전문가 수가 지나치게 적을 경우 모델의 표현력이 제한될 수 있고, 반대로 많은 전문가를 사용할 경우 gating network의 선택 효율이 저하되어 학습 안정성과 모션 품질에 부정적인 영향을 미칠 수 있음을 시사한다. 적절한 전문가 수는 구조적인 희소성을 유지하면서도, 각 전문가가 효과적으로 차별화된 표현을 학습하는 데 기여할 수 있다.

이러한 제거 연구는 제안한 MoE 구조의 구성 요소들이 실제 성능에 영향을 미치며, 올바른 설정을 통해 효율성과 정확도를 모두 달성할 수 있음을 보여준다.

5. 결론 및 향후 계획

본 논문에서는 MoE 구조를 통해 텍스트 기반 3D 모션 생성 모델의 성능을 향상하는 새로운 접근법을 제시하였다. 제안된 모델은 신체 부위별 전문가 모델이 텍스트 의미에 따라 협업적으로 동작을 생성하도록 구성되어, 기존의 단일 표현 기반 접근보다 부위별 정밀성과 전체적인 일관성 모두에서 우수한 결과를 도출하였다. 실험 결과에 따르면 제안 기법은 다양한 텍스트 입력에 대해 자연스럽고 텍스트 의미에 부합하는 모션을 생성함으로써, 기존 방식 대비 모션의 품질과 표현력 측면에서 향상된 성능을 보였다.

또한 본 연구는 MoE 기반 희소 전문가 선택 구조를 통해 계산 자원을 효율적으로 활용함과 동시에, 전문가 수의 확장을 통한 모델 용량 조절 및 새로운 모션 유형에 대한 적응 가능성을 확보하였다. 나아가 본 구조는 다른 생성 모델 (예: diffusion 기반 모델)의 표현 학습 단계에 통합될 수 있는 확장성을 가지며, 부분 제어 능력을 향상시키는 방향으로의 발전도 기대된다.

향후 연구로는 전문가 별로 더욱 명시적인 역할 부여 (예: 특정 관절 그룹 할당) 및 gating 단계에서의 부하 균형 기법 적용 등을 통해 전문가의 활용도를 극대화하는 방안을 모색할 예정이다.

요약하면, 본 연구는 텍스트로부터 사람의 복잡한 동작을 생성함에 있어 모션의 구성 요소를 전문가 모델들에게 분산시켜 학습하게 함으로써, 모델의 표현력과 생성 품질을 향상시켰다. 이러한 MoE 를 활용한 모션 생성 방법은 향후 더욱 풍부하고 정교한 동작 생성을 이루는 데 기여할 것으로 기대된다.

감사의 글

이 논문은 행정안전부 주관 국립과학수사연구원 중장기과학수사감정기법연구개발 (R&D) 사업의 지원을 받아 수행한 연구임 (NFS2025FSA01).
이 논문은 2025 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2025-00564137).

References

[1] D. P. Kingma, and M. Welling. “Auto-encoding variational bayes,” *International Conference on Learning Representations (ICLR),* 2014.

[2] M. Petrovich, M. J. Black, and G. Varol, “TEMOS: Generating diverse human motions from textual descriptions,” *European Conference on Computer Vision (ECCV)*, pp. 480-497, 2022.

[3] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, “Generating Diverse and Natural 3D Human Motions from Text,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5152-5161, 2022.

[4] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” *Advances in Neural Information Processing Systems (NeurIPS)*, 33, pp. 6840–6851, 2020.

[5] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, “Human Motion Diffusion Model,” *arXiv preprint arxiv:2209.14916*, 2022.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is All You Need,” *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[7] X. Chen, B. Jiang,W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu, “Executing Your Commands via Motion Diffusion in Latent Space,” *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR),* 18000-18010, 2023.

[8] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural Discrete Representation Learning,” *Advances in Neural Information Processing Systems (NeurIPS),* 30, 2017.

[9] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, “MoMask: Generative Masked Modeling of 3D Human Motions,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 1900-1910, 2023.

[10] A. Ghosh, N. Cheema, C. Oguz, C. Theobalt, and P. Slusallek, “Synthesis of Compositional Animations from Textual Descriptions,” *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* pp. 1396-1406, 2021.

[11] C. Zhong, L. Hu, Z. Zhang, and S. Xia, “AttT2M: Text-Driven Human Motion Generation with Multi-Perspective Attention Mechanism,” *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* pp. 509-519, 2023.

[12] Q. Zou, S. Yuan, S. Du, Y. Wang, C. Liu, Y. Xu, J. Chen, and X. Ji, “ParCo: Part-Coordinating Text-to-Motion Synthesis,” *European Conference on Computer Vision (ECCV),* pp. 126-143, 2024.

[13] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, “Adaptive Mixtures of Local Experts,” *Neural Computation*, 3, pp. 79–87, 1991.

[14] W. Fedus, B. Zoph, and N. Shazeer, “Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient

Sparsity," *Journal of Machine Learning Research*, 23, 120, pp. 1-39, 2022.
[15] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen, "T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 14730-14740, 2023.
[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *International Conference on Machine Learning (ICML),* pp. 8748–8763, 2021.
[17] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A Skinned Multi-Person Linear Model," *ACM Transactions on Graphics (TOG),* 34, 6, pp. 851-866, 2015.
[18] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *International Conference on Learning Representations (ICLR)*, 2019.
[19] C. Guo, X. Zuo, S. Wang, and L. Cheng, "TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts," *European Conference on Computer Vision (ECCV)*, pp. 580-597, 2022.
[20] M. Zhang, H. Li, Z. Cai, J. Ren, L. Yang, and Z. Liu, "FineMoGen: Fine-Grained Spatio-Temporal Motion Generation and Editing," *Advances in Neural Information Processing Systems (NeurIPS)*, 36, pp. 13981-13992. 2023.

〈 저 자 소 개 〉

• 선 재 영
• 2024년 경희대학교 소프트웨어융합학과 학사
• 2024년 ~ 현재 경희대학교 일반대학원 인공지능학과 석사과정
• 관심 분야: Character Animation, Text-to-motion Generation
• https://orcid.org/0009-0004-7097-7957

• 홍 성 은
• 2023년 경희대학교 소프트웨어융합학과 학사
• 2024년 ~ 현재 고려대학교 일반대학원 컴퓨터학과 석사과정
• 관심 분야: Character Animation
• https://orcid.org/0000-0002-7681-2617

• 김 경 민
• 2023년 경희대학교 소프트웨어융합학과 학사
• 2025년 경희대학교 소프트웨어융합학과 석사
• 2025년 ~ 현재 고려대학교 일반대학원 컴퓨터학과 박사과정
• 관심 분야: Character Animation
• https://orcid.org/0000-0001-5168-9563

• 우 승 우
• 2007년 서강대학교 물리학과 학사
• 2011년 서강대학교 물리학과 석사
• 2011년 ~ 2014년 SK하이닉스 선임연구원
• 2019년 서강대학교 물리학과 박사
• 2014년 ~ 현재 국립과학수사연구원 안전과 화재방화실장
• 관심 분야: AI를 이용한 법과학 감정기법 연구
• https://orcid.org/0000-0001-9649-7776

• 황 효 석
• 2004년 연세대학교 기계공학과 학사
• 2009년 한국과학기술원 로봇공학학제전공 석사
• 2017년 한국과학기술원 전기및전자공학과 박사
• 2009년 ~ 2017년 삼성전자 종합기술원 연구원
• 2018년 ~ 2021년 가천대학교 소프트웨어학과 조교수
• 2021년 ~ 2024년 경희대학교 소프트웨어융합학과 조교수
• 2024년 ~ 현재 경희대학교 소프트웨어융합학과 부교수
• 관심 분야: Domain Generalization, Sim2Real, Reinforcement learning, Robotics
• https://orcid.org/0000-0003-3241-8455

• 강 형 엽
• 2012년 고려대학교 컴퓨터 · 통신공학부 학사
• 2014년 고려대학교 컴퓨터학과 석사
• 2017년 고려대학교 컴퓨터학과 박사
• 2017년 ~ 2018년 고려대학교 컴퓨터학과 연구교수
• 2018년 ~ 2019년 고려대학교 차세대가상증강현실연구소 연구 교수
• 2019년 ~ 2020년 강원대학교 소프트웨어미디어 · 산업공학부 조교수
• 2020년 ~ 2024년 경희대학교 소프트웨어융합학과 조교수
• 2024년 ~ 현재 고려대학교 컴퓨터학과 부교수
• 관심 분야: Computer Graphics, Extended Reality, Agent, World Model, Human-computer
• Interaction, Holography
• https://orcid.org/0000-0001-5292-4342