Journal of Korea Robotics Society (2025) 20(1):077-083
https://doi.org/10.7746/jkros.2025.20.1.077 ISSN: 1975-6291 / eISSN: 2287-3961

CNN을 활용한 sLLM 기반의 한국어 텍스트-모션 검색 시스템

sLLM-based Korean Text-to-Motion Retrieval System Using CNN

최 지 원[1] · 박 광 현†
Jiwon Choi[1], Kwang-Hyun Park†

Abstract: In this paper, we propose a novel approach to improve the performance of Korean text-based 3D human motion retrieval. Previous studies have failed to adequately cover diverse motion representations and have limitations in distinguishing subtle differences between motions. To address these issues, we introduce sLLM to embed comprehensive motion descriptions and utilize CNN to distinguish subtle motion variations. In addition, we created a Korean HumanML3D dataset to adapt previous researches and conduct experiments on Korean. We also developed a new testing protocol and performed additional experiments. The results show that the proposed method outperforms previous researches, with the R@1 metric improving from 4.49 to 6.39 and the MedR metric improving from 35.00 to 24.75. Furthermore, we compared qualitative results and demonstrated that our method performs well.

Keywords: Text-to-motion Retrieval, sLLM, Contrastive Learning

1. 서 론

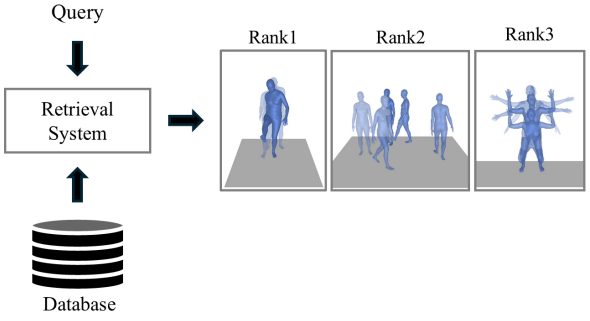
최근 LLM (Large Language Model)의 발전과 함께 크로스 모달(Cross-Modal)과 멀티모달(Multi-Modal)에 대한 연구가 빠르게 진행되고 있다. 특히 이미지 분야에서 CLIP[1], BLIP[2]과 같은 성공적인 모델이 텍스트와 이미지를 연결하며 놀라운 성과를 보였다. 이러한 모델들은 서로 다른 모달리티 간의 의미적 일관성을 학습하여 이미지 설명, 검색 등 다양한 작업에 활용되고 있다. 마찬가지로 모션(Motion) 분야에서도 자연어와 3D 인간 동작 간의 연관성을 학습하여 텍스트에서 동작을 생성하는 T2M-GPT[3], MotionGPT[4]가 있고 검색을 위한 TMR[5]이 있다.

본 논문에서는 사람의 동작을 검색하는 텍스트-모션(Text-to-Motion) 검색을 다룬다. 이는 [Fig. 1]과 같이 모션을 설명하는 자연어 쿼리를 바탕으로 데이터베이스에서 가장 관련성이 높은 3D 인간 동작을 검색하는 것이다.

Received : Oct. 24. 2024; Revised : Dec. 6. 2024; Accepted : Dec. 17. 2024
※ The present research has been conducted by the Excellent researcher support project of Kwangwoon University in 2024.
1. M.S. Student, School of Robotics, Kwangwoon University, Seoul, Korea (jiwonchoi1125@kw.ac.kr)
† Professor, Corresponding author: School of Robotics, Kwangwoon University, Seoul, Korea (akaii@kw.ac.kr)

텍스트-모션 검색 기반의 솔루션은 대규모 모션 캡처 데이터를 자동으로 인덱싱하거나 동작에 대한 라벨링 과정을 초기화하는 데 활용될 수 있다. 또한, 로봇 응용 분야에서 활용 가능성이 크며, 최근 I-CTRL[6] 연구는 인간 동작을 로봇 동작으로 변환하는 방식을 탐구하고 있다. 이처럼 텍스트-모션 검색 기술은 로봇이 자연어 명령을 해석하고, 이에 적합한 동작을 검색하여 실행하는 데 유용하게 적용될 수 있다.

사람의 동작을 텍스트로 표현하고 검색하는 작업은 매우 어려운 일이다. 예를 들어 '사람이 공을 던지고 받는' 동작을 보고 '야구를 한다', '캐치볼을 한다' 등으로 다양하게 텍스트로 표



[Fig. 1] Text-to-motion retrieval: Ranks motions based on their similarity to a query that describes human motion

현할 수 있다. 이처럼 동작을 텍스트로 표현하는 것은 관찰자의 환경과 문화, 관심사에 따라 달라진다. 따라서 같은 동작에 대해서도 표현하는 텍스트가 다양할 수 있으며, 텍스트를 구성하는 단어가 달라질 수 있다. 기존의 연구[5]는 학습 데이터 세트에 없는 단어를 포함한 문장으로 동작을 검색하면 정확도가 떨어진다. 또한 긴 문장을 입력하여 검색하면 원하는 동작을 정확하게 찾아내지 못하는 문제가 있다.

본 논문에서는 이러한 한계점을 극복하기 위해 sLLM (small Large Language Model)을 도입하여 동작에 대한 설명을 임베딩하였다. sLLM은 광범위한 지식을 습득한 언어 모델로서 최근 빠르게 발전하고 있다. sLLM의 지식을 활용하면 학습 데이터에 없는 문장도 효과적으로 의미를 반영하여 동작을 검색할 수 있다. 또한, 동작 인코더에 CNN (Convolutional Neural Network) 기반의 인코더를 추가하여 트랜스포머 기반의 동작 인코더가 비슷한 동작을 같은 위치에 임베딩하는 특성을 보완하였다. CNN 인코더는 동작의 지역적 정보를 추가로 학습함으로써, 동작 간의 세밀한 차이를 효과적으로 표현할 수 있도록 한다. 특히 텍스트-모션 검색 작업에서 긴 문장이 동작의 세부 차이를 설명하는 경향이 있으므로, CNN을 도입하는 것은 긴 문장을 포함한 검색에서도 높은 정확성을 보장한다. 비교 실험을 위해 한국어 HumanML3D[7] 데이터 세트를 제작하고, 기존의 연구를 한국어 데이터 세트로 학습하여 비교 실험을 진행하였다. 본 연구의 기여는 다음과 같다.

1) 다양한 텍스트 표현이 가능한 동작에 대해 설명의 의미를 반영할 수 있도록 sLLM을 통한 임베딩을 도입하였다.
2) 미세한 동작의 변화를 감지하기 위해 CNN 인코더를 추가로 도입하였다.
3) 한국어 HumanML3D[7] 데이터 세트를 제작하고, 기존의 연구를 한국어 데이터 세트로 학습하여 비교 실험을 진행하였다.

2. 관련 연구

2.1 Large Language Model (LLM)

LLM은 대규모 텍스트 데이터를 기반으로 학습된 모델로서, 언어의 문맥과 의미를 효과적으로 이해하고 생성할 수 있으며, 대표적으로 GPT-4[8], LLaMA[9], Qwen[10] 등이 있다. 이러한 모델들은 수십억 개의 매개변수를 통해 방대한 지식을 학습하며 다양한 자연어 처리(NLP) 작업에 활용된다. sLLM (small Large Language Model)은 LLM에 비해 작은 크기의 모델을 의미하며, 작지만 강력한 성능을 제공한다.

최근에는 크로스 모달(Cross-Modal) 또는 멀티모달(Multi-Modal) 분야에서 이미지나 비디오 등의 비언어적 데이터와 연결하는 데 중요한 역할을 하고 있다. 본 연구는 sLLM의 광범위한 지식을 활용하여 3D 동작을 검색하고자 한다.

2.2 크로스 모달 검색(Cross-Modal Retrieval)

크로스 모달 검색은 텍스트, 이미지, 비디오와 같은 서로 다른 모달리티 간의 연관성 정보를 검색하는 기술이다. 비전과 자연어 분야에서 빠르게 발전하였으며, 현재도 널리 사용되고 있는 CLIP[1], BLIP[2], CoCa[11] 같은 이미지 관련 모델들이 있다. 이 모델들은 공통적으로 교차 모달 대조 학습(Cross-Modal Contrastive Learning) 기법을 변형하여 사용한다. 본 연구에서는 InfoNCE[12] 기법을 사용한다.

2.3 텍스트-모션 검색(Text-to-Motion Retrieval)

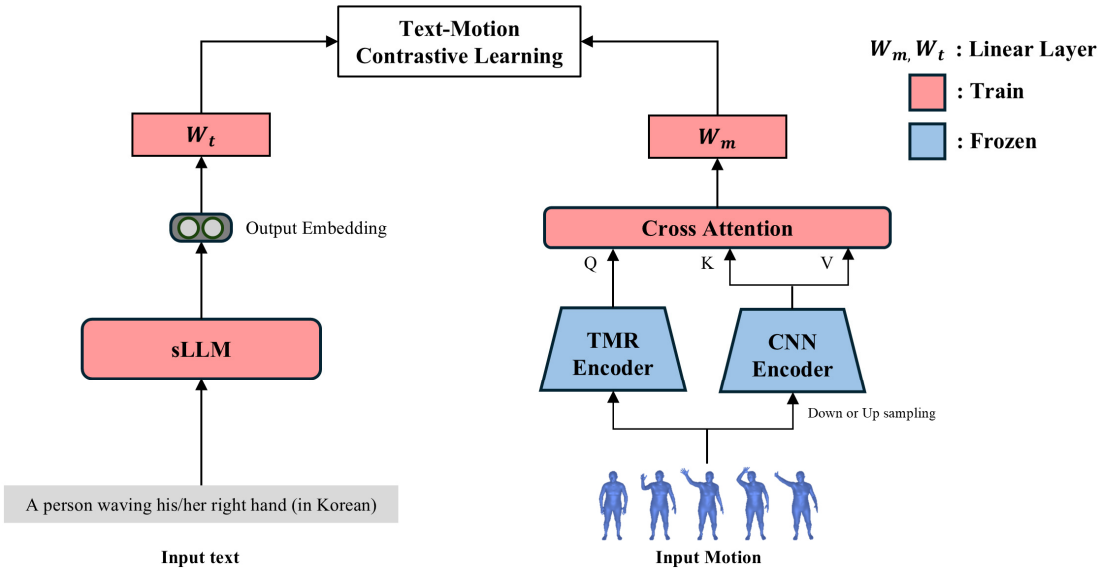
TMR[5]은 텍스트 기반의 동작 검색이라는 새로운 분야를 개척하고 접근 방식을 제시하였다. Nicola[13] 또한 이 분야를 탐구하였으며, 두 연구 모두 CLIP[1]과 같은 대조 학습을 활용하여 텍스트-모션 임베딩 공간을 생성한다. TMR[5]은 텍스트 인코더와 동작 인코더가 모두 트랜스포머 인코더로 구성되며, 텍스트 인코더는 DistilBERT[14] 네트워크에서 텍스트 특징을 입력으로 받는다. 이후, 동작의 평균 토큰과 텍스트의 평균 토큰으로 대조 학습(Contrastive Learning)을 진행한다. 또한, 대조 학습에서 부정 샘플링(Negative Sampling) 방식을 채택하여 비슷한 동작들을 연산에서 제외한다. 본 논문에서도 부정 샘플링을 사용하여 대조 학습을 진행한다.

3. 제안 방법

3.1 sLLM을 활용한 임베딩

기존에 텍스트 인코더로 사용된 DistilBERT[14] 모델을 대신하여 본 연구에서는 sLLM을 활용하여 사람의 동작에 대한 설명을 임베딩하였다. 이전의 연구에서는 학습 데이터 세트에 없는 단어를 사용한 문장으로 동작을 검색하였을 때 정확도가 저하되는 문제가 있다. 이는 사전 지식의 차이에서 비롯된 문제이며, 더 광범위한 사전 지식을 보유한 sLLM을 통해 임베딩을 수행하였다.

[Fig. 2]는 전체 구조를 나타낸다. 입력되는 text에 [EOS] 토큰을 추가하여 마지막 은닉 상태(Hidden State)에서 임베딩을 추출한다. 이렇게 얻은 임베딩은 선형 레이어(Linear Layer)를 통해 텍스트-모션 임베딩 공간으로 맵핑된다. 본 연구에서는 Qwen2 1.5B[15] 모델을 사용하였으며, LoRA[16]를 활용하여 효율적인 학습을 진행하였다.



[Fig. 2] Network architecture: An [EOS] token is appended to the text, and the embedding is extracted from the final hidden state, followed by a Linear Layer. The motion input is processed by two encoders, TMR and CNN, with the frame length fixed through up-sampling or down-sampling before being fed into the CNN encoder. The outputs of the two encoders are combined via Cross Attention and then mapped to the text-motion embedding space via a Linear Layer

3.2 CNN을 활용한 임베딩

트랜스포머 기반의 모델을 사용하는 TMR[5]은 동작의 전체적인 특징을 추출하는 데에는 유리하지만, 구체적이고 미세한 차이를 구별하는 데에는 한계가 있다. 이에 따라 긴 문장이나 여러 동작이 포함된 문장으로 검색할 때 정확도가 떨어진다. 이러한 문제를 해결하기 위해 CNN 인코더를 추가하여 지역적인 정보를 제공함으로써 동작의 미세한 차이를 보다 정확하게 구별하고 차별화된 정보를 제공할 수 있었다. 본 논문에서의 CNN 인코더는 T2M-GPT[3]의 VQ-VAE 구조에서 사용된 CNN 인코더를 활용하였다.

입력되는 동작 데이터의 프레임 수를 맞추기 위해 업샘플링 또는 다운샘플링을 적용하였다. 기준 프레임보다 낮은 동작 데이터 세트는 선형 보간법(Linear Interpolation)을 사용하여 업샘플링, 기준 프레임보다 높은 동작은 적응형 풀링(Adaptive Pooling)을 사용하여 다운샘플링하였다.

TMR[5] 인코더와 CNN 인코더에서 서로 다른 피처를 조합하기 위해 스케일 닷 프로덕트 어텐션(Scaled Dot-Product Attention) 기반의 크로스 어텐션(Cross-Attention)을 사용하였다. TMR 인코더와 CNN 인코더는 학습을 진행하지 않고 크로스 어텐션과 선형 레이어(Linear Layer)만 학습을 진행한다.

3.3 대조 학습(Contrastive Learning)

텍스트와 동작을 텍스트-모션 공간에 맵핑하기 위해 대조 학습을 진행하였다. 식 (1)은 텍스트-모션(Text-to-Motion), 식 (2)는 모션-텍스트(Motion-to-Text)의 infoNCE[12] 손실이다. 최종 손실 식 (3)을 최소화하며 학습을 진행한다. 이는 텍스트-동작 벡터 양성 쌍(Positive Pair) $(x_1, y_1), \ldots, (x_N, y_N)$에 대해 수행되며, $(x_i, y_j), i \neq j$ 쌍은 음성 쌍(Negative Pair)으로 간주한다. 여기서 x_i는 i번째 텍스트, y_j는 j번째 동작 데이터를 의미한다. 식 (2)와 (3)에서 τ는 온도(Temperature)이며, $sim(x, y)$은 코사인 유사도(Cosine Similarity)를 의미한다.

$$L_{t2m} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{\exp(sim(x_i,y_i)/\tau}{\sum_{j=1}^{N}\exp(sim(x_i,y_j)/\tau}\right) \quad (1)$$

$$L_{m2t} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{exp(sim(y_i,x_i)/\tau}{\sum_{j=1}^{N}exp(sim(y_i,x_j)/\tau}\right) \quad (2)$$

$$L_{total} = L_{t2m} + L_{m2t} \quad (3)$$

또한 부정 필터링(Negatives Filtering) 방법을 적용하여 대조 학습을 수행하였다. 동작 캡쳐 데이터 세트는 반복적이거나 유사한 경우가 많다. 예를 들어 '사람이 앞으로 걷는다'와 '누군가 전진하며 걷는다'는 표현은 다르지만 실제로는 동일한 동작을 나타낸다. 이러한 자연어의 유연성과 모호성으로 훈련 과정에 랜덤하게 배치가 선택되면, 모델이 유사한 의미를 가진 두 잠재 벡터를 서로 밀어낼 수 있다. 이는 학습을 방해하게 되어 성능

저하를 초래할 수 있다. 이 문제를 해결하기 위해 기존 연구에서와 같은 방법으로 언어 모델로 문장 유사도 점수를 계산한 후, 일정 임곗값 이상이면 해당 텍스트 쌍을 부정(Negatives)으로 간주하고 손실계산에서 제외하였다. 기존에는 MPNet[17]을 사용하였지만, 이 모델은 영어 기반의 언어 모델이므로 본 연구에서는 다국어 임베딩이 가능한 Multilingual-e5-base[18] 모델을 사용하였으며, 임곗값은 0.95를 사용하였다.

4. 데이터 세트 및 평가 방법

4.1 데이터 세트

본 연구에서는 텍스트와 3D 동작 쌍의 대규모 데이터 세트인 HumanML3D[6] 데이터 세트를 사용하였다. 일상적인 활동(걷기), 운동 등과 같은 다양한 인간 행동이 포함된 데이터 세트로서 14,616개의 모션과 44,790개의 영어 설명으로 구성되어 있다. 한국어 HumanML3D 데이터 세트를 제작하기 위해 영어 설명을 DeepL API를 통해 한국어 문장으로 번역하여, 한국어 데이터 세트를 제작하였다.

4.2 평가 지표 및 프로토콜

평가는 TMR[5]에서 사용한 표준 검색 성능 지표를 사용하였다. 텍스트-모션 및 모션-텍스트 작업 R@1, R@2 등의 랭크(Rank)에서의 재현율(Recall)을 측정하였다. 재현율은 올바른 레이블이 상위 k개의 결과에 포함된 비율을 측정하며, 높을수록 좋다. 평균적으로 정답이 몇 번째 랭크(Rank)에 있는지 측정하는 지표인 중위 랭크(MedR)도 측정하였다. 다음의 (a), (b), (c), (d)는 이전 연구에서 사용한 프로토콜이며 (e)는 추가 실험을 위해 제작한 데이터 세트를 사용한 경우이다.

(a) All: 모든 테스트 세트를 사용하여 평가를 진행한다.

(b) All with threshold: 검색된 동작이 쿼리 텍스트와 임곗값 이상으로 유사하다면 올바른 것으로 간주한다. 예를 들어 '사람이 앞으로 간다'에 대해 '누군가 앞으로 간다'를 올바른 검색 결과로 여긴다. 매우 유사한 텍스트를 제거하기 위해 0.98의 높은 임곗값을 설정하였다.

(c) Dissimilar subset: 텍스트가 최대한 멀리 떨어진 100개의 텍스트-모션 쌍을 샘플링한 데이터 세트를 사용한다.

(d) Small batches: 무작위로 32개의 텍스트-모션 쌍을 선택하여 평균 성능을 측정한다.

(e) Long subset: 단어의 수가 특정 값 이상이면 긴 문장으로 취급한 데이터 세트를 사용한다.

4.3 한국어 기반 TMR 모델 구현

본 연구에서 제안하는 방법론과 기존 방법의 성능을 비교하기 위해 기존 TMR 모델을 한국어 HumanML3D 데이터 세트로 학습하였다. TMR 모델은 텍스트 인코더로 DistilBERT를 사용하며, 이는 영어로 사전 학습된 모델로서 한국어 처리가 불가능하다. 이에 따라 한국어 TMR 모델을 학습하기 위해 텍스트 인코더를 DistilKoBERT로 대체하고, 제작한 한국어 HumanML3D 데이터 세트로 학습을 진행하였다.

5. 실험 결과

[Table 1]과 [Table 2]는 한국어 HumanML3D 데이터 세트를

[Table 1] Text-to-motion retrieval benchmark on HumanML3D: We compare performance across four distinct evaluation sets, and demonstrate improvements over previous approaches

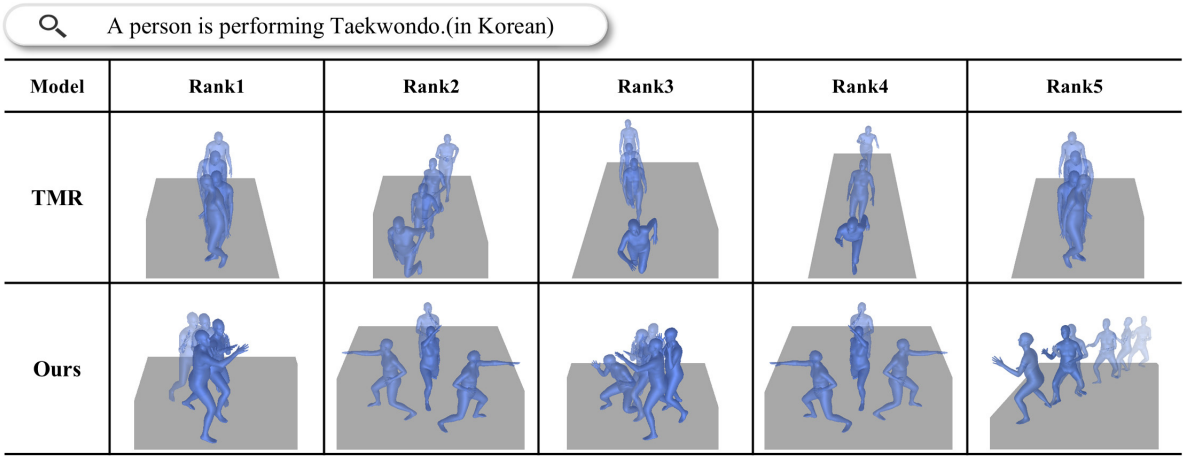
Protocol	Methods	Text-motion retrieval						Motion-text retrieval					
		R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓
(a) All	TMR	4.49	8.60	11.79	17.20	27.14	35.00	7.96	10.08	14.76	19.94	28.47	35.50
	Ours (sLLM)	5.34	9.56	12.68	18.91	28.76	30.50	8.18	10.36	14.58	20.53	30.43	29.75
	Ours	**6.39**	**11.31**	**15.10**	**20.78**	**31.89**	**24.75**	**9.63**	**12.48**	**16.51**	**21.56**	**32.78**	**25.75**
(b) All with threshold	TMR	9.67	13.50	18.50	25.02	35.63	22.00	10.42	13.02	18.77	24.54	33.01	27.50
	Ours (sLLM)	10.52	15.10	20.53	27.65	39.05	19.00	10.76	13.39	19.07	24.83	34.99	23.75
	Ours	**12.89**	**17.04**	**22.17**	**29.93**	**41.90**	**15.00**	**12.52**	**16.33**	**211.1**	**26.92**	**38.05**	**19.50**
(c) Dissimilar subset	TMR	45.00	62.00	72.00	85.00	90.00	2.00	48.00	65.00	74.00	86.00	89.00	2.00
	Ours (sLLM)	53.00	65.00	74.00	85.00	91.00	**1.00**	49.00	68.00	79.00	87.00	92.00	**1.00**
	Ours	**54.00**	**68.00**	**76.00**	**86.00**	**92.00**	**1.00**	**50.00**	**72.00**	**83.00**	**88.00**	**93.00**	**1.00**
(d) Small batches	TMR	63.16	77.49	83.65	89.14	94.18	1.04	63.62	78.42	83.99	89.26	93.98	1.05
	Ours (sLLM)	67.16	81.76	88.74	91.64	94.19	1.01	67.62	81.83	87.48	92.07	94.49	1.01
	Ours	**70.07**	**83.94**	**89.23**	**93.09**	**96.72**	**1.00**	**69.30**	**84.01**	**89.05**	**93.23**	**96.60**	**1.00**

[Table 2] Additional experiments were conducted on long sentences. Sentences with 15 or more, and 20 or more words were classified as long sentences, and performance was evaluated under these conditions

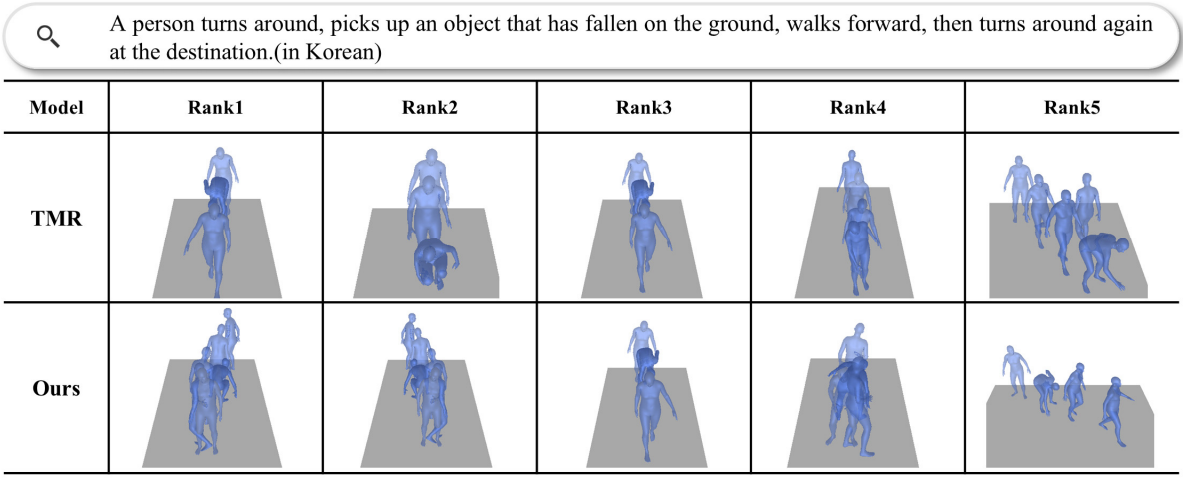
Protocol	Word Count	Methods	Text-motion retrieval						Motion-text retrieval					
			R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓
(e) Long subset	≧15	TMR	11.75	22.94	30.62	41.54	54.21	9.00	18.50	25.62	34.14	42.74	54.58	8.50
		Ours (sLLM)	13.34	25.44	32.38	43.72	57.26	7.50	19.15	27.57	35.71	45.98	58.83	7.00
		Ours	**16.56**	**29.79**	**35.99**	**47.46**	**63.55**	**6.00**	**21.00**	**29.79**	**35.99**	**49.58**	**63.64**	**6.00**
	≧20	TMR	18.20	34.87	41.57	51.53	64.37	5.00	25.10	34.87	41.57	51.53	64.37	5.00
		Ours (sLLM)	19.97	36.06	42.76	53.68	67.05	5.00	26.29	35.93	47.70	56.13	68.97	4.50
		Ours	**23.75**	**39.46**	**46.36**	**57.28**	**69.54**	**4.00**	**28.35**	**37.93**	**49.04**	**58.05**	**70.11**	**4.00**

활용한 정량적 평가 결과를 보여준다. [Table 1]은 프로토콜 (a), (b), (c), (d), [Table 2]는 프로토콜 (e)로 평가한 결과이다. Ours (sLLM)는 CNN 인코더와 크로스 어텐션(Cross-Attention)을 제외하고, sLLM을 통해 임베딩 학습을 진행하였을 때의 결과이며, Ours는 제안하는 전체 학습 방식을 적용했을 때의 결과이다.

[Table 1]에서 Ours (sLLM)와 Ours 모두 기존 연구 대비 우수한 성능을 보인다. 특히, [Table 2]에서 Ours는 Ours (sLLM)보다 더 좋은 성능을 보인다. 이는 CNN 인코더를 통해 모델이



[Fig. 3] When retrieving with the sentence 'A person is performing Taekwondo,' the previous methods retrieve motions like walking or running. In contrast, the proposed method retrieves actions such as punching or performing martial arts



[Fig. 4] The upper section shows an example query, and the lower section visualizes the motions retrieved based on the query. Previous work only includes motions of picking up objects from the ground without turning around. In contrast, our proposedmethod successfully retrieves motions where the subject turns around to pick up an object

긴 동작 설명의 세부적인 특성을 더 잘 포착함을 의미한다.

[Fig. 3]은 '사람이 태권도를 한다'를 검색했을 때의 결과이다. 텍스트-모션 데이터 세트에 포함되지 않은 '태권도' 단어를 사용하여 동작을 검색한 경우, 제안한 방법은 주먹을 뻗거나 무술을 수행하는 동작을 정확하게 검색한다.

[Fig. 4]는 '사람이 뒤돌아서서 땅바닥에 떨어진 물건을 집고 들어 앞으로 걸어가다가 목적지에서 다시 돌아섭니다'와 같은 긴 문장을 검색하였을 때의 결과를 보여준다. 기존 방법은 바닥에 있는 물건을 줍는 동작은 일부 표현하지만 뒤돌아서서 줍는 동작을 반영하지 못한다. 하지만 제안한 방법은 정답(GT) 데이터를 정확히 찾아내며, 복잡하고 구체적인 동작에서도 높은 정확도로 검색을 수행하는 것을 확인할 수 있다. 하지만 모델 크기가 증가함으로 인해 추론 속도의 저하가 발생하였다. RTX 4090 24GB GPU에서 배치 크기 1, 입력 길이 64 토큰을 기준으로 10회의 추론 시간 평균값을 계산한 결과, TMR은 4.2 ms를 기록하였지만, 제안한 모델(Ours)은 15.93 ms로 상대적으로 느려졌다.

6. 결 론

본 연구에서는 한국어를 사용한 3D 동작 검색에서 sLLM과 지역적 정보를 결합한 방식이 효과적임을 확인하였다. sLLM을 활용하여 학습되지 않은 문장도 의미적으로 반영함으로써 검색 성능을 개선할 수 있었으며, CNN 기반의 인코더를 통해 동작의 세부적인 차이를 잘 포착하는 것을 입증하였다. 모델의 크기가 커지면서 추론 속도가 다소 느려졌지만, 향후 더 효율적인 sLLM의 개발이 이루어진다면 성능과 속도 측면에서 추가적인 향상이 기대된다.

References

[1] 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," *arXiv:2103.00020*, 2021, DOI: 10.48550/arXiv.2103.00020.

[2] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," *arXiv:2201.12086*, 2022, DOI: 10.48550/arXiv.2201.12086.

[3] J. Zhang, Y. Zhang, X. Cun, Y. Zhang, H. Zhao, H. Lu, X. Shen, and S. Ying, "Generating human motion from textual descriptions with discrete representations," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, pp. 14730-14740, 2023, DOI: 10.1109/cvpr52729.2023.01415.

[4] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, "Motion GPT: Human motion as a foreign language," *arXiv:2306.14795*, 2023, DOI: 10.48550/arXiv.2306.14795.

[5] M. Petrovich, M. J. Black, and G. Varol, "TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, pp. 9454-9463, 2023, DOI: 10.1109/iccv51070.2023.00870.

[6] Y. Yan, E. V. Mascaro, T. Egle, and D. Lee, "I-CTRL: Imitation to control humanoid robots through constrained reinforcement learning," *arXiv:2405.08726*, 2024, DOI: 10.48550/arXiv.2405.08726.

[7] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3D human motions from text," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 5142-5151, 2022, DOI: 10.1109/cvpr52688.2022.00509.

[8] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, et al., "GPT-4 technical report," *arXiv:2303.08774*, 2024, DOI: 10.48550/arXiv.2303.08774.

[9] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, et al., "LLaMA: Open and efficient foundation language models," *arXiv:2302.13971*, 2023, DOI: 10.48550/arXiv.2302.13971.

[10] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, et al., "Qwen technical report," *arXiv:2309.16609*, 2023, DOI: 10.48550/arXiv.2309.16609.

[11] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive captioners are image-text foundation models," *arXiv:2205.01917*, 2022, DOI: 10.48550/arXiv.2205.01917.

[12] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv:1807.03748*, 2019, DOI: 10.48550/arXiv.1807.03748.

[13] N. Messina, J. Sedmidubsky, F. Falchi, and T. Rebok, "Text-to-motion retrieval: Towards joint understanding of human motion data and natural language," *The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Taipei, Taiwan, pp. 2420-2425, 2023, DOI:10.1145/3539618.3592069.

[14] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv:1910.01108*, 2020, DOI: 10.48550/arXiv.1910.01108.

[15] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, et al., "Qwen2 technical report," *arXiv:2407.10671*, 2024, DOI: 10.48550/arXiv.2407.10671.

[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," *arXiv:2106.09685*, 2021, DOI: 10.48550/arXiv.2106.09685.

[17] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MPNet: Masked and permuted pre-training for language understanding," *arXiv:2004.09297*, 2020, DOI: 10.48550/arXiv.2004.09297.

[18] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 text embeddings: A technical report," *arXiv:2402.05672*, 2024, DOI: 10.48550/arXiv.2402.05672.

최 지 원

2023 광운대학교 로봇학부(학사)

2023~현재 광운대학교 로봇학과 석사과정

관심분야: 자연어 처리, 로보틱스

박 광 현

1994 KAIST 전기및전자공학과(학사)

1997 KAIST 전기및전자공학과(석사)

2001 KAIST 전기및전자공학과(박사)

2008~현재 광운대학교 교수

관심분야: 기계학습, 로봇 소프트웨어