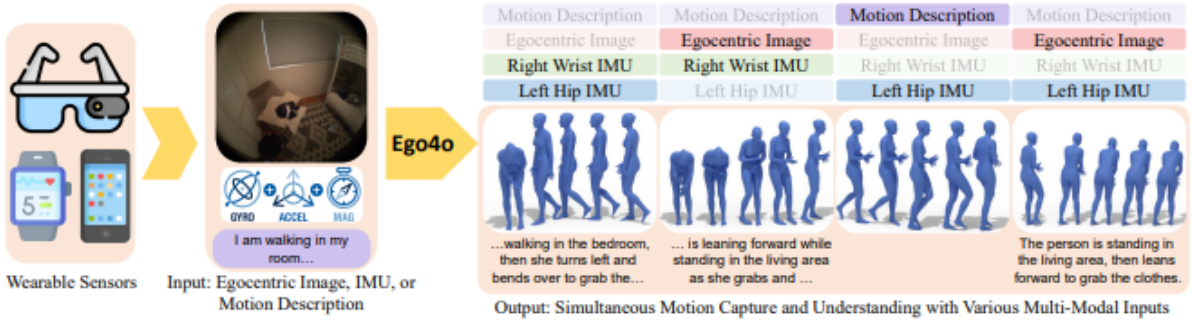
Ego4o: Egocentric Human Motion Capture and Understanding from Multi-Modal Input

https://arxiv.org/abs/2504.08449

0. Introduction

- 웨어러블 기기(예: VR/AR 헤드셋, 스마트 글래스, 스마트워치 등)에서 얻는 여러 센서를 이용해 사람의 움직임을 캡처하고 이해하는 것은 매우 도전적임
- 기존 연구는 보통 단일 입력 모달리티(IMU 또는 비디오)만 사용 → 정보 부족
- 이 논문은 서로 다른 모달리티가 결합되었을 때의 보완적 정보를 활용해 동작 캡처(Motion Capture) 와 동작 이해(Motion Understanding) 를 동시에 수행하는 Ego4o 프레임워크 제안
- 일부 입력만 주어져도 동작 추정 가능하며, 모달리티 조합이 늘어날수록 성능 향상

1. Overview



Wearable Sensors
Input: Egocentric Image, IMU, or Motion Description
Output: Simultaneous Motion Capture and Understanding with Various Multi-Modal Inputs

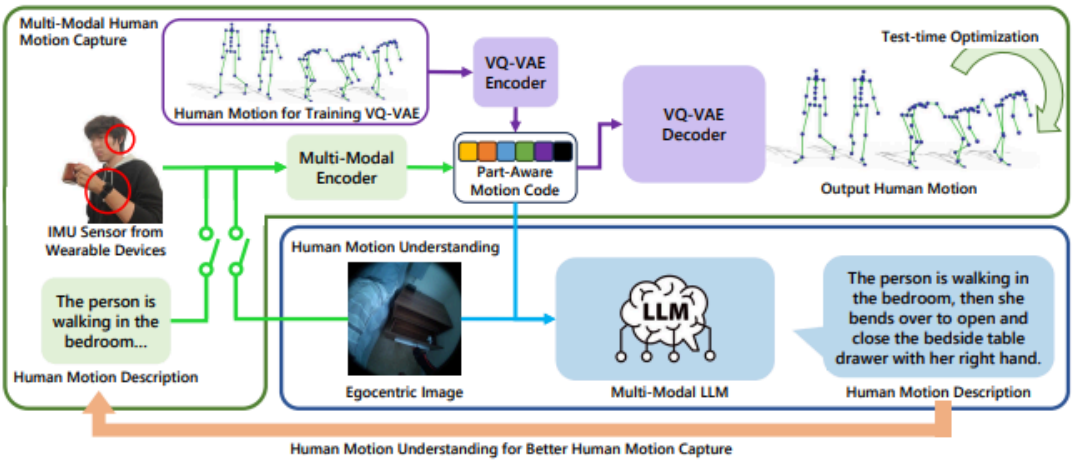
- 입력 모달리티 : IMU 센서, egocentric 카메라 이미지, 텍스트 동작 설명

- 구조 : multi-modal transformer 인코더 → motion codebook 기반 VQ-VAE → 디코더를 통해 모션 예측 및 설명 생성
- 학습 전략 : 다양한 입력 조합을 처리할 수 있도록 랜덤 마스킹 사용
- 설명 생성 : motion description이 없을 경우 LLM과 결합해 설명을 생성하고, 캡처 정확도를 보완하는 피드백 루프 구조 도입

2. Challenges

- 입력 모달리티 불완전성: 일부 센서가 꺼지거나 누락될 수 있음
- 모달리티 간 표현 차이: IMU는 수치, 이미지·텍스트는 시각/언어 표현 → 정렬 필요
- 연속 동작 예측의 정밀도 유지
- 설명 생성과 캡처 정확도의 균형
- 대규모 학습 비용과 모델 복잡성 부담

3. Method



- 인코더 : multi-modal transformer가 IMU, 이미지, 텍스트 입력을 latent vector로 변환
- Motion VQ-VAE & 코드북 : latent vector를 discrete motion codebook에 quantization
- 디코더 / 캡처 : quantized latent를 디코딩해 3D 포즈 및 동작 추정

- 설명 생성 : latent와 이미지 입력을 활용해 motion description 생성
- 랜덤 마스킹 전략 : 학습 중 일부 모달리티 제거해 다양한 조합에 견고하게 대응
- 멀티태스크 학습 : 모션 캡처와 설명 생성을 동시에 최적화

4. Experiments

- 데이터 : IMU + 이미지 기반 egocentric motion dataset
- 비교 대상 : 단일 모달리티 기반 motion capture 모델들
- 평가 지표 : 3D 포즈 예측 정확도, 설명 문장 품질
- 실험 항목 :
  - 입력 모달리티 조합별 성능 비교
  - 설명 포함 여부에 따른 캡처 성능 변화
  - 마스킹 전략 유무에 따른 성능 차이
- 정성적 평가 : 생성된 설명과 동작 일치성 시각화

5. Results

Method	MPJPE (mm)	PA-MPJPE (mm)	Jitter (km/s^3)
DIP-IMU Dataset			
DIP (6 IMU)	73	–	3.01
TransPose (6 IMU)	59	–	0.14
IMUPoser	97	–	0.19
Ego4o-IMU	**84.06**	**63.95**	**0.076**
Nymeria Dataset			
IMUPoser	105.7	72.94	0.054
Ego4o-IMU	95.86	69.03	0.049
Ego4o	**84.82**	**62.33**	**0.048**

Method	Bert(idf)	Bleu@1	Bleu@4	RougeL
TM2T	11.08	40.11	8.99	30.70
MotionGPT	14.09	42.22	**10.31**	32.33
Ego4o	**30.13**	**53.83**	7.46	**38.95**

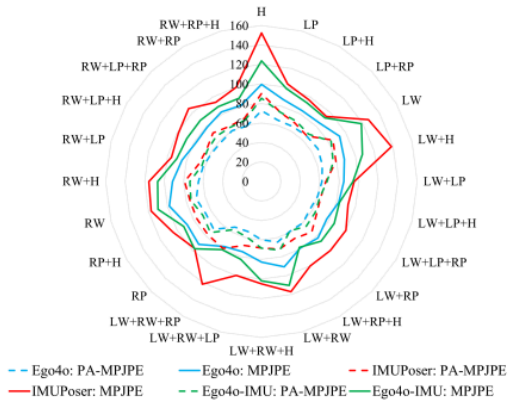


Figure 5. Quantitative results of human motion capture on Nymeria dataset. The result compares our method with IMUPoser under different IMU setups. H, LP, RP, LW, and RW indicate the IMU located on different body parts. H: head, LP: left hip, RP: right hip, LW: left wrist, RW: right wrist.

- 일부 입력만으로도 안정적 모션 캡처 성능 유지

- 다중 모달리티 입력 시 성능 대폭 향상
- 설명 생성이 캡처 정확도 보강에 기여
- 기존 단일 모달리티 기반 방법 대비 전반적 성능 우수
- 동작 캡처와 이해를 통합하는 새로운 방향 제시

6. Insight

- 다양한 모달리티의 상호 보완적 정보 활용이 핵심
- 텍스트 설명과 이미지가 실제 캡처 성능 향상에 기여 가능
- 일부 센서 부재에 대비한 마스킹·대체 전략 필수
- 계산 복잡성 및 실시간 적용 가능성은 여전히 과제
- 후속 연구 방향: 더 다양한 센서 통합, 실시간 응용, 설명의 해석성 강화