장면 및 객체 어포던스 조건을 주입한 확산 기반 인간 행동 예측 시스템 설계

정대원, 염세경*
동국대학교 산업시스템 공학과

2025120165@dgu.ac.kr, *skyoum@dgu.edu

System Design of Scene- and Object-Affordance Conditioning for Diffusion-Based Human Motion Prediction

Daewon Jung, Sekyoung Youm*
*Dongguk Univ., Department of Industrial Systems Engineering.

요 약

인간 행동 예측(HMP)은 관절 오차 최소화를 넘어 이동 가능 및 금지 영역, 지지 구조물, 물체 사용 가능성이 규정하는 가능한 행동의 공간을 반영해야 한다. 본 논문은 장면 및 객체 affordance 를 A_s, A_o로 명시화해 조건 집합 C로 정렬, 압축하고, diffusion denoiser 가 denoising 전 과정에서 cross-attention 으로 C를 반복 주입하도록 구성한 인간 행동 예측 시스템을 제안한다. 또한 관통, 충돌 억제, 이동 가능 영역 준수, 지지 영역 활용, 운동학적 매끄러움 제약과 점유 $O(t)$, MinDist, TTC 기반의 진단 신호 및 정확도/제약 준수 분리 평가 프로토콜을 함께 정의한다.

Ⅰ. 서 론

자율주행, 이동형 자율 시스템, 지능형 감시, 안전, AR/VR 응용에서는 현재 상태 인지만으로 위험을 안정적으로 예방하기 어렵고, 짧은 시간 구간의 인간 동작 변화가 충돌, 접촉, 진입 금지 구역 침범으로 즉시 이어질 수 있다. 따라서 Human Motion Prediction (HMP)은 관절 오차를 줄이는 예측을 넘어, 이동 가능, 금지 영역과 같은 장면 제약, 지지 가능한 구조물, 물체의 기능적 사용 가능성이 규정하는 가능한 행동의 공간을 반영한 미래 동작을 제공해야 한다. 그러나 기존 HMP 는 포즈 시계열 중심이거나 장면 정보를 사용하더라도 단순 특징 결합에 머무르는 경우가 많아, 장애물 관통이나 이동 불가 영역 침범 같은 비현실적 예측을 산출하고 후속 의사결정·위험 평가 단계에서 제약 위반을 유발할 수 있다[1].

최근에는 의도 단서를 고려한 전신 동작 예측, 인간 동작에 diffusion 을 도입해 조건 주입 기반의 생성, 정제를 가능하게 한 프레임워크, 장면 정합을 위해 관통 억제 등 scene-consistency 를 유도한 diffusion 기반 접근, 시선 단서를 통해 3D 장면에서 affordance 와 의도 대상을 결합한 연구가 보고되며 환경 일관성과 조건 주입의 중요성이 강조되고 있다[2-5]. 그럼에도 장면, 객체 affordance 를 부가 입력이 아니라 예측을 규정하는 구조적 조건으로 취급하고, diffusion denoising 전 과정에서 동일 좌표계로 정렬된 조건이 반복 작동하도록 체계화한 설계 관점은 상대적으로 부족하다. 본 논문은 과거 3D 포즈 $X_{1:T}$, 장면 S, 객체 $\{o_i\}$, 의도 단서 I를 입력으로 월드 좌표계로 정렬하고, 장면 affordance 지도 A_s와 객체 affordance 임베딩 A_o를 조건 토큰으로 변환, 압축한 조건 집합 C를 구성한 뒤, cross-attention 기반 denoiser 가 denoising 전 과정에서 C 를 참조하도록 시스템을 설계한다. 또한 예측을 포즈 시계열에 한정하지 않고 점유 $O(t)$, MinDist, TTC 등의 제약 준수 요약 신호를 함께 산출하여, 정확도(MPJPE/FDE)와 제약 준수를 분리 평가하는 프로토콜로 연결한다.

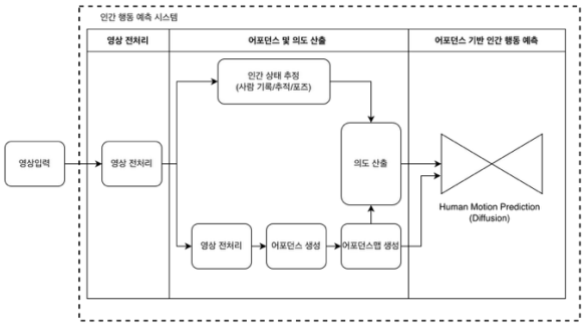

그림 1. 제안 시스템의 전체 개요

Ⅱ. 문제 정의

입력은 과거 T 프레임의 3D 포즈 $X_{1:T}$, 장면 표현 S, 객체 집합 $\{o_i\}$, 의도 단서 I이며, 출력은 미래 H프레임의 3D 전신 포즈 시계열 $X_{T+1:T+H}$ 이다. 본 시스템은 후속 의사결정 단계에서 일관되게 활용될 수 있도록 사람, 객체, 장면 신호를 월드 좌표계로 정렬해 이동과 관절 궤적을 동일 좌표계에서 해석 가능하게 유지한다. 장면 S 는 3D mesh 또는 2D/3D occupancy, semantic map 으로 표현될 수 있고, 각 객체 o_i 는 6-DoF (6 Degrees of Freedom) 포즈, 크기, 카테고리 및 상태로

기술된다. 의도 단서 I는 목표 위치나 관심 영역 등으로 정의하며, 다음 행동의 가능한 범위를 제한하는 약한 조건으로만 사용한다. 안전/제약 지표(MinDist, Collision, TTC)는 예측 인체와 고정 장애물, 금지 영역 간 거리, 충돌을 기준으로 계산하고, 위험 점수 p_{risk}는 충돌, 침범 이벤트의 발생 가능 점수를 확률로 정규화한 값으로 정의한다.

Ⅲ. 방법론

제안 시스템은 환경이 허용하는 가능한 행동의 공간을 장면 affordance A_s와 객체 affordance A_o로 명시화한 뒤, denoiser가 반복적으로 참조할 수 있도록 지도, 토큰 형태로 변환한다. A_s는 이동 가능/불가 영역, 금지, 위험 구역, 지지 가능 영역을 포함하는 affordance map 이며, A_o 는 객체의 기능적 사용 가능성과 상태를 요약한 임베딩이다. Context Encoder 는 과거 포즈를 요약한 pose 토큰, 장면 및 객체 정보를 요약한 scene/object 토큰, 목표, 상태 단서를 요약한 intent 토큰으로 조건 집합 C 를 구성하고, diffusion denoiser 는 잠재 시계열에서 denoising 을 수행하며 cross-attention 으로 C를 매 단계 주입해 예측이 환경 제약을 지속적으로 반영하도록 갱신한다.

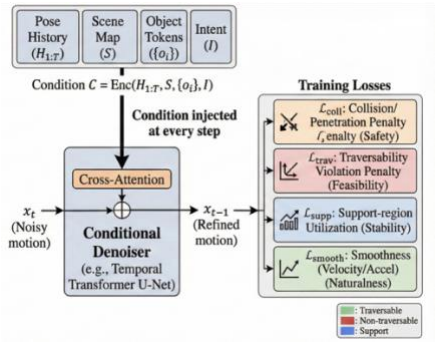

그림 2. 조건부 denoising 과 제약 인지 학습 목표

학습은 기본 denoising 목적에 제약 인지 목표를 결합한다. 관통, 충돌 억제는 예측 인체의 고정 장애물, 금지 영역 침범을 패널티로 제한하고, 이동 가능 영역 준수는 이동 및 접지 위치의 non-traversable 침범을 억제한다. 지지 영역 활용은 불안정 구간에서 지지 가능 영역 근접성을 높이도록 유도하되 정의, 라벨이 제한되면 선택적으로 적용하며, 동작 평활성은 속도·가속도 기반 제약으로 떨림을 완화한다. 출력은 미래 3D 포즈 시계열과 함께 $O(t)$, MinDist, TTC 등 제약 준수 진단 신호를 포함하고, 위험 점수 p_{risk} 는 충돌, 침범 가능 점수를 확률로 정규화해 정의한다. 평가는 정확도(MPJPE/FDE)와 제약 준수(Collision Rate, Violation Rate, Lead Time, ECE)를 분리해 측정한다.

Ⅳ. 평가

본 시스템은 미래 3D 전신 포즈 시계열 $\hat{X}_{T+1:T+H}$ 과 함께, 후속 의사결정, 위험 평가에 활용 가능한 제약 준수 진단 출력을 정의한다. 예측 구간의 시간 누적 점유 $O(t)$ 와 예측 인체-고정 장애물/금지 영역 기준의 MinDist, TTC, Collision Rate, Violation Rate, Lead Time 을 산출하며, 위험 점수 p_{risk} 는 충돌 및 침범 이벤트 가능 점수를 확률로 정규화한 값으로 정의하고 ECE 로 보정을 평가한다. 평가는 정확도(MPJPE, FDE)와 제약 준수를 분리해 측정하며, 비교군은 포즈 기반 HMP(Transformer/GCN), 장면 특징 결합 HMP, affordance 조건을 포함한 diffusion HMP 로 구성한다.

표 1. 평가 지표 및 후속 의사결정/위험평가 활용형 출력

Type	signal	Metric	Definition	Purpose
Accuracy	$\hat{X}_{T+1:T+H}$	MPJPE	평균 관절 위치 오차	fidelity
	$\hat{X}_{T+H}$	FDE	종단 시점 오차	endpoint
Safety	$O(t)$	Collision Rate	충돌 프레임 비율	feasibility
	d_{min}	MinDist	최소 예측 거리	margin
	TTC	TTC	첫 충돌까지 시간	early risk
Warning	τ	Lead Time	경보 시점-이벤트 시점 차	proactive
Calibration	p_{risk}	ECE	위험 확률 보정 오차	reliability
Map	traversability	Violation Rate	% non-trav. violation	consistency

V. 결론

본 논문은 자율 시스템 응용에서 장면, 객체가 제공하는 affordance 를 명시적으로 구성하고 이를 조건으로 주입하는 diffusion 기반 HMP 시스템 설계를 제안하였다. 제안 시스템은 관통, 충돌 억제, 이동 가능 영역 준수, 지지 가능 영역 활용과 같은 제약을 학습 목표와 출력 인터페이스에 통합하여, 환경 제약을 반영한 예측을 지향한다. 향후 연구에서는 대표 HMP 및 장면 융화 동작 데이터셋을 포함한 실제 벤치마크에서 정량 평가를 수행하고, 제안한 제약 및 후속 의사결정 및 위험평가 활용형 출력이 정확도와 장면 일관성에 미치는 영향을 체계적으로 검증할 예정이다. 또한 다양한 환경 표현과 조건 주입 방식에 대한 추가 분석을 통해, 범용 환경에서 안정적으로 동작하는 설계로 확장할 계획이다.

참 고 문 헌

[1] J. Martinez, M. J. Black, and J. Romero, “On Human Motion Prediction Using Recurrent Neural Networks,” *CVPR*, 2017.

[2] P. Kratzer, N. Balachandra Midlagajni, and J. Mainprice, “Anticipating Human Intention for Full-Body Motion Prediction,” *ICRA 2020 Workshop on Long-Horizon Motion Prediction* (workshop paper), 2020.

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

[4] J. Tang, J. Wang, K. Ji, L. Xu, J. Yu, and Y. Shi, “A Unified Diffusion Framework for Scene-aware Human Motion Estimation from Sparse Signals,” *CVPR*, 2024.

[5] T. Yu, Y. Lin, J. Yu, Z. Lou, and Q. Cui, “Vision-Guided Action: Enhancing 3D Human Motion Prediction with Gaze-informed Affordance in 3D Scenes,” *CVPR*, 2025.