Text-Guided Human Motion Generation

2024년도 하계 세미나

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

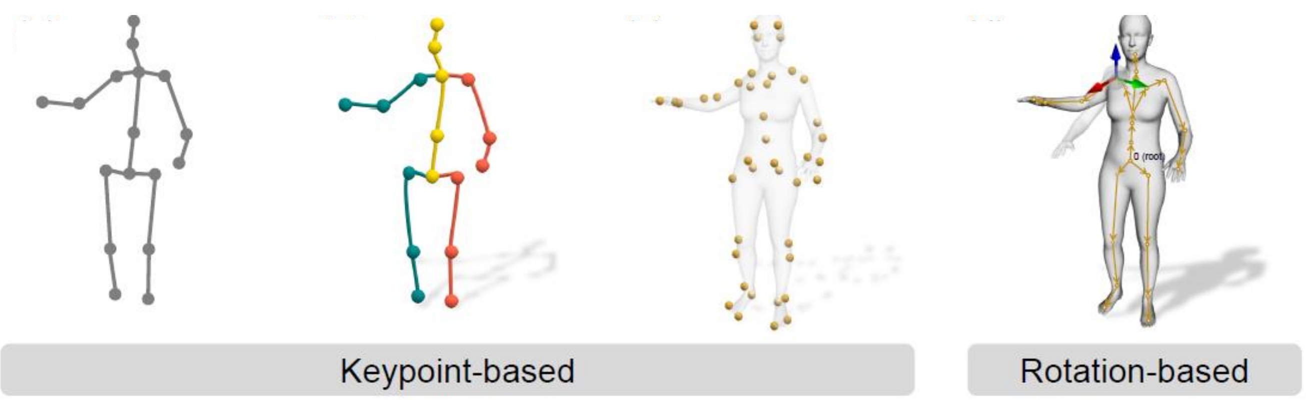
Presented By
안성욱

Outline

- Introduction
  - Human motion generation
- MDM: Human Motion Diffusion Model
  - ICLR 2023 Top-25%
- Move as You Say, Interact as You Can
  - CVPR 2024 Highlight

Introduction to HMG

- Human motion generation (HMG)
  - Goal of HMG
    - 자연스러운 human의 pose sequence 생성
  - Motion Data Representation
    - Keypoint-based
      - 인체 구조에서 구체적인 landmark를 keypoint로 하여, 이들의 집합으로 구성됨
      - Motion capture system에서 직접적으로 얻을 수 있고 해석에 용이함
    - Rotation-based
      - Body joint의 angle에 따라 표현됨
      - SMPL은 joint angle을 통해 human mesh를 모델링하는 대표적인 예시임

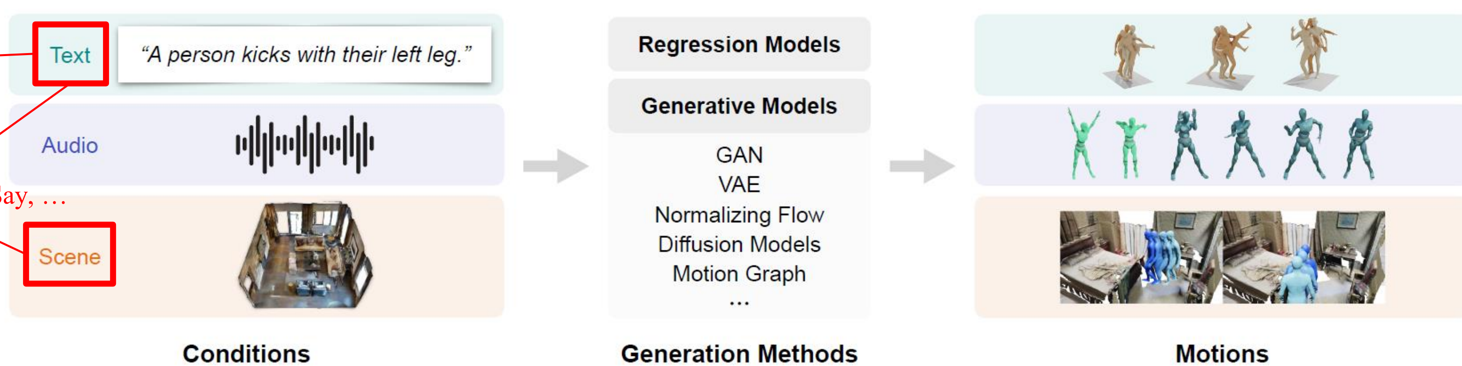


Introduction to HMG

- Typical HMG approaches
  - Text-conditioned motion generation
    - Action-to-motion
      - ‘Walk’, ‘kick’, ‘throw’ 등의 action category에 따라 human motion sequence를 생성
      - Action의 class가 정해져 있어, text-to-motion task에 비해 직관적임
    - Text-to-motion
      - Natural language description에 따라 human motion sequence를 생성
        - ✓Language의 막대한 표현력을 활용함
      - 최근 연구들은 대부분 diffusion model을 사용함

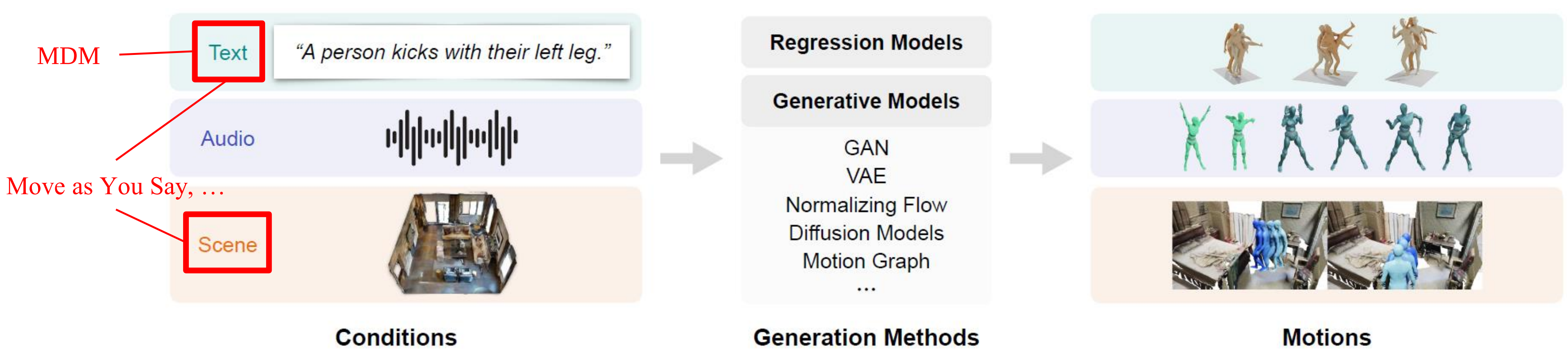
MDM

Move as You Say, …



Introduction to HMG

- Typical HMG approaches
  - Scene-conditioned motion generation
    - Scene representation
      - 대표적으로 point clouds, mesh를 사용하여 3D scene을 표현함
    - Generation pipeline
      - 목표 지점 또는 목표 상호작용 물체를 prediction
      - Path(trajectory)를 planning
      - Planning한 path를 따라 motion infilling



- MDM: Huma Motion Diffusion Model
    - ICLR 2023 Top-25%

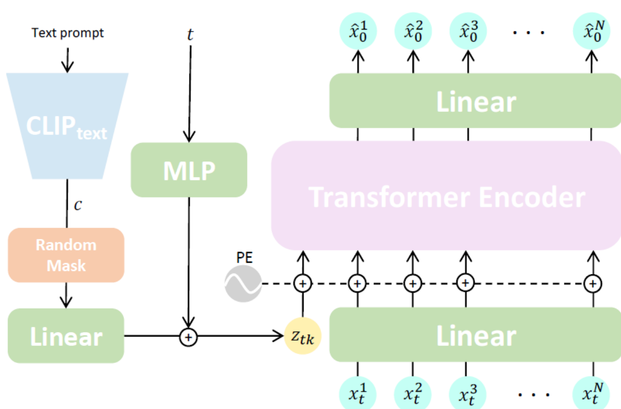
MDM: Human Motion Diffusion Model[1)]

- Introduction
  - Text-guided motion generation의 문제점
    - Text와 motion 사이의 many-to-many problem
      - 하나의 label에 대해, 여러 가지 motion이 존재할 수 있음
      - 반대로, 한 motion에 대해 여러 가지의 설명을 할 수도 있음
    - 사용되는 methods의 one-to-one mapping
      - 기존에 사용되던 auto-encoder나 VAE 기반 HMG는 표현이 한정적임
  - Contributions
    - Human motion diffusion model
      - Diffusion model을 사용하여 human motion을 생성함
    - Fewer GPU resources
      - Trained for 3 days with single NVIDIA GeForce RTX 2080 Ti GPU
    - Geometric losses
      - Motion을 물리적으로 통제

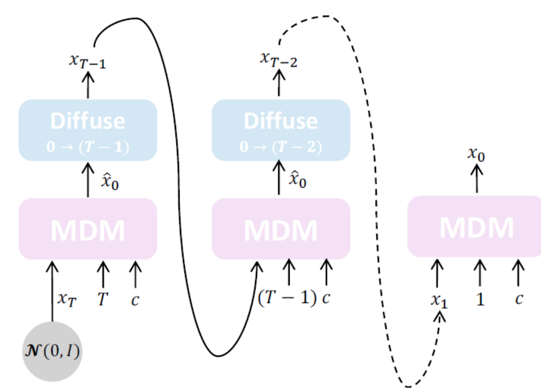
1) TEVET, Guy, et al. Human Motion Diffusion Model. arXiv preprint arXiv:2209.14916, 2022.

MDM: Human Motion Diffusion Model[1)]

- Overview
  - Human motion $x^{1:N}$을 생성하는 것이 목표
    - 주어진 임의의 condition c에 따라 생성함
      - Condition c로는 text 또는 action의 real-world signal이 주어짐
      - $c = \emptyset$인 unconditioned motion generation 또한 가능함
        - ✓In-betweening에서 조건이 주어지지 않을 경우 임의로 motion을 생성함
    - Generated motion $x^{1:N} = \{x^i\}_{i=1}^{N}$은 human pose의 sequence를 나타낸 것
      - $x^i \in \mathbb{R}^{J \times D}$, J is the number of joints, D is the dimension of the joint representation



< MDM block architecture >



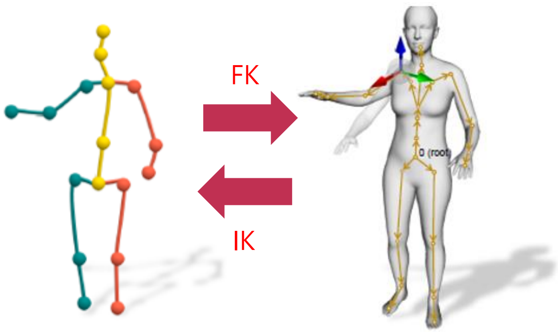
< Overall process >

MDM: Human Motion Diffusion Model[1)]

- Main method
  - Framework
    - Diffusion (Markov noising process $\{x_t^{1:N}\}_{t=0}^{T}$)
      - $x_0^{1:N}$은 data distribution으로부터 얻은 초기의 motion 값
      - 이후에는 다음과 같은 noising process를 거침
        - ✓$q(x_t^{1:N}|x_{t-1}^{1:N}) = \mathcal{N}\left(\sqrt{\alpha_t}x_{t-1}^{1:N}, (1-\alpha_t)I\right)$
          - $\alpha_t \in (0,1)$, constant hyper-parameters
          - α_t가 충분히 작으면 $x_T^{1:N} \sim \mathcal{N}(0,I)$ 를 근사할 수 있음
    - Reversed diffusion process (denoising step)
      - 위에서 생성된 normal distribution 형태의 noise x_T를 점진적으로 denoising
        - ✓Condition에 따라 motion 생성 $p(\boxed{x_0}|c)$ $x_0^{1:N}$과 같음
      - DDPM과 같이 noise ϵ_t를 예측하지 않고, signal 자체를 예측함 $\hat{x}_0 = G(x_t, t, c)$
        - ✓여기서 G는 noise step t와 condition c에 따른 motion generation function
    - Simple loss
      - $\mathcal{L}_{simple} = E_{x_0 \sim q(x_0|c), t \sim [1,T]}[\|x_0 - G(x_t, t, c)\|_2^2]$

MDM: Human Motion Diffusion Model[1)]

- Main method
  - Geometric losses
    - Physical property를 강화하고 artifact를 예방하기 위해 사용
    - 3 geometric losses



- Position loss: body joints의 rotation을 통해 관절의 위치를 optimize
  - ✓ $\mathcal{L}_{pos} = \frac{1}{N}\sum_{i=1}^{N}\left\| \boxed{FK}(x_0^i) - FK(\hat{x}_0^i) \right\|_2^2$

Forward kinematic function
Joint position을 joint rotation으로 변환

- Foot loss: 발이 땅에 닿아 있을 때 velocity를 0으로 설정함으로써 foot-sliding effect를 완화시켜주는 효과가 있음
  - ✓ $\mathcal{L}_{foot} = \frac{1}{N-1}\sum_{i=1}^{N-1}\left\| \left(FK(\hat{x}_0^{i+1}) - FK(\hat{x}_0^i)\right) \cdot \boxed{f_i} \right\|_2^2$

$f_i \in \{0,1\}^J$
foot contact mask

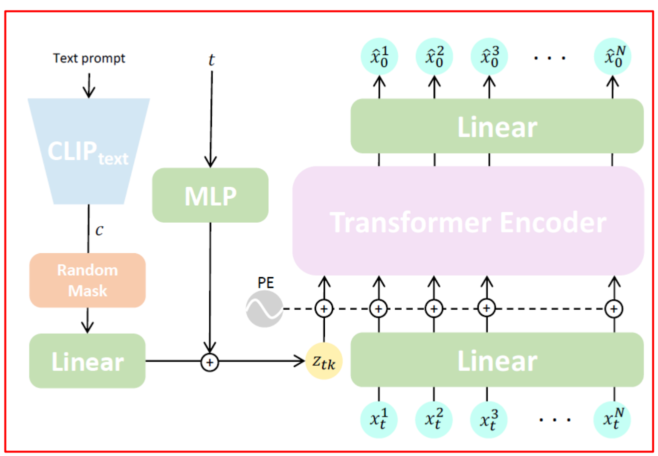
- Velocity loss:
  - ✓ $\mathcal{L}_{vel} = \frac{1}{N-1}\sum_{i=1}^{N-1}\left\| (x_0^{i+1} - x_0^i) - (\hat{x}_0^{i+1} - \hat{x}_0^i) \right\|_2^2$
- Overall training loss
  - ✓ $\mathcal{L} = \mathcal{L}_{simple} + \lambda_{pos}\mathcal{L}_{pos} + \lambda_{foot}\mathcal{L}_{foot} + \lambda_{vel}\mathcal{L}_{vel}$

MDM: Human Motion Diffusion Model[1)]

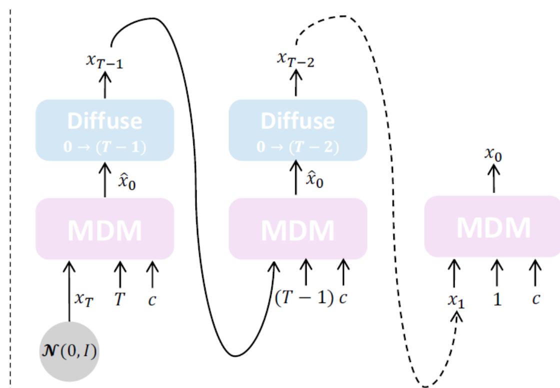
- Main method
  - Model
    - MDM block
      - Noise time step t와 condition code c가 각각의 feed-forward network에 의해 transformer dimension으로 projection된 후, 더해져서 token z_{tk}를 생성함

$x_t^{1:N}$과 같음

      - Noised input $\boxed{x_t}$의 각 frame도 transformer dimension으로 linearly projection됨
      - 이후 encoder로 입력되고, encoder의 output은 다시 motion dimension으로 projection됨
        - ✓Model이 예측한 motion인 $\hat{x}_0$가 생성됨



< MDM block architecture >



< Overall process >

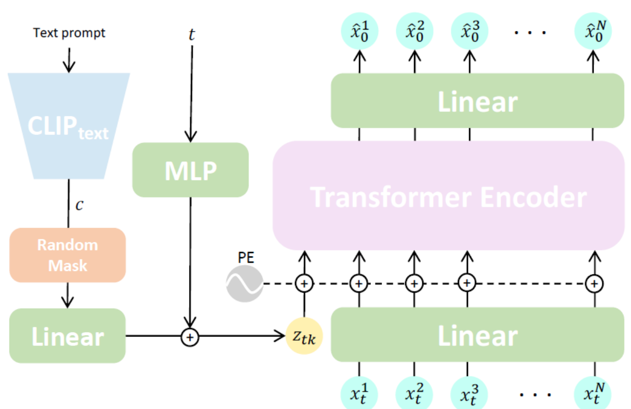
서강대학교 SOGANG UNIVERSITY

VDS LAB

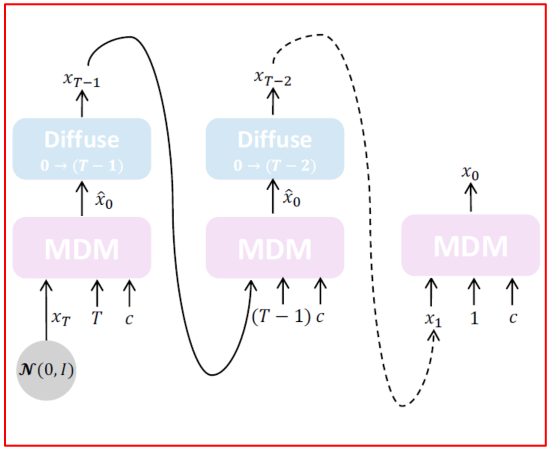
1) TEVET, Guy, et al. Human Motion Diffusion Model. arXiv preprint arXiv:2209.14916, 2022.

MDM: Human Motion Diffusion Model[1)]

- Main method
  - Sampling
    - Overall process
      - $p(x_0|c)$로부터 iterative하게 sampling을 수행함
      - 처음에는 Gaussian noise가 x_T로 입력됨
      - 매 time step t마다 clean sample $\hat{x}_0 = G(x_t, t, c)$를 prediction하고, $\hat{x}_0$ 는 이전보다 한 step 적은 diffusion process로 입력되어 다음 MDM block의 input인 x_{t-1} 을 생성함
      - G는 sample의 10%를 $c = \emptyset$로 setting하여 unconditioned motion도 학습함



< MDM block architecture >

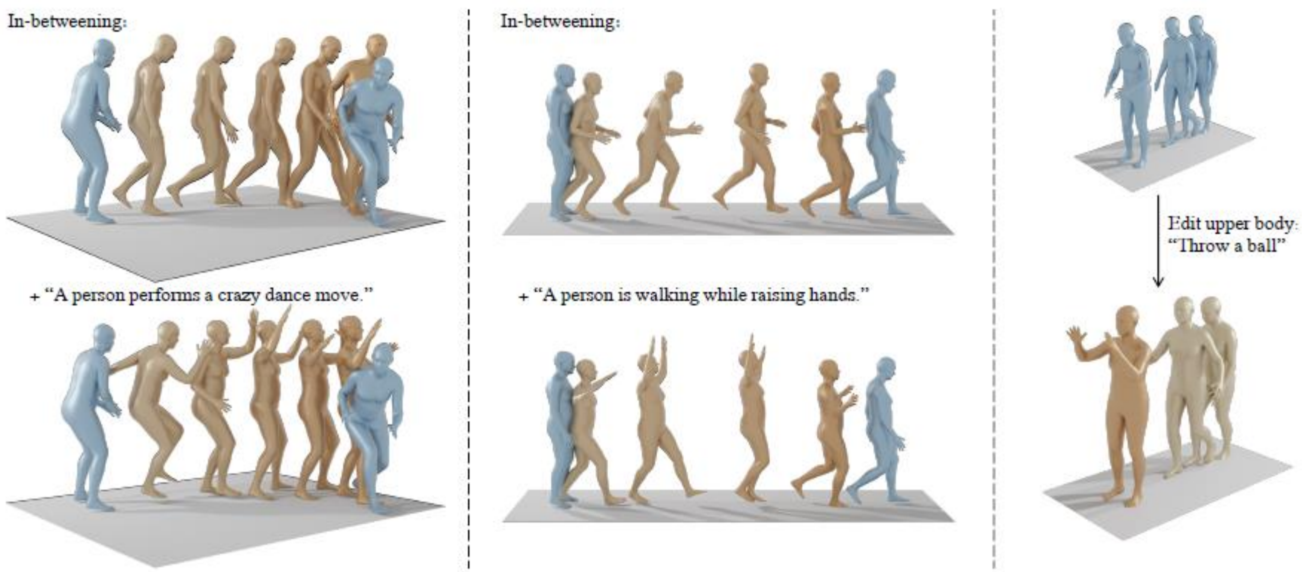


< Overall process >

MDM: Human Motion Diffusion Model[1)]

- Main method
  - Editing
    - Temporal domain에서 motion in-betweening과 spatial domain에서 body part editing이 가능하게 함
      - Diffusion inpainting을 motion data에 적합하게 조정함
      - 별도의 training과정 없이 sampling 과정에서 editing이 가능함

Blue frames: motion input
Bronze frames: generated motion

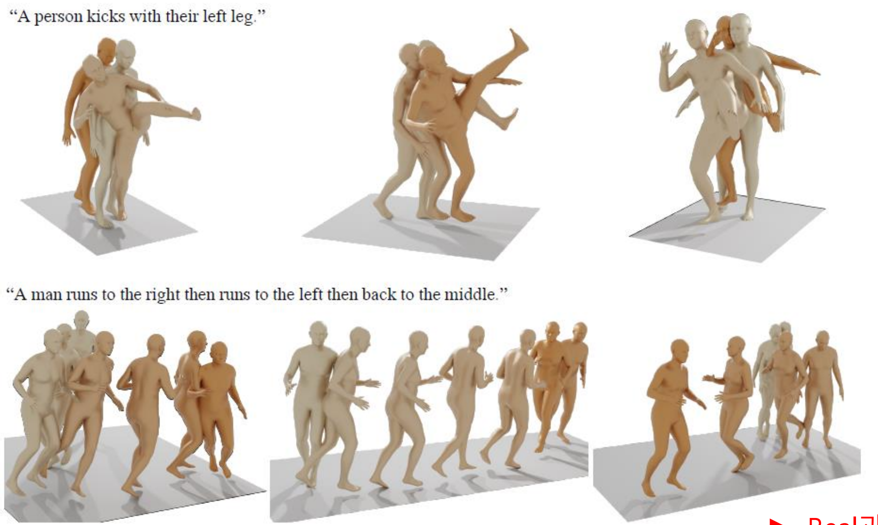


MDM: Human Motion Diffusion Model[1)]

- Result
  - Text-to-motion
    - Input text prompt가 주어진 상황에서 motion을 생성함
    - Datasets
      - KIT, HumanML3D
    - Evaluation metrics
      - R-precision: Cosine similarity 기준 상위 R개의 text, motion pair의 precision
      - FID: Real data distribution과 fake data distribution 사이의 distance
      - Multimodal distribution: 여러 개의 mode(최빈값)를 갖는 연속확률분포
      - Diversity: 생성된 sample의 다양성을 평가함
      - Multimodality: 여러 modality를 처리하는 model의 종합적인 성능

MDM: Human Motion Diffusion Model[1)]

- Result
  - Text-to-motion



Real과 가까울수록 좋은 성능

Method	R Precision (top 3)↑	FID↓	Multimodal Dist↓	Diversity→	Multimodality↑
Real	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
JL2P	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	-
Text2Gesture	$0.345^{\pm.002}$	$7.664^{\pm.030}$	$6.030^{\pm.008}$	$6.409^{\pm.071}$	-
T2M	$\mathbf{0.740^{\pm.003}}$	$1.067^{\pm.002}$	$\mathbf{3.340^{\pm.008}}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
MDM (ours)	$0.611^{\pm.007}$	$\mathbf{0.544^{\pm.044}}$	$5.566^{\pm.027}$	$\mathbf{9.559^{\pm.086}}$	$2.799^{\pm.072}$
MDM (decoder)	$0.608^{\pm.005}$	$0.767^{\pm.085}$	$5.507^{\pm.020}$	$9.176^{\pm.070}$	$\mathbf{2.927^{\pm.125}}$
+ input token	$0.621^{\pm.005}$	$0.567^{\pm.051}$	$5.424^{\pm.022}$	$9.425^{\pm.060}$	$2.834^{\pm.095}$
MDM (GRU)	$0.645^{\pm.005}$	$4.569^{\pm.150}$	$5.325^{\pm.026}$	$7.688^{\pm.082}$	$1.2646^{\pm.024}$

Autoencoder-based (T2M)

KIT test set

Method	R Precision (top 3)↑	FID↓	Multimodal Dist↓	Diversity→	Multimodality↑
Real	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
JL2P	$0.483^{\pm.005}$	$6.545^{\pm.072}$	$5.147^{\pm.030}$	$9.073^{\pm.100}$	-
Text2Gesture	$0.338^{\pm.005}$	$12.12^{\pm.183}$	$6.964^{\pm.029}$	$9.334^{\pm.079}$	-
T2M	$\mathbf{0.693^{\pm.007}}$	$2.770^{\pm.109}$	$\mathbf{3.401^{\pm.008}}$	$\mathbf{10.91^{\pm.119}}$	$1.482^{\pm.065}$
MDM (ours)	$0.396^{\pm.004}$	$\mathbf{0.497^{\pm.021}}$	$9.191^{\pm.022}$	$10.847^{\pm.109}$	$\mathbf{1.907^{\pm.214}}$

HumanML3D test set

서강대학교 SOGANG UNIVERSITY

VDS LAB

MDM: Human Motion Diffusion Model[1)]

- Result
  - Action-to-motion
    - Input action class가 주어진 상황에서 motion을 생성함
    - Datasets
      - HumanAct12, UESTC
    - Evaluation metrics
      - FID: Real data distribution과 fake data distribution 사이의 distance
      - Accuracy: 0~1 사이로 나타내지는 정확도
      - Diversity: 생성된 sample의 다양성을 평가함
      - Multimodality: 여러 modality를 처리하는 model의 종합적인 성능

1) TEVET, Guy, et al. Human Motion Diffusion Model. arXiv preprint arXiv:2209.14916, 2022.

MDM: Human Motion Diffusion Model[1)]

- Result
  - Action-to-motion

Method	FID↓	Accuracy↑	Diversity→	Multimodality→
Real (INR)	$0.020^{\pm.010}$	$0.997^{\pm.001}$	$6.850^{\pm.050}$	$2.450^{\pm.040}$
Real (ours)	$0.050^{\pm.000}$	$0.990^{\pm.000}$	$6.880^{\pm.020}$	$2.590^{\pm.010}$
Action2Motion (2020)	$0.338^{\pm.015}$	$0.917^{\pm.003}$	$6.879^{\pm.066}$	$2.511^{\pm.023}$
ACTOR (2021)	$0.120^{\pm.000}$	$0.955^{\pm.008}$	$\mathbf{6.840^{\pm.030}}$	$2.530^{\pm.020}$
INR (2022)	$0.088^{\pm.004}$	$0.973^{\pm.001}$	$6.881^{\pm.048}$	$2.569^{\pm.040}$
MDM (ours)	$0.100^{\pm.000}$	$\mathbf{0.990^{\pm.000}}$	$6.860^{\pm.050}$	$2.520^{\pm.010}$
w/o foot contact	$\mathbf{0.080^{\pm.000}}$	$\mathbf{0.990^{\pm.000}}$	$6.810^{\pm.010}$	$\mathbf{2.580^{\pm.010}}$

HumanAct12 test set

Method	$\text{FID}_{\text{train}}$ ↓	FID_{test} ↓	Accuracy↑	Diversity→	Multimodality→
Real	$2.92^{\pm.26}$	$2.79^{\pm.29}$	$0.988^{\pm.001}$	$33.34^{\pm.320}$	$14.16^{\pm.06}$
ACTOR (2021)	$20.49^{\pm2.31}$	$23.43^{\pm2.20}$	$0.911^{\pm.003}$	$31.96^{\pm.33}$	$14.52^{\pm.09}$
INR (2022) (best variation)	$\mathbf{9.55^{\pm.06}}$	$15.00^{\pm.09}$	$0.941^{\pm.001}$	$31.59^{\pm.19}$	$14.68^{\pm.07}$
MDM (ours)	$9.98^{\pm1.33}$	$\mathbf{12.81^{\pm1.46}}$	$0.950^{\pm.000}$	$33.02^{\pm.28}$	$\mathbf{14.26^{\pm.12}}$
w/o foot contact	$9.69^{\pm.81}$	$13.08^{\pm2.32}$	$\mathbf{0.960^{\pm.000}}$	$\mathbf{33.10^{\pm.29}}$	$\mathbf{14.06^{\pm.05}}$

UESTC test set

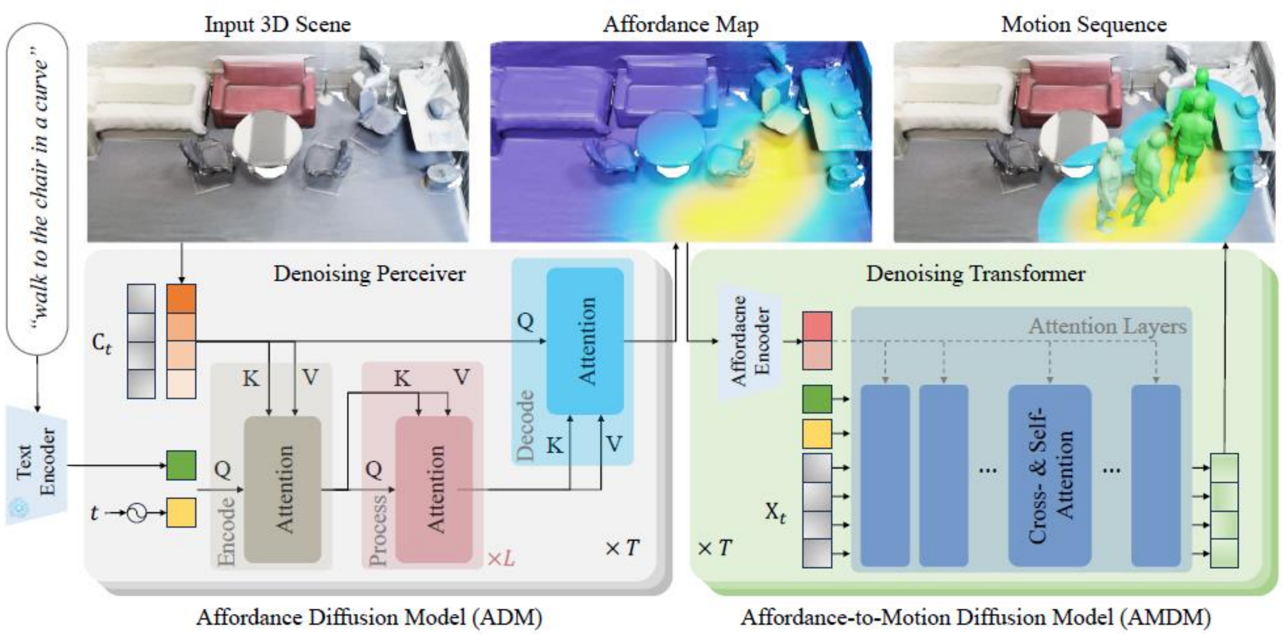
- Move as You Say, Interact as You Can: Language-guided Human Motion Generation with Scene Affordance
  - CVPR 2024 Highlight

Move as You Say, Interact as You Can[1)]

- Introduction
  - 3D environments상에서 수행되는 HMG의 limitation
    - 생성 모델이 language, 3D scene, human motion을 jointly modeling하는 능력이 부족함
    - 높은 품질의 language-scene-motion dataset이 부족함
  - Contributions
    - 3D scene grounding과 conditional motion generation 사이의 gap을 채워주는 intermediate representation역할을 하는 scene affordance를 도입하여 two-stage modeling을 진행함
      - Scene affordance는 간결한 방식으로 3D scene을 표현하면서도, scene과 human motion 사이의 정교한 geometric interplay가 가능하게 함
    - Language-scene-motion data의 결핍에도 불구하고 뛰어난 HMG 성능을 보임

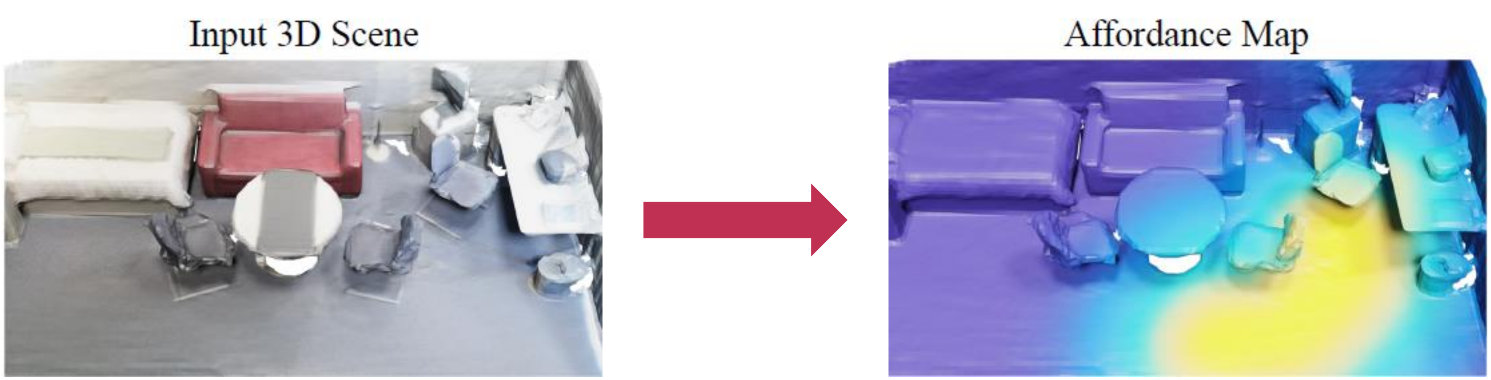
Move as You Say, Interact as You Can[1)]

- Overview
  - Two-stage로 나누어 진행됨
    - Affordance Diffusion Model (ADM)
      - Affordance map을 생성하는 부분
    - Affordance-to-Motion Diffusion Model (AMDM)
      - 생성된 Affordance map을 통해 motion을 생성하는 부분



Move as You Say, Interact as You Can[1)]

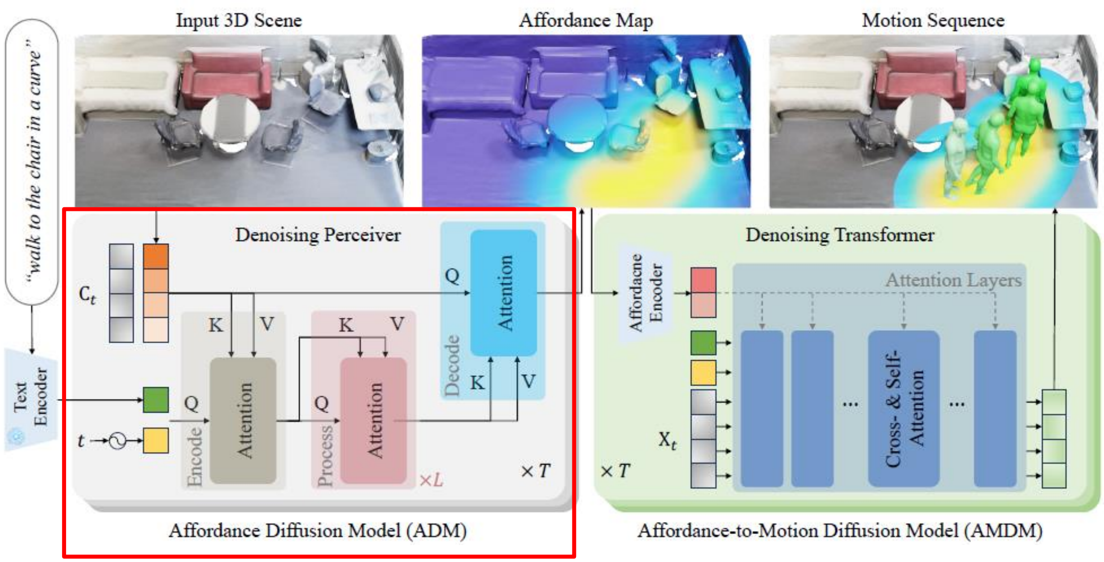
- Main method
  - Affordance Map (ADM의 GT로 사용)
    - 3D indoor scene의 필수적인 detail 정보들을 추출하여 motion 생성을 support하는 역할
    - 3D scene의 point들과 human joint들 사이의 distance field 형태로 표현됨
      - Motion sequence $X = \{x_i\}_{i=1}^{F}, x_i \in \mathbb{R}^{J\times3}$
      - 각 scene point와 각 frame에서의 joint 사이 ℓ_2 distance를 계산하여 per-frame distance field $d \in \mathbb{R}^{N\times J}$를 구함
        - ✓Distance map $c(n,j) = \exp(-\frac{1}{2}\frac{d(n,j)}{\sigma^2})$
        - ✓Affordance map $C = maxpool(c_1, c_2, \dots, c_F)$



Move as You Say, Interact as You Can[1)]

- Main method
  - Affordance Diffusion Model (ADM)
    - 다음과 같은 공식을 통해 affordance map C를 생성함
      - $p_\theta(C_{0:T}|\mathcal{S},\mathcal{L}) = p(C_T)\prod_{t=1}^{T} p_\theta\left(C_{t-1}|C_t,\mathcal{S},\mathcal{L}\right)$

        $\mathcal{S} \in \mathbb{R}^{N\times 6}$: RGB point cloud
        $\mathcal{L} = [w_1, w_2, \ldots, w_M]$: Language description
    - ADM은 perceiver 형태의 architecture로 구성되어 있음
      - Perceiver란 Transformer를 수정하여 만든 신경망으로, Transformer는 language에 대해서만 다룰 수 있었으나 Perceiver는 모든 종류의 입력 데이터를 다룰 수 있음



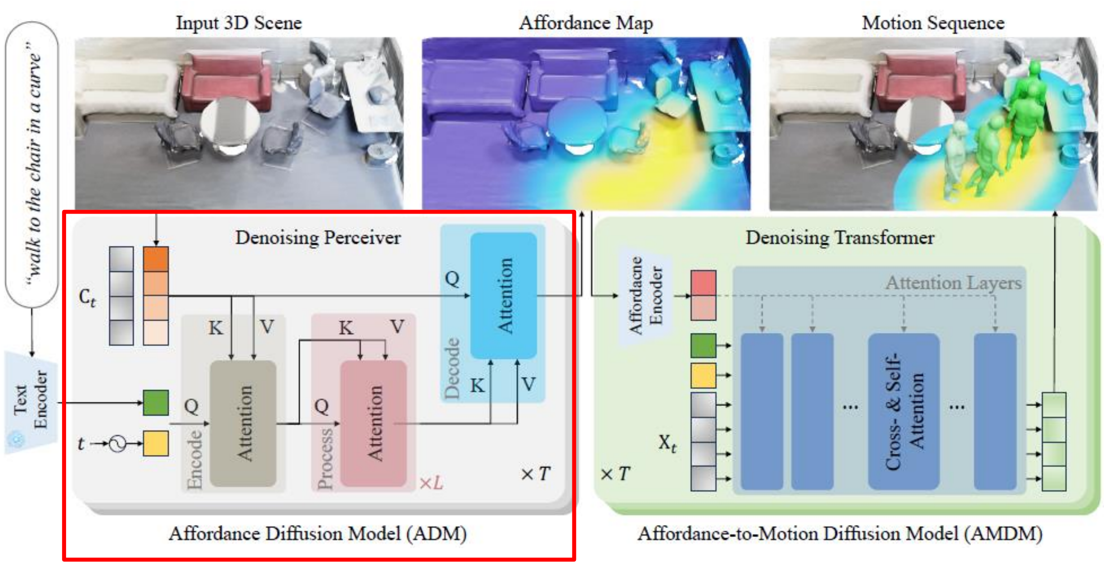
서강대학교 SOGANG UNIVERSITY
VDS LAB

Move as You Say, Interact as You Can[1)]

- Main method
  - Affordance Diffusion Model (ADM)
    - Encode, Process, Decode의 3가지 block으로 구성되어 있음
      - Encode block에서 attention을 수행하여 point feature를 추출함
      - L개의 Process block을 통해 self-attention을 수행하여 latent feature를 개선함
      - Decode block을 통해 또 다른 attention을 수행, per-point feature vector를 추출하고 아래와 같은 수식을 통해 ADM G_θ를 최적화함

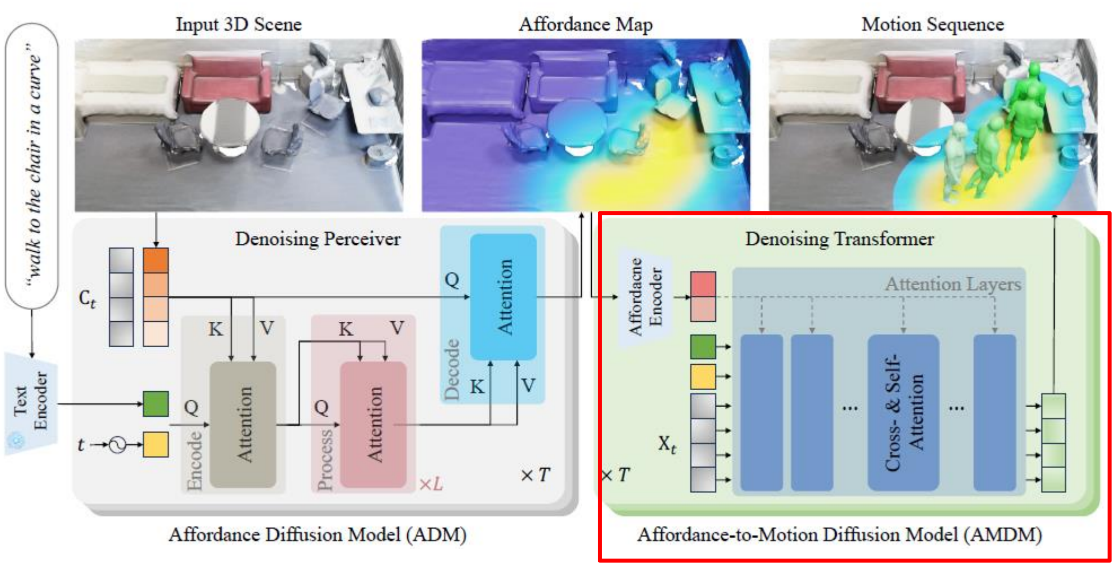
✓$L_{MSE} = E_{C_0,t}[\|C_0 - G_\theta(C_t, t, \mathcal{S}, \mathcal{L})\|_2^2]$

$\mathcal{S} \in \mathbb{R}^{N\times 6}$: RGB point cloud
$\mathcal{L} = [w_1, w_2, \dots, w_M]$: Language description



Move as You Say, Interact as You Can[1)]

- Main method
  - Affordance-to-Motion Diffusion Model (AMDM)
    - Human motion을 생성하는 단계
    - Language description과 affordance map을 사용, 최종적으로 motion sequence를 생성함
      - $p_\phi(X_{0:T}|\mathcal{C},\mathcal{S},\mathcal{L}) = p(X_T)\prod_{t=1}^{T} p_\phi\,(X_{t-1}|X_t,\mathcal{C},\mathcal{S},\mathcal{L})$
    - ADM과 유사하게 아래와 같은 수식으로 AMDM을 최적화함
      - $L_{MSE} = E_{X_0,t}[\|X_0 - G_\phi(X_t,t,\mathcal{C},\mathcal{S},\mathcal{L})\|_2^2]$



Move as You Say, Interact as You Can[1)]

- Result
  - Implementation details
    - Image, text encoder로 CLIP-VIT-B/32를 freeze하여 사용함
    - ADM: A100 GPU 2개 사용, GPU 당 batch size 64
    - AMDM: A100 GPU 4개 사용, GPU 당 batch size 32
  - Datasets
    - HumanML3D, HUMANISE
  - Evaluation metrics
    - R-precision, FID, Multimodal distribution, Diversity, Multimodality
    - Goal distance: 목표 지점까지의 거리를 나타냄
    - Average Pairwise Distance (APD): 모든 data point pair 간의 거리들의 평균
    - Contact: Human과 물체와의 물리적 접촉을 평가함
    - Non-collision: Human motion generation에서 현실적인 움직임을 평가함
    - Quality score: 생성된 motion의 품질을 평가함
    - Action score: 특정 행동이나 동작의 성과를 평가함

1) WANG, Zan, et al. Move as You Say, Interact as You Can: Language-guided Human Motion Generation with Scene Affordance. *arXiv preprint arXiv:2403.18036*, 2024.

Move as You Say, Interact as You Can[1)]

- Result

Model	R-Precision ↑ Top 1	Top 2	Top 3	FID ↓	MultiModal Dist. ↓	Diversity →	MultiModality ↑
Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
Language2Pose [3]	$0.246^{\pm.002}$	$0.387^{\pm.002}$	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	-
T2M [29]	$\mathbf{0.457^{\pm.002}}$	$\mathbf{0.639^{\pm.003}}$	$\mathbf{0.740^{\pm.003}}$	$1.067^{\pm.002}$	$\mathbf{3.340^{\pm.008}}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
MDM [76]	$0.319^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$\mathbf{9.559^{\pm.086}}$	$2.799^{\pm.072}$
Ours	$0.341^{\pm.010}$	$0.514^{\pm.016}$	$0.625^{\pm.011}$	$\mathbf{0.352^{\pm.109}}$	$5.455^{\pm.073}$	$9.772^{\pm.117}$	$\mathbf{2.835^{\pm.075}}$
MDM† [76]	$0.418^{\pm.005}$	$0.604^{\pm.005}$	$0.707^{\pm.004}$	$0.489^{\pm.025}$	$3.631^{\pm.023}$	$\mathbf{9.449^{\pm.066}}$	$\mathbf{2.873^{\pm.111}}$
Ours†	$\mathbf{0.432^{\pm.007}}$	$\mathbf{0.629^{\pm.007}}$	$\mathbf{0.733^{\pm.006}}$	$\mathbf{0.352^{\pm.109}}$	$\mathbf{3.430^{\pm.061}}$	$9.825^{\pm.159}$	$2.835^{\pm.075}$

HumanML3D test set

Model	goal dist.↓	APD↑	contact↑	non-collision↑	quality score↑	action score↑
cVAE [84]	$0.422^{\pm.011}$	$4.094^{\pm.013}$	$84.06^{\pm.716}$	$\mathbf{99.77^{\pm.004}}$	2.25 ± 1.26	3.66 ± 1.38
one-stage @ Enc	$0.326^{\pm.013}$	$\mathbf{5.510^{\pm.019}}$	$76.11^{\pm.684}$	$99.71^{\pm.014}$	2.60 ± 1.24	3.88 ± 1.32
one-stage @ Dec	$0.185^{\pm.014}$	$4.063^{\pm.020}$	$86.43^{\pm.845}$	$99.76^{\pm.006}$	3.09 ± 1.34	4.18 ± 1.16
Ours @ Enc	$\mathbf{0.156^{\pm.006}}$	$2.597^{\pm.008}$	$95.86^{\pm.323}$	$99.69^{\pm.007}$	3.46 ± 1.15	$\mathbf{4.47 \pm 0.84}$
Ours @ Dec	$\mathbf{0.156^{\pm.006}}$	$2.459^{\pm.009}$	$\mathbf{96.04^{\pm.298}}$	$99.70^{\pm.005}$	$\mathbf{3.55 \pm 1.19}$	4.44 ± 0.85

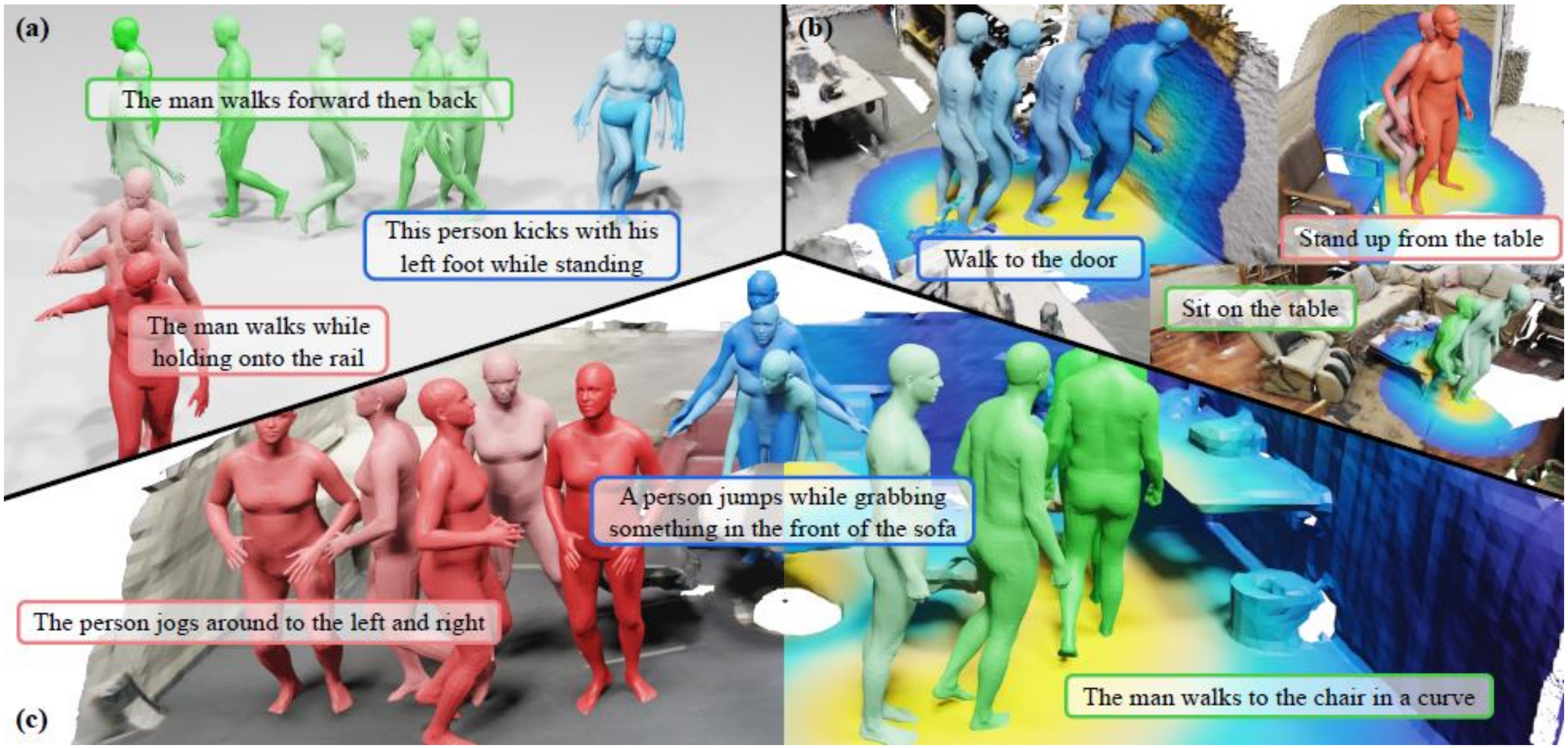
HUMANISE test set

Move as You Say, Interact as You Can[1)]

- Result

HumanML3D

HUMANISE



Unseen scenarios

감사합니다