H-008

Real-time Scene-aware Human Motion Prediction

須藤 雅基
Sudo Masanori
Kanagawa University

清藤 汐音
Kiyofuji Shion
Kanagawa University

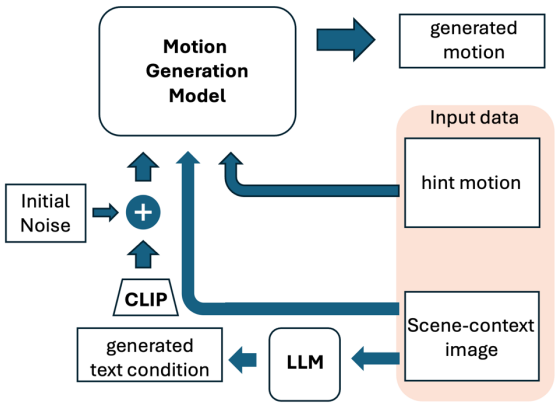
張 善俊
Shanjun Zhang
Kanagawa University

1. Introduction

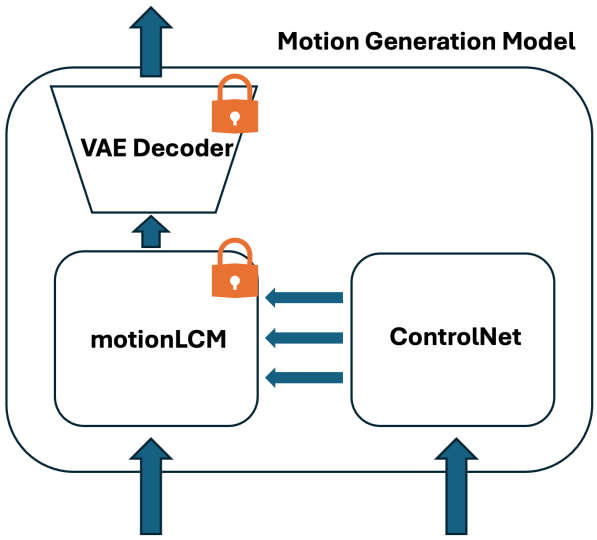
Human Motion Prediction is a task that estimates and generates future human motion based on past motion sequences. In the field of Human–Robot Interaction, accurately predicting human intent enables safe and seamless collaboration with robots. In AR/VR domains, motion prediction can be used for applications such as pre-rendering visual effects to enhance user experience.

While previous studies have proposed many methods that condition generation on text prompts or recent motion sequences, few explicitly incorporate scene context (i.e., surrounding environments). In particular, methods that rely on 3D scan data achieve high accuracy but incur high data collection costs. In contrast, Move-in-2D [1] utilizes 2D still images as scene context, successfully reducing data collection and preprocessing costs while maintaining motion generation quality.

Inspired by this work, we propose a pipeline that takes monocular video of a single person from a fixed camera as input and generates future motion by leveraging both 2D scene context and past motion sequences.



(Fig 1 . Proposed Method D)



(Fig2. Motion Generation Model Architecture)

2. Related Work

2.1 Move-in-2D [1]

Move-in-2D proposes a scene-aware motion generation method that does not require 3D scan data. Instead, it utilizes 2D scene images for conditioning. The method includes a pipeline for collecting video frames from the web and preprocessing them to build a large-scale dataset that pairs 2D key-points of the person with scene images.

1.1 Motion Latent Diffusion (MLD) [2]

Motion Latent Diffusion (MLD) is a latent diffusion model that compresses motion sequences into a latent space and applies diffusion within that space. Compared to diffusion in the original motion space, this approach suppresses high-frequency noise, resulting in improved overall generation quality and computational efficiency.

2.2 MotionLCM [3]

MotionLCM distills knowledge from MLD using Consistency Distillation to enable single-step motion generation. Whereas traditional diffusion models require dozens or hundreds of denoising steps, MotionLCM transfers the learned model into a student model that performs denoising in just 1–3 steps, achieving more than 100× faster inference.

3. Proposed Method

To incorporate scene context while maintaining high inference speed, we propose a pipeline that integrates the following components (Fig 1):

3.1 Base Model：MotionLCM

We freeze the pre-trained MotionLCM model to preserve its high-performance single-step inference capabilities.

3.2 Scene Context Injection: ControlNet

We introduce a custom-trained ControlNet that injects spatial information extracted from 2D scene images into the motion generation pipeline. This is done by combining the spatial context with hint motion (previous motion) and conditioning the generator accordingly.

3.3 High-level Motion Prediction: LLaVA-Mini

We utilize LLaVA-Mini, a lightweight multimodal LLM, to predict the next action of the person in the scene from a scene image and an instruction prompt. The resulting natural language output is then encoded using CLIP and fed into the motion generation model.

4. Dataset

We use the RICH dataset [5] in our experiments.

In addition, inspired by Move-in-2D, we attempted to reproduce a dataset construction method using 4D-Humans [4], which pairs 2D images with SMPL-based motion data. However, due to the processing time bottleneck of 4D-Humans, it was difficult to construct a sufficiently large dataset within a reasonable timeframe for training purposes.

5. Experimental Setup

To evaluate how scene context is incorporated, we compare four variations of our pipeline (Table 1). **A**: No scene context or text condition. **B**: Injects scene context via ControlNet. **C**: Uses LLM to predict future motion in text form. **D**: Combines both ControlNet-based context injection and LLM text (Fig 1).

	Scene-context Injection	Text Prompt
A	✖	✖
B	◯(ControlNet)	✖
C	✖	◯(LLM)
D	◯(ControlNet)	◯(LLM)

(Table1. Comparison among 4 pipeline configurations)

6. Evaluation Metrics

6.1 Frechet Inception Distance (FID)

We map both generated and ground-truth motion sequences into a motion latent space using an autoencoder, then compute the Frechet Distance between their distributions in the latent space.

6.2 Trajectory Error (TE)

We calculate the positional error of the pelvis keypoint between predicted and ground-truth trajectories.

7. Discussion

Mode A generates motion based solely on the initial noise and hint motion, without scene or text conditioning. However, MotionLCM adopts Class-Free Guidance (CFG), allowing it to generate realistic motions along the learned distribution even without prompts.

In Modes C and D, high-level motion prediction is provided by LLMs (LLaVA-Mini), while in Mode B, ControlNet is expected to learn and infer motion transitions directly from spatial context. This allows us to compare explicit (text-based) and implicit (scene-based) motion prediction approaches.

8. References

[1] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Gang Yu, “Motion Latent Diffusion for Efficient Human Motion Generation”, *ICCV*, 2022.

[2] Hsin-Ping Huang, Yang Zhou, Jui-Hsien Wang, Difan Liu, Feng Liu, Ming-Hsuan Yang, Zhan Zu, “Move-in-2D: 2D Scene-Conditioned Human Motion Prediction”, *CVPR,* 2025.

[3] Wenxun Dai, Long-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, Yansong Tang, “motionLCM: Consistency-Distilled Latent Diffusion for One-Step Motion Generation”, *NeurIPS*, 2023.

[4] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, Jitendra Malik, “4D-Humans: Dataset and Benchmark for Realistic Human Motion Capture”, *ICCV, 2023.*

[5] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, Michael J. Black, “Capturing and Inferring Dense Full-Body Human-Scene Contact”, *CVPR*, 2022.