A. SMPL-X Joints to Meshes

We represent the per-frame human pose using SMPL-X [62] body joints, denoted as $\mathbf{x}_i \in \mathbb{R}^{J\times 3}$. Among the available 55 SMPL-X joints, we utilize $J = 22$ joints, excluding those related to the hands, jaw, and eyes. To facilitate visualization and compute physical metrics, we transform the SMPL-X joint positions of motion sequences into meshes through a two-stage optimization process. In the first stage, we employ a pre-trained transformer-based neural network to map the joint sequence to the corresponding SMPL-X parameter sequence, providing an initialization for the subsequent stage. Then, in the second stage, we optimize the SMPL-X parameters using an MSE loss function, aiming to minimize the discrepancy between joints derived from the optimized SMPL-X parameter and the generated joints.

B. Model Architectures

B.1. Details of ADM and AMDM

ADM We first employ a pre-trained Point Transformer [99], which is pre-trained on semantic segmentation task, to extract per-point features from the input 3D scene. We then concatenate the noisy affordance map, per-point feature, and point coordinates; forward the concatenation into the backbone of ADM. We use $J = 6$ to compute the affordance map and extract per-point features with a dimensionality of 32, containing joints of the pelvis, left/right hand, left/right foot, and neck.

AMDM For AMDM, we have implemented both a decoder and an encoder variant for comparison. The main difference between the two architectures is the approach used for fusing affordance features. As shown in Fig. A1, the decoder variant utilizes cross-attention to attend with affordance features, while the encoder variant directly employs self-attention for the fusion process. We utilize the Pytorch implementation of `TransformerEncoderLayer` and `TransformerDecoderLayer`, configuring the latent space dimension to be 512.

B.2. Architecture of ADM Variants

MLP The MLP variant consists of two "SceneAffordMLP" blocks, as depicted in Fig. A2a. In each "SceneAffordMLP" block, we first process the input via a shared MLP layer. Subsequently, a max-pooling layer aggregates information from all points to extract the global feature. The per-point features are then concatenated with the global feature and fed into another shared MLP layer. This model directly operates on the concatenation of per-point features, language features, timestep embedding, and noisy affordance.

Point Transformer The Point Transformer variant employs a U-Net architecture that comprises an encoder and a decoder for extracting point-wise features, as shown in Fig. A2b. The feature encoder comprises four stages designed to gradually downsample the point cloud with transition-down blocks and aggregate point features with point transformer layers. The downsampling rates at each stage are specified as $[1, 4, 4, 4]$, resulting in point cloud cardinalities of $[N, N/4, N/16, N/64]$ after each stage. The decoder, incorporating transition-up blocks and point transformer layers, maps the encoded features onto a higher-resolution point set. This is achieved by concatenating the trilinear-interpolated features from the previous decoder stage with features from the corresponding encoder stage through a skip connection. To enable language-conditioned modeling, we enhance this connection by incorporating a linear layer that fuses language and point features. Ultimately, the per-point feature vectors are forwarded into a linear layer.

C. Implementation Details

C.1. Baseline Models for Motion Generation

cVAE We adopt the model architecture and hyperparameters as proposed by Wang et al. [84] without any alterations. We avoid utilizing the suggested auxiliary loss functions to ensure a fair comparison.

One-stage Baselines Our one-stage baselines directly adopt the architecture illustrated in Fig. A1. Unlike AMDM, the models take input as the scene point coordinates and semantic features instead of the affordance map. We extract the per-point semantic features using a Point Transformer, which is pre-trained on a semantic segmentation task.

C.2. Training Details

In all experiments, we fix the diffusion step of ADM as 500. In the experiments on the HumanML3D dataset, we set the diffusion step of AMDM as 1000, following the MDM [76]. Given the presence of only a floor within the scene in HumanML3D, we opt not to utilize the pre-trained Point Transformer to extract semantic features in ADM. **To mitigate overfitting of the generated results to ground truth scene affordances, we randomly replace half of the ground truth affordances (*i.e.*, 50% proportion) with predicted ones.** We conduct an ablative experiment on the proportion and report the results in Tab. A1. For the HUMANISE benchmark, we set the diffusion step of AMDM as 500, and we directly use the ground truth affordance map during the training in the second stage. We augment each scene point cloud by applying random rotation around the z-axis. For the experiments on our novel evaluation set, we set the diffusion step of ADM and AMDM as 500 and 1000, respectively. We directly utilize the ground truth affordance map in the second-stage training. Tab. A2 shows the ablative results of replacing the ground truth affordance with predicted ones using different proportions.

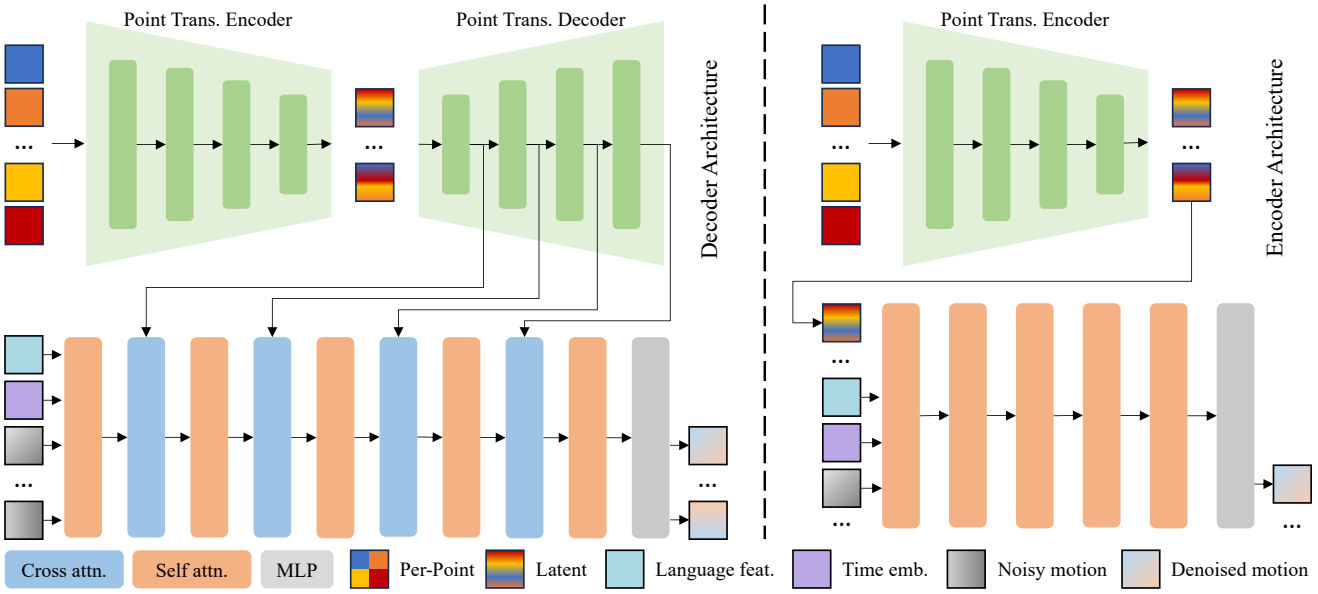
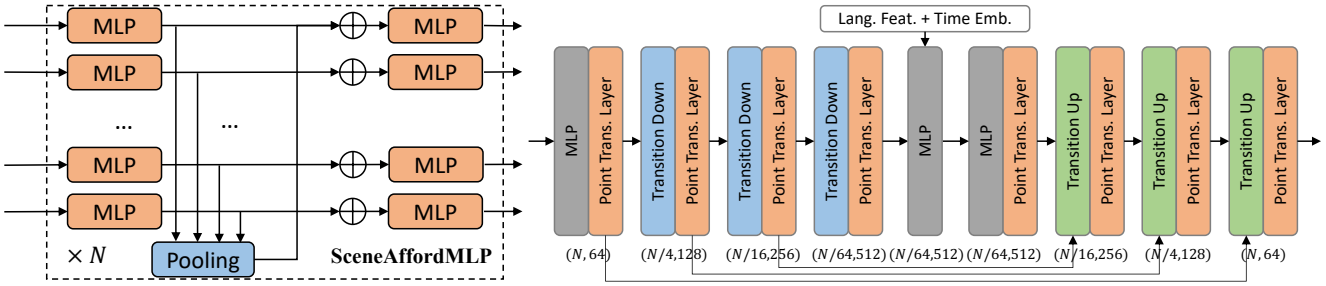


Figure A1. **Illustration of the decoder and encoder variants' architectures.** The left part depicts the architecture of the decoder variant, which stacks self-attention and cross-attention layers alternately to fuse multi-modal conditions effectively. The right part showcases the design of the encoder variant, employing self-attention layers to fuse the language features, affordance features, and noisy motion sequences.



(a) MLP variant (b) Point Transformer variant

Figure A2. **Illustration of MLP and Point Transformer variants of ADM.**

C.3. Data Preprocessing

We curate a dataset that connects language, 3D scene, and motion by incorporating data from several sources, *i.e.*, HUMANISE, HumanML3D, and PROX. For HUMANISE and PROX containing scene context, we crop the scene into $4 \times 4\text{m}^2$ chunks according to the motion range. For HumanML3D, we re-process the data without normalizing the orientation of the first frame pose and randomly add the floor and $3 \sim 4$ furniture scans around the motion sequence. We use scans from ScanObjectNN [79], including the categories of "table", "chair", "bed", "desk" "sofa" "shelf", "door", and "toilet". For each scene, we downsample the point cloud into 8192 points.

C.4. User Study

We performed human perceptual studies to assess the generated results on both HUMANISE and our novel evaluation set. We randomly generated 20 samples for each model and presented them in a disarranged manner. We asked 12 workers to score each sample on a scale from 1 to 5 for the *quality* and *action* score. A higher score indicates a better result. The *quality* score reflects the overall quality of the generation, while the *action* score evaluates the consistency of the generated motions with the provided descriptions.

D. More Qualitative Results

We present the qualitative results on the HumanML3D dataset in Fig. A3. Please visit our project page, where you can find rendered videos showcasing more qualitative results.

E. Novel Evaluation Set

We establish a novel evaluation set where both the scene and language descriptions are unseen during the training. We visualize some evaluation cases in Fig. A4.

Table A1. **Ablation of the proportion about replacing ground truth affordance with predicted ones on HumanML3D.** 0% indicates we train AMDM using ground truth affordance, 100% indicates we use the predicted affordance, and 50% indicates we randomly replace half of the ground truth affordance with predicted ones. We use the Perceiver in the first stage and the encoder-based variant in the second stage.

Proportion	R-Precision ↑			FID ↓	MultiModal Dist. ↓	Diversity →	MultiModality ↑
	Top 1	Top 2	Top 3				
Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
0%	$0.291^{\pm.014}$	$0.434^{\pm.013}$	$0.528^{\pm.016}$	$2.482^{\pm.477}$	$4.784^{\pm.103}$	$8.986^{\pm.103}$	$4.123^{\pm.046}$
50%	$0.432^{\pm.007}$	$0.629^{\pm.007}$	$0.733^{\pm.006}$	$0.352^{\pm.109}$	$3.430^{\pm.061}$	$9.825^{\pm.159}$	$2.835^{\pm.075}$
100%	$0.415^{\pm.010}$	$0.599^{\pm.013}$	$0.703^{\pm.010}$	$0.537^{\pm.218}$	$3.574^{\pm.078}$	$9.730^{\pm.093}$	$3.241^{\pm.042}$

Table A2. **Ablation of the proportion about replacing ground truth affordance with predicted ones on our novel evaluation set.** The proportion ranges from 0.0 to 0.5. We use the Perceiver in the first stage and the encoder-based variant in the second stage.

Proportion	R-Precision ↑			FID ↓	MultiModal Dist. ↓	Diversity →	MultiModality ↑	contact↑	non-collision↑
	Top 1	Top 2	Top 3						
Real	$0.588^{\pm.006}$	$0.784^{\pm.003}$	$0.875^{\pm.002}$	$0.000^{\pm.000}$	$3.342^{\pm.004}$	$9.442^{\pm.301}$	-	-	-
0%	$0.205^{\pm.054}$	$0.343^{\pm.056}$	$0.478^{\pm.069}$	$7.887^{\pm1.189}$	$6.226^{\pm.261}$	$7.935^{\pm.857}$	$5.159^{\pm.356}$	$71.98^{\pm2.542}$	$99.83^{\pm.006}$
30%	$0.238^{\pm.017}$	$0.358^{\pm.023}$	$0.488^{\pm.026}$	$11.457^{\pm1.219}$	$5.896^{\pm.109}$	$8.012^{\pm.378}$	$4.786^{\pm.249}$	$34.03^{\pm.661}$	$99.89^{\pm.024}$
50%	$0.253^{\pm.037}$	$0.415^{\pm.045}$	$0.500^{\pm.042}$	$13.354^{\pm1.281}$	$5.747^{\pm.203}$	$7.976^{\pm.487}$	$4.649^{\pm.337}$	$30.54^{\pm2.296}$	$99.92^{\pm.016}$

A man squats deeply three times while raising both arms in the air as if holding a dumbell

The person is walking forward and then back the other direction

A person jogs forward and semi circles around to the left and then to the right

The person walks in a clockwise circle

A person jumps from side to side right to left

A person waves with his left hand

Figure A3. **Qualitative results on HumanML3D.** In each case, the left figure illustrates the generated motion and the generated affordance maps are depicted in the two figures on the right. The top and bottom figures correspond to the pelvis and left foot joints, respectively.

Evaluation Details Employing joint position representation, we re-split the HumanML3D dataset and follow Guo et al. [29] to utilize the training set to train feature extractors, which supports the computation of metrics like *R-Percision*, *FID*, *MultiModal Dist.*, *Diversity*, and *MultiModality*. For *R-Precision* computation, we form a description pool with one ground truth and 15 randomly selected mismatched descriptions. To evaluate the model's performance on our established novel evaluation set, we use the HumanML3D test set to compute these metrics as a reference (*i.e.*, "Real"). It's worth noting that our evaluation set lacks ground truth motion and differs from the HumanML3D dataset regarding description distribution, *i.e.*, utterances in HumanML3D include fewer scene-related descriptions.

1. a man appears to place a plate on the table

2. a person crouches down to pick up something on the ground and then stands up

3. a person walks to the door and tries to close the door

4. a person cautiously moves while holding a drink in his left hand in the room

5. a person walks to the table and leans against it

1. A person walks slowly backward in the room

2. A man lifts his right hand and points to the table

3. A man is playing violin while sitting on the sofa

4. A man kicks the table with his right foot

5. A man walks to the chair and sits on it

1. a person sits on one of the chairs

2. a person dances in the room gracefully and happily

3. a man kneels on the floor with both of his knees

4. a person is moving their arm around

5. a person sidesteps facing forward and then stands up again

1. a person is pretending to be a chicken in the room

2. a person goes straight and turn left

3. a person walks forward and waves his right arm

4. a person picks an object up off the floor with his left hand

5. a person takes two steps towards the table

Figure A4. **Example cases in novel evaluation set.**