GenM^3: Generative Pretrained Multi-path Motion Model for Text Conditional Human Motion Generation

Supplementary Material

Table 7. Dataset statistics. We calculate the total number of frames and duration and the number of textual descriptions for each dataset.

Dataset	Frame(M)	Duration(h)	Text
BABEL [28]	2.25	20.84	9,742
BEATv2 [20]	9.43	87.29	–
CIRCLE [1]	1.07	9.91	–
EgoBody [44]	0.40	3.70	–
GRAB [30]	0.40	3.74	–
HIMO [21]	1.14	10.59	2,711
HumanML3D [10]	6.09	56.42	29,226
HuMMan [3]	0.69	6.35	6,264
IMHD [46]	0.13	1.21	308
MPI-INF-3DHP [23]	0.13	1.22	–
InterHuman [18]	2.02	18.69	–
Total	23.75	219.96	48,251

In this supplementary material, we provide more information that could not be included in the main manuscript because of space limit. We first provide more details about our constructed dataset and implementations in Sec. 8 and 9. More results and ablation studies are conducted in Sec. 10 Sec. 11 shows the visualization results of generated motion corresponding to different text inputs. Finally, we discuss limitations of our proposed GenM^3 in Sec. 12.

8. Details of the Dataset

We collected and processed 11 high-quality motion capture datasets, including CIRCLE [1], Egobody [44], GRAB [30], HIMO [21], IMHD [46], BABEL [28], HumanML3D [10], HuMMan [3], MPI-INF-3DHP [23], InterHuman [18] and BEATv2 [20]. All datasets were processed at 30 fps, with the duration of each sample ranging from 2 to 10 seconds. Tab. 7 illustrates the number of frames included in each dataset after processing.

Pose Representation We standardized all datasets to align with the format of the HumanML3D dataset. A pose p is represented as a tuple consisting of multiple components: $(r^a, r^x, r^z, r^y, \mathbf{j}^p, \mathbf{j}^v, \mathbf{j}^r, \mathbf{c}^f)$. Here, $r^a \in \mathbb{R}$ represents the root angular velocity along the Y-axis, while $(r^x, r^z \in \mathbb{R})$ describe the root's linear velocities on the XZ-plane. The root height is denoted as $r^y \in \mathbb{R}$. The joint-related attributes include $\mathbf{j}^p \in \mathbb{R}^{3j}, \mathbf{j}^v \in \mathbb{R}^{3j}, \mathbf{j}^r \in \mathbb{R}^{6j}$, which respectively correspond to the positions, velocities, and rotations of joints in the root space, where j indicates the total number of joints. Additionally, $\mathbf{c}^f \in \mathbb{R}^4$ is a set of binary features, derived by thresholding the velocities of the heel and toe joints to highlight ground contact points.

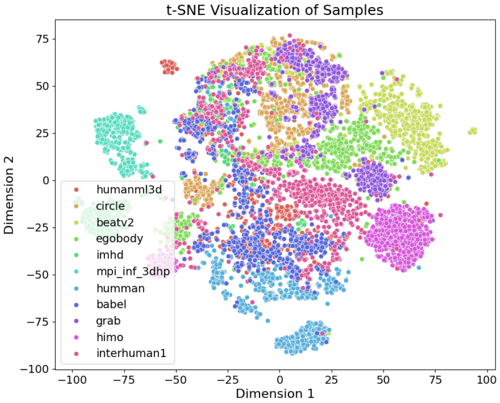


Figure 8. Visualization of data distribution after dimensionality reduction using T-SNE algorithm.

Dataset Distribution We randomly sampled 1,000 motion sequences from each dataset and reduced their dimensionality to two using the T-SNE algorithm. Fig. 8 presents a visualization of the data distribution after dimensionality reduction, which clearly demonstrates significant differences in data distribution across the datasets.

Text Label HumanML3D, HuMMan, and HIMO datasets provide motion sequences paired with corresponding text labels. For these datasets, we adhered to their original splits and utilized the provided textual annotations. HDMI is a dataset focused on human-object interaction. For this dataset, we manually segmented the motion sequences based on video content. During the text annotation process, we described the human motions and the specific interactions with objects in detail. For example: *"The person is holding a pan with their left hand at chest level and a spatula with their right hand, repeatedly stirring. Simultaneously, they sway their body from side to side, with the left foot half a step ahead of the right, and their head looking at* *the pan.”*

The BABEL dataset provides action category information for each subsequence. To process this dataset, we first divided the overlapping subsequences into motion segments ranging from 2 to 10 seconds in duration based on their lengths. Subsequently, we used ChatGLM [7] to automatically generate a concise textual description of the actions in each sequence, arranging the action categories in temporal order. The prompt for ChatGLM was as follows:

The user will give you a sequence of words. Please use these words to form one or a few concise sentences in English to describe the person’s actions. Do not add any unrelated descriptions or overly elaborate embellishments. For example, if given the words walk, scoop, place, turn, walk, output: The person is walking over, scooping up the item, placing it down, turning, and walking away. If the direction is not clear, do not add directional modifiers, but if direction words like ”turn” are present, you can add directional descriptions, such as ”walk over” or ”walk away.”

This process ensures accurate and consistent textual descriptions across datasets.

9. Implementation Details

Our model is implemented using PyTorch 2.0 and trained on Nvidia RTX-4090 GPUs. The architecture consists of two main components: MEVQ-VAE and Multi-path Motion Transformer. In MEVQ-VAE, motion sequences are downsampled by a factor of four during discretization, reducing temporal resolution. Both the encoder and decoder employ a multi-expert architecture with a default of four experts. The codebook is designed with 8,192 entries, each of 32 dimensions, providing a rich discrete representation space for motion features. To address codebook collapse, where large codebooks risk underutilization, we adopt a factorized code approach [38], decoupling code lookup and embedding, and use moving averages for updates while resetting inactive codes to enhance utilization. During training, the loss weight parameter β is set to 1, balancing reconstruction and commitment losses to ensure effective encoding.

The Multi-path Motion Transformer comprises of 18 layers: the first nine layers utilize standard attention mechanisms for initial feature extraction, while the latter nine layers implement a multiway transformer with multiple experts. All attention computations use 16 heads, each with a dimension of 64. For both pretraining and text-conditional training, we use a batch size of 160 and the AdamW optimizer, training the model for 120,000 iterations. The learning rate is set to 0.0002, following a warm-up and cosine annealing decay strategy to facilitate rapid adaptation in early training.

Table 8. Inference speed per sample (on a RTX4090 GPU) with different iteration settings during masked decoding.

Method	FID	Infer. time (ms)	R-Precision	Diversity
GenM^3 (iter.=1)	2.242	**23.73**	0.679	8.477
GenM^3 (iter.=5)	0.054	67.49	0.785	9.375
GenM^3 (iter.=10)	**0.046**	113.22	0.804	9.675
GenM^3 (iter.=15)	0.060	188.12	**0.805**	**9.719**
T2M-GPT	0.160	108.03	0.770	9.653

Table 9. Results on HumanML3D under varying data proportions.

Data Proportion	0%	35%	70%	100%
FID	0.071	0.060	0.050	0.046

Table 10. Results on HumanML3D (annotated by FineMotion).

Method	FID	R.Top1	R.Top2	R.Top3	MMDist.	Diversity
MMM	24.7	0.060	0.113	0.159	5.85	6.01
GenM^3	17.2	0.072	0.135	0.191	5.30	6.00
GenM^{3*}	**12.9**	**0.087**	**0.154**	**0.218**	**5.01**	**6.25**

10. More Results

Inference Speed During decoding, all tokens are decoded in parallel at each iteration. Tab. 8 shows how iteration count (10 for default settings) relates to inference time, and compares this with the AR model T2M-GPT [40].

Influence of Dataset Size We have evaluated our backbone pre-training on varying proportions of the dataset (HumanML3D dataset is fully used). We can conclude that the quality of the generation scales with the size of dataset (see Tab. 9). We believe that integrating additional motion data can unlock further model potential, which will be the focus of our future work.

Zero-shot Generation We added comparisons on unseen HumanML3D annotations from FineMotion[1] (see Tab. 10) to further prove GenM^{3*}’s generalization ability. Since the re-annotated descriptions differ significantly from those in the original HumanML3D, the zero-shot generation performance is generally not ideal. However, our method achieves better results compared to MMM [27]. By comparing the results of GenM^{3*} and GenM^3, it can be concluded that training on a larger dataset (with more text annotations and more motion data) improves the model’s generalization ability.

11. Visualization

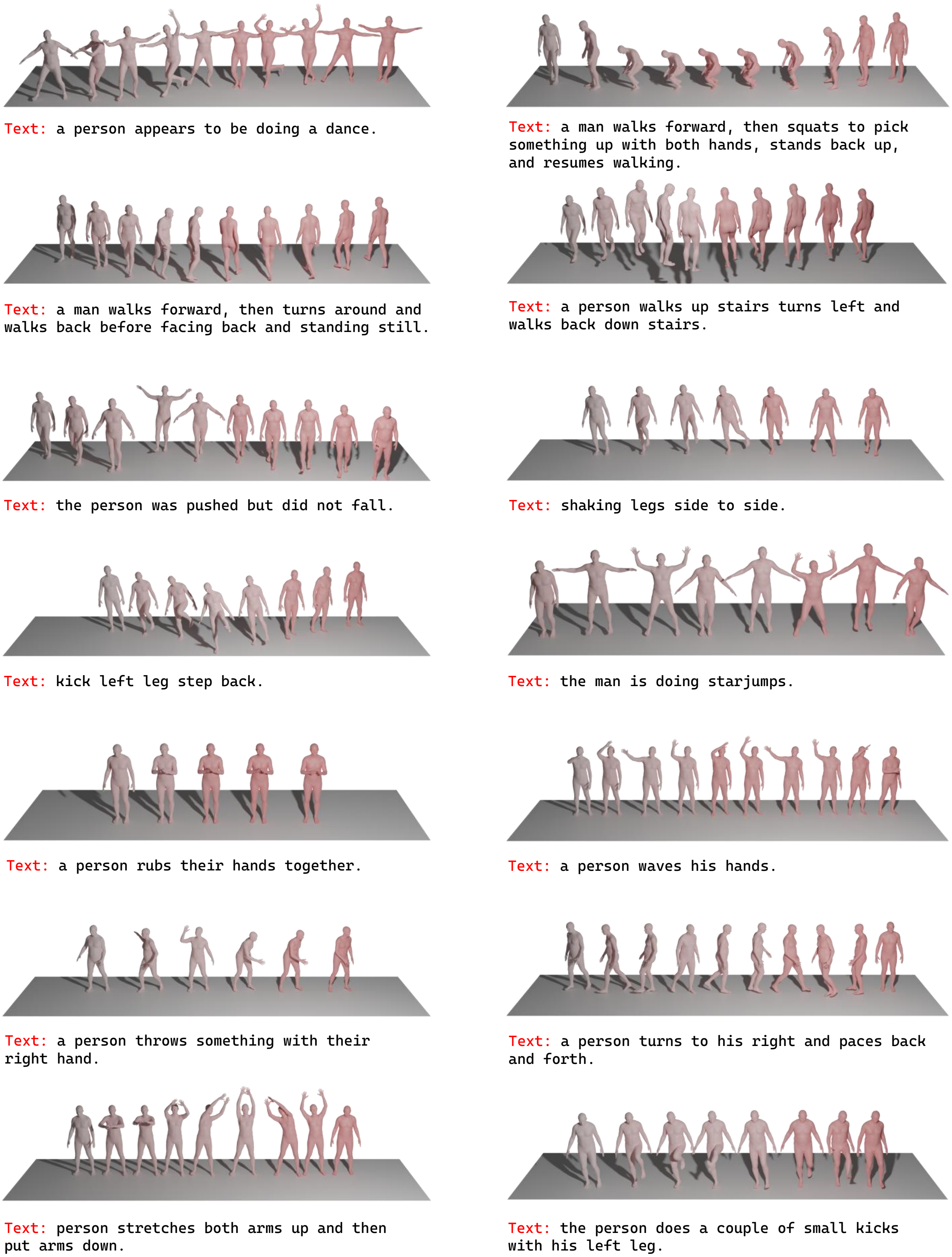
We provide more qualitative results from motion generation experiments using text inputs from the HumanML3D

[1] https://github.com/BizhuWu/FineMotion

test set (Fig. 9), the HuMMan test set (Fig. 10), and GPT-generated textual descriptions (Fig. 11).

12. Limitation

Our approach has two main limitations. First, although we enhanced the dataset by adding text labels for various motion types, the number of these labels remains limited. As a result, our method struggles to generate accurate motions for descriptions that fall outside the dataset's text distribution. In the future, we plan to explore ways to leverage additional text-motion pair data or integrate video-text pairs to enable the model to better comprehend diverse textual descriptions. Second, our current approach focuses primarily on body motion generation, overlooking the generation of full-body motions, such as fine-grained movements of fingers and facial expressions. In future work, we plan to collect and process more comprehensive datasets to enable MMGPT to generate more detailed and precise motions for all human joints.



Text: a person appears to be doing a dance.

Text: a man walks forward, then squats to pick something up with both hands, stands back up, and resumes walking.

Text: a man walks forward, then turns around and walks back before facing back and standing still.

Text: a person walks up stairs turns left and walks back down stairs.

Text: the person was pushed but did not fall.

Text: shaking legs side to side.

Text: kick left leg step back.

Text: the man is doing starjumps.

Text: a person rubs their hands together.

Text: a person waves his hands.

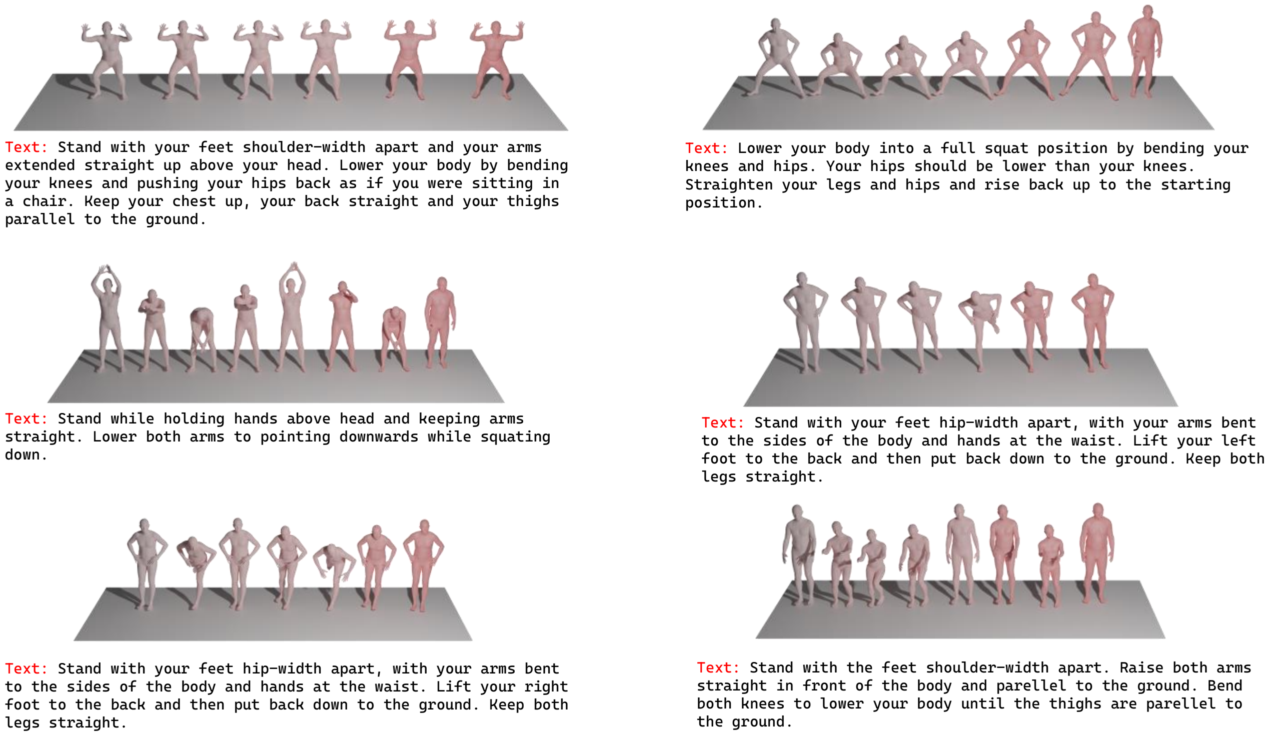
Text: a person throws something with their right hand.

Text: a person turns to his right and paces back and forth.

Text: person stretches both arms up and then put arms down.

Text: the person does a couple of small kicks with his left leg.

Figure 9. Visualizations of generated motion samples on HumanML3D [10] test set.



Text: Stand with your feet shoulder-width apart and your arms extended straight up above your head. Lower your body by bending your knees and pushing your hips back as if you were sitting in a chair. Keep your chest up, your back straight and your thighs parallel to the ground.

Text: Lower your body into a full squat position by bending your knees and hips. Your hips should be lower than your knees. Straighten your legs and hips and rise back up to the starting position.

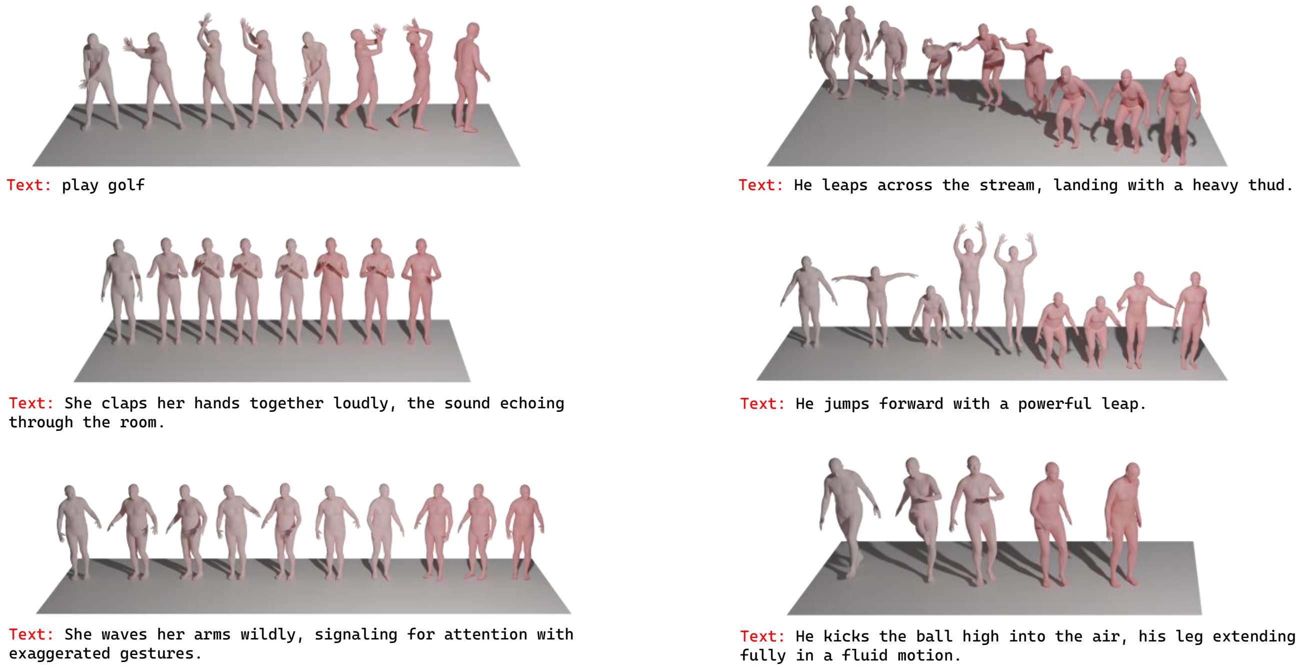
Text: Stand while holding hands above head and keeping arms straight. Lower both arms to pointing downwards while squating down.

Text: Stand with your feet hip-width apart, with your arms bent to the sides of the body and hands at the waist. Lift your left foot to the back and then put back down to the ground. Keep both legs straight.

Text: Stand with your feet hip-width apart, with your arms bent to the sides of the body and hands at the waist. Lift your right foot to the back and then put back down to the ground. Keep both legs straight.

Text: Stand with the feet shoulder-width apart. Raise both arms straight in front of the body and parellel to the ground. Bend both knees to lower your body until the thighs are parellel to the ground.

Figure 10. Visualizations of generated motion samples on HuMMan [3] test set.



Text: play golf

Text: He leaps across the stream, landing with a heavy thud.

Text: She claps her hands together loudly, the sound echoing through the room.

Text: He jumps forward with a powerful leap.

Text: She waves her arms wildly, signaling for attention with exaggerated gestures.

Text: He kicks the ball high into the air, his leg extending fully in a fluid motion.

Figure 11. Visualizations of generated motion samples on GPT-generated textual descriptions.