Go to Zero: Towards Zero-shot Motion Generation with Million-scale Data Supplementary Material

1. Overview

In the following, we provide additional implementation details, including the motion representations in Sec. 1.1, data distribution and prompts in Sec. 2, finegrained scoring criteria in MotionMillion-Eval in Sec. 3, and 126 prompts used by MotionMillion-Eval in Sec. 4. Various generations of out-domain and complex compositional long motions are shown in our demo video.

1.1. Motion Representations

We introduce our motion representation. To mitigate errors introduced by the inverse kinematics process in the HumanML3D format while preserving redundant information (e.g., velocity), we reformulate and refine the motion representation x^i in a manner consistent with previous work on character control [3–5]. Specifically, the i-th pose x^i is defined as a tuple comprising: root linear velocities $(\dot{r}^x, \dot{r}^z \in \mathbb{R})$ on the XZ-plane, root angular velocity $\dot{r}^a \in \mathbb{R}^6$ represented in 6D rotations, local joint positions $p^i \in \mathbb{R}^{3N}$, local velocities $v^i \in \mathbb{R}^{3N}$, and local rotations $r^i \in \mathbb{R}^{6N}$ relative to the root space, where N denotes the number of joints. Formally, this is expressed as:

$$x^i = \{\dot{r}^x, \dot{r}^z, \dot{r}^a, p^i, v^i, r^i\}.$$

A significant advantage of our representation is that it eliminates the need for an inverse kinematics process to obtain SMPL or BVH representations, as required in previous approaches. Moreover, our representation can be losslessly converted to relative rotations akin to those in SMPL. Additionally, because both the rotation and position components are derived from the same skeletal structure, they provide mutual regularization. Notably, the rotation component in the HumanML3D format is erroneous [1], which heavily hinders the applications of downstream tasks. Rather than discarding this flawed rotation component as done in [2], we undertake engineering corrections to rectify it. We hope our representation could correct previous mistakes and guide future development.

2. Data Distribution and Prompts

We further demonstrate the data distribution of motion length, motion velocity, and motion diversity in Fig. 1. We also provide the prompt used during captioning the motions in the web-scale human videos and text rewrite in the inference stage in Fig. 2.

3. Scoring Criteria Details of MotionMillion-Eval

The scoring criteria details for each dimension (Text Alignment, Motion Smoothness, and Physical Plausibility) are defined as follows:

Text Alignment (TA). Score = 4: The generated motion is fully aligned with the textual prompt, accurately depicting all specified elements and details. Score = 3: The motion generally corresponds to the prompt, though minor discrepancies may be present in certain details. Score = 2: The motion exhibits clear misalignment with the prompt, with significant omissions or deviations from the described content. Score = 1: The generated motion is entirely inconsistent with the prompt, displaying substantial inaccuracies in key scenes or actions.

Motion Smoothness (MS). Score = 4: The motion is highly fluid and natural, with smooth and seamless transitions between movements. Score = 3: The motion is generally smooth, though minor unnatural artifacts may occasionally appear in specific segments. Score = 2: The motion lacks fluidity, exhibiting noticeable discontinuities or stuttering. Score = 1: The motion appears highly unnatural, with frequent stuttering and abrupt transitions that disrupt coherence and comprehensibility.

Physical Plausibility (PP). Score = 4: The generated motion adheres to real-world physical laws, accurately simulating object interactions, lighting, shadows, and collision effects. Score = 3: Multiple instances of physically implausible motion, lighting inconsistencies, or unrealistic interactions are observed, though the primary actions maintain a degree of coherence. Score = 2: The generated motion exhibits substantial violations of physical laws, with unrealistic object interactions or lighting effects that diminish realism. Score = 1: The motion is entirely implausible, fea-

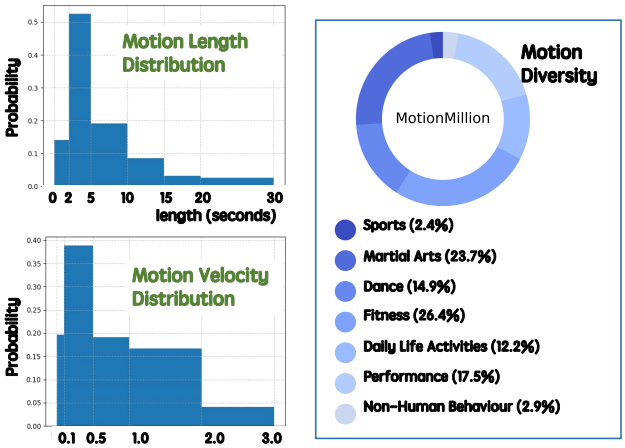

Figure 1. Data Distributions of MotionMillion.

Caption Prompt

[Task Overview]: You will be shown different frames uniformly sampled from a human motion video, along with a general description of the video. Based on these, identify the actions performed by the person within the bounding box and provide a detailed description of their movements.
[Required Information]: **1.** Body Part Involvement: Describe the main limbs involved in the movement. **2.**Action Sequence: Provide a clear, coherent sentence structure that captures the sequence of movements, including joints, body parts, and their logical relationships. **3.**Temporal Progression: Focus on how different body parts move over time in relation to one another. Use appropriate transitions to show the movement sequence.
[Optional Information]: **1.** Physical Characteristics: Age, body shape, and gait of the person, if relevant. **2.**Movement Range or Style: Describe the range or style of movement, e.g., "walking like an elderly person" or "swift, like a martial artist." **3.**Scene Description: Location context, e.g., "indoor," "outdoor," etc. 4.Emotional or Attitudinal Cues: Any emotions or attitudes expressed, such as "appears focused and determined."
The descriptions should be concise, accurate, and based on the provided video frames and the overall scene description.

Augmentation Prompt

You are a helpful assistant. You can rewrite the following sentence and do not change the meaning of the sentence. You can change the subject and the way of description, but do not alter the original meaning.
[Requirements:] **1.** Do not change the meaning of the sentence. **2.** Does not change the logical relationship of the sentences. **3.** The rewritten sentence should be more concise and clear, and write in a pair of brackets. Example: Given the sentence: A man is walking forward for a few steps. The rewritten sentence is: {{A man is walking starightly.}}
Now rewrite the following sentence for 20 times: {sentence}

Figure 2. Prompt used during captioning the motions in the web-scale human videos and text rewrite in inference stage.

turing severe distortions in object dynamics, lighting, or interactions, making the scene difficult to interpret.

4. Prompts in MotionMillion-Eval

. We show all the prompts in the figures below.

References

[1] GitHub discussion. rotation discussion. https://github.com/EricGuo5513/HumanML3D/issues/26, 2023. 2023-03-04.

[2] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation. *arXiv preprint arXiv:2411.16575*, 2024.

[3] Yi Shi, Jingbo Wang, Xuekun Jiang, Bingkun Lin, Bo Dai, and Xue Bin Peng. Interactive character control with auto-regressive motion diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–14, 2024.

[4] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Transactions on Graphics*, 38(6):178, 2019.

[5] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (ToG)*, 41(4):1–13, 2022.

An obese middle-aged male security guard, walking and looking around	Clap hands	Clap once
The boy clutched the flowers tightly with both hands, hiding them behind his back, his body taut and eyes fixed ahead. As the girl approached, his face instantly lit up with a smile, and he quickly strode forward, swiftly extending his right hand to offer the flowers.	Drink water	Turn left
The man stood in the downpour. Upon seeing the woman, his head jerked to the left with a pained look of avoidance, and his body leaned back slightly.	Open the door	Bow lightly
The woman approached the man. Her right hand trembled as she raised it, intending to touch his face.	Snap fingers	Kick forward
The reserved man and the shy woman sat side by side on the edge of the bed. After a furtive glance at the woman, the man coughed lightly, his body gradually edging closer. His hands rubbed on his knees, his movements somewhat awkward.	Bow	Twirl hair
The woman's hands were intertwined, nervously twisting the corner of her clothing, her gaze cast downward.	A teenager jogs slowly in the park, occasionally checking the phone	Knock twice
In the cluttered kitchen, the slightly plump man frantically flipped the pan with his right hand while his left hand scrambled to grab the seasonings. After knocking over the salt shaker, oil splattered in the pan. His eyes widened, his mouth gaped open, and he jumped on the spot with both feet. The spatula flew out of his hand, and he stumbled, looking utterly disheveled.	A young woman quickly folds clothes in a messy living room	Wave goodbye
The boss was speaking in a measured and serious manner, with his left hand behind his back and his right hand occasionally gesturing. Suddenly, upon hearing a sound of flatulence, his expression changed abruptly, and his right hand froze in mid-air.	Pick up a pen from the floor and place it on the table	Spin around
A man of average build who looked lost was walking along the street when a giant pie suddenly hit his head. He clutched his head with both hands and squatted down.	A furious swordsman grips his blade tightly, stomps forward with an angry roar, then slashes diagonally at an invisible foe	Toss a ball
The woman gazed out at the vast ocean, her body swaying as she slowly and heavily made her way towards the precipice.	A confident performer in a flashy costume strikes a dramatic pose, then leaps into a high-flying cartwheel across the stage	Jump high
A strong man in deep sorrow rushed to the scene, running while reaching out and shouting with a voice filled with despair, his steps faltering.	A young boy trudging through knee-deep snow, occasionally pausing to catch his breath	Snap fingers

The emaciated woman sat on the floor, her hands wrapped around her knees as she trembled, her head buried in her arms, sobbing, her body curled up.	A woman practicing yoga, gracefully transitioning from a downward dog position to a cobra pose	Drink water
The woman ran in small quick steps, gradually slowing down, standing on her tiptoes to wave, gazing into the distance with tearful eyes.	students practicing a comedic skit, overacting their gestures and laughing at each other's mistakes	Tie shoelaces
The tall and capable woman adjusted her collar in front of the mirror, took a deep breath, straightened her posture, and smiled confidently. With her right hand holding a bag and her left hand opening the door, she walked out with a light step, her posture upright.	A thin man frantically searching for his keys in a sandstorm, shielding his face with one arm and reaching around with the other	Punch forward
The athletes warmed up, their feet alternating in quick jumps, their hands clenched into fists swinging back and forth. After the whistle blew, they shot off like arrows, their feet rapidly alternating, their arms swinging with power, sprinting towards the finish line with all their might.	A middle-aged couple greeting each other warmly after a long day, wrapping their arms around each other in a tight hug	Slide left
The tall detective held a flashlight, moving slowly with a slight lean forward, turning his head to observe his surroundings. Upon spotting blood, he quickly crouched down, his left hand supporting his knee, and with his right hand, he directed the flashlight's beam, his gaze focused and thoughtful.	A zombie slowly dragging its feet forward, arms outstretched, letting out a low groan	Shake fists
The robust expedition leader held a map in his hand, occasionally looking up to observe the surrounding environment, moving slowly and cautiously. His left hand slightly raised to signal the team to halt, his right index finger pressed to his lips in a gesture for silence. His body was taut, and he strained to listen to any movements around them.	A robot spinning its torso 360 degrees, scanning the environment with glowing eyes, then extending a mechanical arm to pick up an object	Grab handle
A female college student was walking while reading a book when she was suddenly blinded by a dazzling light. She used her right hand to shield her eyes, and then was knocked to the ground by an unidentified object.	Shuffle sideways	Push door
A young girl is skipping rope in the playground.	Gently pat a dog's head	Stomp foot
An old man is slowly walking with a cane in the park.	A grandpa showing a child how to fish by a lake, casting the line then patiently waiting	Twist torso
A strong athlete is lifting heavy weights in the gym.	A teacher writing on the blackboard, pausing occasionally to ask questions	Wink quickly

A cute toddler is crawling on the floor.	Someone standing in the desert, shading their eyes from the sun and scanning the horizon	Bend knees
A middle-aged woman is practicing yoga on a mat.	A chef briskly chopping vegetables, occasionally wiping sweat from his brow	Shrug shoulders
A professional dancer is performing a ballet solo.	A woman nervously tapping her foot while waiting in line, looking at her watch repeatedly	Rest head
A basketball player is dribbling and shooting.	Flip hair back confidently, resting hand on hip	Tap foot
A construction worker is hammering nails.	A furious boxer hitting a punching bag repeatedly, sweat flying with each powerful strike	Look up
A schoolboy is running for the school bus.	A content farmer hoeing the field, wiping his forehead with the back of his hand	Crouch down
A female gymnast is doing somersaults on the balance beam.	A young woman wearing headphones, bobbing her head to the rhythm and tapping her fingers on the table	Pull rope
A waiter is carrying a tray of dishes.	A friend casually leaning against a wall, crossing one leg over the other and scrolling on their phone	Bow deeply
A skateboarder is doing tricks in the skate park.	A young couple having a heated argument in the living room, arms flailing as voices rise	Cross arms
A firefighter is climbing a ladder to rescue people.	Cautiously slide open a window, peering outside with curiosity	Duck quickly
A surfer is riding a big wave.	A man meticulously polishing his car, wiping down every inch with a cloth and stepping back to admire the shine	Rub neck
A tailor is sewing clothes with a sewing machine.	Throw a paper airplane across the room with a flick of your wrist	Scratch chin
A hunter is stalking prey in the forest.	A tall athlete performing a slam dunk on a basketball hoop, shouting in triumph	Walk slowly
A hairdresser is cutting a customer's hair.	Casually flick dust off your shoulder	Point up
A cyclist is racing in a mountain bike competition.	A florist arranging a bouquet, gently snipping stems and adjusting petals	Kick side
A pianist is playing a passionate piece on the piano.	A street performer juggling three brightly colored balls, smiling as the crowd gathers	Arch back

Jump rope	A shy child timidly stepping forward to receive an award, hands clasped together and head lowered	Greet politely
Stand still	A martial artist practicing a high roundhouse kick, exhaling sharply	Cross legs
Raise both hands	A scientist adjusting a microscope carefully, squinting into the eyepiece	Shake head
Kick a ball	A teenage boy throwing his backpack onto a couch and stretching out with a tired groan	Gather items
Wave hello	A woman sipping tea while flipping through a magazine, occasionally glancing out the window	Open door