OmniMotion-X: Versatile Multimodal Whole-Body Motion Generation

Supplementary Material

1. Dataset Text Quality

Tab. 1 presents a comprehensive analysis of the text quality across our *OmniMoCap-X* collection, which comprises over $321,000$ textual descriptions spanning various motion-related tasks. This extensive dataset was created by collecting and curating existing motion datasets, rendering motions to videos, and then leveraging vision-language models (VLMs) to generate rich textual descriptions.

To ensure the quality and accuracy of our caption generation process, we selected GPT-4o [15], the current state-of-the-art closed-source multimodal model, as our primary captioning tool. To maximize caption accuracy and comprehensiveness, we implemented several key optimizations to the captioning pipeline. First, we significantly increased the number of input frames provided to the model, allowing for more comprehensive temporal understanding of the video content. Second, we enhanced the video rendering quality to ensure that visual details are clearly preserved and accurately conveyed to the model. Third, we conducted extensive prompt engineering to design optimal instructions that guide the model toward generating more precise and detailed captions that capture both visual elements and temporal dynamics (see Fig. 1 for representative examples). These technical improvements collectively resulted in substantially improved caption quality compared to baseline approaches [5, 15].

The linguistic diversity of our dataset is evidenced by the Type-Token Ratio (TTR) ranges, with maximum values reaching 1.0 in multiple subsets (AIOZ, Motorica, IDEA400, etc.), indicating exceptional lexical richness. The average sentence length of 276.78 words across the entire collection demonstrates the descriptive depth of our motion annotations, providing detailed accounts of nuanced movements rather than simplistic action labels. Particularly noteworthy is the variety of verbs captured across different motion categories. The dataset exhibits task-specific verb distributions that align with the semantic nature of each motion type: T2M datasets frequently feature locomotion verbs ("walk," "step"), M2D collections emphasize rhythmic movements ("swing," "lift"), HOI datasets contain manipulation verbs ("lift," "hold," "grasp"), and HHI datasets capture interpersonal dynamics ("stand," "extend," "lean"). This semantic richness enables models trained on *OmniMoCap-X* to understand and generate precise motion descriptions across diverse contexts. The comprehensive coverage across six distinct motion-related tasks (T2M, M2D, S2G, HHI, HOI, HSI) makes *OmniMoCap-X* uniquely positioned to support generalizable motion understanding. By integrating multiple motion paradigms under a unified annotation framework, our dataset transcends the limitations of task-specific collections, allowing for transfer learning across motion domains. This textual quality and diversity significantly enhance the capability of motion generation models to comprehend natural language instructions and produce corresponding human-like movements.

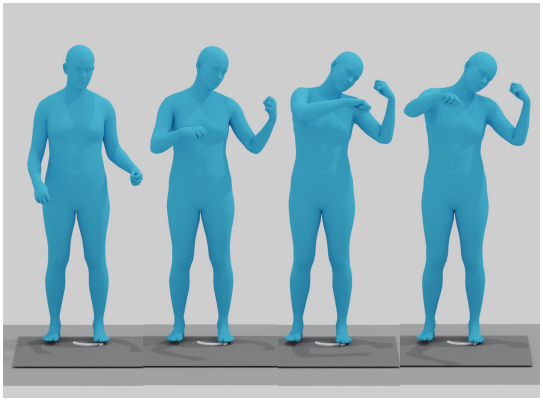
2. Dataset Preprocessing

Format Conversion. Since many datasets are not in the SMPL-series format, we convert all datasets to the SMPL-X format for consistency. For BVH-format datasets such as Mixamo [1], 100Style [29], Motorica [2], and LaFAN1 [13], we first standardize the reference pose of all BVH files to a T-pose to ensure consistent initialization. We then align the root node's coordinate system with that of the SMPL-X [30] model in Blender [7], where the negative Y-axis is defined as the forward direction and the Z-axis as the vertical upward direction. To adapt to the SMPL-X topology, we generate a corresponding BVH skeleton based on a predefined SMPL-X T-pose template. We perform skeleton retargeting in MotionBuilder [4] to map the original animation data onto the SMPL-X hierarchy. Finally, we convert the retargeted BVH files from Euler angles to axis-angle representations and apply Gaussian filtering to smooth both joint rotations and translations over time, yielding stable SMPL-X parameters. For the Choreomaster dataset [8], originally in FBX format, we first convert it to BVH using Blender, and then process it using the same pipeline.

Normalization. To reduce the difficulty of model training, we perform temporal and spatial normalization on SMPL-X motion sequences. Specifically, we standardize all motion sequences in four steps. First, we align the initial frame of each sequence to face the positive Z-axis. Then, we reposition the starting frame to a common location with the feet just touching the ground. Next, we unify the frame rate across datasets to 30 fps. Finally, we segment each motion sequence into 5-second clips (150 frames).

3. Interaction Visualization in Dataset

Our dataset supports various interaction types, including Human-Human Interaction (HHI), Human-Object Interaction (HOI), and Human-Scene Interaction (HSI), as illustrated in Fig. 2. Specifically, human-human interactions encompass a rich spectrum of social scenarios such as conversations, handshakes, embraces, and collaborative activities. Human-object interactions include everyday activities like grasping, manipulation, and tool usage. Human-scene interactions capture the dynamic relationships between in-



GT caption: person is playing the violin.

GT caption refined with GPT-4o: The person gracefully draws the bow across the strings of the violin... **His fingers dance deftly along the fingerboard** …

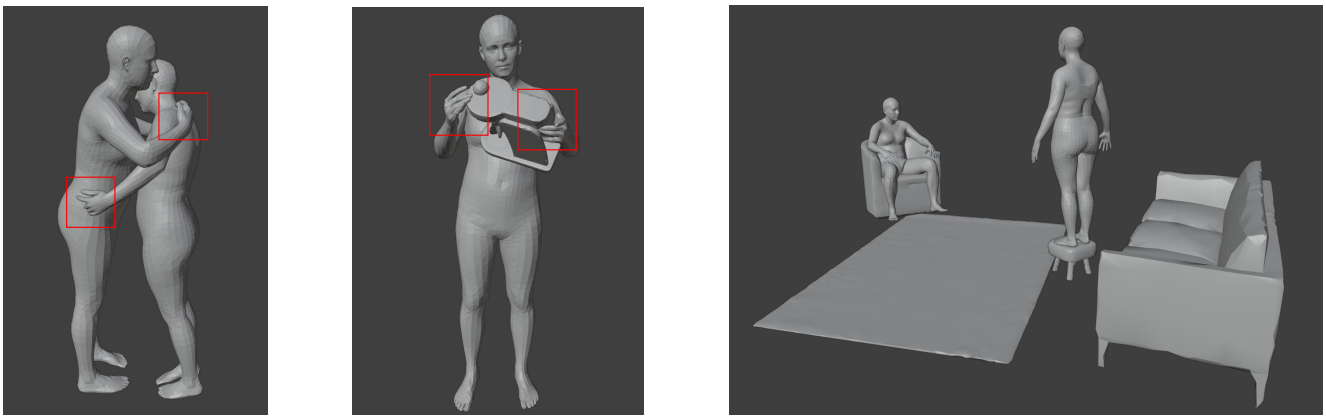
Qwen2.5VL: The person starts by **raising the right arm to hold the violin** and positioning the left hand on the neck…

Ours (GPT-4o): The person **raises the left arm to place the mock violin beneath the chin** and mimics a bowing action with the right arm…

Figure 1. Our caption pipeline produces superior captions. For a mimed violin sequence, our method generates a factually accurate and detailed caption, avoiding the pitfalls of other approaches like the over-simplicity of the ground-truth (GT), the hallucinated details from a baseline GPT-4o refinement, and the factual errors of Qwen2.5VL.

Task	Dataset	Text Count	Avg. Sentence Length	TTR Range	Top-10 Verbs
T2M	MIXAMO [1]	2,350	331.54	0.232-0.600	swing, step, lean, walk, stand, lift, adjust, bend, stabilize, raise
	KIT [33]	5,593	289.67	0.305-0.740	stand, walk, leave, move, turn, step, extend, swing, follow, shift
	OMOMO [21]	5,619	278.76	0.330-0.581	bend, stand, reach, extend, straighten, move, step, lean, indicate, shift
	IDEA400 [25]	14,753	279.69	0.288-1.000	walk, step, lift, raise, stand, coordinate, reach, lean, swing, adjust
	100Style [29]	15,747	394.54	0.225-1.000	leave, move, swing, lean, follow, step, lift, extend, walk, run
	HumanML3D [12]	48,398	232.44	0.139-0.817	stand, bend, walk, lean, turn, control, lift, reach, step, raise
M2D	Choreomaster [8]	2,699	279.87	0.324-0.629	extend, shift, leave, move, raise, weight, lift, bend, transition, lean
	FineDance [23]	5,540	353.40	0.288-0.632	leave, extend, move, shift, weight, raise, step, lift, bend, lean
	PhantomDance [20]	6,389	340.24	0.275-1.000	leave, extend, move, raise, shift, lift, bend, weight, stand, lean
	AIST++ [22]	3,183	405.28	0.260-0.673	leave, extend, swing, move, weight, shift, lean, bend, lift, raise
	Motorica [2]	4,825	269.36	0.289-1.000	swing, lift, lean, bend, raise, step, bring, side, lower, follow
	AIOZ [19]	74,649	262.65	0.311-1.000	extend, shift, leave, weight, move, raise, bend, lift, progress, involve
S2G	BEAT2 [26]	21,576	233.63	0.331-0.942	emphasize, stand, move, shift, indicate, raise, extend, gesture, engage, weight
HHI	HumanSC3D [10]	1,024	279.05	0.339-0.651	stand, bend, move, lower, raise, lean, extend, lift, hold, indicate
	InterHuman [24]	23,072	285.82	0.206-0.756	extend, leave, stand, step, move, shift, raise, lean, involve, bend
	InterX [35]	10,460	219.28	0.282-0.693	stand, extend, lean, raise, move, indicate, step, shift, bend, involve
HOI	ARCTIC [9]	1,596	270.38	0.338-1.000	extend, move, stand, focus, reach, adjust, indicate, hold, lean, simulate
	TACO [27]	3,391	279.06	0.295-0.733	lift, skim, hold, smear, reach, scoop, control, grasp, stabilize, dip
	FIT3D [11]	1,987	290.21	0.317-1.000	stand, extend, bend, control, move, lower, focus, reach, return, lift
	BEHAVE [6]	3,089	309.72	0.294-0.577	stand, bend, lean, leave, move, shift, extend, step, lift, control
	CHAIRS [16]	8,476	282.23	0.325-0.599	stand, lean, sit, bend, seat, extend, move, adjust, shift, lower
	HOIM3 [38]	7,717	346.58	0.209-0.909	leave, stand, lean, adjust, move, shift, walk, indicate, extend, seat
	Oaklnkv2 [36]	24,854	266.42	0.280-1.000	focus, stand, extend, adjust, move, indicate, lean, involve, control, suggest
	NeuralDome [37]	2,333	360.82	0.240-0.599	leave, extend, adjust, move, lift, stand, shift, stabilize, control, bend
HSI	EMDB [18]	237	311.78	0.320-0.542	leave, walk, swing, extend, move, stand, lift, pkl, lean, step
	RICH [14]	598	307.25	0.313-0.566	extend, stand, bend, move, lean, leave, lower, reach, focus, control
	LAFAN [13]	3,297	316.25	0.301-1.000	leave, extend, move, swing, step, lean, walk, stand, shift, lift
	Trumans [17]	8,231	323.08	0.299-0.667	stand, extend, bend, lean, reach, move, straighten, leave, shift, control
	CIRCLE [3]	10,284	278.17	0.298-0.917	reach, extend, stand, lean, bend, shift, indicate, leave, focus, involve
All	*OmniMoCap-X*	321,967	276.78	0.139-1.000	stand, lean, bend, lift, raise, adjust, step, swing, control, walk

Table 1. Text quality statistics across motion datasets. The third column shows the number of text samples per dataset. The fourth column displays the average sentence length (in word count). The fifth column presents the Type-Token Ratio (TTR) range, indicating lexical diversity (higher values represent richer vocabulary). The sixth column lists the ten most frequent verbs in each dataset, reflecting the predominant actions described in the motion text. The datasets are grouped by their primary task: Text-to-Motion (T2M), Music-to-Dance (M2D), Speech-to-Gesture (S2G), Human-Human Interaction (HHI), Human-Object Interaction (HOI), and Human-Scene Interaction (HSI).



(a) Human-Human (b) Human-Object (c) Human-Scene

Figure 2. Visualization of interaction types in our dataset. From left to right: (a) Human-Human Interaction (HHI), (b) Human-Object Interaction (HOI), and (c) Human-Scene Interaction (HSI).

dividuals and their environments, such as sitting on indoor sofas with feet resting on carpets, and other contextual behavioral patterns. Each interaction type is accompanied by detailed spatio-temporal annotations and semantic descriptions, providing comprehensive training samples for multimodal whole-body motion generation.

4. More Implementation Details

Implementation Details of the Unified Architecture. Fig. 3 illustrates the unified architecture of our model, *OmniMotion-X*, which is designed for versatile multimodal whole-body motion generation. The core of the model is a Transformer Encoder that processes a long sequence of concatenated embeddings. As shown at the bottom of the figure, this input sequence is formed by amalgamating various conditional inputs with the target noisy motion. Specifically, all conditioning signals—including a Time Embedding for the diffusion step, Text and Global Motion Embeddings for contextual and spatial guidance, Speech and Music Embeddings for audio-driven synthesis, and a Reference Motion Embedding for stylistic control—are first transformed into sequences of feature tokens. These conditional tokens are then concatenated with the Noise Motion Embedding tokens, which represent the motion sequence to be denoised. The Transformer Encoder processes this entire unified sequence, allowing it to learn complex cross-modal relationships through its self-attention mechanism. The final Motion Representation is then produced by extracting the output tokens corresponding to the positions of the initial Noise Motion Embedding, as depicted at the top of the figure.

Implementation Details of Encoders. Here, we provide detailed descriptions of the architectures for the modality-specific encoders illustrated in Fig. 4. All encoders project their respective inputs into a shared latent space of dimension D.

a) Timestep Embedder. The timestep t is first transformed into a sinusoidal position embedding, resulting in a vector of dimension D. This vector is then processed by a small Multi-Layer Perceptron (MLP) consisting of two linear layers with a SiLU activation in between, as shown in the figure. The input and output dimensions of this MLP are both D, allowing it to refine the time embedding without altering its shape.

b) Text Encoder. We utilize a pre-trained T5 text encoder to extract high-level semantic features from the input text. The raw text is tokenized and fed into the T5 encoder, which outputs a sequence of feature vectors of dimension D_t. A subsequent linear projection layer maps these features from D_t to our model's latent dimension D. Following standard practice, the T5 encoder's weights are kept frozen during training.

c) Global Pose Encoder. This encoder processes a sequence of global motion information, represented by the 3D coordinates of N key joints over L_g frames. The input tensor of shape $(B, L_g, N \times 3)$ is flattened along the feature dimension and then passed through a single linear layer to project it into the latent space, resulting in a tensor of shape (B, L_g, D).

d) Speech Encoder. The raw speech waveform, with a sampling rate of 16kHz, serves as the input. We employ WavEncoder module [26], which is a deep 1D convolutional neural network composed of a stack of residual blocks. This network progressively downsamples the temporal dimension while increasing the channel dimension, effectively extracting meaningful acoustic features from the raw audio. The final output of this encoder is a sequence of feature vectors of dimension D.

e) Music Encoder. The process for music involves two stages. First, we extract a set of acoustic features from the raw music waveform using the `librosa` library. Specifi-

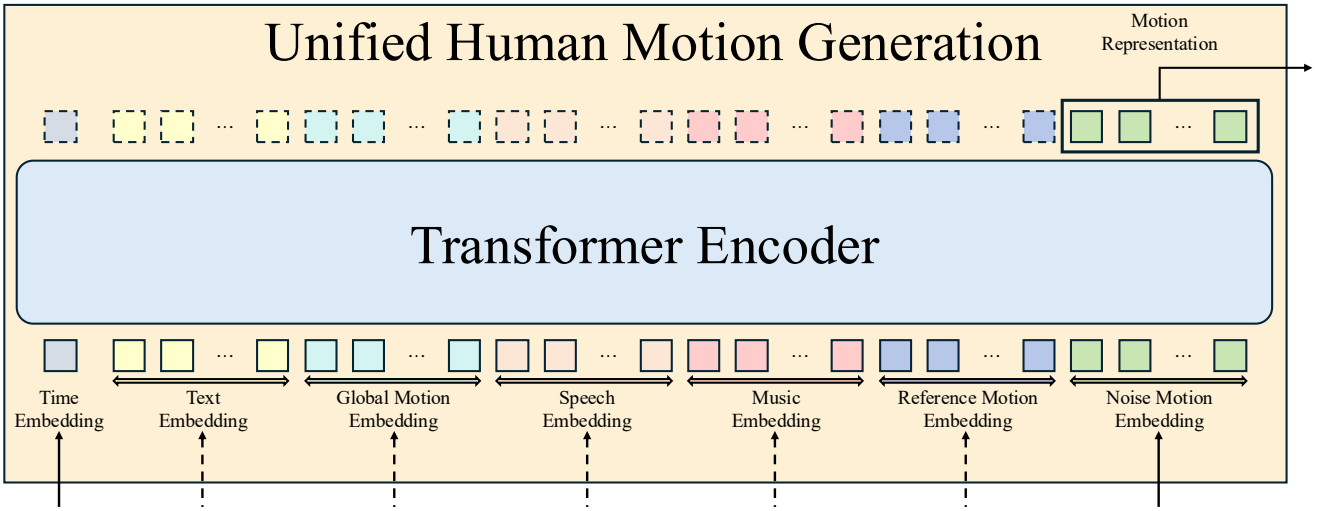


Figure 3. **The Unified Architecture of *OmniMotion-X*.** Our model unifies various motion generation tasks within a single framework. All conditioning inputs, including timestep, text, global motion, speech, music, and a reference motion, are first encoded into sequences of embeddings. These conditional embeddings are then concatenated with the noisy motion embedding sequence and fed into a central Transformer Encoder to produce the final motion representation.

cally, we compute a 35-dimensional feature vector for each time step by concatenating five components: 1D onset strength (envelope), 20D Mel-Frequency Cepstral Coefficients (MFCCs), 12D Chroma features, and two 1D one-hot vectors indicating the locations of onset peaks and tracked beats. In the second stage, this sequence of 35-dimensional features is fed into a linear projection layer, which maps it to the final music embedding of dimension D.

Implementation Details of Body-wise Pose Encoder and Decoder. As illustrated in Fig. 5, our model adopts a body-wise strategy to process the high-dimensional motion representation. This approach allows for fine-grained feature learning and control.

a) Body-wise Pose Encoder. The input to the encoder is a raw motion representation tensor of shape (B, L_m, D_p), where D_p is the full pose feature dimension. This tensor is first split along the last dimension into multiple segments, each corresponding to a specific body part (e.g., head, face, left hand). Each segment is then fed into its own dedicated linear layer, which projects the part-specific features into a smaller, fixed-size latent dimension. Finally, the outputs from all these linear layers are concatenated (denoted by 'C' in the figure) to form the final motion embedding of shape (B, L_m, D).

b) Body-wise Pose Decoder. The decoder performs the inverse operation. It takes the motion representation from the Transformer's output, with a shape of (B, L_m, D), as input. This tensor is first split into the same body-part-specific segments. Each segment is then passed through its dedicated linear decoder layer, which maps it back from the latent dimension to its original feature dimension. The outputs of all decoder layers are then concatenated to reconstruct the full-body motion representation of shape (B, L_m, D_p). This reconstructed motion is subsequently used for the final output or the next diffusion step.

Text and Motion Feature Extractors. Existing research typically relies on pre-trained motion and text feature extractors for evaluation. However, due to significant differences in dataset scale, distribution, and motion representation compared to existing works, these pre-trained extractors are not directly applicable. Therefore, following the approach [28], we re-train motion and text feature extractor using a contrastive learning framework tailored to this dataset. The text feature extractor is based on the Transformer encoder architecture [32], which encodes raw text into a semantic vector s_t. The motion feature extractor also uses a Transformer encoder to encode motion sequences of up to 150 frames into a semantic vector s_m. Both encoders include additional semantic tokens, with a structure similar to the encoder in ACTOR [31], but without involving probabilistic distribution modeling. In implementation, the text encoder takes as input text features extracted from a pre-trained and frozen DistilBERT network, while the motion encoder directly processes raw motion sequence data. In the contrastive learning framework, we optimize the feature space such that matching text-motion feature pairs (s_t, s_m) are brought closer in the embedding space, while ensuring that non-matching pairs are separated by at least a distance d. This optimization objective is achieved through the following contrastive loss function:

$$D_{s_t, s_m} = ||s_t - s_m||_2, \tag{1}$$

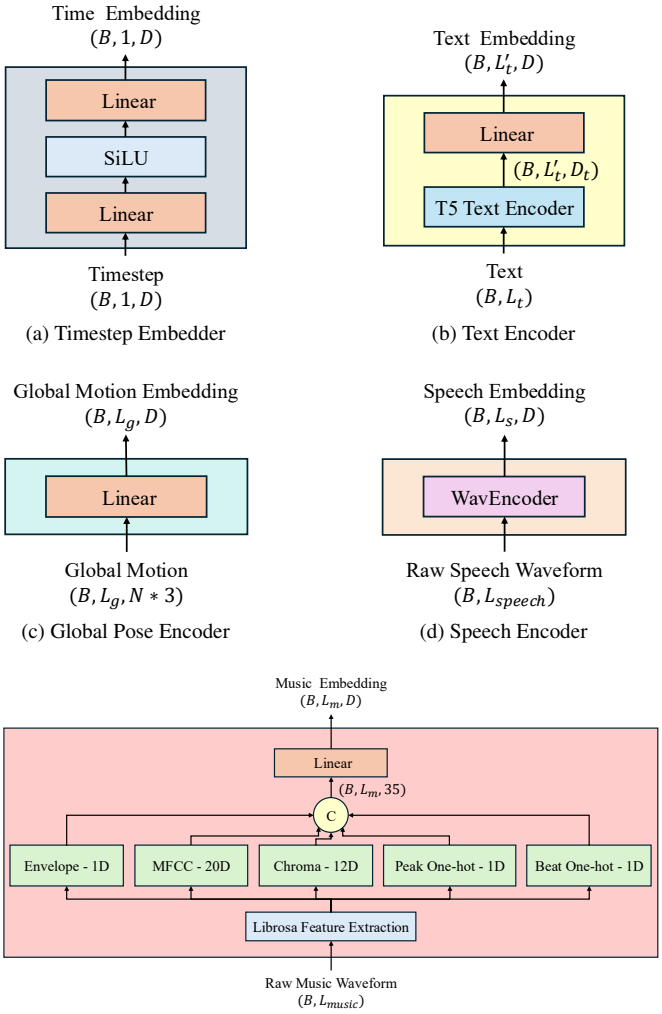


Figure 4. **Architectures of Different Encoders in *OmniMotion-X*.** Key encoder components for processing various inputs: (a) Timestep, (b) Text, (c) Global Motion, (d) Speech, and (e) Music.

$$\mathcal{L}_{Cta} = (1-y)(D_{s_t,s_m})^2 + (y)\{max(0, d - D_{s_t,s_m})\}^2, \quad (2)$$

where, y is a binary label: $y = 0$ if s_t and s_m come from a matching text-motion pair, and $y = 1$ otherwise. $m > 0$ is the margin for non-matching pairs, which is set to 10 in our experiments.

Mask. Our model leverages three specialized masking mechanisms: (1) Component attention mask, the standard source key padding mask in Transformers—regulates attention across different conditional inputs. In our DiT-based architecture, input conditions may be partially missing; this mask zeroes out attention weights for missing inputs, allowing the model to focus only on valid ones. (2) Global task-dependent mask, a spatiotemporal dynamic mask tailored to tasks like motion prediction, interpolation, completion, and trajectory-guided generation—distinguishes between

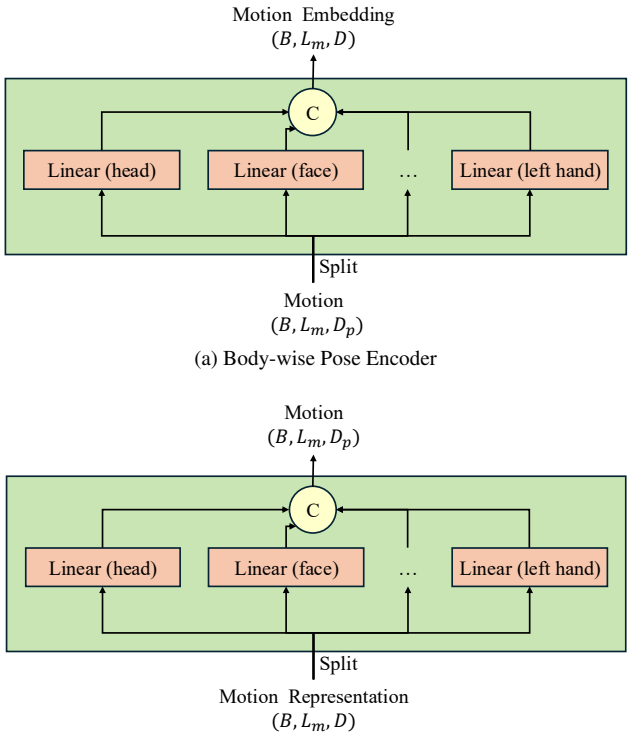


Figure 5. **Architecture of the Body-wise Pose Encoder and Decoder.** Our model processes motion data by first encoding pose features for different body parts into a latent space (a), and then decoding them back to the full pose representation (b).

observed and unobserved motion regions based on task requirements. (3) Motion representation reconstruction mask filters out reconstruction errors of joints without ground-truth annotations during training. For instance, it suppresses hand-joint errors in datasets lacking hand annotations, effectively addressing cross-dataset joint count mismatches.

5. Condition Conflicting

As shown in Fig. 6, we constructed adversarial examples to analyze the model's handling of conflicting conditions between textual prompts and global motion trajectories. In a representative case, the prompt "The man is walking backward" was supplied alongside a global trajectory specifying a forward-then-rightward path, which was activated at the start and conclusion of the video sequence. We observed that in the absence of the global condition's activation, the model's synthesis is guided by the textual prompt. However, when the global condition is activated, it overrides the textual description, compelling the generated motion to follow the specified trajectory. This result suggests that when conditions conflict, the model adheres to the stronger constraint.

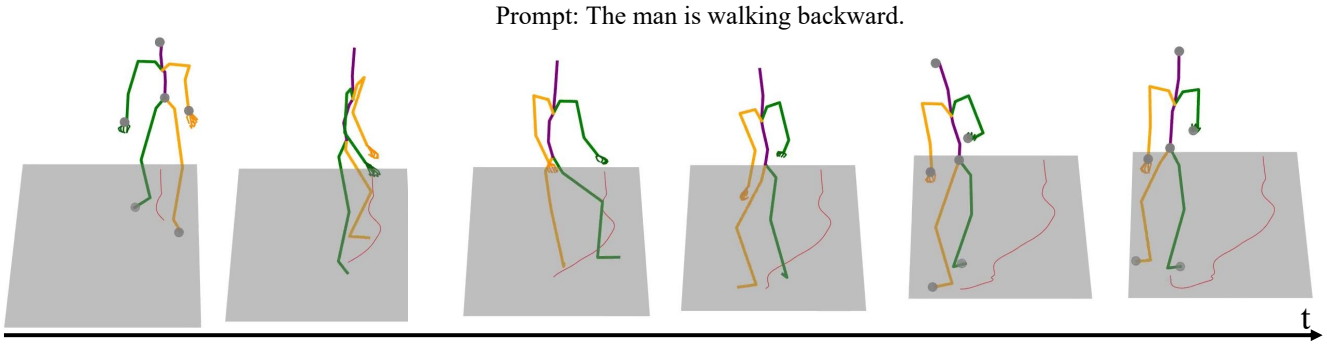


Figure 6. An example of conflicting text and global trajectory conditions. Although the prompt is "The man is walking backward," the model's generated motion overrides the text and follows the predefined global trajectory (pink path) when it is activated, resulting in a forward walk.

6. Speed

In contrast to text-only DiT-based methods such as MDM, our OmniMotion-X natively supports a diverse set of conditioning modalities, including text, reference motion, global trajectories, audio, and music. At a comparable parameter count, our approach achieves a significant performance improvement, elevating the AIT score from 0.65 to 2.14. Furthermore, as detailed in Table 2, scaling our model from 27.22M to 355.13M parameters results in only a marginal increase of 0.86s in inference time. This suggests that the primary bottleneck for inference speed is the processing of conditional inputs, rather than the scale of the model itself.

Table 2. Comparison of model size and average inference time (AIT). The AIT is measured in seconds per example. A lower AIT is better.

Method	**Parameters (M)**	**AIT (s)** ↓
MDM [34]	26.42	0.65
Ours (Small)	27.22	2.14
Ours (Large)	355.13	3.00

References

[1] Adobe. Mixamo. https://www.mixamo.com. 1, 2

[2] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph.*, 42(4):44:1–44:20, 2023. 1, 2

[3] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21211–21221, 2023. 2

[4] Autodesk Inc. Autodesk MotionBuilder. 1

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[6] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 2

[7] Blender Foundation. Blender. 1

[8] Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 1, 2

[9] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 2

[10] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1343–1351, 2021. 2

[11] Mihai Fieraru, Mihai Zanfir, Silviu-Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[12] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, pages 5152–5161, 2022. 2

[13] Félix G. Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. 39(4), 2020. 1, 2

[14] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings*

IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pages 13274–13285, 2022. 2

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[16] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. *ICCV*, 3, 2022. 2

[17] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024. 2

[18] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[19] Nhat Le, Thang Pham, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Music-driven group choreography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8673–8682, 2023. 2

[20] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1272–1279, 2022. 2

[21] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM TOG*, 42(6):1–11, 2023. 2

[22] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021. 2

[23] Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *ICCV*, pages 10234–10243, 2023. 2

[24] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *IJCV*, pages 1–21, 2024. 2

[25] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *NeurIPS*, 2024. 2

[26] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via masked audio gesture modeling. *arXiv e-prints*, pages arXiv–2401, 2023. 2, 3

[27] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21740–21751, 2024. 2

[28] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *ICML*, 2024. 4

[29] Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(1):1–18, 2022. 1, 2

[30] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 1

[31] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, pages 10985–10995, 2021. 4

[32] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *ICCV*, pages 9488–9497, 2023. 4

[33] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2

[34] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *ICLR*, 2022. 6

[35] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *CVPR*, pages 22260–22271, 2024. 2

[36] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 445–456, 2024. 2

[37] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023. 2

[38] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 516–526, 2024. 2