MotionCraft: Crafting Whole-Body Motion with Plug-and-Play Multimodal Controls

Yuxuan Bian[1], Ailing Zeng[2*], Xuan Ju[1], Xian Liu[1], Zhaoyang Zhang[1], Wei Liu[2], Qiang Xu[1*]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Tencent, Guangdong Province, China

Abstract

Whole-body multimodal motion generation, controlled by text, speech, or music, has numerous applications including video generation and character animation. However, employing a unified model to process different condition modalities presents two main challenges: motion distribution drifts across different tasks (e.g., co-speech gestures and text-driven daily actions) and the complex optimization of mixed conditions with varying granularities (e.g., text and audio). In this paper, we propose *MotionCraft*, a unified diffusion transformer that crafts whole-body motion with plug-and-play multimodal control. Our framework employs a coarse-to-fine training strategy, starting with the text-to-motion semantic pre-training, followed by the multimodal low-level control adaptation. To effectively learn and transfer motion knowledge across different distributions, we design *MC-Attn* for parallel modeling of static and dynamic human topology graphs. To overcome the motion format inconsistency of existing benchmarks, we introduce *MC-Bench*, the first available multimodal whole-body motion generation benchmark based on the unified SMPL-X format. Extensive experiments show that *MotionCraft* achieves state-of-the-art performance on various standard motion generation tasks.

Code — https://cure-lab.github.io/MotionCraft

Introduction

Whole-body human motion generation with multimodal controls (Zhang et al. 2024b), which produces coherent human movements based on multimodal conditions, has numerous applications, including human video generation (Hu 2024) and character animation (Zhang et al. 2023a).

Recent advancements in single-conditioned motion generation can generate realistic motion from conditions with varying granularities, including text descriptions (Guo et al. 2022; Zhang et al. 2023c), music clips (Siyao et al. 2022; Li et al. 2023), and speech segments (Liu et al. 2024a; Chen et al. 2024). However, extending them to whole-body motion generation with multimodal controls within a unified model introduces several significant challenges:

- ***Motion distribution drifts:*** Motion patterns exhibit significant variations across different contexts (Zhang et al. 2024b; Ling et al. 2023). Text-to-motion (T2M) primarily generate torso movements based on text descriptions (Guo et al. 2022; Lin et al. 2023), while speech-to-gesture (S2G) synthesizes gestures and facial expressions from first-person audio (Liu et al. 2024a; Yi et al. 2023). Music-to-dance (M2D) involves more complex correlations between third-person music and full-body movements (Li et al. 2023). Due to these distinct distributions, prior works have typically focused on individual tasks to avoid challenges in cross-task generalization.
- ***Optimization challenges under mixed conditions:*** Existing multimodal motion generation approaches attempt to unify diverse control signals (E.g., texts, music, and speech) by projecting them into a shared space. This unification is achieved through methods like transformer token embeddings (Zhou, Wan, and Wang 2023) and ImageBind's feature space (Girdhar et al. 2023), and it faces two key limitations: cross-modal alignment inconsistencies and optimization difficulties when handling conditions at varying granularities (Team et al. 2023).
- ***Non-uniform whole-body motion format and evaluation:*** Finally, there are no high-quality multimodal whole-body human motion generation benchmarks with unified motion representation and evaluation pipelines.

In this work, we propose a unified motion diffusion transformer, **MotionCraft**, that crafts whole-body motion with plug-and-play multimodal controls, generating fine-grained motions aligned with given texts, speech, or music. It also supports simultaneously generating motion with multiple conditions, such as text combined with speech or music.

To enable precise control over specific modalities while preserving semantic coherence and avoiding training conflicts, *MotionCraft* implements a hierarchical two-stage multimodal generation, progressing from coarse to fine control. The first stage develops core motion generation capabilities using broad textual guidance, while the second stage introduces fine-grained control modules on top of the frozen first-stage backbone. To address motion distribution drifts across various generation scenarios, we analyze human motion kinematics and distribution using t-distributed stochastic neighbor embedding. We find that motion distributions corresponding to different control signals can be decomposed into static human topology structures and dynamic topology relationships that can be generalizable across dif-

*Corresponding authors.

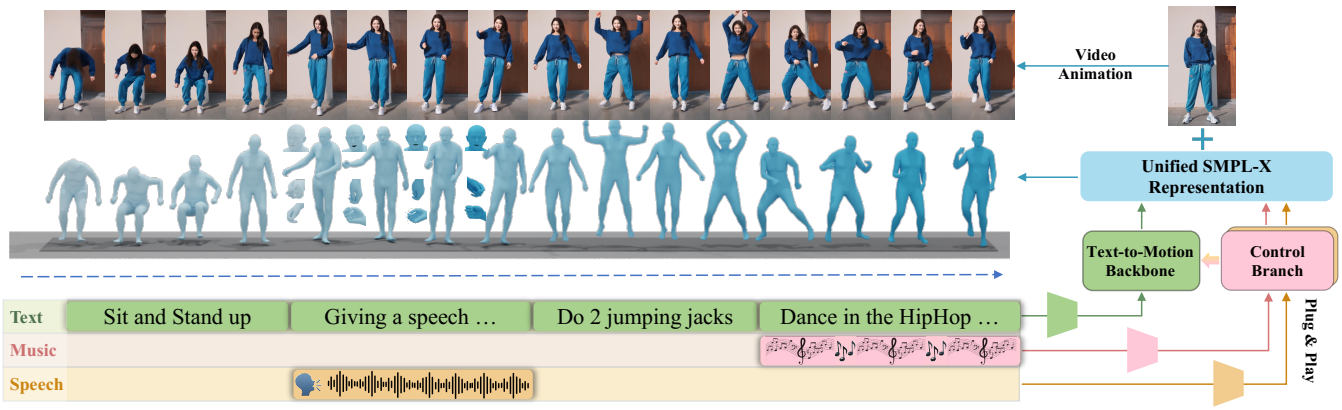

Figure 1: We propose *MotionCraft*, a diffusion transformer that crafts whole-body motion with plug-and-play multimodal controls, encompassing robust motion generation abilities including Text-to-Motion, Speech-to-Gesture, and Music-to-Dance.

ferent scenarios. To model these human-centric spatiotemporal properties with current limited and unscalable motion data, we propose **MC-Attn**. This architecture incorporates a spatial branch that transfers topological motion knowledge via parallel processing of static and dynamic human topology graphs, complemented by a temporal branch that captures sequential motion dependencies.

To address the inconsistent motion representations across existing benchmarks like Rot6D (Guo et al. 2022), SMPL (Loper et al. 2015), and SMPL-X (Pavlakos et al. 2019), we introduce **MC-Bench**, a unified benchmark based on the SMPL-X whole-body format. Our framework provides comprehensive data processing and evaluation protocols. Extensive experiments show that *MotionCraft* achieves strong performance across text-to-motion, speech-to-gesture, and music-to-dance generation tasks. Through detailed ablation studies, we provide key insights into architectural design choices and scaling effects for future multimodal whole-body motion generation models.

In summary, our contributions are as follows:

- We propose **MotionCraft**, a two-stage, coarse-to-fine multimodal motion generation framework that supports control signals at different granularities, enabling efficient plug-and-play multimodal motion generation.
- We design **MC-Attn**, the first attempt to achieve modeling of static and dynamic human topology against motion distribution drifts in multimodal motion generation.
- We create **MC-Bench**, the first publicly available multimodal whole-body motion generation benchmark with a unified whole-body motion representation SMPL-X.

Related Work

Human Motion Generation Models

Conditioned human motion generation have made significant progress, including text-to-motion (T2M) (Liu et al. 2023; Zhang et al. 2024a, 2023c; Liang et al. 2024), speech-to-gesture (S2G) (Yi et al. 2023; Chen et al. 2024), and music-to-dance (M2D) (Li et al. 2023; Tseng, Castellon, and Liu 2023). Recently, increasing attention has been paid to multimodal motion generation (Ling et al. 2023; Zhang et al. 2024b; Luo et al. 2024). Motion-Verse (Zhang et al. 2024b) implements dynamic attention for inter-body-part relationships but neglects global human topology, limiting its generalization capabilities. Moreover, its ImageBind-based (Girdhar et al. 2023) mixed training approach struggles with multi-granular conditions and lacks flexibility for new control signals. Although MCM (Ling et al. 2023) addresses mixed training challenges using ControlNet (Zhang, Rao, and Agrawala 2023), its disregard for human topology constrains cross-scenario generalization. In contrast to existing approaches (Tab. 1), *MotionCraft* achieves plug-and-play whole-body motion generation across diverse control signals by leveraging *MC-Attn* to model both static topology and domain-specific skeletal dynamics, while implementing control branches and coarse-to-fine training.

Human Motion Generation Benchmarks

Various conditioned human motion generation benchmarks have been constructed in recent years. For T2M, researchers have curated datasets encompassing action categories (Chung et al. 2021; Trivedi, Thatipelli, and Sarvadevabhatla 2021), sequential action labels (Zhang et al. 2022; Guo et al. 2020), and arbitrary natural language descriptions (Lin et al. 2023; Guo et al. 2022; Tang et al. 2023). For M2D, AIST++ (Li et al. 2021) reconstructs 5 hours of dance based on SMPL (Loper et al. 2015) format from videos. Finedance (Li et al. 2023) collects dances of 14.6 hours across 22 genres and supplements the dataset with detailed gestures using the SMPL-H (Pavlakos et al. 2019) format. For S2G datasets (Liu et al. 2024a, 2022; Yi et al. 2023), BEAT2 (Liu et al. 2024a) and BEAT (Liu et al. 2022) have emerged as the most popular benchmarks, celebrated for their diverse range of motion and extensive data volume. BEAT2, built upon BEAT, utilizes SMPL-X and FLAME (Kim, Kim, and Choi 2023) to achieve higher-quality unified mesh-level data. Despite these developments, no publicly available benchmark supports unified representation for multimodal whole-body motion generation.

Model	Text2Motion	Music2Dance	Speech2Gesture	Static Body Prior	Dynamic Body Adaption	Whole Body	Unified Representation	Plug-and-Play
FineMoGen (Zhang et al. 2023c)	✓	✗	✗	✓	✗	✗	✗	✗
HumanTomato (Lu et al. 2023)	✓	✗	✗	✓	✗	✓	✗	✗
FineDance (Li et al. 2023)	✗	✓	✗	✗	✗	✗	✗	✗
Bailando (Siyao et al. 2022)	✗	✓	✗	✓	✗	✗	✗	✗
EMAGE (Liu et al. 2024a)	✗	✗	✓	✓	✗	✓	✗	✗
TalkShow (Yi et al. 2023)	✗	✗	✓	✓	✗	✓	✗	✗
MCM (Ling et al. 2023)	✓	✓	✓	✗	✗	✗	✗	✓
Motion-Verse (Zhang et al. 2024b)	✓	✓	✓	✗	✓	✓	✗	✗
MotionCraft	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: **Comparison of *MotionCraft* with previous motion generation methods.** *MotionCraft* jointly models the static human skeleton structure and dynamic human topology relationships to achieve flexible motion knowledge transfer across various whole-body generation scenarios, supporting plug-and-play with any new control signal modality.

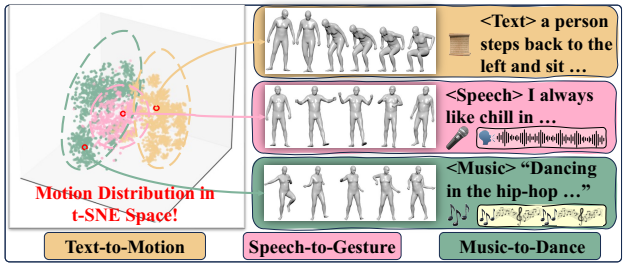

Figure 2: **The t-SNE latent space of motion in different generation tasks.** It illustrates the motion distribution drifts across different generation scenarios.

Motivation

The key challenge in achieving whole-body human motion generation with multimodal controls is addressing motion distribution drifts across different generation scenarios (Zhang et al. 2024b) and the efficient learning of control signals at varying granularities (Ling et al. 2023).

Motion distribution drifts solution. Motion generation models typically excel in single-scenario applications but face challenges with distribution drifts across different contexts (Zhou, Wan, and Wang 2023). As illustrated in Fig. 2, distinct motion patterns emerge across datasets: T2M features primarily torso movements, S2G encompasses detailed hand gestures and facial expressions with static lower body, while M2D exhibits diverse limb movements with minimal hand activity. Recent human-centric research (Zeng et al. 2021; Ma, Bai, and Zhou 2022) demonstrates that representing human skeletal structure as a directed weighted graph, with body segments as vertices, effectively incorporates kinematic priors and enhances robustness to distribution drifts. Human kinematic (Loper et al. 2015; Pavlakos et al. 2019) naturally support decomposing the skeletal structure into static and dynamic topological components. The hip joint consistently maintains a strong influence over connected limbs, with symmetric relationships between arm pairs across scenarios. In S2G specifically, while limb correlations diminish, hand-facial expression connections strengthen. Therefore, dual modeling of static and dynamic topologies enables effective cross-task knowledge transfer, even with limited data and significant distribution variations.

Efficient learning of conditions at varying granularities. Motion generation tasks require handling conditions at multiple granularity levels. While text prompts provide coarse-grained control at the sequence level, speech and music inputs enable fine-grained frame-level control (Liu et al. 2024a; Li et al. 2023). Combining all conditional modalities into a unified learning space introduces modality alignment challenges and prevents effective granularity-specific optimization (Zhang, Rao, and Agrawala 2023; Ling et al. 2023). Drawing inspiration from image and video generation approaches (Rombach et al. 2022; Liu et al. 2024b), we decouple the learning process across different conditional modalities. By using text-to-motion as the foundational pre-training task, we establish robust generative capabilities that enable precise multimodal control.

Proposed Method

MotionCraft Framework

The overview of *MotionCraft* is described in Fig. 3. Aimed at decoupling the conditioned generation learning at varying granularities, we adopt a two-branch architecture consisting of a text-to-motion branch and a plug-and-play low-level control branch, along with a two-stage coarse-to-fine training strategy to efficiently grasp the motion topology across different scenarios with various control signal modalities.

Stage ❶ Text-to-Motion Semantic Pre-training. The main branch $f_m(\cdot)$ is optimized in Stage I, text-to-motion semantic pre-training, using text-to-motion paired data collected from diverse scenarios in *MC-Bench*. We choose text as the shared condition among various unimodal datasets, allowing *MotionCraft* to acquire sequence-level generation and coarse-grained text-guidance following abilities between text $\mathbf{H}_{text} \in \mathbb{R}^{B\times F_t\times D_t}$ and motion $\mathbf{H}_{motion} \in \mathbb{R}^{B\times F_m\times D_m}$. Overall, text guidance pre-training in diverse generation scenarios helps follow fine-grained controls of other low-level conditions in Stage II.

Stage ❷ Multimodal Low-level Control Adaptation. During the low-level control adaptation fine-tuning stage, we aim to model the correlation between various condition signals $\mathbf{H}_c \in \mathbb{R}^{B\times T_c\times D_c}$ and motion sequences $\mathbf{H}_{motion} \in \mathbb{R}^{B\times F_m\times D_m}$. All main branch $f_m(\cdot)$ are frozen to maintain coarse-grained motion generation and semantic text following abilities. A copy of the main branch $\hat{f_m}(\cdot)$ is then used to initialize the control branch, connecting them with a zero-initialized linear layer $\mathbf{W}_p \in \mathbb{R}^{D_m\times D_m}$ to prevent early training collapse. Then the condition signals (speech, music, or other low-level control signals) are fed into the control

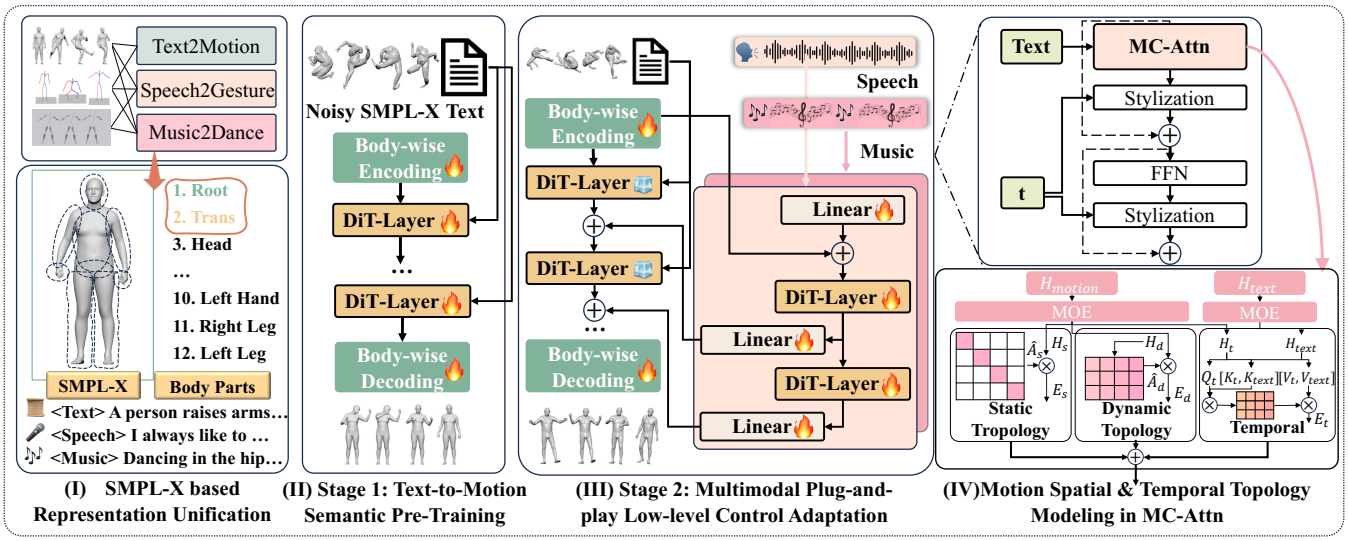

Figure 3: **Architecture of *MotionCraft*.** *MotionCraft* is a transformer-based diffusion model. In the first stage, *MotionCraft* uses text as a semantic control guide to learn coarse-grained cross-scenario motion knowledge across multiple datasets; in the second stage, *MotionCraft* freezes the backbone while adding a plug-and-play control branch to learn the different low-level control signals. The core of *MotionCraft* is *MC-Attn*, which optimizes the representation of motion token sequences by capturing the spatial properties of static and dynamic human topology graphs and learning temporal relationships in parallel.

branch. The output of each control branch layer is directly added to the corresponding main branch layer input through the zero bridge linear layer, allowing new control signals to guide frame-level human motion generation.

MC-Attn Design

MotionCraft is built upon *MC-Attn*, which simultaneously processes static and dynamic human topology graphs, enhancing the transferability of motion topology knowledge across diverse generation scenarios while addressing distribution drifts. *MC-Attn* has three key components: a static skeleton graph learner and a dynamic topology relationship graph learner to parallel model spatial properties of motion, and temporal attention to model the frame-level dynamics of each body part over time. These modules share the same input $\mathbf{H}_m \in \mathbb{R}^{B\times F_m\times D_m}$, the output of the last *MC-Attn* layer refined further by a MOE(Shazeer et al. 2017).

The static-skeleton graph learner generates vertex representations $\mathbf{H}_s \in \mathbb{R}^{B\times F_m\times N_b\times D_b}$ for N_b body parts and initializes a diagonal unit matrix $\mathbf{A}_s \in \mathbb{R}^{N_b\times N_b}$ as the adjacency matrix for the initial topology graph $\mathcal{G}_s$. Each vertex connects only to itself to prevent training instability. Through optimization, $\hat{\mathbf{A}}_s$ learns the input-independent human topology, enabling robust structural understanding even with limited training data. The module produces $\mathbf{E}_s = \hat{\mathbf{A}}_s \cdot \mathbf{H}_s$. While static topology graphs effectively capture fundamental human structure and accelerate adaptation to new distributions, they lack the flexibility to accommodate dynamic contexts, leading to potential underfitting (Zhang et al. 2024b) To overcome this limitation, we propose a dynamic-topology relationship graph learner that adapts to distribution shifts based on control signals, complementing the static structure. The dynamic learner represents body parts as vertices $\mathbf{H}_d \in \mathbb{R}^{B\times F_m\times N_b\times D_b}$ and employs attention-based edge weights $\mathbf{A}_d \in \mathbb{R}^{B\times F_m\times N_b\times N_b}$ in the dynamic topology graph $\mathcal{G}_d$, enabling adaptive spatial modeling. The final output is $\mathbf{E}_d = \mathbf{A}_d \cdot \mathbf{H}_d$.

Research has demonstrated that standard attention adequately captures temporal dependencies (Nie et al. 2022; Bian et al. 2024). We therefore represent each body part as a unit $\mathbf{H}_t \in \mathbb{R}^{B\cdot N_b\times F_m\times D_b}$ and employ attention (Vaswani et al. 2017) to model inter-frame temporal relationships. We incorporate sequential texts into our attention computation for better temporal coherence: $\hat{\mathbf{E}}_t = Softmax(\mathbf{Q}_{H_t} \cdot [\mathbf{K}^T_{H_t}, \mathbf{K}^T_{H_{text}}]/\sqrt{D_b}) \cdot [\mathbf{V}^T_{H_t}, \mathbf{V}^T_{H_{text}}]$, where $\mathbf{H}_{text}$ represents text features and $[,]$ denotes concatenation. Additional sequential controls, such as speech and music, are processed in the control branch. The final output $\mathbf{E} = \mathbf{E}_s + \mathbf{E}_d + \mathbf{E}_t$ of *MC-Attn* integrates the spatial skeleton representations with the temporal dynamics of individual body parts.

MC-Bench Construction

To prevent the information loss when aligning different motion formats, we select HumanML3D (Guo et al. 2022) in SMPL format for T2M, FineDance (Li et al. 2023) in SMPL-H Rot-6D format for M2D, and BEAT2 (Liu et al. 2024a) in SMPL-X format for S2G from public datasets, as they are the most representative unimodal datasets in their respective areas. To enable whole-body multimodal control of human motion generation, we converted all data to the SMPL-X format. Key operations include filling in missing facial information in HumanML3D and FineDance with average expressions and converting FineDance from SMPL-H Rot-6D format to axis-angle representation for efficient alignment

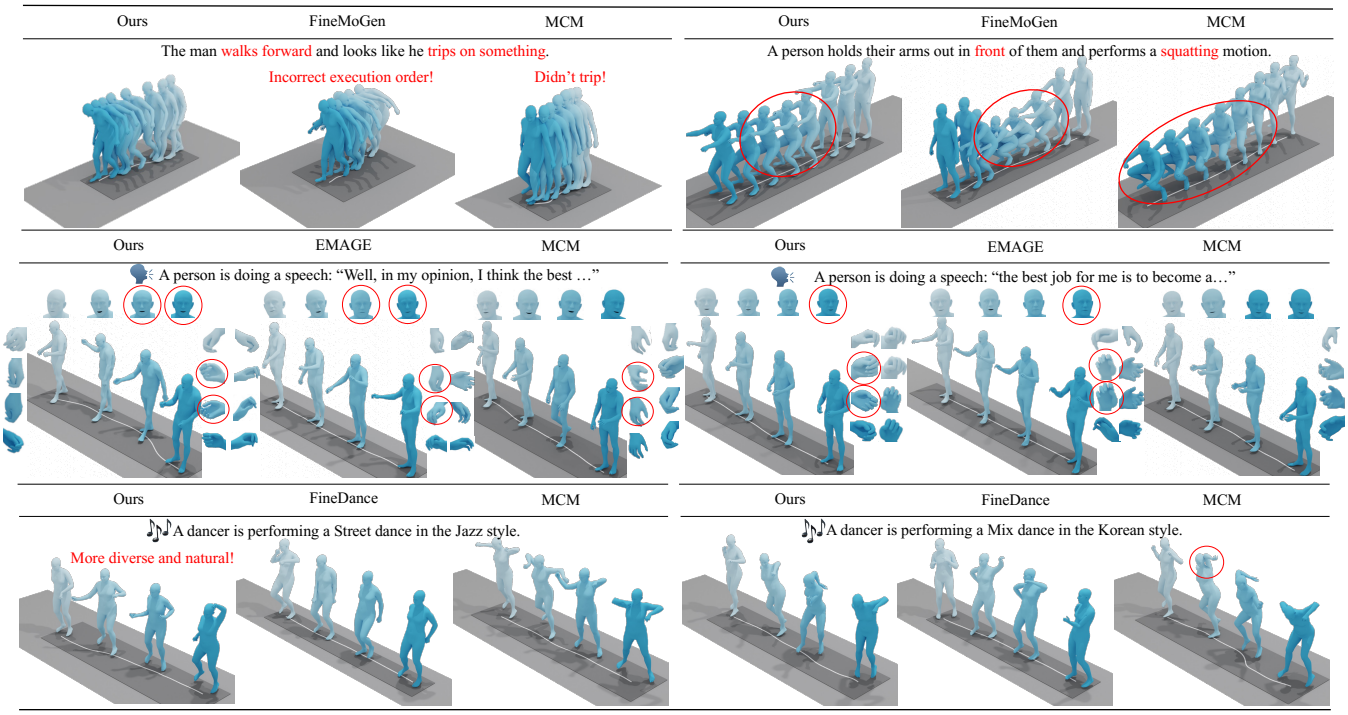

Figure 4: The qualitative results of *MotionCraft* and other state-of-the-art baselines on three representative tasks, text-to-motion, speech-to-gesture, and music-to-dance. More detailed visualization comparisons are in our supplementary.

with SMPL-X parameters and minimal alignment errors compared to the official body-retargeting method. We then pre-train a motion encoder and a text encoder by aligning text and motion contrastively with a retrieval optimization goal (Lu et al. 2023) for a unified evaluation of the SMPL-X motion representation. For FineDance and BEAT2, which lack corresponding textual information, we generate pseudo-captions such as "A dancer is performing a street dance in the Jazz style to the rhythm of the wildfire" and "A person is giving a speech, and the content is ...".

Experiments

Implementation Details

We designed two model variants for the first stage of Text-to-Motion backbone training, *MotionCraft*-Basic and *MotionCraft*-Mix, which were trained on the HumanML3D subset in *MC-Bench* and the entire *MC-Bench*, respectively. In the second stage, we used BEAT2 (Liu et al. 2024a), a large dataset for speech gesture synthesis, and FineDance (Li et al. 2023), a high-quality choreography dataset, to train control branches for Speech-to-Gesture and Music-to-Dance. *MotionCraft*-Basic and *MotionCraft*-Mix share the same 4-layer transformer backbone configuration, dividing the body topology into 12 parts, each with a body-part hidden encoding dimension of 64. *MC-Bench* used a unified whole-body motion format SMPL-X (Pavlakos et al. 2019) in the form of axis-angle, instead of the joint positions or 6D rotation. Thus we retrained the motion and text encoder based on SMPL-X using OpenTMR (Lu et al. 2023) for evaluation.

Evaluation Metrics

Text-to-Motion. We use Fréchet Inception Distance (**FID**) to measure the distribution distance between generated motion and the ground truth, and diversity (**Div**) to measure the average pairwise Euclidean distance among random pairs of generated motion. Furthermore, we use **R-Precision** to measure how often the top-k closest motions to their corresponding captions are achieved within a 32-sample batch. Finally, we employ Multi-Modal Distance (**MM Dist**) to quantify the average Euclidean distance between motion representations and their corresponding text features.

Speech-to-Gesture. We use FID_H, FID_B, and Div for quality and diversity measurement. FID_H represents the difference between the hand motion distribution and the ground truth gesture distribution, while FID_B focuses on the distance between the distributions of whole-body motion. Moreover, we use the Beat Alignment Score (Li et al. 2021) to measure the alignment between the motion and speech beats and employ L2 Loss to measure the difference between generated and real expressions.

Music-to-Dance. Similar to Speech-to-Gesture, we use FID_H, FID_B, and Div to measure the quality of music-to-motion generation for hand and whole-body movements, as well as the diversity of the generated motions.

Quantitative and Qualitative Results

We evaluate *MotionCraft* on three representative tasks: ① Text-to-Motion, ② Speech-to-Gesture, and ③ Music-to-

Method	R Precision			FID ↓	Div ↑	MM Dist↓
	Top-1 ↑	Top-2 ↑	Top-3 ↑			
GT	$0.663^{\pm0.006}$	$0.807^{\pm0.002}$	$0.864^{\pm0.002}$	$0.000^{\pm0.000}$	$36.423^{\pm0.183}$	$15.567^{\pm0.036}$
T2M-GPT(Zhang et al. 2023b)	$0.529^{\pm0.004}$	$0.652^{\pm0.003}$	$0.732^{\pm0.003}$	$10.457^{\pm0.108}$	$36.114^{\pm0.098}$	$17.029^{\pm0.039}$
MDM(Tevet et al. 2023)	$0.383^{\pm0.010}$	$0.527^{\pm0.012}$	$0.604^{\pm0.009}$	$18.671^{\pm0.370}$	$36.156^{\pm0.103}$	$18.785^{\pm0.054}$
MotionDiffuse(Zhang et al. 2024a)	$0.525^{\pm0.004}$	$0.675^{\pm0.009}$	$0.743^{\pm0.009}$	$9.982^{\pm0.379}$	$36.187^{\pm0.160}$	$17.314^{\pm0.066}$
FineMoGen(Zhang et al. 2023c)	$0.565^{\pm0.001}$	$0.710^{\pm0.004}$	$0.775^{\pm0.004}$	**7.323**$^{\pm0.143}$	$36.324^{\pm0.069}$	$16.679^{\pm0.029}$
MCM(Ling et al. 2023)	$0.407^{\pm0.002}$	$0.559^{\pm0.003}$	$0.636^{\pm0.001}$	$15.540^{\pm0.443}$	$35.813^{\pm0.137}$	$18.673^{\pm0.029}$
MotionCraft-Basic	**0.590**$^{\pm0.003}$	**0.743**$^{\pm0.002}$	**0.804**$^{\pm0.004}$	$8.477^{\pm0.102}$	**36.210**$^{\pm0.089}$	**16.252**$^{\pm0.035}$
MotionCraft-Mix	**0.600**$^{\pm0.003}$	**0.747**$^{\pm0.004}$	**0.812**$^{\pm0.006}$	**6.707**$^{\pm0.081}$	**36.419**$^{\pm0.047}$	**16.334**$^{\pm0.059}$

Table 2: **Results of Text-to-Motion in HumanML3D of *MC-Bench*.** We compare the results of text-to-motion between ours and the SOTA methods. **Red** and **Blue** colors indicate the best and second-best results respectively.

Dance, analysing both quantitative and qualitative results.[1] More visualization comparisons are in our supplementary.

Comparison on Text-to-Motion Generation. In the text-to-motion task, we compare *MotionCraft* with current SOTA baselines (Zhang et al. 2023c; Ling et al. 2023; Zhang et al. 2024b,a; Tevet et al. 2023; Zhang et al. 2023b) in two benchmarks: the HumanML3D subset with whole-body format SMPL-X of *MC-Bench* in Tab. 2 and the original HumanML3D (Guo et al. 2022) with the tensor-only format (The results are in supplementary due to page limit). In both benchmarks, *MotionCraft* achieved better text-guided generation capability, diversity, and motion generation quality. Notably, in the HumanML3D subset of *MC-Bench*, the inadequate evaluation abilities in the original HumanML3D benchmark with torso-only representation were significantly improved, providing a more comprehensive and objective comparison. This is because the whole-body SMPL-X representation requires the model to generate the torso movements, gestures, and expressions rather than the only torso. Additionally, we found that *MotionCraft*-Mix trained on the *MC-Bench* has a significant advantage over *MotionCraft*-Basic. This is because *MotionCraft*-Mix can efficiently transfer human topology knowledge against distribution drifts in various generation scenarios. Visualization is in Fig. 4, and *MotionCraft* can follow diverse textual descriptions with fine-grained control.

Comparison on Speech-to-Gesture Generation. In Tab. 3, we compared *MotionCraft* with MCM (Ling et al. 2023), Talkshow (Yi et al. 2023), and EMAGE (Liu et al. 2024a). Our model achieved good quality and diversity in both hand and whole-body motion generation and excelled in aligning with the rhythm of first-perspective speech. This is credited to our coarse-to-fine training strategy and the robust topology knowledge learned from the static and dynamic human topology graphs. However, in expressions, *MotionCraft*-Mix performs slightly worse than EMAGE and Talkshow. This arises from origin dataset limitations in HumanML3D and FineDance, where the face was filled with random or average expressions, confusing the first training stage that affects the following S2G generation. Still, we find *MotionCraft*-Mix possesses a notable performance boost against *MotionCraft*-Basic, further confirming that *MC-Attn* learned robust topology knowledge that can be generalized across different generation scenarios. Qualitative results in Fig. 4 clearly show that *MotionCraft* can effectively follow the beats and generate reasonable gestures and lip movements.

Comparison on Music-to-Dance Generation. *MotionCraft* achieves performance comparable to the SOTA baselines, as shown in Tab. 3. Both variants of our model perform well in diversity, attributed to the first stage of coarse text-to-motion generation training. This equips the model with extensive motion topology knowledge across various scenarios. However, *MotionCraft*-Mix has an increase in **FID** compared to *MotionCraft*-Basic. This is likely due to the FineDance dataset's lack of necessary text descriptions, leading to identical pseudo-captions for different segments of the same song during the first stage of training. This one-to-many generation mode confuses when the model incorporates corresponding music information for each segment in the second stage, attempting to learn many-to-many relationships. Qualitative results in Fig. 4 show that *MotionCraft* can generate natural dances according to the music beats.

Ablation Study

We conducted ablation explorations about the necessity of *MC-Attn* design and scaling up influences in Tab .4.

Different motion topology modeling designs. We have three key observations about decoupling the static and dynamic human topology graph learning. ❶ ***Only modeling static topology decreases performance in T2M but significantly improves performance in S2G and M2D.*** We attribute this to the static topology ensuring the model grasps basic spatial relations between body parts, enhancing generalization across various generation scenarios. However, the additional learnable spatial structure module, unrelated to input, increases learning difficulty in the T2M task. ❷ ***Only modeling dynamic topology nearly brings no benefit.*** This is because the initial optimization of the input-adaptive dynamic topology adjacency matrix is complex, especially for transferring topology knowledge against distribution drifts, making it hard to converge to the correct dynamic topology graph (Zhang et al. 2024b). ❸ ***Joint modeling of static and dynamic topologies effectively captures motion knowl-***

[1]Missing expressions are filled with zero for generated motion and ground truth to avoid affecting the evaluation results.

S2G-Method	$FID_H \downarrow$	$FID_B \downarrow$	Face L2 Loss ↓	Beat Align Score ↑	Div ↑	M2D-Method	$FID_H \downarrow$	$FID_B \downarrow$	Div ↑
Talkshow	26.713	74.824	**7.791**	6.947	**13.472**	Edge	93.430	108.507	13.471
EMAGE	39.094	90.762	**7.680**	7.727	13.065	Finedance	10.747	**72.229**	13.813
MCM	23.946	71.241	16.983	7.993	**13.167**	MCM	4.717	78.577	14.890
MotionCraft-Basic	**18.486**	**27.023**	10.097	**8.098**	10.334	*MotionCraft*-Basic	**3.858**	76.248	**16.667**
MotionCraft-Mix	**12.882**	**25.187**	8.906	**8.226**	12.595	*MotionCraft*-Mix	**2.849**	**67.159**	**18.483**

Table 3: **Results of Speech-to-Gesture in BEAT2 and Music-to-Dance in FineDance of *MC-Bench*.** We respectively evaluate the FID_H and FID_B, Face L2 Loss$\times 10^{-8}$, Beat Align Score$\times 10^{-1}$, and diversity for S2G and the FID_H, FID_B, and the diversity for M2D. **Red** and **Blue** colors indicate the best and second-best results respectively.

Method		HumanML3D (Text-to-Motion)					BEAT2 (Speech-to-Gesture)					Finedance (Music-to-Dance)		
Dynamic-Spatial	Static-Spatial	Top-1 ↑	Top-2 ↑	Top-3 ↑	FID ↓	Div ↑	$FID_H \downarrow$	$FID_B \downarrow$	Face L2 ↓	Beat Align Score ↑	Div ↑	$FID_H \downarrow$	$FID_B \downarrow$	Div ↑
✗	✗	0.583	0.729	0.794	8.911	35.954	15.587	31.839	12.448	7.908	11.752	7.088	150.733	17.984
✗	✓	0.557	0.706	0.772	9.041	36.101	12.929	27.928	12.287	8.077	12.230	5.104	112.186	**18.503**
✓	✗	0.582	0.732	0.798	8.455	36.241	15.517	28.631	12.544	7.708	11.313	4.972	102.103	16.385
✓	✓	**0.600**	**0.747**	**0.812**	**6.707**	**36.419**	**12.882**	**25.187**	**8.906**	**8.226**	**12.595**	**2.849**	**67.159**	18.483
MotionCraft-Tiny-(4, 64, 77M)		0.600	0.747	0.812	6.707	**36.419**	**12.882**	**25.187**	8.906	**8.226**	**12.595**	**2.849**	67.159	**18.483**
MotionCraft-Small-(4, 128, 130M)		**0.653**	**0.794**	**0.847**	**5.593**	36.264	15.346	27.140	**8.322**	8.023	**11.906**	**2.370**	**59.471**	**17.036**
MotionCraft-Small-(8, 64, 145M)		0.635	0.779	0.802	6.193	36.311	15.702	28.094	8.589	8.031	11.824	3.749	**66.958**	16.478
MotionCraft-Medium-(8, 128, 250M)		**0.647**	**0.785**	**0.854**	**5.670**	**36.384**	**14.937**	**23.498**	**8.125**	**8.089**	10.962	3.904	75.412	16.507
MotionCraft-Large-(16, 128, 478M)		0.604	0.744	0.809	7.872	36.169	15.964	27.476	9.036	7.969	10.625	4.837	77.341	16.426

Table 4: **Ablation Study. (a) Ablation on model design (Upper half).** The results suggest that jointly modeling dynamic and static human skeleton topologies significantly improves performance since this provides robust topology knowledge against distribution drifts. **(b) Ablation on scaling up impacts (Lower half).** We design four scaling variants, where **-(a, b, c) denotes model ** with a transformer layer, b body-part encoding dimension, and total c parameters. We observe a rise-then-fall performance trend as the model size increases. **Red** and **Blue** colors indicate the best and second-best results respectively.

edge against distribution drifts, as in human-centric research (Zeng et al. 2021). The static topology learns basic human structure, providing foundational spatial knowledge across tasks, while the dynamic topology adjusts according to specific motion distributions and control signals.

Scaling up impacts. Based on the acknowledgment of the scalability of transformer models, we explored the impact of model size on task performance. We increased the size of *MotionCraft*-Mix from $77M$ to $478M$, observing a rise-then-fall performance trend across three types of tasks as the model size increased with limited data. This verifies that increasing the model's parameter size can enhance generative capabilities, but without a corresponding increase in high-quality data, model performance may decline.

Application: Multimodal Video Generation

To demonstrate the downstream application, in Fig. 5, we present two animation videos driven by *MotionCraft* in M2D and S2G. Our generated motion sequences can be combined with any off-the-shelf human video generation framework, such as MimicMotion (Zhang et al. 2024c), AnimateAnyone (Hu et al. 2023), and VividPose (Wang et al. 2024), enabling users to customize videos of any character based on specific control signals, such as speech or music. Notably, unlike the traditional 2D keypoints estimated from videos, our generated 3D motion approach allows for flexible adjustment of camera parameters to project different visible body regions (e.g., full body or upper body, as in Fig. 5). More detailed visualizations are in our supplementary.

Figure 5: Multimodal video generation application with our generated motions conditioned on music (upper row) or speech (lower row). We project them to 2D images to serve as motion conditions for MimicMotion (Zhang et al. 2024c).

Conclusion

We present *MotionCraft*, a unified diffusion transformer for whole-body human motion synthesis that supports plug-and-play multimodal control across diverse generative distributions and control signal granularities. Through a coarse-to-fine training strategy, *MotionCraft* enables precise, modular control across multiple modalities (text, speech, and music) while avoiding the computational overhead of joint training. Our core design is *MC-Attn*, which effectively captures and transfers motion patterns by simultaneously modeling static and dynamic human topology graphs. We introduce *MC-Bench*, the first comprehensive benchmark for multimodal whole-body motion generation using the unified SMPL-X representation. Extensive experiment results demonstrate that *MotionCraft*achieves state-of-the-art performance across standard motion generation tasks.

Acknowledgments

This work was supported in part by the CUHK Strategic Seed Funding for Collaborative Research Scheme under Grant No. 3136023.

References

Bian, Y.; Ju, X.; Li, J.; Xu, Z.; Cheng, D.; and Xu, Q. 2024. Multi-Patch Prediction: Adapting Language Models for Time Series Representation Learning. In *Forty-first International Conference on Machine Learning*.

Chen, J.; Liu, Y.; Wang, J.; Zeng, A.; Li, Y.; and Chen, Q. 2024. DiffSHEG: A Diffusion-Based Approach for Real-Time Speech-driven Holistic 3D Expression and Gesture Generation. In *CVPR*.

Chung, J.; Wuu, C.-h.; Yang, H.-r.; Tai, Y.-W.; and Tang, C.-K. 2021. Haa500: Human-centric atomic action dataset with curated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13465–13474.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15180–15190.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5152–5161.

Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2021–2029.

Hu, L. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8153–8163.

Hu, L.; Gao, X.; Zhang, P.; Sun, K.; Zhang, B.; and Bo, L. 2023. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. *arXiv preprint arXiv:2311.17117*.

Kim, J.; Kim, J.; and Choi, S. 2023. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021. Learn to Dance with AIST++: Music Conditioned 3D Dance Generation. arXiv:2101.08779.

Li, R.; Zhao, J.; Zhang, Y.; Su, M.; Ren, Z.; Zhang, H.; Tang, Y.; and Li, X. 2023. FineDance: A Fine-grained Choreography Dataset for 3D Full Body Dance Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10234–10243.

Liang, H.; Bao, J.; Zhang, R.; Ren, S.; Xu, Y.; Yang, S.; Chen, X.; Yu, J.; and Xu, L. 2024. Omg: Towards open-vocabulary motion generation via mixture of controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 482–493.

Lin, J.; Zeng, A.; Lu, S.; Cai, Y.; Zhang, R.; Wang, H.; and Zhang, L. 2023. Motion-X: A Large-scale 3D Expressive Whole-body Human Motion Dataset. *Advances in Neural Information Processing Systems*.

Ling, Z.; Han, B.; Wong, Y.; Kangkanhalli, M.; and Geng, W. 2023. Mcm: Multi-condition motion synthesis framework for multi-scenario. *arXiv preprint arXiv:2309.03031*.

Liu, H.; Zhu, Z.; Becherini, G.; Peng, Y.; Su, M.; Zhou, Y.; Zhe, X.; Iwamoto, N.; Zheng, B.; and Black, M. J. 2024a. EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Expressive Masked Audio Gesture Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1144–1154.

Liu, H.; Zhu, Z.; Iwamoto, N.; Peng, Y.; Li, Z.; Zhou, Y.; Bozkurt, E.; and Zheng, B. 2022. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*, 612–630. Springer.

Liu, J.; Dai, W.; Wang, C.; Cheng, Y.; Tang, Y.; and Tong, X. 2023. Plan, posture and go: Towards open-world text-to-motion generation. *arXiv preprint arXiv:2312.14828*.

Liu, Y.; Zhang, K.; Li, Y.; Yan, Z.; Gao, C.; Chen, R.; Yuan, Z.; Huang, Y.; Sun, H.; Gao, J.; He, L.; and Sun, L. 2024b. Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. arXiv:2402.17177.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16.

Lu, S.; Chen, L.-H.; Zeng, A.; Lin, J.; Zhang, R.; Zhang, L.; and Shum, H.-Y. 2023. HumanTOMATO: Text-aligned Whole-body Motion Generation. *arxiv:2310.12978*.

Luo, M.; Hou, R.; Chang, H.; Liu, Z.; Wang, Y.; and Shan, S. 2024. M$\hat{3}$-GPT: An Advanced Multimodal, Multi-task Framework for Motion Comprehension and Generation. *arXiv preprint arXiv:2405.16273*.

Ma, J.; Bai, S.; and Zhou, C. 2022. Pretrained diffusion models for unified human motion synthesis. arXiv 2022. *arXiv preprint arXiv:2212.02837*.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 10975–10985.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Siyao, L.; Yu, W.; Gu, T.; Lin, C.; Wang, Q.; Qian, C.; Loy, C. C.; and Liu, Z. 2022. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11050–11059.

Tang, Y.; Liu, J.; Liu, A.; Yang, B.; Dai, W.; Rao, Y.; Lu, J.; Zhou, J.; and Li, X. 2023. FLAG3D: A 3D Fitness Activity Dataset with Language Instruction. In *CVPR*.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*.

Trivedi, N.; Thatipelli, A.; and Sarvadevabhatla, R. K. 2021. NTU-X: an enhanced large-scale dataset for improving pose-based recognition of subtle human actions. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, 1–9.

Tseng, J.; Castellon, R.; and Liu, K. 2023. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 448–458.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Q.; Jiang, Z.; Xu, C.; Zhang, J.; Wang, Y.; Zhang, X.; Cao, Y.; Cao, W.; Wang, C.; and Fu, Y. 2024. VividPose: Advancing Stable Video Diffusion for Realistic Human Image Animation. *arXiv preprint arXiv:2405.18156v1*.

Yi, H.; Liang, H.; Liu, Y.; Cao, Q.; Wen, Y.; Bolkart, T.; Tao, D.; and Black, M. J. 2023. Generating Holistic 3D Human Motion from Speech. In *CVPR*.

Zeng, A.; Sun, X.; Yang, L.; Zhao, N.; Liu, M.; and Xu, Q. 2021. Learning skeletal graph neural networks for hard 3d pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11436–11445.

Zhang, J.; Yan, H.; Xu, Z.; Feng, J.; and Liew, J. H. 2023a. MagicAvatar: Multi-modal Avatar Generation and Animation. In *arXiv:2308.14748*.

Zhang, J.; Zhang, Y.; Cun, X.; Huang, S.; Zhang, Y.; Zhao, H.; Lu, H.; and Shen, X. 2023b. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024a. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, M.; Jin, D.; Gu, C.; Hong, F.; Cai, Z.; Huang, J.; Zhang, C.; Guo, X.; Yang, L.; He, Y.; et al. 2024b. Large motion model for unified multi-modal motion generation. *arXiv preprint arXiv:2404.01284*.

Zhang, M.; Li, H.; Cai, Z.; Ren, J.; Yang, L.; and Liu, Z. 2023c. FineMoGen: Fine-Grained Spatio-Temporal Motion Generation and Editing. *NeurIPS*.

Zhang, S.; Ma, Q.; Zhang, Y.; Qian, Z.; Kwon, T.; Pollefeys, M.; Bogo, F.; and Tang, S. 2022. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, 180–200. Springer.

Zhang, Y.; Gu, J.; Wang, L.-W.; Wang, H.; Cheng, J.; Zhu, Y.; and Zou, F. 2024c. MimicMotion: High-Quality Human Motion Video Generation with Confidence-aware Pose Guidance. *arXiv preprint arXiv:2406.19680*.

Zhou, Z.; Wan, Y.; and Wang, B. 2023. A unified framework for multimodal, multi-part human motion synthesis. *arXiv preprint arXiv:2311.16471*.