DanceEditor: Towards Iterative Editable Music-driven Dance Generation with Open-Vocabulary Descriptions

Supplementary Material

A. Overview

The supplementary material includes the subsequent components.

- Details of Automatic Dataset Construction WorkFlow
- Details of Methodology and Experiment.
  - Architecture Details.
  - Explanation on the evaluation metrics.
  - Clarification of iterative editing.
  - Detailed ablations about editability.
- Visualization of the ablation studies.
- Demo Video

B. Dataset

Inspired by the recent success of GPT-based models for motion-annotation tasks [6, 21], we propose a data collection methodology that integrates the capabilities of existing large language models (LLMs), such as ChatGPT [12], and multimodal large language models (MLLMs), such as Gemini [18], with the text-motion retrieval model TMR [14] to facilitate editable dance data collection.

B.1. Construction of Our DanceRemix Dataset

In this section, we give a detailed explanation of the data processing pipeline of our DanceRemix dataset. We summarize the acquisition, processing, captioning and final generation of our DanceRemix dataset in two procedures: **Retrieval of Similar Dance Pairs** and **Editing Description Generation**.

B.2. Dataset sources and dance genres.

Our DanceRemix dataset incorporates AIST++, a subset of Motion-X, as well as videos from YouTube and Bilibili. It covers 20 dance genres, including C-pop, K-pop, J-pop, house dance, ballet, Latin dance, jazz, tap dance, folk dance, modern dance, and 10 street dance styles from AIST (Break, Pop, Lock, Waack, Middle Hip-hop, LA-style Hip-hop, House, Krump, Street Jazz, and Ballet Jazz).

B.3. Retrieval of Similar Dance Pairs.

To build a high-quality dance dataset that supports multi-turn editing, we must ensure it features natural transitions between dance sequences, provides plausible editing instructions, and maintains precise alignment of all movements with carefully tailored music.

Creating such a dataset—precisely aligning iterative motions with music while providing accurate textual descriptions—presents significant challenges. The inherently subjective nature of dance complicates consistent description, while the labor-intensive process of manual annotation further drives up costs. Consequently, we opted to develop an automated data collection workflow.

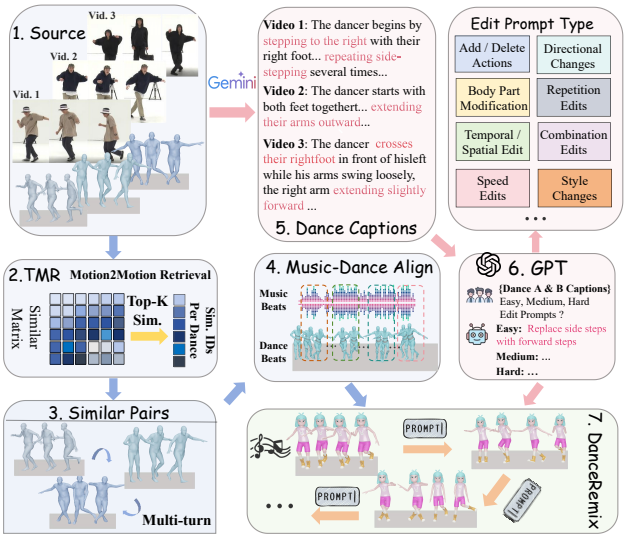

Figure 1. The workflow of DanceRemix dataset construction. Firstly, we perform motion-to-motion retrieval to obtain similar dance motion pairs. Then, we align the motion beats of the edited dance with the music beats. For aligned dance pairs, we use Gemini to generate dense dance captions for the dance videos. Next, based on the generated captions, we leverage ChatGPT to generate edit instructions. Through several motion pair retrievals, we obtain music, seed dance, a series of edit prompts, and corresponding edited dance motions. In this way, we construct the first large-scale, multi-modal, multi-turn editable dance dataset.

Collection of Dance Videos and Motions: To harness advanced general-purpose multimodal large models—renowned for their powerful comprehension and captioning capabilities—to generate detailed descriptions of human dance videos, we need a substantial collection of dance footage. We draw on existing dance datasets from previous work, which include original dance videos, and supplement them with additional online videos. For each dance motion in our collection, we search for the most similar ones from our collection to create pairs of similar dances as candidates for subsequent editing.

Motion-to-Motion Retrieval. The first challenge in our automatic data curation pipeline is obtaining similar dance movements. We need to ensure that the dance motion pairs we select are sufficiently alike to allow for a concise, meaningful, and plausible edit text describing their differences. At the same time, the transition from one motion to the other should feel natural. This requirement means the differences between the pairs cannot be too large, to avoid forcing the LLM annotators to produce awkward edits. Motivated by the application of CLIP [16] in image similarity, we employ text-motion retrieval approach TMR [14] to perform motion-to-motion retrieval in a similar fashion. We find that the recent TMR motion encoder effectively captures semantic information as well as sufficient detail for body dynamics. Thanks to its contrastive and generative training on large-scale motion data, we were pleasantly surprised to find that it also performs effectively for dance movements. We extract TMR motion embeddings using 5-second sliding windows and compute pairwise similarities, since using shorter ones usually yields motion pairs that have a high probability to be almost identical. Then, we compute the pairwise embedding similarities and filter out all motion pairs with a similarity score ≥ 0.99 to eliminate identical motions. Inspired by [1], we apply a top-k selection strategy to identify the the top-k most similar dance motions. Additionally, we find that dance pairs retrieved from top-2 to top-5 are sufficiently similar yet distinct enough to support further editing.

Music-Dance Align. To ensure that the similar motion pairs retrieved through TMR remain aligned with the same piece of music, we apply Dynamic Time Warping (DTW) to match the musical beats B_m with the visual motion beats B_v. The goal is to minimize the total Euclidean distance between corresponding beats. During the search for the minimum-distance path, we use the Rabiner-Juang step pattern [15]. We identify motion beats by detecting sudden decelerations in the motion sequence as a heuristic for "visual beats.", similar to [20]. Once the beat alignment is established, we warp the motion sequences according to the resulting warping curve. In this manner, movements that are not in harmony with the beat of the music are discarded and then manually verified by visualizing them alongside the audio. Finally, we apply first-frame canonicalization to ensure that the motions face the same direction and have identical initial global positions, similar to the approaches in [1, 13].

Post Processing. To further ensure the quality and reliability of the motion data, we implement a multi-stage post-processing pipeline that integrates rule-based filtering, motion smoothing, and manual verification.

B.4. Editing Description Generation.

Leveraging advanced MLLMs to generate captions and transformation scripts for similar dance pairs is both practical and valuable, given the inherently subjective nature of dance, which makes it difficult to describe.

Why Choose Dense Dance Caption to generate edit texts? We have obtained similar dance pairs that align well with musical beats and rhythms. Now, we need to annotate the transformation from one movement to another using MLLMs. To achieve this, we compare five different approaches for generating editing descriptions:

- **Input Dense Dance Caption:** We utilize dense captions, well-aligned with the original dance videos generated by MLLMs, to enable ChatGPT [12] to produce effective transformation scripts. This approach significantly enhances editing quality, delivering impressive results.
- **Input Key Step Description:** Inspired by [21], we provide dense dance captions to ChatGPT-4 as if it were watching the dance videos. It then breaks down complex and detailed dance descriptions into clear and concise dance steps, summarizing key dance movements,
- **Input Dance Pair Videos :** We directly input paired dance videos into MLLMs [18] and carefully design prompts to generate precise edit instructions.
- **Input Videos & Captions:** We first experiment with inputting videos alongside captions to provide the most comprehensive contextual information, ensuring a detailed and accurate understanding of the dance movements.
- **Input Videos & Step Descriptions:** Since inputting both videos and captions results in excessive tokens, we also experiment with using videos and key step descriptions to facilitate the generation of edit descriptions.

Through extensive comparative experiments, we found that inputting dense captions yields the best results, while directly inputting videos performs the worst. This may be because MLLMs struggle with fine-grained analysis of differences between two dance videos due to a lack of domain-specific training.

Additionally, using videos combined with supplementary text (captions or step descriptions) also produces suboptimal results. One possible reason is that the excessive token input makes it difficult for the model to extract task-relevant information from redundant details. Another potential factor is the reduced reasoning capability of MLLMs compared to LLMs.

Dense Dance Caption Generation. The core insight of our work is leveraging advanced MLLMs to generate dense dance captions that closely align with human perception.

Recent advancements in MLLMs have demonstrated the effectiveness of their captions for this task.

After evaluating various open-source and closed-source models, we select Gemini-1.5-Pro [18] for its optimal balance of cost and performance. Additionally, we carefully design prompts to prioritize dynamic body movements over static pose descriptions. For caption granularity, we take inspiration from the detailed dance descriptions in Ego-Exo4D [3].

We carefully design our caption generation prompts by referencing a wide range of existing motion datasets [3, 4, 6–9, 21] and choreography-related literature [11]. Our goal is to prompt Gemini to focus on dynamic body movements rather than static keyframe pose descriptions. The concurrent study [9] inputs detailed pose descriptions from PoseScript [2] into ChatGPT-4o-mini, converting frame-level captions into sentence-level captions. However, we find that these so-called sentence-level captions primarily describe transitions such as "converted from one pose to another" rather than capturing full dynamics of dance movements.

Our carefully designed prompts, as illustrated in Figure 3, guide Gemini to generate detailed descriptions by emphasizing actions and specific body movements rather than merely describing poses. Here, we prompt Gemini to indicate the frame index for each dance motion, allowing us to later verify whether the captions align with the video. Additionally, we specifically emphasize summarizing the movements into $1 \sim 3$ steps to encourage Gemini to capture the semantic essence of the dance. Excessive step segmentation may cause Gemini to emphasize static poses over dynamic transitions, resulting in poorer performance.

Since dance movements change rapidly, using too few video frames can result in the omission of many expressive dance motions. To address this, we experiment with different frame rates (FPS) for input images. Drawing inspiration from traditional film standards (24 FPS), we compare FPS 4, FPS 6, and FPS 12 while keeping the prompt constant. Our results show that FPS 6 yields the best performance. At FPS 4, many crucial movements are not captured in the input, leading to incomplete descriptions. Conversely, at FPS 12, the excessive visual information introduces redundancy, making it harder for the model to focus on key motion details. Ultimately, we choose an input frame rate of 6 FPS for the image sequences fed into Gemini.

Edit Descriptions Generation. Through the previous processing steps, we obtained similar dance pairs and their corresponding dense dance captions. Based on these, we generate edit instructions by providing them to ChatGPT-4, utilizing detailed dance captions to create three level edit descriptions with different granularity. For example:

- **Easy:** "Extend arms upward instead of waving."
- **Medium:** "Raise the right arm diagonally while lowering the left arm as a mirror."
- **Hard:** "Rotate the torso to the right, extend the left arm in an arc, lift the right knee, extend the right leg forward, and sweep the arms fluidly."

Here, the easy, medium, and hard versions of the prompts correspond to different levels of edit granularity, ranging from prominent body movements to more detailed limb-specific modifications. The prompt used in our workflow is shown is Figure 2.

Prompts for Edit Description

You have a source dance A and target dance B, both provided through text descriptions. Please generate an edit prompt that transforms Dance A into Dance B, as if you are watching the dance video. The edit prompt should be 3-25 words, focusing on specific body movement modifications, such as adding left hand on the floor, kicking leg higher, swinging arms up and down, etc. Use words indicative of edit texts, e.g., 'instead','while', 'higher/lower', 'same/opposite'.

Dance A is {xxxxx}, Dance B is {xxxxx}.

Generate three versions of the edit prompt as follows:
Easy version: Contains only the main movement edits, simple and easy to understand.
Medium version: Adds some details, slightly more complex.
Hard version: Includes more details and continuity of movements, complex and challenging.

JSON template:
{
"Easy": Add hands reaching floor,
"Medium": Change steps to running in place, bend arms and swing knees,
"Hard": "Convert deep squat into lunge, shift weight onto left leg, extend right leg forward, swing arms overhead, pivot into body roll.'
Only output the edit prompt",
}

Figure 2. **Prompts for Edit Description Generation Based on Captions.**

Diverse Editing Types. As shown in Figure 1, The editing instructions in our DanceRemix are highly diverse, thanks to Gemini's powerful captioning capabilities for human motion videos and GPT-4's strong reasoning abilities. The most common types of edits in our DanceRemix Dataset are as follows:

- **Add / Delete Actions.** "Add both hands to the floor."
- **Specific Body Part Editing.**"Arch right arm overhead, extend left leg forward."
- **Spatial / Temporal Changes.**"Swing arms wide", "Start in crouch."
- **Directional Changes.** "Extend left leg instead of right."
- **Repetition Edits.**"Open arms wider, step right three times."
- **Style Changes.**"Swing arms freestyle, jump from crouch."
- **Speed Edits.** "Circle arms faster, end with playful arm

raise."

- **Combination Edits.** "Change to small run, extend arms forward, then bounce." "Bend knees, widen stance, jump with arms overhead."

DanceRemix and DanceRemix-X Dataset. Benefiting from the proposed automatic data collection workflow, we constructed **DanceRemix**, the first large-scale, high-quality editable dance dataset. It features over **25.3M** dance frames and **84.5K** editing prompt pairs. To further enhance model training and accommodate more complex, fine-grained edit descriptions, we developed an extended version, **DanceRemix-X**. This dataset retrieves similar dance motion pairs to build multi-turn editable dance sequences with three-level edit descriptions, enabling more refined and progressive editing.

Dance Caption Prompt

Dance Motion Description Task

You will be provided with 30 frames uniformly sampled from a 5-second human dance video. Combine all the frames and the given general description as if you are actually watching the video. Describe the **key actions** in clear, concise, and dynamic steps. Each step must specify the **frame range** (e.g., *frame 0~6*) and highlight **specific body movements**. Focus on transitions, flow, and spatial relationships, ensuring distinctions between **left** and **right limbs**.

Instructions

1. **Frame Ranges**

- Specify the **frame range** for each step (e.g., *frame 0~6*). Make sure each range represents a meaningful transition or unique phase of movement.

2. **Action-Oriented Descriptions**

- Focus on **observable and dynamic movements**, avoiding vague or overly static phrases.
  - **Good Example:** "The dancer spins counterclockwise, extending their left arm outward."
  - **Avoid:** "The dancer stands still and prepares."

3. **Summarize in 1~3 Steps**

- Break the sequence into **1~3 key steps**, each step capturing a significant part of the motion.
- Highlight **unique actions** or transitions in each step, avoiding redundancy.

4. **Detailed Movements**

- Include:
  - **Upper body:** Describe arm swings, hand gestures, shoulder movements, or torso twists.
  - **Lower body:** Focus on steps, jumps, kicks, squats, or weight shifts.
  - **Whole-body coordination:** Highlight how different body parts, including the head, limbs, shoulders, and torso, work together. For example:
    - *"As the right arm arcs forward, the left leg steps back, and the head follows the arm's motion."*
  - **Torso rotation:** If the torso twists or rotates significantly to the left or right, describe it clearly. For example:
    - *"The torso rotates sharply to the left as the arms swing outward in a wide arc."*
    - *"The dancer's torso twists slightly to the right, adding a flowing motion to the step."*

5. **Dynamic Flow and Transitions**

- Emphasize **how movements connect smoothly** and maintain rhythmic flow.
  - Example: "The dancer pivots on their right foot, transitioning into a leap with arms extended."
- Note if movements are abrupt, smooth, explosive, or controlled to reflect the rhythm or energy of the dance.

6. **Spatial Details**

- Specify movement direction, body positioning, and interaction with space.
  - Example: "The dancer steps diagonally backward, leaning slightly to the left while extending their arms forward."
- Consider how the dancer uses their space dynamically, such as stepping outward, inward, or rotating to face a new direction.

7. **Energy and Rhythm**

- Highlight **speed, intensity, or rhythm** of movements.
  - Example: "The dancer performs a sharp turn, quickly shifting weight from the left foot to the right with a burst of energy."
- If the rhythm changes within the sequence, note how the movements respond (e.g., slowing down, speeding up, becoming more fluid, or more staccato).

8. **Left/Right Limb Clarity**

- Always distinguish **left** and **right limbs** for precise and accurate descriptions. For example:
  - *"The dancer's right arm extends upward while the left knee bends slightly."*
- Avoid ambiguity by clearly relating the motion of limbs to their position in space or their interaction with other body parts.

Output Format

1. Use the format `Step [Number]: *frame [start]~[end]*`.
2. Write clear, concise, action-based sentences focusing on **movement dynamics**, **flow**, and **spatial details**.

Examples

- **Step 1:** *frame 0~9* The dancer steps back with the right foot while moving both hands in a smooth circular motion. Their head tilts slightly forward, following the motion of the arms.
- **Step 2:** *frame 10~17* The dancer jumps, rotating their torso to the left, while pulling left hands close to the chest. Their hips twist slightly to the right as they land softly.
- **Step 3:** *frame 18~29* The dancer lowers into a squat, pushing their right hands downward, and rises gracefully with arms extended overhead. Their chest expands as their head follows the upward motion of the arms.

Figure 3. **Prompts for Dense Dance Generation from Videos.** We show prompts we employ to obtain detailed captions focusing on dynamic body movements.

C. Details of Models and Training

C.1. Architecture details.

Music Encoder. Inspired by [10, 19], the backbone of our music encoder consists of a 2-layer Transformer encoder with RoPE [17], and the latent dimension is 1024. We also find that adding a Dense FiLM layer between each layer of our basic transformation block, which takes the concatenation of music embeddings from the music encoder and timesteps as input, effectively models the relationship between music and dance. This enhancement leads to more vivid dance movements that align rhythmically with the music.

Text Encoder. We use CLIP's text encoder to extract the embedding of the edit text and experiment with Conv1D, Linear layers, and a Learnable Query Transformer with a fixed number of query tokens. We expand the embedding from CLIP from 77 to 150, matching the sequence lengths of our music and motion features. Finally, we choose Conv1D for its higher efficiency.

The Injection of Dance Fusion Features. The dance fusion features come from our novel Cross-Modality Editing Module (CEM). We explore three injection methods into the original transformer block: AdaIN, addition, and concatenation. Benefiting from the strong style transfer capability of AdaIN, the generated dance motions not only maintain the rhythmic alignment of the initial prediction with the music but also achieve fine-grained semantic coherence with the edit prompts.

C.2. Explanation on the evaluation metrics

In Section 4.1, we provide a brief overview of the evaluation metrics used in our quantitative analysis. Through a comprehensive survey, we found that while metrics like **FID** and **Diversity** are commonly used in dance and motion generation research, their implementations vary, particularly in the motion feature extraction process. Additionally, we also adopt two commonly used metrics in dance generation: the Beats Align Score (BAS) and the Physical Foot Contact (PFC) score. We further propose and explain a new metric, namely the Motion-Editing Text Align Score (MEAS).

- **Motion-Editing Text Align Score (MEAS):** Existing metrics for evaluating text-motion alignment, such as Multi-modal Distance [5], are not suitable for our task. These metrics focus on describing motions, whereas our text primarily emphasizes motion edits, which are not directly related to the motions themselves. Therefore, we train a contrastive learning model using initial dance and edited dance pairs along with edit instructions. Inspired by [7], we measure the distance between dance motion pairs and editing texts using a custom-trained CLIP-based model. A lower score indicates better alignment between the edited dance and the editing descriptions.

Table 1. Ablation of editing branch (**1-round editing**) && Comparison between direct (combining three prompts into one) and multi-turn editing.

Models	FID ↓	BAS ↑	Diversity ↑	MEAS ↓
w/o Editing Brach	3.95	0.2514	2.32	1.351
Editing Branch w/o CEM	3.68	0.2537	2.69	1.024
Editing Branch w/o inital motion	3.51	0.2548	2.74	0.988
Editing Branch w/o music	3.34	0.2539	2.71	0.932
DanceEditor (1-round editing)	**2.85**	**0.2553**	**3.16**	**0.784**
direct editing	3.21	0.2519	2.92	0.925
3-round editing	**3.04**	**0.2524**	**3.35**	0.793

C.3. Clarification of iterative editing

In Table 3, the multi-round editing results refer to outputs generated after several rounds of editing prompts. We also compare the results of three-round editing with directly editing (combining multiple prompts into one) in Tab. 1.

The more editing, the more complex the motions will be. Despite slightly worse motion quality, user studies and visualizations show that the results remain satisfactory. This minor decline is also commonly observed in multi-turn editing tasks [22] as editing grows more complex.

C.4. Detailed ablations about editability

We conduct additional studies on the editing branch's input components: (1) excluding initial motion features, and (2) removing music integration.

As shown in Tab. 1, our complete editing branch achieves optimal performance, demonstrating that our DanceEditor framework achieves SOTA results while maintaining an excellent trade-off between editability and motion quality.

Furthermore, we present a comparison between directly using complex edit prompts—which combine three individual prompts into one—and performing multi-turn editing with simple edit prompts in each turn. This comparison demonstrates the superiority of achieving complex edits through the multi-turn approach.

D. Visualization of the ablation studies

Here, we present additional visualization results from our ablation study. As illustrated in Figure 4, the complete version of our framework produces vivid and natural dance movements that closely adhere to the semantic details of the edit prompt.

References

[1] Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J Black, and Gül Varol. Motionfix: Text-driven 3d

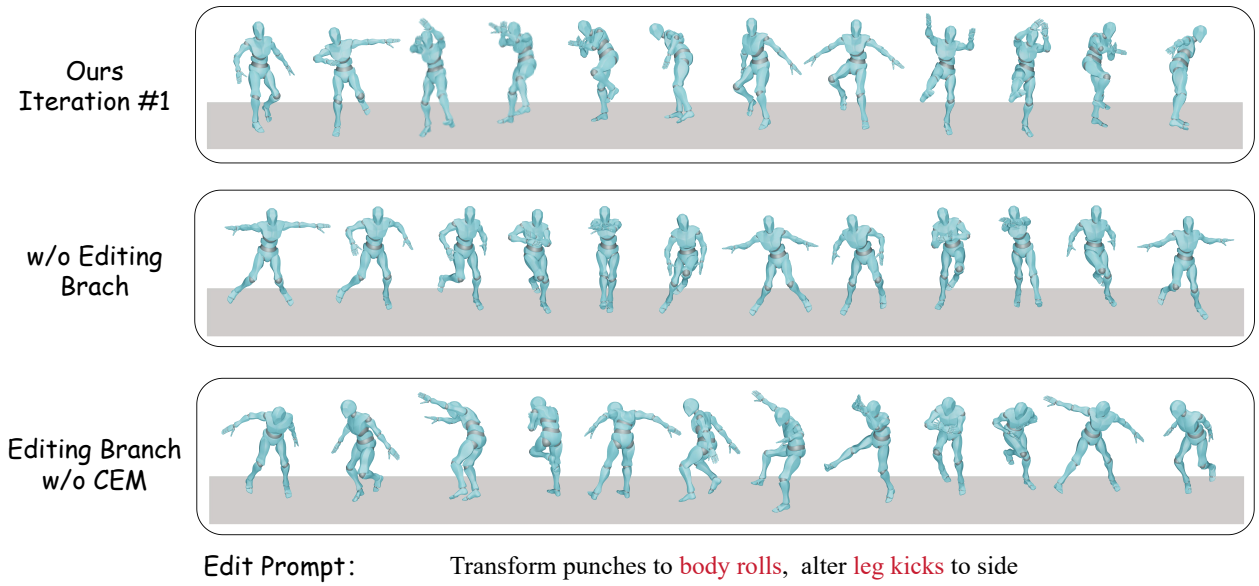

Figure 4. Given the same music piece and edit prompt, we compare the results generated by our full framework with those produced by its ablated versions.

human motion editing. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2

[2] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *European Conference on Computer Vision*, pages 346–362. Springer, 2022. 3

[3] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 3

[4] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022. 3

[5] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 6

[6] Biao Jiang, Xin Chen, Chi Zhang, Fukun Yin, Zhuoyuan Li, Gang Yu, and Jiayuan Fan. Motionchain: Conversational motion controllers via multimodal prompts. In *European Conference on Computer Vision*, pages 54–74. Springer, 2024. 1, 3

[7] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9):3463–3483, 2024. 6

[8] 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. *Advances in Neural Information Processing Systems*, 36: 25268–25280, 2023.

[9] Zeyu Ling, Bo Han, Shiyang Li, Hongdeng Shen, Jikang Cheng, and Changqing Zou. Motionllama: A unified framework for motion synthesis and comprehension. *arXiv preprint arXiv:2411.17335*, 2024. 3

[10] Zhenye Luo, Min Ren, Xuecai Hu, Yongzhen Huang, and Li Yao. Popdg: Popular 3d dance generation with popdanceset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26984–26993, 2024. 6

[11] Sandra Cerny Minton. *Choreography: a basic approach using improvisation*. Human Kinetics, 2017. 3

[12] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 1, 2

[13] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 2

[14] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9488–9497, 2023. 1, 2

[15] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993. 2

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya

Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[17] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 6

[18] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2, 3

[19] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 6

[20] Han Yang, Kun Su, Yutong Zhang, Jiaben Chen, Kaizhi Qian, Gaowen Liu, and Chuang Gan. Unimumo: Unified text, music and motion generation. *arXiv preprint arXiv:2410.04534*, 2024. 2

[21] Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1357–1366, 2024. 1, 2, 3

[22] Zijun Zhou, Yingying Deng, Xiangyu He, Weiming Dong, and Fan Tang. Multi-turn consistent image editing. *arXiv preprint arXiv:2505.04320*, 2025. 6