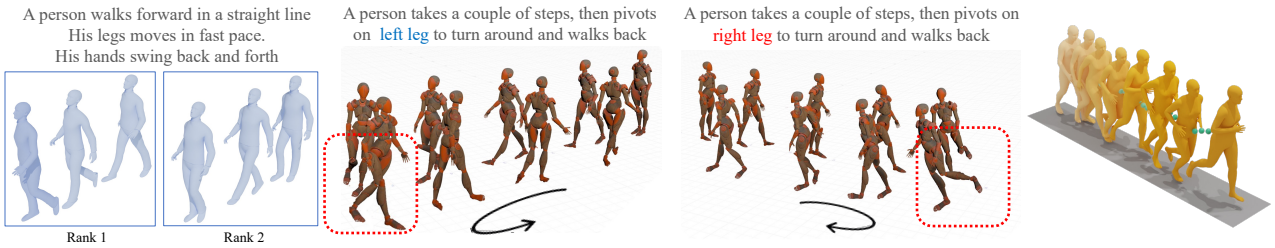
KinMo: Kinematic-aware Human Motion Understanding and Generation

Pengfei Zhang[1*] Pinxin Liu[2*] Pablo Garrido[4] Hyeongwoo Kim[3] Bindita Chaudhuri[4†]
[1] University of California, Irvine, [2] University of Rochester, [3] Imperial College, London, [4] Flawless AI
[1]pengfz5@uci.edu, [2]pliu23@u.rochester.edu,
[3]hyeongwoo.kim@imperial.ac.uk, [4]{pablo.garrido, bindita.chaudhuri}@flawlessai.com



(a) Text-to-motion Retrieval (b) Text-to-motion Generation (c) Text-Motion Editing (d) Motion Trajectory Control

Figure 1. **We present KinMo**, a method that achieves fine-grained motion understanding for (a) effective text-motion retrieval, and text-aligned motion (b) generation, (c) editing, and (d) trajectory control on local kinematic body parts.

Abstract

Current human motion synthesis frameworks rely on global action descriptions, creating a modality gap that limits both motion understanding and generation capabilities. A single coarse description, such as "run", fails to capture details such as variations in speed, limb positioning, and kinematic dynamics, leading to ambiguities between text and motion modalities. To address this challenge, we introduce ***KinMo****, a unified framework built on a hierarchical describable motion representation that extends beyond global actions by incorporating kinematic group movements and their interactions. We design an automated annotation pipeline to generate high-quality, fine-grained descriptions for this decomposition, resulting in the KinMo dataset and offering a scalable and cost-efficient solution for dataset enrichment. To leverage these structured descriptions, we propose Hierarchical Text-Motion Alignment that progressively integrates additional motion details, thereby improving semantic motion understanding. Furthermore, we introduce a coarse-to-fine motion generation procedure to leverage enhanced spatial understanding to improve motion synthesis. Experimental results show that KinMo significantly improves motion understanding, demonstrated by enhanced text-motion retrieval performance and enabling more fine-grained motion generation and editing capabilities. Project Page:* https://andypinxinliu.github.io/KinMo

1. Introduction

Controlling human motion through natural language is a rapidly expanding area within computer vision, enabling interactive systems to generate or modify 3D human motions based on textual input. This technology has a wide range of applications, including robotics[16], digital avatar[13, 18, 45], and automatic animation [20, 44, 50, 62, 63], where human-like motion is crucial for user interaction and immersion. Despite efforts to generate general motion [9, 38, 42, 47, 52, 60], fine-grained control over individual body parts remains largely an unsolved challenge. Current models are proficient in producing coherent whole-body movements from global action descriptions but struggle when tasked with controlling local body parts independently. This limitation prevents them from achieving precision and adaptability for real-world applications.

Recent advances [14, 36] have introduced more refined approaches by incorporating controllability into motion generation. However, these models are limited to processing simple instructions and lack the compatibility required for scenarios where multiple body parts must coordinate to perform complex actions. Similarly, generative models for motion synthesis [9, 38] present innovative methods but do not directly address the issue of controlling specific body parts with specific textual descriptions.

This challenge stems from the inherent ambiguity of motion text descriptions in existing datasets. For example, multiple phrases (such as *pick up an object from the ground* and *bend down to reach something*) can describe the same motion. In contrast, a single term (such as *running*) can en-

[1]* Work done as interns at Flawless AI. These authors contributed equally to this work. † Corresponding Author.

compass a wide range of variations, depending on factors such as speed, arm movement, or direction. This many-to-many mapping problem [24] hinders existing models from handling the multiplicity of natural language or motions, often resulting in inconsistent or unnatural outcomes when trying to generate or edit specific body parts.

To solve this problem, we introduce a novel motion representation based on six fundamental kinematic components: torso, head, left arm, right arm, left leg, and right leg. Unlike existing methods that treat the body as a whole, our approach explicitly models each component and its interactions, enabling a more detailed representation of global action through localized body movements. For instance, a *sneaking* motion should involve coordinated torso and leg movement, while the arms are used for balance. Based on this, we propose a kinematic-aware formulation that opens new possibilities for text-motion understanding, fine-grained motion generation, and editing. Building on this insight, we propose a kinematic-aware formulation, which enables improved text-motion understanding, fine-grained motion generation, and editing capabilities.

To achieve this, we reformulate existing motion representations and enhance the widely used HumanML3D [8] dataset by introducing a semi-supervised annotation system that enriches motion data with body-part-specific descriptions, forming our KinMo dataset. We investigate the retrieval capability of these body-part-specific descriptions, demonstrating that our proposed Hierarchical Text-Motion Alignment effectively integrates these semantics to further enhance text-motion understanding. In addition, we demonstrate how this enhanced understanding benefits motion generation by extending the MoMask [9] model to support fine-grained body-part generation and editing, enabling greater control and manipulation of motion sequences. Our contributions can be summarized as follows:

1. We introduce **KinMo**, a novel framework that decomposes human motion through a three-level hierarchy: global actions, local kinematic groups, and group interactions. This hierarchical representation significantly bridges the gap between text and motion. We further develop a semi-supervised annotation pipeline for generating our dataset.
2. We propose a Hierarchical Text-Motion Alignment method that leverages enriched textual descriptions by progressively encoding them and integrating hierarchical semantics. This approach enhances retrieval capabilities, showing notable improvements in semantic motion understanding within spatial contexts.
3. We extend **Motion Generation** process into a coarse-to-fine procedure, which enhances motion understanding by transitioning from global actions to joint groups and their interactions. This supports various generative and editing applications with fine-grained control.

2. Related Work

Text-to-Motion Understanding. Similar to other modality alignments [21, 23, 54, 58], alignment/retrieval between text and motion modalities is the key indicator of motion understanding. PoseScript [4] uses fine-grained text descriptions to represent various human poses. MotionCLIP [51] and TMR [35] enhance the alignment from single poses to motion sequences. MotionLLM [2] creates a large corpus of Motion-QA for text-motion understandings. However, these methods only focus on global action descriptions, ignoring the extent of local kinematic movements, which are essential for alignment and motion understanding.

Text-to-Motion Generation. Diffusion Models have been the main trend for various modalities [5, 11, 12, 17, 26–29, 31] and demonstrated notable success in motion generation [22, 52, 60]. T2M-GPT [59] and MotionGPT [40] represent motions as discrete tokens and leverage autoregressive models to improve motion generation quality. Masking Motion Models [9, 23, 38] further improve motion generation quality with a bidirectional masking mechanism.

Large Language Models (LLMs) have empowered various understanding and generation with fine-grained control [19, 48, 49]. LGTM [46] and FG-MDM [43] use LLMs to generate additional texts to describe local motions to assist the generation process. Although these methods partially focus on fine-grained local motion control, they cannot address the core ambiguity problem between the two modalities. To address this problem, we take a step further in reformulating a linguistically describable motion representation from global actions to groups and interactions.

Trajectory Control and Editing. Diffusion-based models [1, 52, 60] can perform zero-shot editing by infilling specific joints. TLControl [56], OmniControl [57], and CoMo [14] can control arbitrary joints at any time by combining spatial and temporal control together. However, none of these methods adopt a fine-grained approach that allows local editing while ensuring overall motion compatibility.

3. Kinematic-aware Human Motion

In this section, we first introduce the core principle of KinMo, which bridges the gap between text and motion by linearly transforming motions into **linguistically describable representations** in 3D space (Sec. 3.1). We then propose an LLM-based pipeline to annotate these representations and present the Kinematic-aware Motion-Text (KinMo) dataset (Sec. 3.2). Additionally, we propose an alignment method to achieve spatial understanding given the enriched textual descriptions (Sec. 3.3).

3.1. Describable Motion Representations

Existing Motion Representations. Current text-motion alignment research [8, 9, 34, 35, 37, 38] represents motion

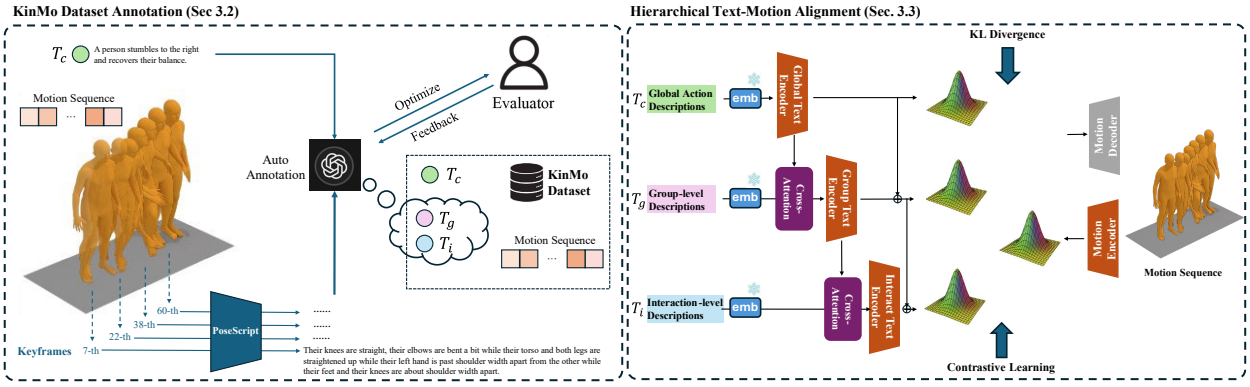


Figure 2. **KinMo Framework**. *Left:* We extract pose descriptions of the keyframes and feed them into an LLM to produce group- and interaction-level descriptions, which generate KinMo dataset together with original motion sequences and global action texts. *Right*: We apply encoders with the same architecture (brown) to process the features of global action, group-level descriptions, and interaction-level descriptions extracted from a pretrained model (blue: emb). The cross-attention layer (purple) is employed to combine embeddings of different levels to enable hierarchical representation learning, with contrastive learning at each level for modality alignment.

as the time evolution of each body joint $j \in J$ of the human body, characterized by its position $\mathbf{p}_j(t)$, axis-angle rotation relative to its parent in the kinematic tree $\mathbf{r}_j(t)$, and relative angular velocity $\mathbf{v}_j(t)$ with respect to the center joint.[1] However, this representation is hard to describe in natural language. Besides, global action descriptions struggle to represent local movements. To solve this problem, we create an intermediate representation of a motion that is explicitly describable in natural language. Specifically, we reformulate motion representations by organizing joints into a set of **kinematic groups** following kinematic tree, defined as $G = \{\text{Torso, Neck, Left Arm, Right Arm, Left Leg, Right Leg}\}$, where each group $g \in G$ consists of joints $J_g \subseteq J$.

Kinematic-Group Representations. For each group g at time t, we define the Group Position $\mathbf{P}_g(t)$ as the average position $\mathbf{p}_j(t)$ of the joints within that group:

$$\mathbf{P}_g(t) = \frac{1}{|J_g|} \sum_{j \in J_g} \mathbf{p}_j(t). \tag{1}$$

We then define the *Limb Angles* $\Theta_g(t)$ as the collection of joint rotations $\mathbf{r}_j(t)$ within the group, and the *Group Velocity* $\mathbf{V}_g(t)$ as the average velocity $\mathbf{v}_j(t)$ of the joints:

$$\mathbf{\Theta}_g(t) = \{\mathbf{r}_j(t) \mid j \in J_g\}, \quad \mathbf{V}_g(t) = \frac{1}{|J_g|} \sum_{j \in J_g} \mathbf{v}_j(t). \tag{2}$$

Group-Interaction Representations. Human motion also involves the relationships between each pair of groups $(g, h) \in G \times G$.

$$\begin{bmatrix} \Delta\mathbf{P}_{g,h}(t) \\ \Delta\mathbf{\Theta}_{g,h}(t) \\ \Delta\mathbf{V}_{g,h}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{P}_h(t) - \mathbf{P}_g(t) \\ \mathbf{\Theta}_{h \cap g}(t) \\ \mathbf{V}_h(t) - \mathbf{V}_g(t); \mathbf{v}_{h \cap g}(t) \end{bmatrix}, \tag{3}$$

[1] W.l.o.g. we omit some boundary conditions, e.g., foot contact and center joint selection, as they do not affect the core conclusions. The supplementary document (Appendix G) shows a detailed computation of existing motion representations.

where $\Delta\mathbf{P}_{g,h}(t)$ denotes the difference in position, $\Delta\mathbf{\Theta}_{g,h}(t)$ represents the angles at the connecting joint (if exists), and $\Delta\mathbf{V}_{g,h}(t)$ is the relative angular velocity and the angular velocity at the connecting joint (if exists) between two groups.

The proposed formulation of motion representations is a linear transformation of the existing formulation used in current text-motion alignment methods [9, 34, 35, 37, 38] and can be transformed back to existing representations, as detailed in Appendix G. Additionally, our formulation is inherently compatible with natural language descriptions, capturing both the movements of individual kinematic groups and their interactions. With this formulation, the key task is to annotate our proposed linguistically describable motion representations to collect textual descriptions of kinematic groups and group interactions.

3.2. KinMo Dataset

Kinematic-aware Joint-Motion Text Annotation. Good annotations of the proposed motion representations must capture both spatial and temporal details of each kinematic group and their interactions. We employ a two-step LLM-based strategy to ensure high-quality automatic annotation, as shown in Fig. 2. Using this strategy, we enhance the HumanML3D dataset [8] with fine-grained annotations.

Spatial-Temporal Motion Processing. A motion consists of a sequence of pose frames over time. Existing human motion understanding models [2, 8, 40] struggle to capture subtle local movements due to the complex interplay of spatial and temporal dynamics. To address this, we propose a two-stage disentanglement approach, first resolving spatial dynamics and then temporal dynamics.

To extract detailed spatial information, we adopt PoseScript [4] to generate annotations for each pose frame, capturing precise angular rotations of joints for any given human pose. To obtain fine-grained temporal information, we

propose a keyframe selection pipeline. We use sBERT [39] to extract embeddings for the per-frame pose descriptions. We assess the similarity of poses across the time frames by calculating the cosine similarity between text embeddings. If the cosine similarity falls below a user-defined threshold, we label that frame as a keyframe. The pose differences within each kinematic group over a specified time window are used to approximate local temporal motions during that period.

Semi-supervised Annotation. We then design an automatic annotator using GPT-4o-mini [32] to generate textual descriptions of kinematic groups and their interactions based on keyframe pose annotations. To refine the prompt, two human evaluators iteratively assess and improve the annotation process. We begin by randomly sampling 20 preprocessed motion sequences and using GPT-4o-mini to infer textual descriptions based on keyframe pose annotations. The two evaluators independently review the generated descriptions, documenting errors made by the LLM. These insights are then fed back into the model to refine the prompt design. This iterative process continues until the evaluators reach a consensus, achieving a kappa statistic above 0.8. At this point, we ensure that subsequent LLM-generated descriptions for the remaining dataset align with our objectives. Additional details on the text prompt and example descriptions can be found in Appendix B.

In summary, the KinMo dataset provides three types of descriptions for each motion: 1) Global action descriptions; 2) Spatial and temporal per-group descriptions of the movement and dynamics of each kinematic group g; 3) Spatial and temporal per-group-pair descriptions of the relative movements and dynamics between each pair of kinematic groups h and g. We will refer to them as *global action descriptions* (T_c), *group-level descriptions* (T_g), and *interaction-level descriptions* (T_i), respectively.

3.3. Hierarchical Text-Motion Alignment (HTMA)

Existing text-motion alignment methods typically encode global actions and parse additional descriptions directly [15, 43, 61], which limits spatial understanding. Rich contextual dependencies and explicit positional relationships among descriptions make it challenging to capture high-quality semantics when directly encoding additional descriptive levels [24, 64]. Instead, we propose a hierarchical framework that first encodes the global action as **coarse embeddings** and then progressively incorporates group- and interaction-level descriptions to refine the embeddings, achieving hierarchical alignment, as illustrated in Figure 2.

Modality Encoders. To align motion and text modalities, we follow TMR [35] to map textual descriptions and motion into a shared co-embedding space. Motion and text encoders are Transformer-based [55] with additional learnable distribution parameters, as in the VAE-based ACTOR model [33]. We follow a probabilistic approach that utilizes two prefix tokens for each text or motion sequence to learn the (μ, Σ) of a Gaussian distribution $\mathcal{N}$, from which a latent vector $z \in \mathbb{R}^d$ is sampled.

Motion sequences are processed directly by the motion encoder. For the textual inputs, we first extract text features from a pre-trained and frozen RoBERTa [25] or DistilBERT [41] to obtain global action, group-level, and interaction-level text descriptions, and then encode these features with cross-attention in a hierarchical process:

$$\mathbf{h_c} = E_c\big(\text{emb}(T_c)\big), \tag{4}$$

$$\mathbf{h_g} = E_g\big(\text{CrossAttn}(\text{emb}(T_g), \mathbf{h_c})\big), \tag{5}$$

$$\mathbf{h_i} = E_i\big(\text{CrossAttn}(\text{emb}(T_i), \mathbf{h_g})\big), \tag{6}$$

where $\text{emb}(\cdot)$ represents a pre-trained RoBERTa or DistilBERT model for feature extraction, and $E_{c,g,i}$ are the VAE-based ACTOR models used as text encoders for each level. $\text{CrossAttn}(\cdot,\cdot)$ denotes a single cross-attention layer with a residual connection. The cross-attention mechanism allows each level to build upon previous embeddings, creating progressively refined semantic embeddings. The group-level embeddings incorporate global action, while the interaction-level embeddings incorporate the group-level.

Contrastive Learning. We use contrastive learning to align the embeddings of motion and text modalities in each hierarchy [35]. For simplicity, we denote any level of text and motion pair as $(z^T, z^M), T \in \{T_c, T_g, T_i\}$. For a batch of N positive pairs $(z_1^T, z_1^M), \ldots, (z_N^T, z_N^M)$, any pair (z_i^T, z_j^M) where $i \neq j$ is considered a negative sample. The similarity matrix S computes the pairwise cosine similarities for all pairs in the batch, defined as $S_{ij} = \cos(z_i^T, z_j^M)$. We apply an InfoNCE loss [54], as follows:

$$\mathcal{L}_{\text{NCE}} = \frac{-1}{2N} \sum_T \sum_i \left(\log \frac{\exp S_{ii}/\tau}{\sum_j \exp S_{ij}/\tau} + \log \frac{\exp S_{ii}/\tau}{\sum_j \exp S_{ji}/\tau} \right), \tag{7}$$

where τ represents a temperature parameter.

To maximize the proximity between the two modalities, we follow TMR [35] to construct a weighted sum of 3 losses: (a) Kullback–Leibler divergence loss $\mathcal{L}_{\text{KL}}$, (b) cross-modal embedding similarity loss $\mathcal{L}_{\text{E}}$, and (c) motion reconstruction loss $\mathcal{L}_{\text{R}}$ for each semantic hierarchy.

4. Motion Understanding and Generation

KinMo framework provides new insight into motions. By bridging the gap between text and motion through our proposed motion representations and alignment method, we will show how KinMo achieves motion understanding[2] and

[2]Since *motion understanding* currently lacks diverse downstream tasks, with only text-motion retrieval widely used, our claim of *motion understanding* may be an overstatement. Here, we specifically explore its impact on text-motion retrieval.

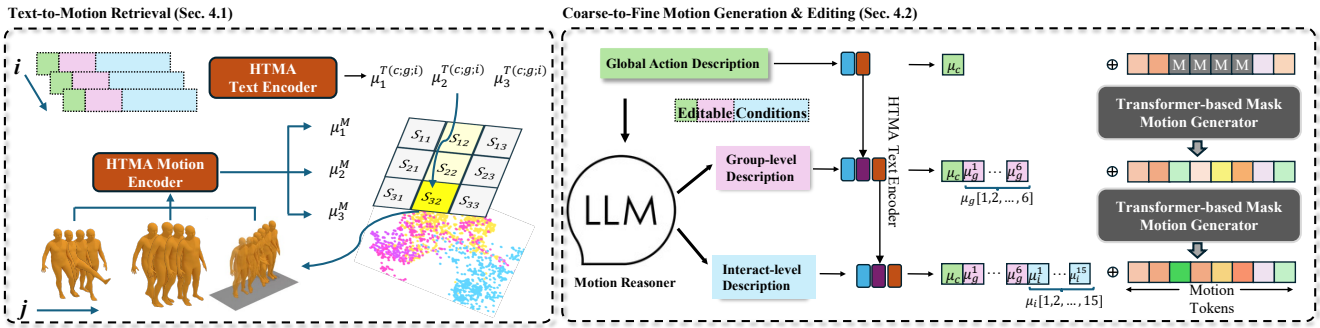


Figure 3. **Motion Retrieval and Generation.** *Left*: Overview of Text-to-Motion Retrieval, where we compute the similarity matrix defined between text and motion embeddings. Here, we present a batch of three samples with an example. To retrieve the most similar motion to the 2nd text T_{c_2}, we use *Motion Reasoner* to generate corresponding group- and interaction-level descriptions, producing aligned embeddings $\mu_2^{T(c;g;i)}$. We then check the similarity matrix against the motion embeddings, where S_{32} has the highest value, indicating that the 3rd motion is the closest match to the 2nd text. *Right:* Given a global action, *Motion Reasoner* generates group- and interaction-level descriptions. The aligned text embeddings at each level are prepended to the motion tokens and collectively fed into the motion generator through a coarse-to-fine generation process. Editing can be performed on the texts either before or after the Motion Reasoner stage. *HTMA Text/Motion Encoder (Sec. 3.3)* aligns text and motion into a shared embedding space to obtain their respective aligned representations.

generation, as well as downstream applications (Sec. 5.4).

Motion Reasoner. Our goal is to generate group- and interaction-level descriptions based on global action inputs. We finetune LLaMA-3 [6] on the KinMo Dataset to serve as a motion reasoner. The model is trained using the standard next-token prediction loss, with an added conditioning on global action T_c to output the corresponding group-level descriptions T_g and interaction-level descriptions T_i:

$$\mathcal{L}_{reasoner} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i | T_c, T_{<i}), \tag{8}$$

where y_i and $\hat{y}_i$ represent the ground truth and predicted tokens at position i, respectively. $T_{<i}$ denotes the previously generated tokens. This loss encourages the model to generate motion descriptions at different granularity levels.

Text Alignment. Given the additional descriptions of Motion Reasoner, we obtain their corresponding text embeddings μ_c, μ_g, μ_i based on hierarchical text encoders from Sec. 3.3 for various applications.

4.1. Text-Motion Retrieval

As shown in Fig. 3, for a given motion M and its corresponding text descriptions T_c, T_g, T_i, the mean token of the output motion parameters μ_j^M serves as aligned motion embedding, while the mean token of the output text parameters $\mu_i^{T(c;g;i)} = [\mu_c; \mu_g; \mu_i]$ acts as aligned text embedding in the co-embedding space at different hierarchical levels. For retrieval, we directly compare these embeddings, identifying the best match by maximizing the cosine similarity between motion and text embeddings.

4.2. Coarse-to-Fine Motion Generation

For motion generation, we adopt MoMask [9] as the base architecture. Specifically, a VQ-VAE [53, 59] is trained to convert a motion sequence into discrete tokens. Then, given a text T_c prepended to the discrete motion tokens as input condition, a Transformer-based generator is trained with a masking-based strategy for token generation. The generated token will be used to query VQ-VAE to generate motions. To incorporate KinMo into the existing motion generation framework, we leverage Motion Reasoner to output group-level T_g and interaction-level T_i descriptions given input global action descriptions and then encode them into μ_c, μ_g, μ_i. In addition, we propose a coarse-to-fine generation procedure conditioned on the text hierarchy.

Hierarchical Generation. As shown in Fig. 3, we extend MoMask [9] from generation under condition CLIP(T_c), into condition μ_c, μ_g, μ_i, which are aligned embeddings of $[T_c; T_g; T_i]$, through a coarse-to-fine generation process. Specifically, after the initial motion tokens are generated conditioned on μ_c, they will be re-fed into the generator to output intermediate tokens conditioned on μ_g. Then, the intermediate tokens are re-fed into the generator to produce the final tokens conditioned on μ_i. The final tokens are used to query motions into a trained VQ-VAE. For efficiency, the generator shares weights with the same logit classification loss functions that were used to reconstruct motion tokens for the three levels of text conditioning. Other configurations are the same as in MoMask [9].

5. Experiments

We conduct experiments on the motion-text benchmark dataset, HumanML3D [8], which collects 14,616 motions from AMASS [30] and HumanAct12 [7] datasets, with each motion described by 3 text scripts, totaling 44,970 descriptions. We adopt their pose representation and augment the dataset using mirroring, followed by a 80/5/15 split for training, validation, and testing, akin to previous

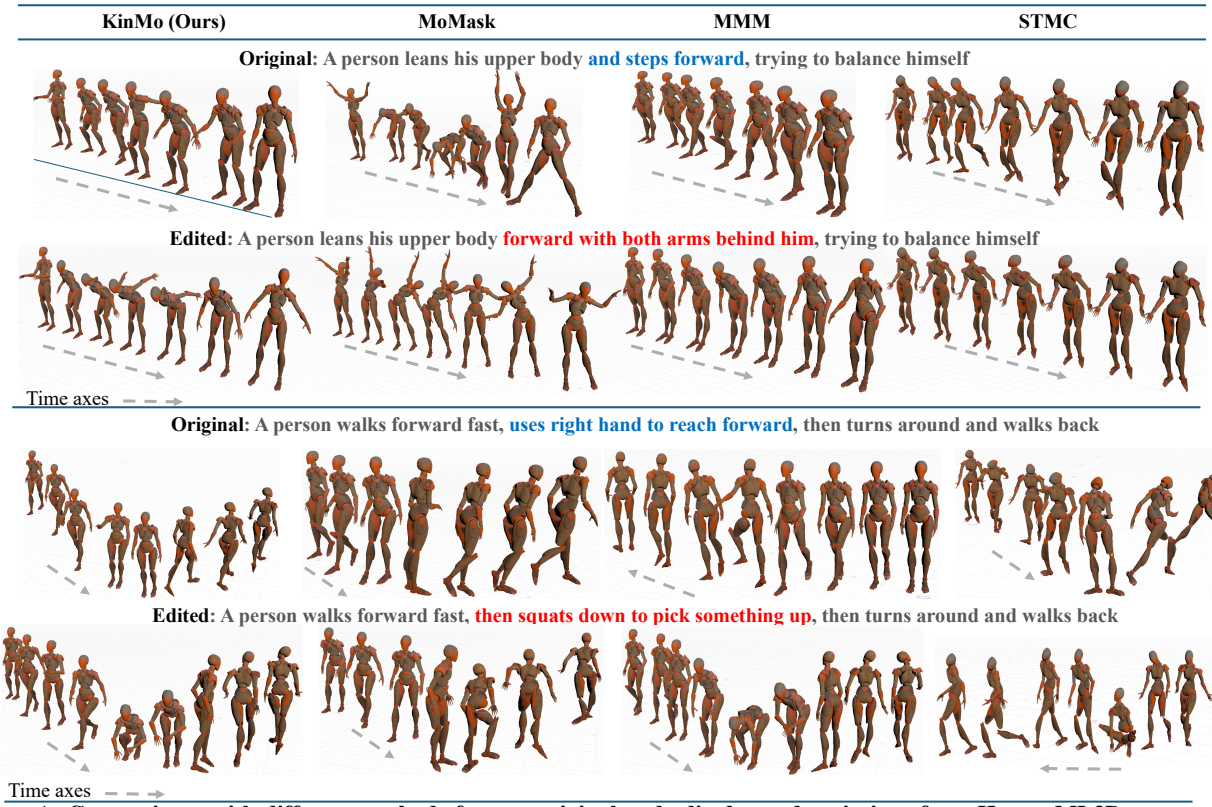


Figure 4. **Comparisons with different methods for two original and edited text descriptions from HumanML3D test set**. Unlike previous methods, our results match the input text descriptions better and show the ability to edit specific body parts.

Table 1. **Text-to-motion retrieval benchmark on HumanML3D.** Evaluation protocols with decreasing difficulty from (a) to (d).

Protocol	Methods	Text-motion retrieval						Motion-text retrieval					
		R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	R@5 ↑	R@10 ↑	MedR ↓
(a) All	TEMOS [34]	2.12	4.09	5.87	8.26	13.52	173.0	3.86	4.54	6.94	9.38	14.00	183.25
	HumanML3D [8]	1.80	3.42	4.79	7.12	12.47	81.00	2.92	3.74	6.00	8.36	12.95	81.50
	TMR [35]	5.68	10.59	14.04	20.34	30.94	28.00	**9.95**	12.44	17.95	23.56	32.69	28.50
	Ours (distilbert)	8.13	14.16	19.69	27.07	39.18	18.00	9.29	15.51	20.61	28.29	40.25	18.00
	Ours (RoBERTa)	**9.05**	**15.23**	**20.47**	**28.62**	**41.60**	**16.00**	9.01	**15.92**	**21.42**	**29.50**	**41.43**	**16.00**
(b) All with threshold	TEMOS [34]	5.21	8.22	11.14	15.09	22.12	79.00	5.48	6.19	9.00	12.01	17.10	129.0
	HumanML3D [8]	5.30	7.83	10.75	14.59	22.51	54.00	4.95	5.68	8.93	11.64	16.94	69.50
	TMR [35]	**11.60**	15.39	20.50	27.72	38.52	19.00	**13.20**	15.73	22.03	27.65	37.63	21.50
	Ours (distilbert)	10.82	18.49	25.33	33.89	46.54	**12.00**	12.25	**19.69**	24.98	32.70	44.04	**14.00**
	Ours (RoBERTa)	11.39	**19.18**	**25.73**	**34.76**	**47.94**	**12.00**	11.65	19.54	**25.45**	**33.67**	**45.08**	**14.00**
(c) Dissimilar subset	TEMOS [34]	33.00	42.00	49.00	57.00	66.00	4.00	35.00	44.00	50.00	56.00	70.00	3.50
	HumanML3D [8]	34.00	48.00	57.00	72.00	84.00	3.00	34.00	47.00	59.00	72.00	83.00	3.00
	TMR [35]	47.00	61.00	71.00	80.00	86.00	2.00	48.00	63.00	69.00	80.00	84.00	2.00
	Ours (distilbert)	45.73	62.80	70.73	79.88	**90.85**	2.00	46.95	62.80	70.12	82.93	**91.46**	2.00
	Ours (RoBERTa)	**57.73**	**78.35**	**81.44**	**86.60**	90.72	**1.00**	**63.92**	**80.41**	**82.47**	**87.63**	90.72	**1.00**
(d) Small batches [8]	TEMOS [34]	40.49	53.52	61.14	70.96	84.15	2.33	39.96	53.49	61.79	72.40	85.89	2.33
	HumanML3D [8]	52.48	71.05	80.65	89.66	96.58	1.39	52.00	71.21	81.11	89.87	96.78	1.38
	TMR [35]	67.16	81.32	86.81	91.43	95.36	1.04	67.97	81.20	86.35	91.70	95.27	1.03
	Ours (distilbert)	72.28	85.42	**90.15**	**94.01**	**97.09**	**1.00**	72.21	85.19	90.00	**94.42**	**97.04**	**1.00**
	Ours (RoBERTa)	**72.88**	**85.54**	89.91	93.46	96.68	**1.00**	**73.00**	**85.64**	**90.17**	93.70	96.49	**1.00**

work [9, 38]. Our KinMo dataset is built on this dataset with each motion described by 6 group-level and 15 interaction-level descriptions scripts in accordance with human kinematics (see Sec. 3.2). An example is presented in the supplementary material (Appendix B), along with additional details on dataset collection. All experiments are performed in such settings, as shown in Fig. 3.

5.1. Text-Motion Retrieval

We first evaluate whether the introduction of group- and interaction-level motion descriptions reduces any ambiguity for the text-motion retrieval problem and improves the overall motion understanding.

Evaluation Metrics. We adopt TMR settings [35] to measure retrieval performance using recall scores at various

Table 2. **Comparison of text-to-motion generation on HumanML3D.** For each metric, we repeat the evaluation 20 times and report the average with 95% confidence interval. The right arrow (→) indicates that the closer the result is to real motion, the better.

Methods	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity →	MModality ↑
	Top-1 ↑	Top-2 ↑	Top-3 ↑				
Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
MDM [52]	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$\mathbf{2.799}^{\pm.072}$
GuidedMotion [15]	$0.503^{\pm.002}$	$0.691^{\pm.002}$	$0.788^{\pm.002}$	$0.057^{\pm.006}$	$3.040^{\pm.012}$	$9.864^{\pm.077}$	$2.473^{\pm.096}$
KP [24]	0.496	-	-	0.275	-	9.975	2.218
FG-MDM [43]	$0.374^{\pm.003}$	$0.582^{\pm.003}$	$0.709^{\pm.005}$	$0.618^{\pm.009}$	$5.274^{\pm.048}$	$9.563^{\pm.0.097}$	-
FineMoGen [61]	$0.504^{\pm.002}$	$0.690^{\pm.002}$	$0.784^{\pm.002}$	$0.151^{\pm.008}$	$2.998^{\pm.008}$	$9.263^{\pm.094}$	$\underline{2.696}^{\pm.079}$
MotionLCM [3]	$0.504^{\pm.002}$	$0.698^{\pm.003}$	$0.796^{\pm.002}$	$0.304^{\pm.003}$	$3.012^{\pm.007}$	$9.634^{\pm.064}$	$2.267^{\pm.082}$
ParCo [64]	$0.515^{\pm.003}$	$0.706^{\pm.003}$	$0.801^{\pm.002}$	$0.109^{\pm.005}$	$2.927^{\pm.008}$	$9.576^{\pm.088}$	$1.382^{\pm.060}$
MMM [38]	$0.504^{\pm.003}$	$0.696^{\pm.003}$	$0.794^{\pm.002}$	$0.080^{\pm.003}$	$2.998^{\pm.007}$	$9.411^{\pm.058}$	$1.164^{\pm.041}$
MoMask [9]	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$\underline{0.045}^{\pm.003}$	$2.958^{\pm.008}$	$9.678^{\pm.052}$	$1.241^{\pm.040}$
Ours (CLIP)	$\underline{0.529}^{\pm.003}$	$\underline{0.722}^{\pm.002}$	$\underline{0.817}^{\pm.002}$	$0.050^{\pm.003}$	$\underline{2.907}^{\pm.009}$	$9.684^{\pm.063}$	$1.313^{\pm.041}$
Ours (HTMA)	$\mathbf{0.532}^{\pm.002}$	$\mathbf{0.724}^{\pm.003}$	$\mathbf{0.821}^{\pm.003}$	$\mathbf{0.039}^{\pm.003}$	$\mathbf{2.901}^{\pm.010}$	$9.674^{\pm.058}$	$1.321^{\pm.039}$

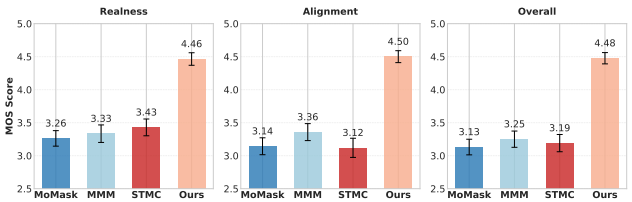


Figure 5. **User Study**. We generate 80 videos for each method to assess *Realness*, *T2M Alignment*, and *Overall Impression*.

Table 3. **Comparisons with other methods**. Motion Generation with CLIP-embed additional texts as conditions of MoMask.

Method	FID↓	R-Prec(Top 3)↑	MM-Dist ↓	MModality ↑
MoMask(base)	**0.045**	0.807	2.958	1.241
MoMask+LGTM	0.057	0.801	2.963	1.123
MoMask+FinMoGen	0.062	0.799	2.998	1.223
Ours+Parco (2-group)	0.077	0.793	3.232	1.101
Ours (MoMask+KinMo) (6-group)	0.050	**0.817**	**2.907**	**1.313**

ranks (*e.g.*, R@1, R@2) and the median rank (MedR) of our results. MedR represents the median ranking position of the ground-truth result, with lower values indicating more precise retrievals. The four evaluation protocols used in our experiments are outlined below: (i) *All* uses the complete test dataset, though similar negative pairs can affect precision; (ii) *All with threshold* sets a similarity threshold of 0.8 to determine accurate retrievals; (iii) *Dissimilar subset* uses 100 distinctly different sampled pairs measured by sBERT [39] embedding difference; and (iv) *Small batches* evaluates performance on random batches of 32 motion-text pairs.

Evaluation Results. We benchmark KinMo against [8, 34, 35]. In Tab. 1, our model outperforms existing baselines, particularly in setting (a). This improvement is primarily due to our annotated descriptions, which help to resolve ambiguities in action-level text-motion correspondence. By providing finer-grained details, our approach enhances the discrimination of motions with subtle local movement differences but similar global action descriptions. Our proposed formulation and annotations contribute significantly to motion understanding by capturing intricate local movement details throughout the motion sequence. Motion understanding can be further enhanced using RoBERTa [25] as a stronger text encoder for additional descriptions.

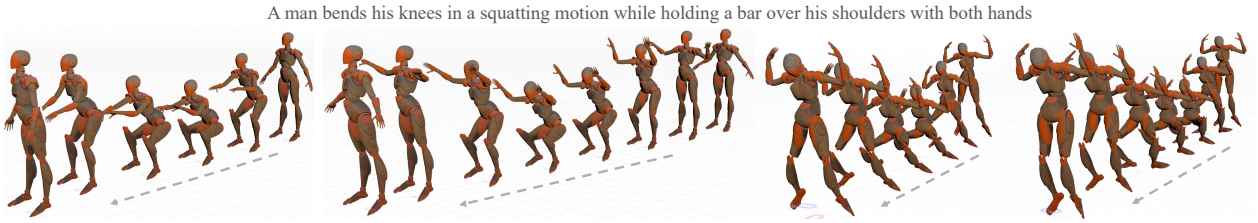
5.2. Text-Motion Generation

Evaluation Metrics. We adopt (1) *FID* [10] as an overall motion quality metric to measure the difference between generated and real motion distributions; (2) *R-Precision (R-Prec)* and *multimodal distance (MM-Dist)* to quantify the semantic alignment between text and generated motions; and (3) *Multimodality (MModality)* to assess the diversity of motions generated from the same text, as in T2M [8].

Evaluation Settings. To present a fair comparison, we consider each T2M method as the whole system. For KinMo, we apply a different random seed during inference. Motion Reasoner generates the group- and interaction-level motion descriptions based on the provided global action descriptions and feeds them into the generator.

Evaluation Results. Tab. 2 compares KinMo with various methods for T2M generation [3, 9, 15, 24, 34, 38, 43, 52, 60, 61, 64]. Our method attains the best motion generation quality with the highest text alignment score (R-Prec and MM-Dist). Thanks to the introduction of explicit group- and interaction-level descriptions, we observe that KinMo generates better aligned motions for any given dense and fine-grained text descriptions shown in Fig. 4, while other baseline methods fail to capture local body part movements.

User Study. To assess the quality of our results, we conduct a user study involving 20 participants and 320 samples, 80 from KinMo, MoMask [9], MMM [38], and STMC [37], respectively. Each participant was presented the video clips in a random order and asked to rate the results between 1 (lowest) and 5 (highest) based on (1) *realness*, (2) correctness of text-motion *alignment*, and (3) *overall impression*. Fig. 5 shows that, unlike other baseline methods, KinMo achieves higher Mean Opinion Scores (MOS) overall.



(a) T_c (b) $T_c + T_g$ (c) $T_c + T_g + T_i$ (d) Ground truth

Figure 6. **Additional Descriptions resolve the ambiguity of motion generation**. We create motion using (a) only global action, (b) global + group-level descriptions, and (c) global, group-level, and interaction-level descriptions, and compare them with (d) the ground truth.

Table 4. **Effect of Additional Descriptions for Text-Motion Alignment.** Different strategies for incorporating descriptions generated by Motion Reasoner on motion- and text-retrieval tasks.

Motion Semantic	Text-motion retrieval				Motion-text retrieval			
	R@1 ↑	R@2 ↑	R@3 ↑	MedR ↓	R@1 ↑	R@2 ↑	R@3 ↑	MedR ↓
global	3.67	7.17	10.32	40.00	8.08	11.56	17.23	38.00
+ group	7.58	13.16	16.97	22.00	8.58	14.51	19.21	21.00
+ interact	**9.05**	**15.23**	**20.47**	**16.00**	**9.01**	**15.92**	**21.42**	**16.00**
- cross	7.63	13.13	16.94	22.00	8.60	14.54	19.21	21.00

Table 5. **Effect of Hierarchical Text-Motion Alignment.** Comparisons are conducted for Motion Generator with RQ base layer.

Embedder	Global	Joint	Inter	FID↓	R-Prec(Top 3)↑	MM-Dist ↓	MModality ↑
CLIP	✓	–	–	0.115	0.499	2.999	1.221
	✓	✓	–	0.096	0.503	2.953	**1.308**
	✓	✓	✓	0.098	0.512	2.912	**1.308**
HTMA	✓	–	–	0.056	0.512	2.969	1.232
	✓	✓	–	0.051	0.525	2.911	1.292
	✓	✓	✓	**0.044**	**0.527**	**2.904**	1.305

5.3. Ablation Study

Effect of Additional Descriptions generated by Motion Reasoner at each level. Tab. 4 summarizes several strategies for incorporating text-motion alignment: (1) only global action (global), (2) + group-level (+ group), (3) + group + interaction-level (+ interact), and (4) without cross-attention (- cross). We observe that adding extra descriptions generated by Motion Reasoner enhances motion understanding. Cross-attention improves the connectivity of descriptions from different hierarchy levels. As shown in Fig. 6, both group- and interaction-level descriptions are beneficial for resolving global action ambiguity and generating local body parts (e.g., the hands and arms in the figure). Refer to Appendix D for further analysis of the order of the different descriptions and additional design choices.

Effect of Hierarchical Text-Motion Alignment (HTMA). We provide additional quantitative results and comparisons for various text encoders. As demonstrated in Tab. 5, the CLIP encoder, as used in previous work [9, 38], shows superior text-motion alignment after complete training. Our HTMA method enhances motion smoothness and naturalness, as evident by a significantly lower FID. For the motion generation procedure, it can be seen that both of these text encoders benefit from our coarse-to-fine generation approach. Further analysis of training is given in Appendix D.

Text Granularity and Motion Decomposition. Several methods, including LGTM [46], FG-MDM [43], and FineMoGen [61], employ LLMs to generate supplementary motion descriptions to improve generation. KinMo formulates the generated supplementary descriptions using insight of motion components (position, angle, velocity) with natural language to enhance text-motion alignment. We validate this advantage via quantitative experiments (Tab. 3) and ensure fairness using MoMask as the generator across all methods, with only additional descriptions replaced. Moreover, we compare with ParCo [64] which decomposes motion into 2 parts (upper and lower body) as opposed to our proposed 6 parts based on kinematic knowledge. KinMo outperforms these approaches, indicating that (1) the formulated text descriptions, instead of random ones, improve model performance and (2) the proposed linguistically describable motion representation based on kinematic parts (Sec 3.1) and corresponding descriptions are necessary.

5.4. Applications

Text-to-Motion Editing. *Motion Reasoner* enables precise action-level edits (e.g., changing *running* to *jumping*) or local joint adjustments (e.g., *slightly raising the hands*). Our method uses a coarse-to-fine approach, assisted by a masking mechanism, to perform these edits at varying levels of granularity. Please refer to Appendix D for evaluation.

Motion Trajectory Control. We employ ControlNet [57] to condition the motion generator using the provided trajectory of the target joint during the generation, with the descriptions adjusted by the *Motion Reasoner*. We defer the technical details to Appendix C.

6. Conclusion

We present **KinMo**, a framework that represents human motion as kinematic parts movements and interactions, thereby enabling fine-grained text-to-motion understanding, generation, editability, and control. Our method progressively encodes global actions with kinematic descriptions and leverages these descriptions to achieve enhanced alignment and understanding, thus generating coarse-to-fine motions. The KinMo dataset is publicly available to the scientific community. Extensive comparisons with state-of-the-art methods show that **KinMo** improves text-motion alignment and body part control.

Acknowledgements. We thank the anonymous reviewers for their constructive feedback. Special thanks to Brian Burritt and Avi Goyal for helping with visualizations, and Kyle Olszewski and Ari Shapiro for valuable discussions.

References

[1] Nikos Athanasiou, Alpár Ceske, Markos Diomataris, Michael J. Black, and Gül Varol. MotionFix: Text-driven 3d human motion editing. In *SIGGRAPH Asia*, 2024. 2

[2] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. MotionLLM: Understanding Human Behaviors from Human Motions and Videos. *arXiv preprint arXiv:2405.20340*, 2024. 2, 3

[3] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. MotionLCM: Real-time Controllable Motion Generation via Latent Consistency Model. *arXiv preprint arXiv:2404.19759*, 2024. 7

[4] 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, 2022. 2, 3

[5] Daiheng Gao, Shilin Lu, Shaw Walters, Wenbo Zhou, Jiaming Chu, Jie Zhang, Bang Zhang, Mengxi Jia, Jian Zhao, Zhaoxin Fan, et al. EraseAnything: Enabling Concept Erasure in Rectified Flow Transformers. *International Conference on Machine Learning*, 2025. 2

[6] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. 5

[7] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3D human motions. In *ACM International Conference on Multimedia*, pages 2021–2029, 2020. 5

[8] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 2, 3, 5, 6, 7

[9] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3D human motions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2, 3, 5, 6, 7, 8

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 7

[11] Chao Huang, Susan Liang, Yunlong Tang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Scaling Concept With Text-Guided Diffusion Models. *arXiv preprint arXiv:2410.24151*, 2024. 2

[12] Chao Huang, Susan Liang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. High-quality visually-guided sound separation from diverse categories. *arXiv preprint arXiv:2308.00122*, 2024. 2

[13] Chao Huang, Dejan Markovic, Chenliang Xu, and Alexander Richard. Modeling and Driving Human Body Soundfields through Acoustic Primitives. *arXiv preprint arXiv:2407.13083*, 2024. 1

[14] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. CoMo: Controllable Motion Generation through Language Guided Pose Code Editing. In *European Conference on Computer Vision*, 2024. 1, 2

[15] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Runyi Yu, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Local Action-Guided Motion Diffusion Model for Text-to-Motion Generation. *arXiv preprint arXiv:2407.10528*, 2024. 4, 7

[16] Hema Swetha Koppula and Ashutosh Saxena. Anticipating human activities for reactive robotic response. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Tokyo, 2013. 1

[17] Leyang Li, Shilin Lu, Yan Ren, and Adams Wai-Kin Kong. Set You Straight: Auto-Steering Denoising Trajectories to Sidestep Unwanted Concepts. *arXiv preprint arXiv:2504.12782*, 2025. 2

[18] Yong-Lu Li, Xiaoqian Wu, Xinpeng Liu, Zehao Wang, Yiming Dou, Yikun Ji, Junyi Zhang, Yixing Li, Xudong Lu, Jingru Tan, et al. From isolated islands to pangea: Unifying semantic space for human action understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16582–16592, 2024. 1

[19] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Language-guided joint audio-visual editing via one-shot adaptation. *arXiv preprint arXiv:2410.07463*, 2024. 2

[20] Pinxin Liu, Luchuan Song, Daoan Zhang, Hang Hua, Yunlong Tang, Huaijin Tu, Jiebo Luo, and Chenliang Xu. GaussianStyle: Gaussian Head Avatar via StyleGAN. *arXiv preprint arXiv:2402.00827*, 2024. 1

[21] Pinxin Liu, Haiyang Liu, Luchuan Song, and Chenliang Xu. Intentional Gesture: Deliver Your Intentions with Gestures for Speech. *arXiv preprint arXiv:2505.15197*, 2025. 2

[22] Pinxin Liu, Luchuan Song, Junhua Huang, and Chenliang Xu. GestureLSM: Latent Shortcut based Co-Speech Gesture Generation with Spatial-Temporal Modeling. *arXiv preprint arXiv:2501.18898*, 2025. 2

[23] Pinxin Liu, Pengfei Zhang, Hyeongwoo Kim, Pablo Garrido, Ari Shapiro, and Kyle Olszewski. Contextual Gesture: Co-Speech Gesture Video Generation through Context-aware Gesture Representation. In *ACM International Conference on Multimedia*, 2025. 2

[24] Xinpeng Liu, Yong-Lu Li, Ailing Zeng, Zizheng Zhou, Yang You, and Cewu Lu. Bridging the gap between human motion and action semantics via kinematic phrases. *arXiv preprint arXiv:2310.04189*, 2023. 2, 4, 7

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019. 4, 7

[26] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. TF-ICON: Diffusion-Based Training-Free Cross-Domain Image

Composition. In *IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023. 2

[27] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. TF-ICON: Diffusion-Based Training-Free Cross-Domain Image Composition. In *IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023.

[28] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. MACE: Mass Concept Erasure in Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024.

[29] Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Wai-Kin Kong. Robust watermarking using generative priors against image editing: From benchmarking to advances. *arXiv preprint arXiv:2410.18775*, 2024. 2

[30] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. In *IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019. 5

[31] Mang Ning, Mingxiao Li, Jianlin Su, Haozhe Jia, Lanmiao Liu, Martin Beneš, Wenshuo Chen, Albert Ali Salah, and Itir Onal Ertugrul. DCTdiff: Intriguing Properties of Image Generative Modeling in the DCT Space. *arXiv preprint arXiv:2412.15032*, 2024. 2

[32] OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2024. 4

[33] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-Conditioned 3D Human Motion Synthesis with Transformer VAE. In *IEEE/CVF International Conference on Computer Vision*, 2021. 4

[34] 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, 3, 6, 7

[35] Mathis Petrovich, Michael J Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *IEEE/CVF International Conference on Computer Vision*, pages 9488–9497, 2023. 2, 3, 4, 6, 7

[36] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gul Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[37] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Multi-Track Timeline Control for Text-Driven 3D Human Motion Generation. In *Workshop on Human Motion Generation, IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3, 7

[38] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. MMM: Generative masked motion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2024. 1, 2, 3, 6, 7, 8

[39] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*, 2019. 4, 7

[40] Jose Ribeiro-Gomes, Tianhui Cai, Zoltán Á Milacski, Chen Wu, Aayush Prakash, Shingo Takagi, Amaury Aubel, Daeil Kim, Alexandre Bernardino, and Fernando De La Torre. MotionGPT: Human Motion Synthesis with Improved Diversity and Realism via GPT-3 Prompting. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5070–5080, 2024. 2, 3

[41] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv preprint arXiv:1910.01108*, 2019. 4

[42] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 1

[43] Xu Shi, Wei Yao, Chuanchen Luo, Junran Peng, Hongwen Zhang, and Yunlian Sun. FG-MDM: Towards Zero-Shot Human Motion Generation via ChatGPT-Refined Descriptions. *arXiv preprint arXiv:2312.02772*, 2024. 2, 4, 7, 8

[44] Luchuan Song, Pinxin Liu, Lele Chen, Guojun Yin, and Chenliang Xu. Tri^2-plane: Thinking Head Avatar via Feature Pyramid. *arXiv preprint arXiv:2401.09386*, 2024. 1

[45] Luchuan Song, Pinxin Liu, Guojun Yin, and Chenliang Xu. Adaptive Super Resolution for One-Shot Talking-Head Generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4115–4119, 2024. 1

[46] Haowen Sun, Ruikun Zheng, Haibin Huang, Chongyang Ma, Hui Huang, and Ruizhen Hu. LGTM: Local-to-Global Text-Driven Human Motion Diffusion Model. In *ACM SIGGRAPH*, 2024. 2, 8

[47] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic 3D facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–9, 2024. 1

[48] Yunlong Tang, Junjia Guo, Hang Hua, Susan Liang, Mingqian Feng, Xinyang Li, Rui Mao, Chao Huang, Jing Bi, Zeliang Zhang, et al. VidComposition: Can MLLMs Analyze Compositions in Compiled Videos? *arXiv preprint arXiv:2411.10979*, 2024. 2

[49] Yunlong Tang, Daiki Shimada, Jing Bi, Mingqian Feng, Hang Hua, and Chenliang Xu. Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding. *arXiv preprint arXiv:2403.16276*, 2024. 2

[50] Yunlong Tang, Junjia Guo, Pinxin Liu, Zhiyuan Wang, Hang Hua, Jia-Xing Zhong, Yunzhong Xiao, Chao Huang, Luchuan Song, Susan Liang, et al. Generative AI for Cel-Animation: A Survey. *arXiv preprint arXiv:2501.06250*, 2025. 1

[51] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 2

[52] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human Motion Diffusion Model. *arXiv preprint arXiv:2209.14916*, 2022. 1, 2, 7

[53] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. *arXiv preprint arXiv:1711.00937*, 2018. 5

[54] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2019. 2, 4

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 4

[56] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. TLControl: Trajectory and Language Control for Human Motion Synthesis. *arXiv preprint arXiv:2311.17135*, 2024. 2

[57] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. OmniControl: Control Any Joint at Any Time for Human Motion Generation. In *International Conference on Learning Representations*, 2024. 2, 8

[58] Xiangpeng Yang, Linchao Zhu, Xiaohan Wang, and Yi Yang. DGL: Dynamic Global-Local Prompt Tuning for Text-Video Retrieval. In *AAAI Conference on Artificial Intelligence*, pages 6540–6548, 2024. 2

[59] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5

[60] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 1, 2, 7

[61] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. FineMoGen: Fine-Grained Spatio-Temporal Motion Generation and Editing. *arXiv preprint arXiv:2312.15004*, 2023. 4, 7, 8

[62] Pengfei Zhang and Deying Kong. Handformer2T: A Lightweight Regression-based Model for Interacting Hands Pose Estimation from A Single RGB Image. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6248–6257, 2024. 1

[63] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3D parametric guidance. In *European Conference on Computer Vision*, 2024. 1

[64] Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. ParCo: Part-Coordinating Text-to-Motion Synthesis. *arXiv preprint arXiv:2403.18512*, 2024. 4, 7, 8