

# Bridging the Gap between Human Motion and Action Semantics via Kinematic Phrases

Xinpeng Liu<sup>1</sup>, Yong-Lu Li<sup>1\*</sup>, Ailing Zeng<sup>2</sup>, Zizheng Zhou<sup>1</sup>, Yang You<sup>3</sup>,  
and Cewu Lu<sup>1\*</sup>

<sup>1</sup> Shanghai Jiao Tong University

<sup>2</sup> Tencent

<sup>3</sup> Stanford University

{xinpengliu0907,ailingzengzzz}@gmail.com,  
{yonglu\_li,zhou\_zz,lucewu}@sjtu.edu.cn, yangyou@stanford.edu

**Abstract.** Motion understanding aims to establish a reliable mapping between motion and action semantics, while it is a challenging many-to-many problem. An abstract action semantic (i.e., *walk forwards*) could be conveyed by perceptually diverse motions (walking with arms up or swinging). In contrast, a motion could carry different semantics w.r.t. its context and intention. This makes an elegant mapping between them difficult. Previous attempts adopted direct-mapping paradigms with limited reliability. Also, current automatic metrics fail to provide reliable assessments of the consistency between motions and action semantics. We identify the source of these problems as the **significant gap** between the two modalities. To alleviate this gap, we propose Kinematic Phrases (KP) that take the objective kinematic facts of human motion with **proper abstraction, interpretability, and generality**. Based on KP, we can unify a motion knowledge base and build a motion understanding system. Meanwhile, KP can be *automatically* converted from motions to text descriptions with no subjective bias, inspiring Kinematic Prompt Generation (KPG) as a **novel white-box motion generation benchmark**. In extensive experiments, our approach shows superiority over other methods. Our project is available at <https://foruck.github.io/KP/>.

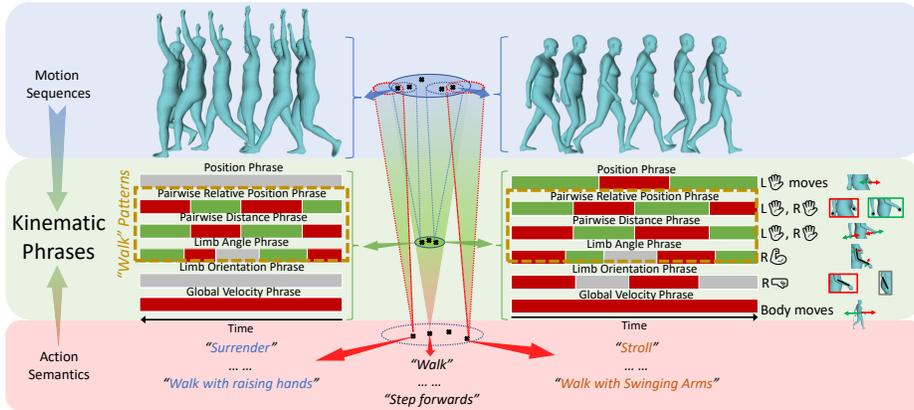
**Keywords:** Motion Representation · Motion Understanding · Text-to-Motion Benchmark

## 1 Introduction

Human motion understanding has a wide range of applications, including activity understanding [31], autonomous driving [36], robotics [27], and automatic animation [49], making it increasingly attractive. The core is to establish a mapping between the motion space and the action semantics space. The motion space indicates a space of sequential 3D human representations, e.g., 3D pose or SMPL [33]/SMPL-X [37] parameter sequence, while the action semantic space

---

\* Corresponding authors.



**Fig. 1:** The huge gap between motion and action semantics results in the *many-to-many* problem. We propose Kinematic Phrases (KP) as an intermediate to bridge the gap. KPs objectively capture human kinematic cues. It properly abstracts diverse motions with interpretability. As shown, the Phrases in the yellow box could capture key patterns of *walk* for diverse motions.

can be represented as action categories or sentences described by natural language.

Recently, a growing focus has been on generative mapping from semantics to motion, including action category-based generation [38] and text-based generation [7, 14, 35, 39, 48, 54, 55]. Most of them typically build a mapping that links motion and semantics either directly or via motion latent, with understated concerns for intermediate motion-semantic structures. However, these models suffer from inferior reliability. They cannot guarantee they generated correct samples without human filtering. Additionally, the existing evaluation of motion generation is problematic. Widely adopted FID and R-Precision rely on the latent space from a black-box pre-trained model, which might fail to out-of-distribution (OOD) and over-fitting cases. There is a long-standing need for an evaluation method that can cheaply and reliably assess whether a generated motion is consistent with particular action semantics. We identify the essence of these as the significant gap between raw human motion and action semantics, which makes direct mapping hard to learn.

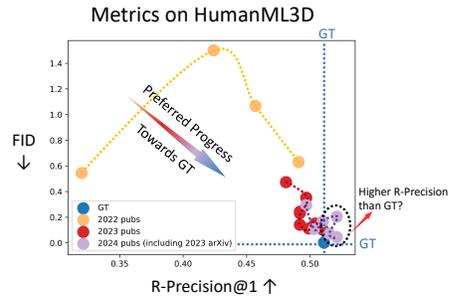
As in Fig. 1, an action semantics can correspond to diverse motions. For instance, a person could *walk* in countless ways with diverse motions, either with arms up or swinging, while action semantics tend to abstract these away from a walking motion. Additionally, they are robust against small perturbations, while motion is more specific and complex, with representations changing vastly when perturbed or mis-captured. Moreover, a motion sequence could have diverse semantics w.r.t. contexts. Modeling this many-to-many mapping between motion and semantics is challenging.

To bridge this gap between motion and action semantics, we propose Kinematic Phrases (KP), an interpretable intermediate representation. KP focuses on the objective kinematic facts, which are usually omitted by general action semantics, like `left-hand moving forwards then backward`. KP is designed as qualitative categorical representations of these facts. For objectivity and actuality, KP captures **sign changes** with minimal pre-defined standards. Inspired by previous studies on kinematic human motion representation [5, 28], KP is proposed as six types shown in Fig. 1, covering **joint positions**, **joint pair positions** and **distances**, **limb angles** and **directions**, and **global velocity**. Note that, although KP can be described by natural language, a major difference is that KP is strictly dedicated to objective kinematic facts instead of coarse actions such as *surrender* or fine-grained actions like *raise both hands*.

We highlight three advantages of KP. First, KP offers **proper abstraction**, which disentangles motion perturbations and semantics changes, easing the learning process. Even though the motion differs significantly, KP manages to capture *walk* patterns easily. Second, KP is **interpretable**, as it can be viewed as instructions on executing the action, making it easily understandable to humans. Finally, KP is **general**, as it can be automatically extracted from different modalities of human motion, including skeleton and SMPL parameters. The conversion from KP to text is also effortless.

With KP as an intermediate representation, a unified large-scale motion knowledge base is constructed. Then, to fully exploit KP and the knowledge base, we build a motion understanding system with KP mediation. In detail, we learn a motion-KP joint latent space for multiple motion understanding applications, including motion interpolation, modification, and generation. Moreover, leveraging the interpretability of KP, we propose a benchmark called Kinematic Prompts Generation (KPG), which generates motion from text prompts converted from KPs. Thanks to the consistency and convenience of the KP-to-text conversion, KPG enables efficient white-box motion generation evaluation. Given the bottle-necked evaluation of the current motion generation benchmarks [14, 41] as shown in Fig. 2, KPG could be a promising *complementary* benchmark. Results in Tab. 2 show that KPG brings novel insights into the current text-to-motion progress.

Our contributions are: (1) We propose KP as an intermediate representation to bridge the gap between motion and action semantics. (2) We build a novel



**Fig. 2:** HumanML3D [14] metric values are approaching even surpassing GT-level while being increasingly indistinguishable. The higher-than-GT R-Precision could be insufficient as a semantic consistency indicator.

motion understanding system using KP and the aggregated large-scale knowledge base. (3) We propose KPG as a benchmark for efficient white-box motion generation evaluation. Promising results are achieved on motion interpolation and generation tasks. Moreover, extensive user studies are conducted, verifying the efficacy of our methods, and the consistency between KPG evaluation and human perception.

## 2 Related Works

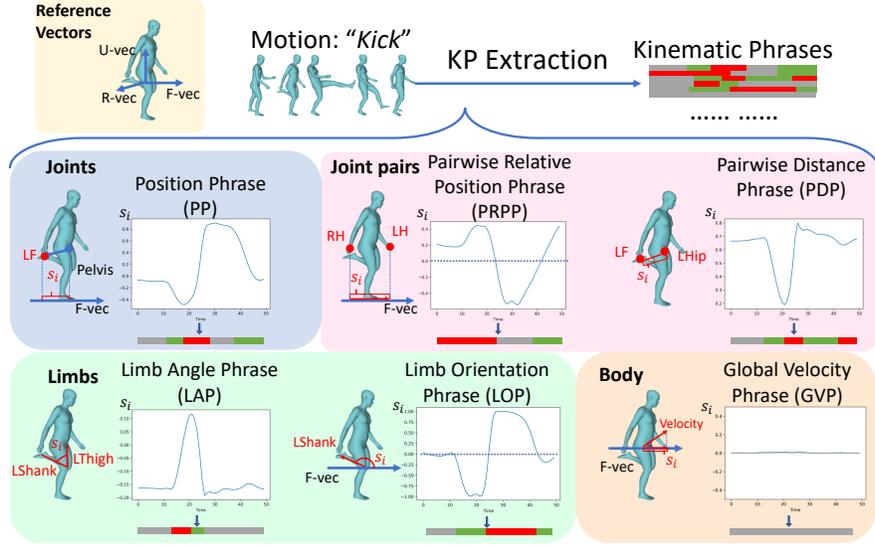
**Motion Representation** An intuitive motion representation is a sequence of static pose representations, like joint locations and limb rotations. Efforts are paid to address the discontinuity of rotation for deep-learning methods [6, 59]. Recent works on parametric body models [33, 37] enable a more realistic body representation. Meanwhile, [42] proposed Posebits, representing pose with boolean geometric part relationships. [9, 10] translates Posebits into text descriptions. These abstract representations are flexible and insensitive to little perturbations, but their static nature ignores motion dynamics. [46] acquire similar fine-grained descriptions from human annotation, while [3, 51] adopted large-scale language models. However, few recognize their potential in bridging the low-level motion and the high-level action semantics. Phase functions [20], Labanotations [28], and learned Motion Words [1] were also explored, though limited to specific actions like locomotion and dancing.

**Motion Generation** can be conditioned by its prefix/suffix [2, 17, 19], action categories [16, 38, 52], or audio [29, 30]. Text-based motion generation has developed rapidly with the proposal of text-motion datasets [14, 43]. Early works [14, 39, 44] used VAEs, while some [21, 32, 47] extended the CLIP [45] space to motion. Diffusion models [8, 48, 50, 55] were also adopted. MAA [4] adopted a U-Net structure. Some [40, 56] explored retrieval-based methods. [25] aimed at controllable generation, while [53] introduced physical constraints. However, most approaches still suffer from the gap between motion and action semantics. Multiple methods [7, 15, 26, 35, 54, 58, 61] adopted (VQ-)VAE-compressed motion representation as mediation, while in the current data-limited situation, we identify that this single-modality compression might be sub-optimal. Instead, KP could alleviate this by introducing explicit semantic-geometric correlation.

## 3 Kinematic Phrase Base

### 3.1 Kinematic Phrases

Kinematic Phrases abstract motion into objective kinematic facts like **left-hand moves up** qualitatively. We take inspiration from previous kinematic motion representations [28] and qualitative static pose representations [9, 11, 42], proposing six types of KP to comprehensively represent motion from different kinematic hierarchies: For **joint movements**, there are 34 Position Phrases (PPs). For **joint**



**Fig. 3:** Six types of KP from four kinematic hierarchies are extracted from a motion sequence. A scalar indicator  $s_i$  is calculated per Phrase *per frame*. Its sign categorizes the corresponding Phrase.

**pair movements**, there are 242 Pairwise Relative Position Phrases (PRPPs) and 81 Pairwise Distance Phrases (PDPs). For **limb movements**, there are 8 Limb Angle Phrases (LAPs) and 24 Limb Orientation Phrases (LOPs). For **whole-body movements**, there are 3 Global Velocity Phrases (GVPs). KP extraction is based on a skeleton sequence  $X = \{x_i | x_i \in \mathcal{R}^{n_k \times 3}\}_{i=1}^t$ , where  $n_k$  is the number of joints ( $n_k = 17$  here),  $x_i$  is the joint coordinates at  $i$ -th frame, and  $t$  is the sequence length. Note that  $x_i^0$  indicates the pelvis/root joint. For each Phrase, a scalar indicator sequence is calculated from the skeleton sequence. Phrases are extracted as per-frame categorical representations w.r.t. indicator signs. Unlike previous efforts [9, 42], we limit the KP criteria as the indicator signs to minimize human-defined biases (e.g., numerical criteria on the closeness of two joints) for objectivity and actuality. Fig. 3 illustrated the procedure.

**Reference Vectors** are first constructed, indicating right, upward, and forward directions from a human *cognitive view*. We aim at the *egocentric* reference frames that human tends to use when performing actions. The negative direction of gravity is adopted as upward vector  $r^u$ , the vector from left hip to right hip is adopted as right vector  $r^r$ , and the forward vector is calculated as  $r^f = r^u \times r^r$ . These vectors of each frame are denoted as  $R = \{r_i\}_{i=1}^t$ .

**Position Phrase (PP)** focuses on the movement direction of joint  $x^j$  w.r.t. reference vector  $R$ . The indicator for PP at  $i$ -th frame is calculated as

$$s_i^{(j,\cdot)} = \langle (x_i^j - x_i^0), r_i^{\cdot} \rangle - \langle (x_{i-1}^j - x_{i-1}^0), r_{i-1}^{\cdot} \rangle. \quad (1)$$

The sign of  $s_i^{(j,\cdot)}$  categorizes PP into moving **along/against**  $R$ , or **relatively static** along  $R$  for indicators with small amplitudes. 12 joints (except the pelvis, eyes, and hips) and 3 reference vectors resulted in originally 36 PPs. Specifically, shoulders are excluded for the leftward vector, and hips are excluded for the upward vector, resulting in 36-2=34 PPs.

**Pairwise Relative Position Phrase (PRPP)** describes the relative position of a pair of joints  $(x^j, x^k)$  w.r.t. reference vector  $R$ . PRPP indicator at  $i$ -th frame is

$$s_i^{(j,k,\cdot)} = \langle (x_i^j - x_i^k), r_i \rangle. \quad (2)$$

For (L-Hand, R-Hand) and forward vector  $R^f$ , PRPP could be L-Hand **behind** /in front of R-Hand according to the sign of  $s_i^{(j,k,\cdot)}$ . 136 joint pairs and 3 reference vectors composed 408 Phrases originally. 24 joint pairs linked by limbs are filtered out. 30 Eye-related pairs are filtered out except (left eye, right eye) due to the others are covered by head-related pairs. 4 triplets that the relationship barely changes along the reference vector (e.g., right knee, left hip, and leftward vector) are filtered out. These result in 408-24\*3-30\*3-4=242 PRPPs.

**Pairwise Distance Phrase (PDP)** describes how the L2 distance between a pair of joints  $(x^j, x^k)$  changes. The indicator for PDP is calculated as

$$s_i^{(j,k)} = \|x_i^j - x_i^k\|_2 - \|x_{i-1}^j - x_{i-1}^k\|_2. \quad (3)$$

The sign of  $s_i^{(j,k)}$  categorizes PDP into moving **closer/away**, or **relatively static**. Among the original 105 joint pairs, 24 pairs that are in the skeleton topology are excluded, such as the hand and elbow, resulting in 81 PDPs.

**Limb Angle Phrase (LAP)** targets at the change of bend angle between two connected limbs  $(x^j, x^k)$  and  $(x^j, x^l)$ . The LAP indicator is calculated as

$$s_i^{(j,k,l)} = \arccos(\langle x_i^k - x_i^j, x_i^l - x_i^j \rangle) - \arccos(\langle x_{i-1}^k - x_{i-1}^j, x_{i-1}^l - x_{i-1}^j \rangle). \quad (4)$$

8 LAPs describe the limb chain  $(x^j, x^k)$ - $(x^j, x^l)$  as **bending** or **unbending**.

**Limb Orientation Phrase (LOP)** describes the orientation of the limb  $(x^j, x^k)$  w.r.t.  $R$ , with  $x^k$  being the distal limb. The LOP indicator is

$$s_i^{(j,k,\cdot)} = \langle x_i^k - x_i^j, r_i \rangle. \quad (5)$$

The sign of  $s_i^{(j,k,\cdot)}$  categorizes the LOP  $(x^j, x^k)$  pointing **along/against**  $R$ , or a placeholder category for those with little magnitude. 4 arms, 4 leg limbs, head, collarbones, hips, torsos, and upper body, paired with the 3 reference vectors, result in 45 LOPs in total. 21 pairs that barely change along the reference vector (e.g., left-right hip, leftward vector) are filtered out, resulting in 45-21=24 LOPs.

**Global Velocity Phrase (GVP)** describes the direction of global velocity with respect to  $R$ . The indicator is calculated as

$$s_i = \langle x_{i+1}^0 - x_i^0, r_i \rangle. \quad (6)$$

The three categories are moving **along/against**  $R$ , or **static** along  $R$  according to the sign of  $s_i$ .

These result in 392 Phrases in total, covering motion diversity and distribution from various levels. While we clarify that these Phrases do not rule out the possibility of other possible useful potentials.

### 3.2 Constructing Kinematic Phrase Base

KP enables us to unify motion data with different formats to construct a large knowledge base containing motion, text, and KP thanks to its generality. Motion sequences of different representations are collected, including 3D skeleton sequences and SMPL [33] /SMPL-X [37] parameter sequences. The sequences are first re-sampled to 30Hz and rotated so that the negative direction of the z-axis is the gravity direction. Then, the sequences are converted into 3D skeleton sequences for KP extraction as in Sec. 3.1. Text annotations attached to the sequences are directly saved. For sequences with action category annotation, the category name is saved. For those with neither text nor action category, the text information is set from its attached additional information, like objects for SAMP [18]. Finally, we collect 140K motion sequences from 11 datasets. More details are included in the appendix.

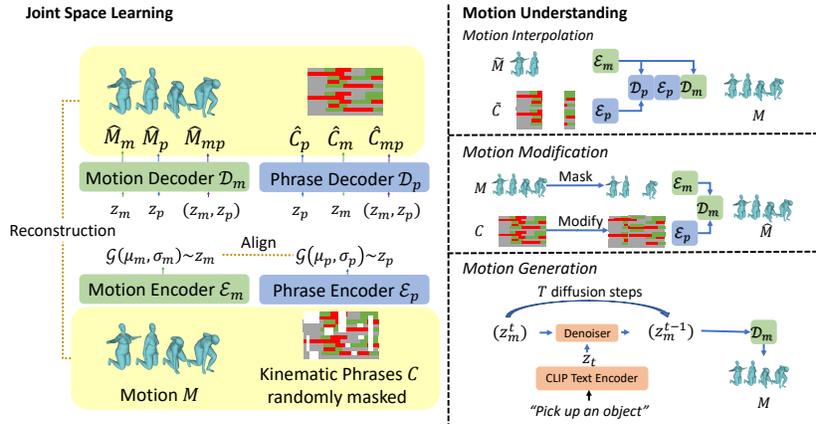
## 4 Applications of KP

### 4.1 Motion Understanding with KP

An intuitive application of KP is to leverage it for better motion understanding. To achieve this, we first learn a motion-KP joint space, which exploits the clarity and interpretability of KP. Then, we adopt the learned latent space for multiple motion understanding tasks. An overview is illustrated in Fig. 4.

**Preliminaries.** Motion is represented as a pose sequence with  $n$  frames as  $M = \{m_i\}_{i=1}^n$ . In detail, SMPL [33] pose parameters are transformed to the 6D representation [59], then concatenated with the velocity of the root, resulting in a 147-dim representation per frame. KP is represented by indicator signs as  $C = \{c_i\}_{i=1}^n, c_i \in \{1, 0, -1\}^{392}$ .

**Joint Space Learning.** The joint space is learned with Motion VAE  $\{\mathcal{E}_m, \mathcal{D}_m\}$  and KP VAE  $\{\mathcal{E}_p, \mathcal{D}_p\}$ . All VAEs follow the transformer-based architecture from [38]. Motion encoder  $\mathcal{E}_m$  takes motion  $M$  and two distribution tokens  $m_\mu, m_\sigma$  as input, and the outputs corresponding to the distribution tokens are taken as the  $\mu_m$  and  $\sigma_m$  of the Gaussian distribution. Then, motion decoder  $\mathcal{D}_m$  takes  $z_m \sim \mathcal{G}(\mu_m, \sigma_m)$  as  $K, V$ , and a sinusoidal positional encoding of the expected duration as  $Q$ . The output is fed into a linear layer to reconstruct motion  $\hat{M}$ . KP VAE  $\{\mathcal{E}_p, \mathcal{D}_p\}$  operates similarly, with the sign of  $\mathcal{D}_p$  output as the predicted KP  $\hat{C}$ . Notice that the  $\mathcal{D}_m, \mathcal{D}_p$  could take arbitrary combinations of  $z_m, z_p$  as input, outputting  $\hat{M}, \hat{C}$ . To learn the joint space, besides the conventional loss



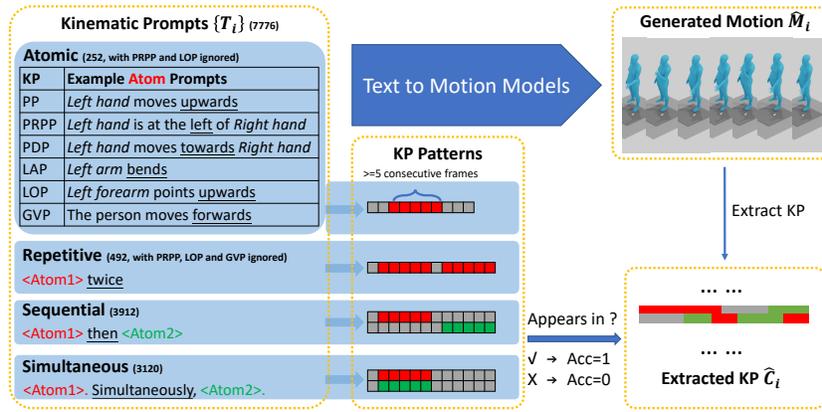
**Fig. 4:** We train a motion-KP joint latent space to exploit the clarity and interpretability of KP. The space is then applied to multiple tasks, including motion interpolation, modification, and generation.

of reconstruction and KL divergence, we also align the two VAEs similarly to [39]. In this way, we guide the learned motion latent space with the structure of KP. Meanwhile, we randomly corrupt samples during training by setting a small portion of KP as 0. The idea is that the overall representation should be robust with a small portion of KP unknown. Also, the missing Phrases should be recovered from existing Phrases. This mines the correlation among Phrases and increases the robustness of the space. More details are in the supplementary.

**Motion Understanding Tasks with Joint Space.** With the joint space, we can perform both low-level and high-level motion understanding tasks. We introduce three applications to show the capability of KP, as shown in Fig. 4. For **motion interpolation** given a corrupted motion sequence  $\tilde{M}$ , we extract its corresponding KP sequence  $\tilde{C}$ , then feed them to encoders  $\mathcal{E}_m, \mathcal{E}_p$  and decoder  $\mathcal{D}_p$ , resulting in the estimated KP sequence  $\hat{C}$ .  $\hat{C}$  and  $\tilde{M}$  are fed into  $\mathcal{E}_m, \mathcal{E}_p$  and  $\mathcal{D}_m$ , resulting in interpolated  $\hat{M}$ . For **motion modification** on motion  $M$ , modifications could be made on the extracted KP sequence  $C$  resulting in  $\tilde{C}$ . Motion frames to modify are then masked, getting  $\tilde{M}$ .  $\tilde{M}, \tilde{C}$  are fed into  $\mathcal{E}_m, \mathcal{E}_p$  and  $\mathcal{D}_m$ , getting the interpolated  $\hat{M}$ . For **motion generation**, we simply replace the motion latent space in MLD [7] with the latent space of our Motion VAE. Experiments show that KP manages to boost MLD with a considerable margin.

## 4.2 Motion Benchmarking with Kinematic Prompt Generation

A more interesting and important application is benchmarking motion generation with the interpretability and objectivity of KP.



**Fig. 5:** Kinematic Prompt Generation. 7,796 prompts are converted from KP with templates, corresponding to certain KP patterns. We calculate the generation accuracy by checking the appearance of the patterns in KP extracted from the generated motion.

Before that, we first analyze current benchmarks. A crucial aspect of motion generation evaluation is motion-semantic consistency. The gold standard is user study. However, it is expensive and inefficient to scale. Early metrics like MPJPE (Mean Per Joint Position Error) and MAE (Mean Angle Error) mechanically calculate the error between the generated and GT samples. These metrics fail to reveal the real ability of generative models: What if the models memorize GT samples? Or what if the samples are diverse from GT but also true? FID (Fréchet Inception Distance) is adopted to mitigate this issue. However, it provides a macro view of the quality of all generated samples without guarantees for individual samples. [14] proposed R-Precision, using a pre-trained text-motion matching model to examine whether the generated samples carry true semantics. They both rely on the latent space from a black-box pre-trained model, which is not credible. Besides, models might learn short paths to over-fit the pre-trained model. Meanwhile, since automatic mapping from motion to semantics is still an unsettled problem, adopting it to evaluate motion generation is not a decent choice. Moreover, most current motion generation evaluations are performed on datasets [14, 22, 41] with considerable complex everyday actions, further increasing the difficulty. As shown in Fig. 2 and Tab. 1, the current evaluation metrics are becoming increasingly indistinguishable. Also, the higher-than-GT R-Precision could be insufficient as an indicator of semantic consistency.

To this end, we propose a novel benchmark: Kinematic Prompts Generation (KPG). Instead of previous benchmarks focusing on everyday activities or sports, we step *back* in the complexity of action semantics. Based on KP, KPG focuses on evaluating *whether the models could generate motion sequences consistent with specific kinematic facts given text prompts*. An overview is given in Fig. 5.

In detail, we convert KP into 7,796 text prompts with templates as in Fig. 5. The prompts could be categorized into 4 groups. First, each **atomic prompt**

involves only 1 KP, such as “left hand moves upwards”, which evaluates the basic semantic consistency of text-to-motion models. Notice that PRPP and LOP are excluded since they represent static states, resulting in 252 prompts. Second, each **repetitive prompt** involves a KP being repeated two or three times., like “left arm bends twice”, evaluating the basic counting ability. PRPP, LOP, and GVP are excluded, bringing 492 prompts. Third, each **sequential prompt** involves two independent KP to be sequentially executed, like “left arm bends, then right arm bends”. 3,912 sequential prompts are randomly sampled. Finally, each **simultaneous prompt** involves two independent KPs to be simultaneously executed, like “left arm bends, and simultaneously, right arm bends.” 3,120 prompts are randomly sampled. The sequential and simultaneous prompts examine whether the model can understand the composition of simple actions. Given prompt  $T_i \in T$  from Phrase  $c_i$  (or Phrases  $(c_i^1, c_i^2)$ ), the model generates motion  $\hat{M}_i$ , along with extracted KP  $\hat{C}_i$ . Then, whether  $T_i$  is hit could be inferred via designed protocols by identifying certain KP patterns in  $\hat{C}_i$ . In detail, the atomic prompt is identified as hit if  $c_i$  appears for more than 5 consecutive frames in  $\hat{C}_i$ . The repetitive prompt is identified as hit if and only if  $c_i$  appears for at least 5 consecutive frames in  $\hat{C}_i$  for corresponding times. The sequential prompt is identified as hit if  $c_i^2$  appears for at least 5 consecutive frames after (maybe not immediately)  $c_i^1$  does the same. The simultaneous prompt is identified as hit if  $c_i^1$  and  $c_i^2$  appear simultaneously for at least 5 consecutive frames. The calculation involves no black-box model thanks to KP, presenting a fully white-box evaluation. Also, with the effortless motion-to-KP conversion, the computation could be conducted automatically. More details are in the appendix.

## 5 Experiment

**Implementation Details.** HumanML3D [14] test split is held out for evaluation, with the rest of KPB for training. The batch size is set as 288, and an AdamW optimizer with a learning rate of 1e-4 is adopted. We randomly corrupt less than 20% of the Phrases for a sample. The Motion-KP joint space is trained for 6,000 epochs. While the text-to-motion latent diffusion model is trained for 3,000 epochs, with the joint space frozen. All experiments are conducted in 4 NVIDIA RTX 3090 GPUs. More details are in the appendix.

### 5.1 Motion Interpolation

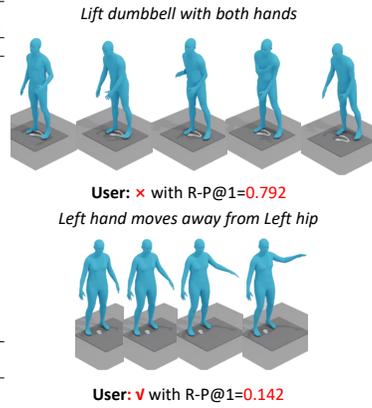
Following [23], 50% frames are randomly masked for interpolation. In Tab. 1, our method provides better FID. With KPB, Diversity is increased.

### 5.2 Motion Generation

**Results on conventional text to motion** are shown in Tab. 1. We adopt the HumanML3D test set [14] for conventional text-to-motion evaluation. The evaluation model from [14] is adopted to calculate R-Precision, FID, Diversity,

Methods	Interpolation			Generation		
	FID↓	Diversity→	R-P@1↑	FID↓	Diversity→	Multimodality
GT	0.002	9.503	0.511	0.002	9.503	-
TEMOS [39]	-	-	0.424	3.734	8.973	0.368
T2M [14]	-	-	0.455	1.067	9.188	2.090
MotionDiffuse [55]	-	-	0.491	0.630	9.410	0.730
MDM [48]	2.698	8.420	0.320	0.544	9.559	2.799
TM2T [15]	-	-	0.424	1.501	8.589	2.424
MLD [7]	-	-	0.481	0.473	9.724	2.413
T2M-GPT [54]	-	-	0.492	0.141	9.722	1.831
Fg-T2M [50]	-	-	0.492	0.243	9.278	1.614
M2DM [26]	-	-	0.497	0.352	9.926	3.587
AttT2M [58]	-	-	0.499	0.112	9.700	2.452
ReMoDiffuse [56]	-	-	0.510	0.103	9.018	1.795
MotionGPT [23]	0.214	<b>9.560</b>	0.492	0.232	<b>9.528</b>	2.008
FineMoGen [57]	-	-	0.504	0.151	9.263	2.696
GraphMotion [24]	-	-	0.504	0.116	9.692	2.766
AvatarGPT [60]	-	-	0.510	0.168	9.624	-
GUESS [12]	-	-	0.503	0.109	9.826	2.430
MDD [34]	-	-	0.515	0.072	9.683	1.869
MoMask [13]	-	-	<b>0.521</b>	<b>0.045</b>	-	1.241
Ours	<b>0.192</b>	9.623	0.496	0.275	9.975	2.218
Ours w/ KPB	0.214	10.142	0.471	0.641	10.262	2.891

**Table 1:** Results on HumanML3D. R-P@1 is short for R-Precision@1. Most methods show comparable performance. Also, the higher-than-GT performance indicates that R-Precision might be insufficient as a semantic consistency indicator.



and Multimodality. We run the evaluation 20 times and report the average value. Our method is competitive without KPB. However, there exist some unsatisfying observations. First, KPB brings a counter-intuitive performance drop. Second, the metric values of most techniques are rather close (e.g., difference  $<0.03$  for R-P@1), and no substantial advantage is perceived for one method over others. Also, as discovered by MAA [4], despite its relatively poor R-Precision and FID, MAA manages to achieve better user reviews, indicating the potential loss of efficacy of these conventional metrics. This is further revealed in Fig. 6, where a contradiction is observed between human judgment and R-precision. Moreover, some methods show higher-than-GT R-P@1, which indicates that R-Precision might be insufficient as a semantic consistency indicator. Given these, we further switch to KPG and pursue further insights.

**Results on KPG** are shown in Tab. 2. Besides accuracy, diversity and R-precision are reported. We run the evaluation 20 times and report the average metric value. The white-box designation of KPG evaluation makes it possible to evaluate the different aspects of the models. Meanwhile, KPG is considered an easier task than conventional text-based motion generation since it is targeted at much less complex action semantics. Results show that previous methods are not performing as well as expected. Also, R-Precision fails to model the performance on KPG with trivially low values. Most methods perform well for atomic prompts, but when it comes to compositions, the accuracy drops. Early efforts like HMDM [48] and MotionDiffuse [55] are surprisingly competitive, even surpassing several more recent works. T2M-GPT [54] outperforms other methods by a considerable margin, but it fails for repetitive prompts. An unexpected

Methods	Accuracy				Diversity R-P@1		
	Atomic	Repetitive	Sequential	Simultaneous			Overall
HMDM [48]	93.25	23.98	56.11	30.90	45.16	5.923	0.034
MotionDiffuse [55]	93.21	22.15	51.54	27.76	41.49	5.257	0.051
T2M-GPT [54]	<u>97.22</u>	1.02	<u>70.24</u>	<u>42.47</u>	<u>55.59</u>	6.153	0.042
MLD [7]	92.06	22.76	48.26	28.01	39.94	5.526	0.061
ReMoDiffuse [56]	53.17	3.05	32.16	31.06	30.56	5.809	0.048
MotionGPT [23]	92.06	16.87	55.98	31.76	44.96	5.776	0.039
MoMask [13]	96.43	<b>28.86</b>	59.10	32.66	47.79	5.965	0.060
Ours	<b>98.80</b>	<u>26.02</u>	<b>71.32</b>	<b>43.08</b>	<b>57.86</b>	6.048	0.053

**Table 2:** Results on KPG demonstrate the various strengths of different methods.

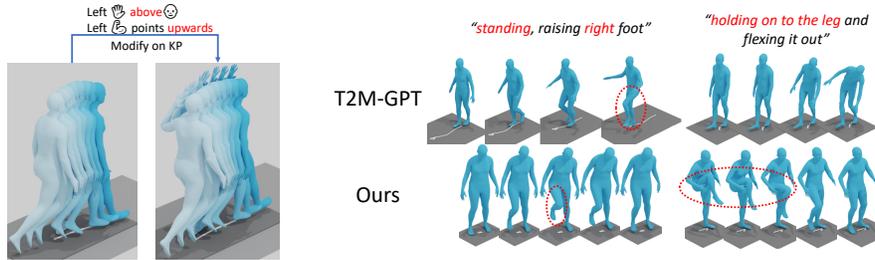
phenomenon is the failure of ReMoDiffuse [56], which might be constrained by its retrieval-based methodology since most prompts in KPG are not available in the retrieval database. Our method, which replaces the latent space of MLD [7], delivers substantial improvements, but the accuracy remains below 60%, which is far from satisfying. A considerable gap exists between existing efforts and ideal motion generation models. More analyses are in the appendix.

**User Study.** To reach a convincing performance comparison, we conduct user studies on both HumanML3D and KPG. Our user study is different from previous efforts in two aspects. First, instead of testing a small set of text prompts (less than 50 in previous works [7, 48]), we randomly select 600 sentences from the HumanML3D test set. By scaling up, the result is convincing in reflecting the ability to generate motion for diverse text inputs. Second, without asking the volunteers to give a general rating for each sample or to choose between different samples, we ask them two questions: 1) Do the motion and the text match? and 2) Is the motion natural? For Q1, three choices are given as “No, Partially, Yes”. For Q2, two choices are given as “Yes, No”. In this way, we explicitly evaluate two aspects of text-to-motion tasks: semantic consistency and naturalness, corresponding to R-Precision and FID. For each prompt, we generate one sample considering the annotation cost. We claim that the models should generate natural text-matching motion most of the time so that the one-sample setting would not hurt the fidelity of our user study. 36 volunteers are invited, each reviewing 200 sequences. Thus each sequence receives 3 user reviews. Also, we compute R-precision@1 of the generated sequences.

Besides our model, we select MDM [48] as a naive baseline, T2M-GPT [54] as the SOTA model given the results in Tab. 1 and Tab. 2. MLD [7] is also selected as a direct baseline of our proposed method. Results on HumanML3D [14] are shown in Fig. 7. Though our method is not superior in R-Precision, we receive competitive user reviews, showcasing the efficacy of our motion-KP joint space. Also please notice that our model only contains **45.1M** parameters, while T2M-GPT contains **228M** parameters. T2M-GPT and MLD present similar R-Precision, but only T2M-GPT manages to keep a good performance with user reviews. Moreover, the discrepancy between R-Precision and user reviews is revealed in both absolute value and trends, also in Fig. 6.



**Fig. 7:** User study on HumanML3D, with “Y” for Yes and “P” for partially. **Fig. 8:** User study on KPG, with “Y” for Yes and “P” for partially.



**Fig. 9:** Our model supports fine-grained modification on motion via KP. **Fig. 10:** Visualization of generated samples. Compared to T2M-GPT, our method provides a better response to prompts with explicit constraints on specific body parts.

A similar user study on KPG is conducted in Fig. 7 with 100 randomly selected prompts from KPG involving T2M-GPT and our model. Fig. 8 demonstrates that KP-inferred Accuracy and user reviews share similar trends. Meanwhile, KP and user study give the same reviews for **84%** of the samples. We believe KPG could thus be a first step towards white-box automatic motion generation evaluation for semantic consistency. More analyses are in the appendix.

### 5.3 Visualization

A modification sample is shown in Fig. 9. By modifying KP, we could edit arbitrary motion at a fine-grained level. Fig. 10 compares generated samples of T2M-GPT and our methods. Our method properly responds to text prompts with constraints on specific body parts thanks to KP, which explicitly models the action semantics with body part kinematics cues. Note that T2M-GPT generates redundant motion for simple prompts, while our method provides more concise and precise results. More visualizations are in the appendix.

### 5.4 Ablation Studies

KPG ablation study results are in Tab. 3.

**KP latent.** By using a latent space without KP, a substantial performance drop is perceived, revealing the significance of KP in conveying action semantics.

**Different KP sets.** We examine the contribution of different KP sets: joint KP (PP), joint pair KP (PRPP, PDP), limb KP (LAP, LOP), and body KP (GVP). A leave-one-out evaluation shows the elimination of joint KP and joint pair KP results in notable performance degradation, while the influence of the rest is relatively subtle.

Methods	Acc.% $\uparrow$	Diversity
Ours	<b>57.86</b>	6.048
w/o KP	39.94	5.526
w/o Joint KP	50.03	5.685
w/o Joint Pair KP	47.24	5.772
w/o Limb KP	55.92	5.934
w/o Body KP	56.84	5.871

**Table 3:** KPG ablation results.

## 6 Discussion

**Limitations. First**, the current KP criteria guarantee objectivity but overlook kinematic information like amplitude and speed. Also, due to the granularity of the adopted skeleton, fine-grained information on fingers is not well-preserved. These would be promising for further exploration. **Second**, KPG could be extended to datasets with other modalities, like 2D pose and egocentric action datasets. Although these modalities provide incomplete 3D information, we could extract KP that is credibly accessible across modalities. **Third**, with the convenient KP-to-text conversion, auxiliary motion descriptions could be automatically generated via KP. **Fourth**, KPG could be extended by paraphrasing existing prompts and combining different Phrases. **Finally**, the proposed motion generation pipeline is devised for quick evaluation, while the motion-KP latent space could be incorporated with more advanced algorithms for better exploitation.

**Broader Impacts.** The proposed KP and KPG provide a white-box instrument to diagnose text-to-motion algorithms, which could help develop reliable motion tools. Better motion synthesis could advance fields including animation production, sports training, human-robot interaction, and rehabilitation. However, it could also be used for illegal media content or false information.

## 7 Conclusion

In this paper, we proposed an intermediate representation to bridge human motion and action semantics as the Kinematic Phrase. KP achieved proper abstraction, interpretability, and generality by focusing on objective kinematic facts of human motion. A motion understanding system based on KP was proposed and proven effective in motion interpolation, modification, and generation. Moreover, Kinematic Prompt Generation is proposed as a white-box text-to-motion benchmark, bringing new insights to the literature. We believe that KP has great potential to advance motion understanding.

## Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grants No.62306175, No.62302296, National Key Research and Development Project of China (No.2022ZD0160102, No.2021ZD0110704), Shanghai Artificial Intelligence Laboratory, XPLOER PRIZE grants.

## References

1. Aristidou, A., Cohen-Or, D., Hodgins, J.K., Chrysanthou, Y., Shamir, A.: Deep motifs and motion signatures. *ACM Trans. Graph.* **37**(6), 187:1–187:13 (Nov 2018). <https://doi.org/10.1145/3272127.3275038>, <http://doi.acm.org/10.1145/3272127.3275038>
2. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Teach: Temporal action composition for 3d humans. In: 2022 International Conference on 3D Vision (3DV). pp. 414–423. IEEE (2022)
3. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: SINC: Spatial composition of 3D human motions for simultaneous action generation. In: ICCV (2023)
4. Azadi, S., Shah, A., Hayes, T., Parikh, D., Gupta, S.: Make-an-animation: Large-scale text-conditional 3d human motion generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15039–15048 (October 2023)
5. Bartlett, R.: Introduction to Sports Biomechanics. Introduction to Sports Biomechanics, E & FN Spon (1997), <https://books.google.com.tw/books?id=-6Db8mgxsqQC>
6. Brégier, R.: Deep regression on manifolds: a 3d rotation case study. In: 2021 International Conference on 3D Vision (3DV). pp. 166–174. IEEE (2021)
7. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023)
8. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9760–9770 (June 2023)
9. Delmas, G., Weinzaepfel, P., Lucas, T., Moreno-Noguer, F., Rogez, G.: Posescript: 3d human poses from natural language. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI. pp. 346–362. Springer (2022)
10. Delmas, G., Weinzaepfel, P., Moreno-Noguer, F., Rogez, G.: Posefix: Correcting 3d human poses with natural language. arXiv preprint arXiv:2309.08480 (2023)
11. Fang, H.S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.L., Lu, C.: Alpha-pose: Whole-body regional multi-person pose estimation and tracking in real-time. TPAMI (2023)
12. Gao, X., Yang, Y., Xie, Z., Du, S., Sun, Z., Wu, Y.: Guess: Gradually enriching synthesis for text-driven human motion generation. *IEEE Transactions on Visualization and Computer Graphics* (2024)
13. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. arXiv preprint arXiv:2312.00063 (2023)

14. Guo, C., Zuo, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (June 2022)
15. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022)
16. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)
17. Guo, W., Du, Y., Shen, X., Lepetit, V., Alameda-Pineda, X., Moreno-Noguer, F.: Back to mlp: A simple baseline for human motion prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4809–4819 (2023)
18. Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.: Stochastic scene-aware motion prediction. In: Proceedings of the International Conference on Computer Vision 2021 (Oct 2021)
19. Hernandez, A., Gall, J., Moreno-Noguer, F.: Human motion prediction via spatio-temporal inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7134–7143 (2019)
20. Holden, D., Kanoun, O., Perepichka, M., Popa, T.: Learned motion matching. *ACM Transactions on Graphics (TOG)* **39**(4), 53–1 (2020)
21. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. arXiv preprint arXiv:2205.08535 (2022)
22. Ji, Y., Xu, F., Yang, Y., Shen, F., Shen, H.T., Zheng, W.S.: A large-scale rgb-d database for arbitrary-view human action recognition. In: Proceedings of the 26th ACM International Conference on Multimedia. pp. 1510–1518 (2018)
23. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. arXiv preprint arXiv:2306.14795 (2023)
24. Jin, P., Wu, Y., Fan, Y., Sun, Z., Yang, W., Yuan, L.: Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. *Advances in Neural Information Processing Systems* **36** (2024)
25. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2151–2162 (October 2023)
26. Kong, H., Gong, K., Lian, D., Mi, M.B., Wang, X.: Priority-centric human motion generation in discrete latent space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14806–14816 (October 2023)
27. Koppula, H.S., Saxena, A.: Anticipating human activities for reactive robotic response. In: IROS. p. 2071. Tokyo (2013)
28. von Laban, R., Lange, R.: *Laban’s Principles of Dance and Movement Notation*. Macdonald & Evans (1975), <https://books.google.com.tw/books?id=-VrOAAAAAAAJ>
29. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13401–13412 (2021)
30. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Learn to dance with aist++: Music conditioned 3d dance generation (2021)

31. Li, Y.L., Wu, X., Liu, X., Wang, Z., Dou, Y., Ji, Y., Zhang, J., Li, Y., Lu, X., Tan, J., et al.: From isolated islands to pangea: Unifying semantic space for human action understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16582–16592 (2024)
32. Lin, J., Chang, J., Liu, L., Li, G., Lin, L., Tian, Q., Chen, C.W.: Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23222–23231 (June 2023)
33. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* **34**(6), 1–16 (2015)
34. Lou, Y., Zhu, L., Wang, Y., Wang, X., Yang, Y.: Diversemotion: Towards diverse human motion generation via discrete diffusion. *arXiv preprint arXiv:2309.01372* (2023)
35. Lucas, T., Baradel, F., Weinzaepfel, P., Rogez, G.: Posegpt: quantization-based 3d human motion generation and forecasting. In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*. pp. 417–435. Springer (2022)
36. Paden, B., Čáp, M., Yong, S.Z., Yershov, D., Frazzoli, E.: A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on intelligent vehicles* **1**(1), 33–55 (2016)
37. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. pp. 10975–10985 (2019)
38. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10985–10995 (2021)
39. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*. pp. 480–497. Springer (2022)
40. Petrovich, M., Black, M.J., Varol, G.: Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 9488–9497 (October 2023)
41. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. *Big data* **4**(4), 236–252 (2016)
42. Pons-Moll, G., Fleet, D.J., Rosenhahn, B.: Posebits for monocular human pose estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2337–2344 (2014)
43. Punakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: Babel: bodies, action and behavior with english labels. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 722–731 (2021)
44. Qian, Y., Urbanek, J., Hauptmann, A.G., Won, J.: Breaking the limits of text-conditioned 3d motion synthesis with elaborative descriptions. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 2306–2316 (October 2023)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

- natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
46. Tang, Y., Liu, J., Liu, A., Yang, B., Dai, W., Rao, Y., Lu, J., Zhou, J., Li, X.: Flag3d: A 3d fitness activity dataset with language instruction. arXiv preprint arXiv:2212.04638 (2022)
  47. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII. pp. 358–374. Springer (2022)
  48. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
  49. Van Welbergen, H., Van Basten, B.J., Egges, A., Ruttkay, Z.M., Overmars, M.H.: Real time animation of virtual humans: a trade-off between naturalness and control. In: Computer Graphics Forum. vol. 29, pp. 2530–2554. Wiley Online Library (2010)
  50. Wang, Y., Leng, Z., Li, F.W.B., Wu, S.C., Liang, X.: Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22035–22044 (October 2023)
  51. Xiang, W., Li, C., Zhou, Y., Wang, B., Zhang, L.: Language supervised training for skeleton-based action recognition. arXiv preprint arXiv:2208.05318 (2022)
  52. Xu, L., Song, Z., Wang, D., Su, J., Fang, Z., Ding, C., Gan, W., Yan, Y., Jin, X., Yang, X., Zeng, W., Wu, W.: Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2228–2238 (October 2023)
  53. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16010–16021 (October 2023)
  54. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. arXiv preprint arXiv:2301.06052 (2023)
  55. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiandiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022)
  56. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 364–373 (October 2023)
  57. Zhang, M., Li, H., Cai, Z., Ren, J., Yang, L., Liu, Z.: Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems* **36** (2024)
  58. Zhong, C., Hu, L., Zhang, Z., Xia, S.: Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 509–519 (October 2023)
  59. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5745–5753 (2019)
  60. Zhou, Z., Wan, Y., Wang, B.: Avatargpt: All-in-one framework for motion understanding, planning, generation and beyond. arXiv preprint arXiv:2311.16468 (2023)

61. Zhou, Z., Wang, B.: Ude: A unified driving engine for human motion generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5632–5641 (June 2023)