BRIDGING THE GAP BETWEEN HUMAN MOTION AND ACTION SEMANTICS VIA KINEMATIC PHRASES

Xinpeng Liu[1], **Yong-Lu Li**[1]*, **Ailing Zeng**[2], **Zizheng Zhou** [1], **Yang You**[1], **Cewu Lu**[1]*
[1]Shanghai Jiao Tong University, [2]International Digital Economy Academy (IDEA)
xinpengliu0907@gmail.com, yonglu_li@sjtu.edu.cn,
zengailing@idea.edu.cn, {zhou_zz,qq456cvb,lucewu}@sjtu.edu.cn

ABSTRACT

The goal of motion understanding is 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 (walk with arms up or swinging), while 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** characteristics. Based on KP as a mediator, we can unify a motion knowledge base and build a motion understanding system. Meanwhile, KP can be **automatically** converted from motions and to text descriptions with no subjective bias, inspiring Kinematic Prompt Generation (KPG) as a novel automatic motion generation benchmark. In extensive experiments, our approach shows superiority over other methods. Our code and data would be made publicly available here.

1 INTRODUCTION

Human motion understanding has a wide range of applications, including autonomous driving (Paden et al., 2016), robotics (Koppula & Saxena, 2013), and automatic animation (Van Welbergen et al., 2010), making it increasingly attractive. The core of human motion understanding 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 (Loper et al., 2015)/SMPL-X (Pavlakos et al., 2019) parameter sequence, while the action semantic space 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 (Petrovich et al., 2021) and text-based generation (Petrovich et al., 2022; Guo et al., 2022a; Lucas et al., 2022; Zhang et al., 2022; Tevet et al., 2022b; Chen et al., 2023; Zhang et al., 2023a). 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

*Yong-Lu Li and Cewu Lu are the corresponding authors.

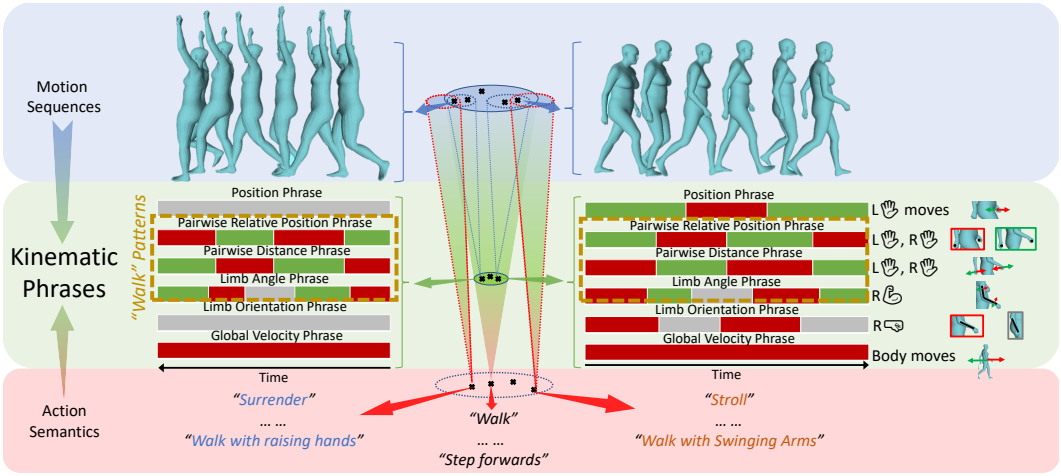

Figure 1: The huge gap between motion and action semantics results in the *many-to-many* problem. We propose Kinematic Phrases (KP) as an intermediate to bridge the gap. KPs objectively capture human kinematic cues. It properly abstracts diverse motions with interpretability. As shown, the Phrases in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

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 (von Laban & Lange, 1975; Bartlett, 1997), 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, we first construct a unified large-scale motion knowledge base. 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 in a self-supervised manner and then adopt it 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 reliable and efficient motion generation evaluation.

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 motion understanding system using KP and the aggregated large-scale knowledge base. (3) We propose KPG as a benchmark for reliable and efficient 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, also 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 (Zhou et al., 2019; Brégier, 2021). Recent works on parametric body models (Loper et al., 2015; Pavlakos et al., 2019) enable a more realistic body representation. Meanwhile, Pons-Moll et al. (2014) proposed Posebits, representing pose with boolean geometric part relationships. Delmas et al. (2022; 2023) translates Posebits into text descriptions. These abstract representations are flexible and insensitive to little perturbations, but their static nature ignores motion dynamics. Tang et al. (2022) acquire similar fine-grained descriptions from human annotation, while Xiang et al. (2022); Athanasiou et al. (2023) adopted large-scale language models. However, few recognize their potential in bridging the low-level motion and the high-level action semantics. Phase functions (Holden et al., 2020), Labanotations (von Laban & Lange, 1975), and learned Motion Words (Aristidou et al., 2018) were also explored, though limited to specific actions like locomotion and dancing.

Motion Generation can be conditioned by its prefix/suffix (Hernandez et al., 2019; Athanasiou et al., 2022; Guo et al., 2023), action categories (Petrovich et al., 2021; Guo et al., 2020; Xu et al., 2023), or audio (Li et al., 2021a;b). Text-based motion generation has developed rapidly with the proposal of text-motion datasets Punnakkal et al. (2021); Guo et al. (2022a). Petrovich et al. (2022); Guo et al. (2022a); Qian et al. (2023) used VAEs, while Tevet et al. (2022a); Hong et al. (2022); Lin et al. (2023b) extended the CLIP (Radford et al., 2021) space to motion. Recently, attention has been paid to diffusion models (Zhang et al., 2022; Tevet et al., 2022b; Dabral et al., 2023; Wang et al., 2023). Azadi et al. (2023) adopted a U-Net structure. Zhang et al. (2023b); Petrovich et al. (2023) explored retrieval-based methods. Karunratanakul et al. (2023) aimed at controllable generation, while Yuan et al. (2023) introduced physical constraints. However, most approaches still suffer from the gap between motion and action semantics. Lucas et al. (2022); Guo et al. (2022b); Zhang et al. (2023a); Chen et al. (2023); Zhou & Wang (2023); Zhong et al. (2023); Kong et al. (2023) 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 (von Laban & Lange, 1975) and qualitative static pose representations (Delmas et al., 2022; Pons-Moll et al., 2014), proposing six types of KP to comprehensively represent motion from different kinematic hierarchies: For **joint movements**, there are 36 Position Phrases (PPs). For **joint 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 33 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 (Pons-Moll et al., 2014; Delmas et al., 2022), we limit the criteria of KP as the indicator signs to minimize the need for human-defined standards (e.g., numerical criteria on the closeness of two joints) for objectivity and actuality. Fig. 2 illustrated the extraction 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 $r^f = r^r \times r^r$. These vectors of each frame are denoted as $R^{\cdot} = \{r_i^{\cdot}\}_{i=1}^{t}$.

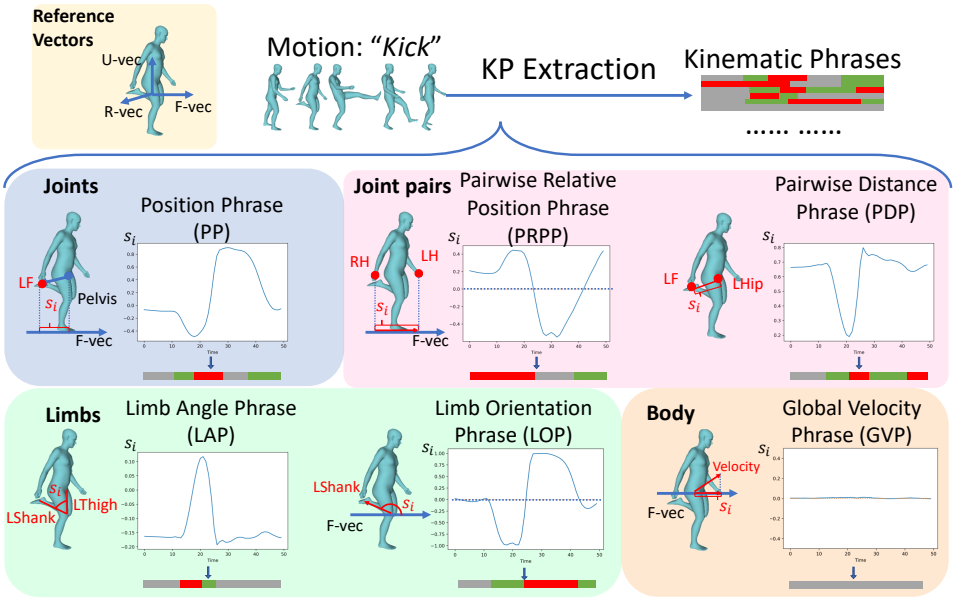

Figure 2: Six types of KP from four kinematic hierarchies are extracted from a motion sequence. A scalar indicator s_i is calculated per Phrase *per frame*. Its sign categorizes the corresponding Phrase.

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

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

The sign of $s_i^{(j,\cdot)}$ categorizes PP into moving `along/against` $R^{\cdot}$, or `relatively static` along $R^{\cdot}$ for indicators with small amplitudes. After filtering, 36 different PPs are extracted.

Pairwise Relative Position Phrase (PRPP) describes the relative position between a pair of joints (x^j, x^k) w.r.t. reference vector $R^{\cdot}$. PRPP indicator at i-th frame is $s_i^{(j,k,\cdot)} = \langle (x_i^j - x_i^k), r_i^{\cdot} \rangle$. 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 $s_i^{(j,k,\cdot)}$. After filtering, 242 PRPPs are extracted.

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

$$s_i^{(j,k)} = \|x_i^j - x_i^k\|_2 - \|x_{i-1}^j - x_{i-1}^k\|_2. \tag{2}$$

The sign of $s_i^{(j,k)}$ categorizes PDP into moving `closer/away`, or `relatively static`. After dropping joint pairs in the skeleton topology, such as the hand and elbow, 81 PDPs are extracted.

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 indicator for LAP is calculated as

$$s_i^{(j,k,l)} = arccos(\langle x_i^k - x_i^j, x_i^l - x_i^j \rangle) - arccos(\langle x_{i-1}^k - x_{i-1}^j, x_{i-1}^l - x_{i-1}^j \rangle). \tag{3}$$

LAP describes the limb chain (x^j, x^k)-(x^j, x^l) as `bending` or `unbending`. 8 LAPs are extracted.

Limb Orientation Phrase (LOP) describes the orientation of the limb (x^j, x^k) w.r.t. $R^{\cdot}$, note that x^k is the distal limb. The scalar indicator for LOP is calculated as $s_i^{(j,k,\cdot)} = \langle x_i^k - x_i^j, r_i^{\cdot} \rangle$. The sign of $s_i^{(j,k,\cdot)}$ categorizes the LOP into limb (x^j, x^k) pointing `along/against` $R^{\cdot}$, or a placeholder category for those with little magnitude. 33 LOPs are extracted.

Global Velocity Phrase (GVP) describes the direction of global velocity with respect to $R^{\cdot}$. The indicator is calculated as $s_i^{\cdot} = \langle x_{i+1}^0 - x_i^0, r_i^{\cdot} \rangle$. The three categories are moving `along/against` $R^{\cdot}$, or `static along` $R^{\cdot}$ according to the sign of $s_i^{\cdot}$.

These result in 403 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-scale knowledge base containing motion, text, and KP. Motion sequences of different representations are collected, including 3D skeleton sequences and SMPL (Loper et al., 2015)/SMPL-X (Pavlakos et al., 2019) 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 (Hassan et al., 2021). Finally, we collect 87k motion sequences from 11 datasets. Detailed statistics are shown in Tab. 1. More details are included in the appendix.

Dataset	Mot. Rep.	#Seqs	#Actions	Text
AMASS (Mahmood et al., 2019)	SMPL-X	26k	260	✓
GRAB (Taheri et al., 2020)	SMPL-X	1k	4	✓
SAMP (Hassan et al., 2021)	SMPL-X	0.2k	N/A	✓*
Fit3D (Fieraru et al., 2021)	SMPL-X	0.4k	29	✓
CHI3D (Fieraru et al., 2020)	SMPL-X	0.4k	8	✓
UESTC (Ji et al., 2018)	SMPL	26k	40	✓
AIST++ (Li et al., 2021a)	SMPL	1k	N/A	✓*
BEHAVE (Bhatnagar et al., 2022)	SMPL	0.3k	N/A	✓*
HuMMan (Cai et al., 2022)	SMPL	0.3k	339	✓
GTAHuman (Cai et al., 2021)	SMPL	20k	N/A	x
Motion-X(Lin et al., 2023a)	SMPL-X	65k	N/A	✓
Sum	-	**140k**	**680+**	-

Table 1: Statistics of Kinematic Phrase Base. *Mot. Rep.* indicates motion representation. "✓*" means texts are generated from the attached additional information instead of human annotation.

4 MOTION UNDERSTANDING VIA KP

By motion understanding, we mean both low-level understanding like interpolation and modification, and high-level understanding like generative mapping from text to motion. To achieve this, we first learn a motion-KP joint space with less ambiguity and more interpretability. Then, with this space, we introduce its application to both low-level and high-level motion-semantics understanding.

4.1 PRELIMINARIES

We first introduce the representation for motion and KP. **Motion** is represented as a human pose sequence with n frames as $M = \{m_i\}_{i=1}^{n}$. In detail, SMPL (Loper et al., 2015) pose parameters are transformed from axis-angle format to the 6D continuous representation (Zhou et al., 2019), then concatenated with the velocity of the root joint, resulting in a 147-dimensional representation per frame. **KP** is represented by signs of the indicators.

4.2 JOINT SPACE LEARNING

Model Structure. An overview of our model is illustrated in Fig. 3. **Motion VAE** is a transformer-based VAE adapted from Petrovich et al. (2021). The encoder $\mathcal{E}_m$ takes motion M and two distribution tokens m_μ, m_σ as input, and the outputs corresponding to the distribution tokens are taken as the μ_m and σ_m of the Gaussian distribution. Then, the transformer 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 obtain the reconstructed motion sequence $\hat{M}$. **KP VAE** with encoder $\mathcal{E}_p$ and decoder $\mathcal{D}_p$ resembles Motion VAE. The sign of $\mathcal{D}_p$ output is adopted as the predicted KP $\hat{C}$. Notice that the decoders $\mathcal{D}_m, \mathcal{D}_p$ could take arbitrary combinations of z_m, z_p as input, outputting $\hat{M}_\cdot, \hat{C}_\cdot$.

Self-supervised Training. With the VAEs, we propose a self-supervised training strategy to learn motion-KP joint space. As a coherent representation, the overall representation should not change drastically with a small portion of KP unknown. Even more, the missing Phrases should be recovered from existing Phrases. In this view, we randomly corrupt samples during training by setting a small portion of KP as 0. The training is thus executed in a self-supervised manner. This helps mine the correlation among different Phrases while also effectively increasing the robustness of the joint

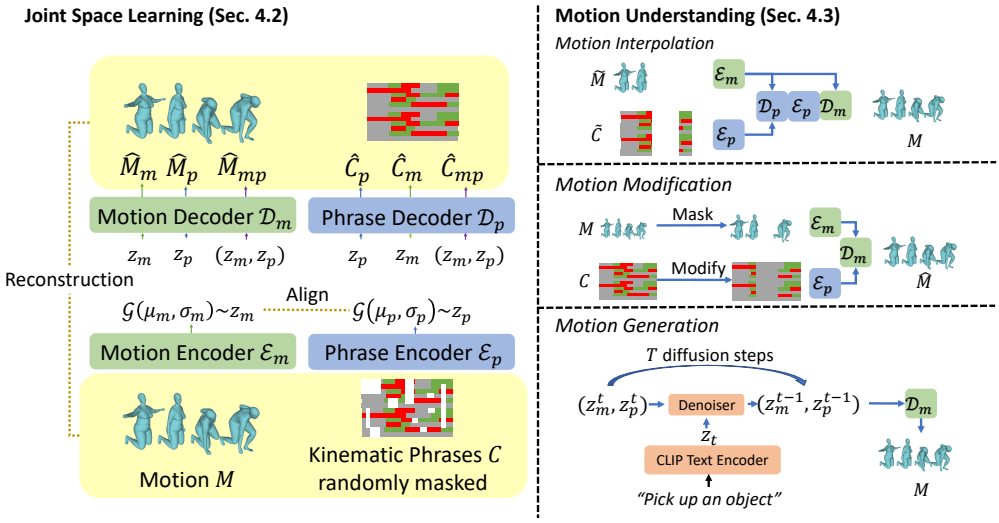

Figure 3: We train motion-KP joint latent space in a self-supervised training manner. KP is randomly masked during training. Reconstruction and alignment losses are adopted. The joint space could be applied for multiple tasks, including motion interpolation, modification, and generation.

space. Similar to TEMOS (Petrovich et al., 2022), four losses are adopted: reconstruction loss, KL divergence loss, distribution alignment loss, and embedding alignment loss.

4.3 KP-mediated Motion Understanding

With the joint space, we can perform both low-level and high-level motion understanding with KP mediation. We introduce three applications to show the capability of KP, as shown in Fig. 3.

KP-mediated 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 $\hat{M}$.

Motion Modification Motion modification functions similarly. Motion M is first extracted into KP sequence C. Modifications could be made on C resulting in $\tilde{C}$. Modified motion frames 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 $\hat{M}$.

KP-mediated Motion Generation. Given text t, to generate a motion sequence from it, we first encode it into latent z_t with CLIP text encoder $\mathcal{E}_t$. Direct mapping could be achieved by training the motion decoder $\mathcal{D}_m$ for $\hat{M} = \mathcal{D}_m(z_t)$. We show that the direct mapping could be impressively improved with our joint space in Sec. 6.4. With KP, we could perform a novel KP-mediated motion generation. We adopt a vanilla latent diffusion paradigm for KP-mediated text-to-motion tasks. An extra denoiser is trained to denoise a random noise z_p^T to KP latent $z_p = z_p^0$ with T diffusion steps. We then decode KP sequence $\hat{C}$ from z_p with $\mathcal{D}_p$. Then, $\hat{C}$ is encoded by $\mathcal{E}_p$, getting distribution $\mathcal{G}(\mu_p, \sigma_p)$. z_p is sampled and sent to $\mathcal{D}_m$ to generate a motion sequence. Experiments show that KP could be a promising stepping stone to mitigate the huge gap from action semantics to motion.

5 Kinematic Prompt Generation

With the interpretability and objectivity of KP, we propose a new motion generation benchmark.

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 (Frechet 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. Guo et al. (2022a) 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. Moreover, since automatic mapping from motion to semantics across their huge gap 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 (Guo et al., 2022a; Plappert et al., 2016; Ji et al., 2018) with considerable complex everyday actions, further increasing the difficulty.

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

In detail, we convert KP into text prompts with templates as in Tab. 2, resulting in 840 text prompts. Given prompt $T_i \in T$ from Phrase c_i, the model generates motion $\hat{M}_i$, along with extracted KP $\hat{C}_i$. We calculate Accuracy as $Acc = \frac{1}{|T|}\sum_{T_i \in T} 1[c_i \in \hat{C}_i]$, where $1[\cdot] = 1$ if the expression in $[\cdot]$ is True, otherwise 0. Note that, for $c_i \in \hat{C}_i$, c_i should keep for more than 5 consecutive frames to avoid trivial perturbations. Accuracy examines whether the Phrase corresponding to the given prompt appears in the KP sequence converted from generated motion. The calculation involves no black-box model thanks to KP, presenting a fully reliable evaluation pipeline. Also, with the effortless motion-to-KP conversion, the computation could be conducted automatically. More details are in the appendix.

KP	Text prompt samples
PP	**Left hand** moves `forwards`.
PRPP	**Left hand** is `below` **head** then `above` **head**.
PDP	**Left hand** moves `away from` **head**.
LAP	**Left arm** `bends`.
LOP	**Left forearm** points `forwards` then `backward`.
GVP	The person moves `forwards`.

Table 2: Text prompts converted from KP. **Joint/limb names**, *prepositions, verbs, and adverbials* could be replaced w.r.t. specific Phrases.

6 Experiment

Implementation Details. HumanML3D (Guo et al., 2022a) test split is held out for evaluation, with the rest of KPB for training. During training, the motion sequences are canonicalized by eliminating the rotation along the z-axis in the first frame, and the same counter-rotation is applied to the following frames. Sequences are sampled to 15 FPS and randomly clipped into short clips with lengths between 30 frames and 150 frames. 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 provided in the appendix.

6.1 Motion Interpolation

Following Jiang et al. (2023), 50% frames are randomly masked for interpolation evaluation. FID and Diversity are also evaluated. We adopt MDM (Tevet et al., 2022b) as the baseline. In Tab. 3, our method provides better FID. While with additional KPB, the Diversity is increased.

6.2 Motion Generation

Settings. We adopt the HumanML3D test set (Guo et al., 2022a) for conventional text-to-motion evaluation. The evaluation model from Guo et al. (2022a) is adopted to calculate R-Precision, FID, Diversity, and Multimodality. KPG is also adopted, with the proposed Accuracy. Also, Diversity is computed as a reference. We run the evaluation 20 times and report the average metric value. Details are given in the appendix.

Results on conventional text to motion are shown in Tab. 3. Our method is competitive without KPB. However, KPB brings a counter-intuitive performance drop. To evaluate this, we further conduct a user study to make human volunteers judge the motions instead of a proxy neural network.

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 (Tevet et al., 2022b; Chen et al., 2023)), we randomly

Methods	Motion Interpolation FID↓	Diversity→	Motion Generation R-P@1↑	FID↓	Diversity→	Multimodality
GT	0.002	9.503	0.511	0.002	9.503	-
TEMOS (Petrovich et al., 2022)	-	-	0.424	3.734	8.973	0.368
T2M (Guo et al., 2022a)*	-	-	0.455	1.067	9.188	2.090
MDM (Tevet et al., 2022b)*	2.698	8.42	0.320	0.544	9.559	2.799
TM2T (Guo et al., 2022b)*	-	-	0.424	1.501	8.589	2.424
MLD (Chen et al., 2023)*	-	-	0.481	0.473	9.724	2.413
T2M-GPT (Zhang et al., 2023a)*	-	-	**0.492**	**0.141**	9.722	1.831
MotionGPT (Jiang et al., 2023)*	0.214	**9.560**	**0.492**	0.232	**9.528**	2.008
Ours*	**0.197**	9.772	0.475	0.412	10.161	2.065
Ours	0.226	10.022	0.434	0.631	10.372	2.584

Table 3: Result Comparison of motion interpolation and generation on HumanML3D. R-P@1 is short for R-Precision@1. * indicates the model is trained on the HumanML3D train set only.

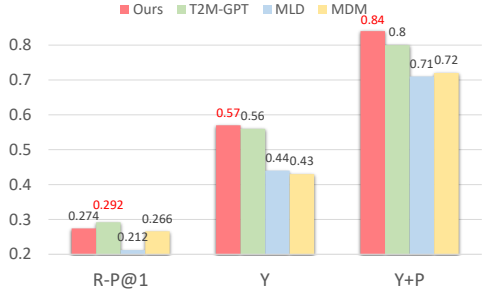

Figure 4: User study on HumanML3D, with "Y" for Yes and "P" for partially.

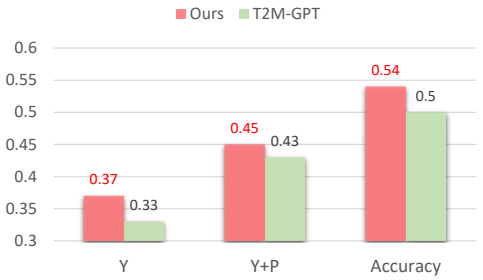

Figure 5: User study on KPG, with "Y" for Yes and "P" for partially.

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, neither asking the volunteers to give a general rating for each sample nor 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 decouple the evaluation of text-to-motion into 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 for reference. MDM (Tevet et al., 2022b), T2M-GPT (Zhang et al., 2023a), MLD (Chen et al., 2023), and our method are evaluated.

User study results are shown in Fig. 4. Though our method is not superior in R-Precision, we receive better user reviews, showcasing the efficacy of our KP-mediated generation strategy. Recent 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. More results and analysis are given in the appendix.

Results on KPG are shown in Tab. 4. KPG is considered an easier task than conventional text-based motion generation since it is targeted at action semantics with much less complexity. However, previous methods are not performing as well as expected. Though we managed to deliver substantial improvements, the accuracy remains below 60%, which is far from satisfying. There is a considerable gap between existing methods and ideal motion generation models.

Furthermore, given the discrepancy between automatic metrics and user study as shown in Fig. 4, we conducted a similar user study with 100 randomly selected prompts from KPG involving T2M-GPT and our model. Fig. 5 demonstrates that KP-inferred Accuracy and user reviews share similar trends. We also calculate their consistency, showing KP and user study give the same reviews for **84%** of the samples. We believe KPG could thus be a first step towards reliable automatic motion generation evaluation. More analyses are given in the appendix.

6.3 VISUALIZATION

We first present a modification sample in Fig. 6. By modifying KP, we could edit arbitrary motion at a fine-grained level. Also, We compare generated samples of T2M-GPT and our methods in Fig. 7.

Methods	Acc.%↑	Diversity
HMDM (Tevet et al., 2022b)	44.40	5.725
MLD (Chen et al., 2023)	44.76	5.901
T2M-GPT (Zhang et al., 2023a)	47.86	6.593
Ours	**52.14**	6.017

Table 4: Results on Kinematic Prompt Generation.

Methods	Acc.%↑	Diversity
Ours	**52.14**	6.017
w/o KP mediation	50.43	5.616
Direct mapping	42.28	5.379
w/o Joint KP	51.03	5.765
w/o Joint Pair KP	48.24	5.596
w/o Limb KP	51.92	5.804
w/o Body KP	52.44	5.903

Table 5: Ablation results on KPG.

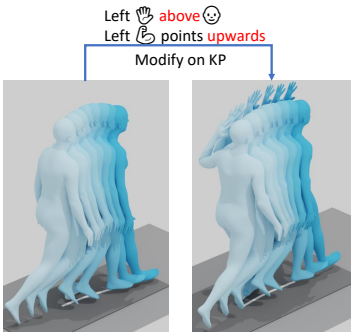

Figure 6: Our model supports fine-grained modification on motion via modification on KP.

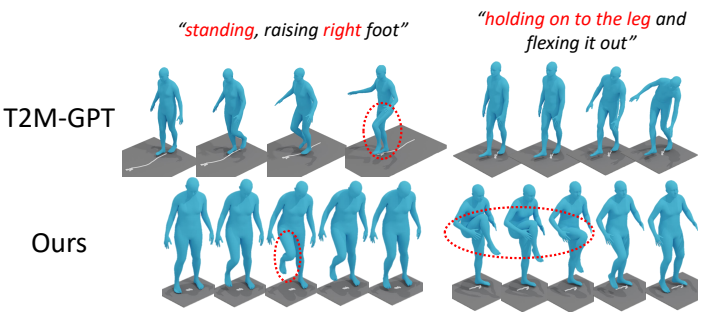

Figure 7: Visualization of generated samples. Compared to T2M-GPT, our method provides a better response to prompts with explicit constraints on specific body parts.

Our method properly responds to text prompts with constraints on specific body parts. This could be attributed to KP mediation, which explicitly decomposes the action semantics into kinematics cues of body parts. Note that T2M-GPT might generate redundant motion for simple prompts, while our method provides more concise and precise results. More visualizations are in the appendix.

6.4 Ablation Studies

Ablation study results on KPG are shown in Tab. 5.

KP mediation. By using our joint space without KP mediation, we still present a competitive result, showing the efficacy of motion-KP joint space.

Direct mapping. By directly mapping with no KP involved, we present a similar performance compared to previous methods. It demonstrates 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 style 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.

7 Discussion

Here, we discuss the limitations and prospects of KP and KP-based applications. **First**, KP could be extended beyond its current criteria of sign. These criteria guarantee objectivity but overlook important kinematic information like movement amplitude and speed. Also, due to the granularity of the adopted skeleton, fine-grained kinematic information on fingers is not well-preserved. The exploration of amplitude/speed/finger-based KP would be a promising goal to pursue. **Second**, KPB could be extended to datasets with other modalities, like 2D pose and egocentric action datasets. Though these modalities provide incomplete 3D information, we could extract KP that is credibly accessible across modalities. **Third**, with the convenient conversion from KP to text, auxiliary text descriptions could be automatically generated for motions via KP. **Fourth**, KPG could be extended by paraphrasing existing prompts and combining different Phrases.

8 Conclusion

In this paper, we proposed an intermediate representation to bridge human motion and action semantics as the Kinematic Phrase. By focusing on objective kinematic facts of human motion, KP

achieved proper abstraction, interpretability, and generality. A motion understanding system based on KP was proposed and proven effective in motion interpolation, modification, and generation. Moreover, a novel motion generation benchmark Kinematic Prompt Generation is proposed. We believe that KP has great potential for advancing motion understanding.

REFERENCES

Andreas Aristidou, Daniel Cohen-Or, Jessica K. Hodgins, Yiorgos Chrysanthou, and Ariel Shamir. Deep motifs and motion signatures. *ACM Trans. Graph.*, 37(6):187:1–187:13, November 2018. doi: 10.1145/3272127.3275038. URL http://doi.acm.org/10.1145/3272127.3275038.

Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*, pp. 414–423. IEEE, 2022.

Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. SINC: Spatial composition of 3D human motions for simultaneous action generation. In *ICCV*, 2023.

Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, and Sonal Gupta. 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.

R. Bartlett. *Introduction to Sports Biomechanics*. Introduction to Sports Biomechanics. E & FN Spon, 1997. ISBN 9780419208402. URL https://books.google.com.tw/books?id=-6Db8mgxsqQC.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022.

Romain Brégier. Deep regression on manifolds: a 3d rotation case study. In *2021 International Conference on 3D Vision (3DV)*, pp. 166–174. IEEE, 2021.

Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, and Ziwei Liu. Playing for 3d human recovery. *arXiv preprint arXiv:2110.07588*, 2021.

Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 557–577, Cham, 2022. Springer Nature Switzerland.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. 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.

Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. 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.

Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. 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.

Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. Posefix: Correcting 3d human poses with natural language. *arXiv preprint arXiv:2309.08480*, 2023.

Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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

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 *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2021–2029, 2020.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5152–5161, June 2022a.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 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, 2022b.

Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. 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.

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *Proceedings of the International Conference on Computer Vision 2021*, October 2021.

Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7134–7143, 2019.

Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. Learned motion matching. *ACM Transactions on Graphics (TOG)*, 39(4):53–1, 2020.

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022.

Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. 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.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795*, 2023.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. 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.

Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, and Xinchao Wang. 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.

Hema Swetha Koppula and Ashutosh Saxena. Anticipating human activities for reactive robotic response. In *IROS*, pp. 2071. Tokyo, 2013.

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

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13401–13412, 2021b.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *arXiv preprint arXiv:2307.00818*, 2023a.

Junfan Lin, Jianlong Chang, Lingbo Liu, Guanbin Li, Liang Lin, Qi Tian, and Chang-Wen Chen. 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 2023b.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. 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.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.

Brian Paden, Michal Čáp, Sze Zheng Yong, Dmitry Yershov, and Emilio Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on intelligent vehicles*, 1(1):33–55, 2016.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 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.

Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10985–10995, 2021.

Mathis Petrovich, Michael J Black, and Gül Varol. 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.

Mathis Petrovich, Michael J. Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9488–9497, October 2023.

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

Gerard Pons-Moll, David J Fleet, and Bodo Rosenhahn. Posebits for monocular human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2337–2344, 2014.

Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. 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.

Yijun Qian, Jack Urbanek, Alexander G. Hauptmann, and Jungdam Won. 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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. URL https://grab.is.tue.mpg.de.

Yansong Tang, Jinpeng Liu, Aoyang Liu, Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, and Xiu Li. Flag3d: A 3d fitness activity dataset with language instruction. *arXiv preprint arXiv:2212.04638*, 2022.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 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, 2022a.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022b.

Herwin Van Welbergen, Ben JH Van Basten, Arjan Egges, Zs M Ruttkay, and Mark H Overmars. Real time animation of virtual humans: a trade-off between naturalness and control. In *Computer Graphics Forum*, volume 29, pp. 2530–2554. Wiley Online Library, 2010.

R. von Laban and R. Lange. *Laban's Principles of Dance and Movement Notation*. Macdonald & Evans, 1975. ISBN 9780712116480. URL https://books.google.com.tw/books?id=-Vr0AAAAMAAJ.

Yin Wang, Zhiying Leng, Frederick W. B. Li, Shun-Cheng Wu, and Xiaohui Liang. 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.

Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Language supervised training for skeleton-based action recognition. *arXiv preprint arXiv:2208.05318*, 2022.

Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, Wenjun Zeng, and Wei Wu. 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.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16010–16021, October 2023.

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. *arXiv preprint arXiv:2301.06052*, 2023a.

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.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 364–373, October 2023b.

Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. 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.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 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.

Zixiang Zhou and Baoyuan Wang. 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.

APPENDIX

A OVERVIEW

This supplementary material presents more details and additional results not included in the main paper due to page limitation. The list of items included are:

- More details about *kinematic phrase* in Sec. B.
- More details about *kinematic phrase base* in Sec. C.
- Our method details in Sec. D.
- Additional experimental details in Sec. E.

B KINEMATIC PHRASE DETAILS

This section lists the details of the six defined types of KP. During extraction, the indicator is set as zero if it is smaller than 1e-4.

B.1 POSITION PHRASE

There are 36 phrases, corresponding to 36 interested $\langle joint,\ reference\ vector \rangle$ pairs like $\langle left\ hand,\ forward\ vector \rangle$. Combinations whose joint and reference vector are directly correlated are filtered out, like $\langle lefthip, forward \rangle$.

B.2 PAIRWISE RELATIVE POSITION PHRASE

There are 242 phrases corresponding to 242 interested $\langle joint,\ joint, reference\ vector \rangle$ triplets like $\langle left\ hand,\ right\ hand, forward\ vector \rangle$.

B.3 PAIRWISE DISTANCE PHRASE

Joint pairs that are connected by human body topology are filtered out, like hand-elbow and shoulder-hip. There are 81 phrases corresponding to 81 interested $\langle joint,\ joint \rangle$ pairs like $\langle left\ hand,\ right\ hand \rangle$.

B.4 LIMB ANGLE PHRASE

There are 8 phrases corresponding to 8 interested limbs, listed in the file `KP/lap.txt`.

B.5 LIMB ORIENTATION PHRASE

There are 33 phrases corresponding to 33 interested $\langle limb,\ reference\ vector \rangle$ pairs like $\langle left\ shank,\ right\ vector \rangle$.

B.6 GLOBAL VELOCITY PHRASE

There are 3 phrases corresponding to the velocity direction with respect to the three reference vectors.

C KINEMATIC PHRASE BASE DETAILS

Over **140 K** motion sequences are collected to construct the Kinematic Phrase Base, including **9 M** frames (in 30 FPS) with **48 K** different sentences, covering a vocabulary size of **7,418**. Here, we illustrate the distribution of the collected database represented in motion, KP, and text in Fig. 8. Besides, a word cloud visualization of the texts in the database is illustrated in Fig. 9.

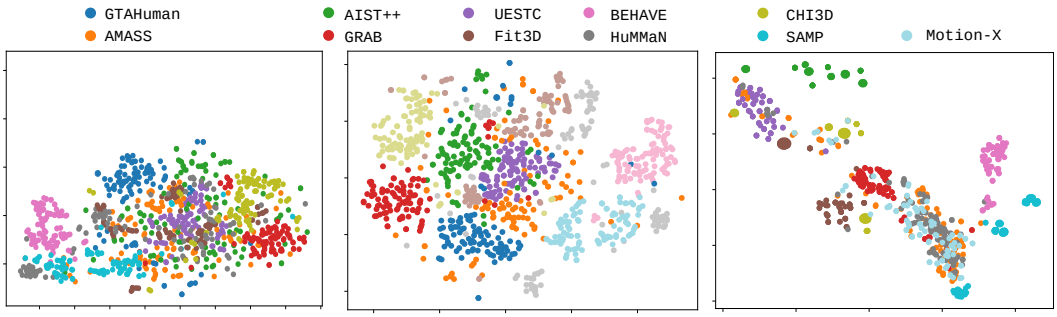

Figure 8: Motion, KP, and text distribution of Kinematic Phrase Base.

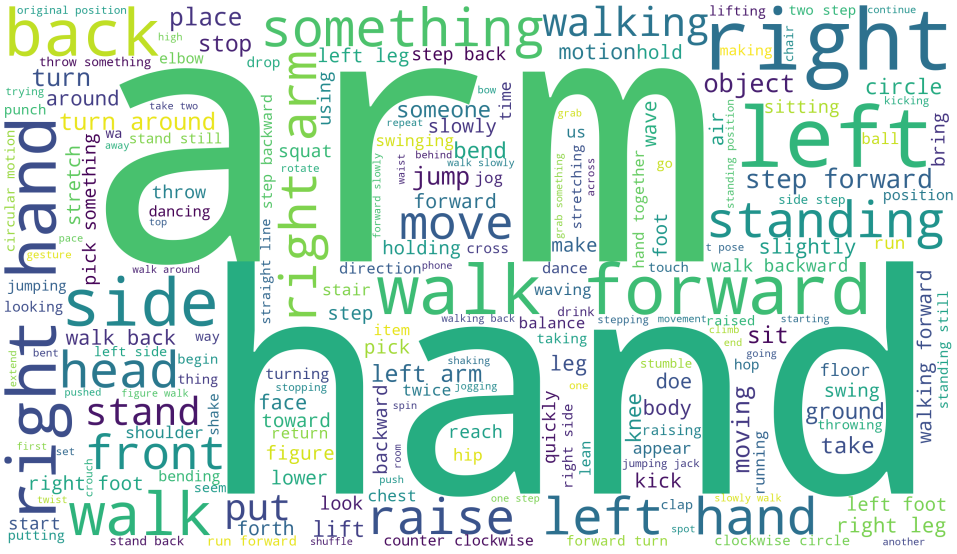

Figure 9: Word cloud visualization of the texts in Kinematic Phrase Base.

D METHOD DETAILS

D.1 LOSSES FOR JOINT SPACE LEARNING

Reconstruction loss $\mathcal{L}_{rec}$ compares the GT with the outputs of the VAEs. L1 losses are calculated for the motion representation $M, \hat{M}$, KP $C, \hat{C}$, the skeleton joints $J, \hat{J}$, the down-sampled mesh vertices $V, \hat{V}$, and the joint accelerations $A, \hat{A}$.

$$\mathcal{L}_{rec} = \sum_{\cdot \in \{m,p,mp\}} ||M_\cdot - \hat{M}_\cdot||_1 + ||C_\cdot - \hat{C}_\cdot||_1 + ||J_\cdot - \hat{J}_\cdot||_1 + ||V_\cdot - \hat{V}_\cdot||_1 + ||A_\cdot - \hat{A}_\cdot||_1. \quad (4)$$

KL divergence loss $\mathcal{L}_{KL}$ encourages each distribution to be similar to a normal distribution $\pi = \mathcal{G}(0, I)$ by minimizing the Kullback-Leibler(KL) divergence between the normal distribution and the learned motion and KP distributions. The loss is calculated as

$$\mathcal{L}_{KL} = KL(\phi_m, \pi) + KL(\phi_p, \pi). \quad (5)$$

Distribution alignment loss $\mathcal{L}_{da}$ encourages the distributions of motion and KP to resemble each other by minimizing the KL divergence between them. The loss is calculated as

$$\mathcal{L}_{da} = KL(\phi_m, \phi_p) + KL(\phi_p, \phi_m). \quad (6)$$

Embedding alignment loss $\mathcal{L}_{emb}$ encourages the sampled latent vectors to be aligned by minimizing their L1 distance. The loss is calculated as

$$\mathcal{L}_{emb} = ||z_m - z_p||_1. \quad (7)$$

The overall loss is calculated as

$$\mathcal{L} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{KL} + \lambda_3 \mathcal{L}_{da} + \lambda_4 \mathcal{L}_{emb}, \quad (8)$$

where $\{\lambda_i\}_{i=1}^4$ are weighting coefficients.

E EXPERIMENT DETAILS

E.1 IMPLEMENTATION DETAILS

The Motion VAE and KP VAE share the same structure: a 4-layer transformer encoder, a 4-layer transformer decoder, and a fully connected layer for final outputs. The denoiser adopted for text-to-motion is designed as a 4-layer transformer decoder. The latent size is set to 256. $\{\lambda_i\}_{i=1}^4$ are set as 1. The learning rate is decayed at 4,000 epochs for joint space training and at 2,000 epochs for text-to-motion latent diffusion model training.

E.2 MOTION GENERATION SETTINGS

For HumanML3D (Guo et al., 2022a), motion sequences are generated for 10 seconds given a text prompt. For KPG, the models are required to generate 120 frames given a text prompt.

R-Precision is calculated in a similar way to Guo et al. (2022a). For each generated motion, its text description is mixed with 31 randomly selected mismatched descriptions from the test set. The cosine distances between the motion feature and text features are computed. The average accuracy at the top-1 place is reported.

FID is adopted to measure the divergence between the GT motion distribution and the generated motion distribution in the latent space.

Diversity measures the variance of the generated motion sequences. It is calculated as the average latent distance between two randomly sampled generated motion sets. The set size is set as 300 in this paper.

Multimodality measures the variance of the generated motion sequences within each text prompt. For each description, two subsets of motion sequences with the same size are generated, and then the Multimodality is calculated as the average distance between the two sets of motions in the latent space. The size of each subset is set as 10 in this paper.

FID = 0.544 R-P@1 = 0.266		Semantic consistency Yes	 Partially	 No	Sum
Naturalness	Yes	0.40	0.18	0.10	0.68
	No	0.03	0.11	0.18	0.32
Sum		0.43	0.29	0.28	1

(a) MDM (Tevet et al., 2022b).

FID = 0.212 R-P@1 = 0.473		Semantic consistency Yes	 Partially	 No	Sum
Naturalness	Yes	0.34	0.13	0.04	0.51
	No	0.10	0.14	0.25	0.49
Sum		0.44	0.27	0.29	1

(b) MLD (Chen et al., 2023).

FID = 0.141 R-P@1 = 0.292		Semantic consistency Yes	 Partially	 No	Sum
Naturalness	Yes	0.50	0.16	0.05	0.71
	No	0.06	0.08	0.15	0.29
Sum		0.56	0.24	0.20	1

(c) T2M-GPT (Zhang et al., 2023a).

FID = 0.631 R-P@1 = 0.274		Semantic consistency Yes	 Partially	 No	Sum
Naturalness	Yes	0.52	0.21	0.02	**0.75**
	No	0.05	0.06	0.14	0.25
Sum		**0.57**	**0.27**	0.16	1

(d) Ours.

Table 6: Detailed user study results on HumanML3D.

Accuracy = 50%		Semantic consistency Yes	 Partially	 No	Sum
Naturalness	Yes	0.29	0.09	0.53	0.91
	No	0.04	0.01	0.04	0.09
Sum		0.33	**0.10**	0.57	1

(a) T2M-GPT Zhang et al. (2023a).

Accuracy = **54%**		Semantic consistency Yes	 Partially	 No	Sum
Naturalness	Yes	0.33	0.07	0.51	**0.92**
	No	0.04	0.01	0.04	0.08
Sum		**0.37**	0.08	0.55	1

(b) Ours.

Table 7: Detailed user study results on KPG.

E.3 User Study Details

E.3.1 User Study Design

As stated in the main text, we adopt a direct Q&A-style user study instead of a popular preference test or ratings. Here we clarify the reason for this design choice. First, this design is more suitable in evaluating **semantic consistency**, which we identify as categorical instead of continuous at the sample level. That is, it is hard to tell whether a motion is more `raising left-hand up` than another. Instead, there is only whether a motion is `raising left-hand up` or not. Therefore, we chose to present a direct question on semantic consistency. Second, this design explicitly decouples the evaluation of text-to-motion into semantic consistency and naturalness, corresponding to R-Precision and FID. When rating motions or choosing between two motions, it is hard to guarantee the users make choices according to the expected standard. Therefore, we explicitly ask decoupled binary questions for decomposition. Third, it helps reduce annotation costs. For preference testing, the complexity is $O(N^2)$, while with our user-study protocol, the complexity is only $O(N)$. In consideration of our primary focus on semantic consistency, we adopt this protocol. We also admit this protocol is sub-optimal in naturalness evaluation, which is a continuous factor. We present the results on naturalness as a reference in the following sections.

E.3.2 User Study on Conventional Text-to-Motion

Detailed user study results on HumanML3D are demonstrated in Tab. 6. As shown, both FID and R-P@1 are not totally consistent with the user reviews, indicating these black-box-based metrics might be sub-optimal for motion generation evaluation. Meanwhile, the four evaluated methods present a similar positive correlation between semantic consistency and naturalness. Moreover, it shows that generating natural motions is a little harder than generating partially semantic-consistent motions, which might be a potential direction to advance motion generation.

E.3.3 User Study on KPG

Detailed user study results on KPG are demonstrated in Tab. 7. Our proposed Accuracy shares a similar trend with user-reviewed semantic consistency between the two methods. Both methods receive good naturalness reviews, which could result from the simple prompt structure of KPG.

		User Reviewed			Sum
		Yes	Partially	No	
KP-Inferred	Yes	0.32	0.08	0.12	0.52
	No	0.03	0.01	0.44	0.48
Sum		0.35	0.09	0.56	1

Table 8: Detailed consistency statistics between KP-inferred Accuracy and user-reviewed semantic consistency.

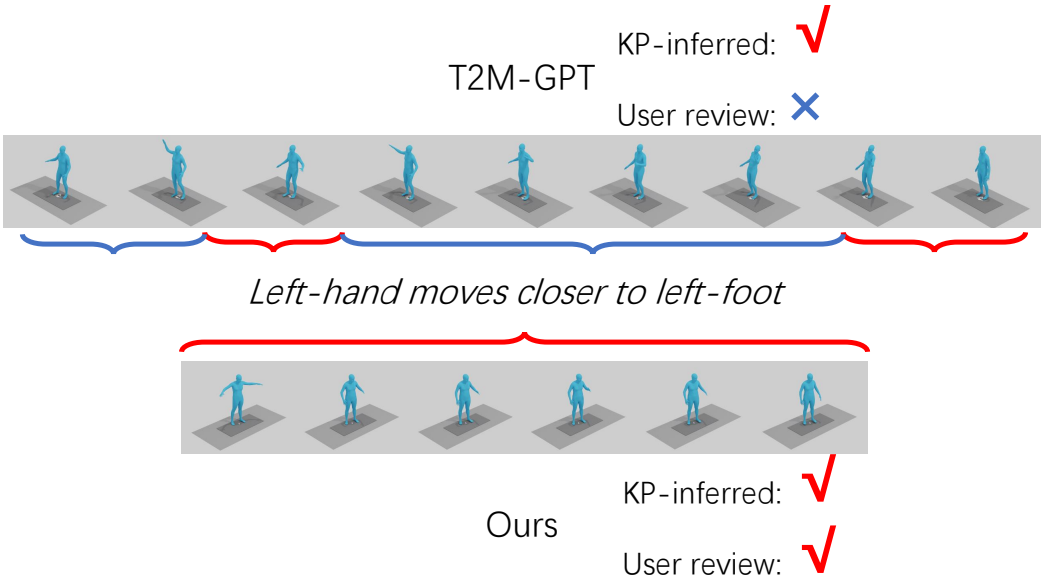

Figure 10: For KPG, we generate more concise motion than T2M-GPT (Zhang et al., 2023a).

Furthermore, we provide detailed consistency statistics between KP-inferred Accuracy and user-reviewed semantic consistency in Tab. 8. Samples generated from T2M-GPT and our method are included. KP and users provide similar reviews for over 80% of the samples, showing good consistency. With respect to user reviews, KP-inferred Accuracy has a higher false positive rate (0.12 / 0.52 = 0.2308) than a false negative rate (0.04 / 0.48 = 0.0833). We find there are two typical false positive scenarios. First, the generated motion results in rather small indicators, close to the 1e-4 threshold. KP captures this, however, it is hard for humans to notice such subtle movements. Second, as shown in Fig. 10, the generated motions sometimes tend to be redundant compared to the given prompts. Users might be distracted, overlooking the targeted semantics. We find this happens more for T2M-GPT generated samples (in Fig. 7, extra walking motion; in Fig. 10, extra right-hand waving motion), while our method manages to provide more concise responses. We think this could partially explain the higher Diversity of T2M-GPT in Tab. 4. For the first scenario, we think an adaptive threshold w.r.t. the overall motion intensity would be helpful, since to human perception, the relative amplitude is usually more important than the absolute amplitude. Also, as stated in Sec. 7, extending KP to amplitude might also help. The second scenario urges us to rethink the current text-to-motion task setting. For a "matched" motion-text pair, should the text semantics be a subset of motion semantics, or strictly match? Also, is it expected to increase diversity by introducing redundant motions? We identify these questions as interesting points of attack and leave them for future exploration.