Supplementary Materials for "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[3], and Cewu Lu[1⋆]

[1] Shanghai Jiao Tong University
[2] Tencent
[3] Stanford University
{xinpengliu0907,ailingzengzzz}@gmail.com,
{yonglu_li,zhou_zz,lucewu}@sjtu.edu.cn, yangyou@stanford.edu

1 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.

1.1 Position Phrase

There are 34 phrases, corresponding to 34 interested $\langle joint,\ reference\ vector\rangle$ pairs like $\langle left\ hand,\ forward\ vector\rangle$. The pairs are listed in the file `KP/pp.txt`.

1.2 Pairwise Relative Position Phrase

There are 242 phrases corresponding to 242 interested $\langle joint,\ joint, reference\ vector\rangle$ triplets like $\langle left\ hand,\ right\ hand, forward\ vector\rangle$, listed in the file `KP/prpp.txt`.

1.3 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 $\langle left\ hand,\ right\ hand\rangle$, listed in the file `KP/pdp.txt`.

1.4 Limb Angle Phrase

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

1.5 Limb Orientation Phrase

There are 24 phrases corresponding to 24 interested $\langle limb,\ reference\ vector\rangle$ pairs like $\langle left\ shank,\ right\ vector\rangle$, listed in the file `KP/lop.txt`.

⋆ Corresponding authors.

Dataset	Mot. Rep.	#Seqs	#Actions	Text
AMASS [14]	SMPL-X	26k	260	✓
GRAB [16]	SMPL-X	1k	4	✓
SAMP [9]	SMPL-X	0.2k	N/A	✓*
Fit3D [6]	SMPL-X	0.4k	29	✓
CHI3D [5]	SMPL-X	0.4k	8	✓
UESTC [11]	SMPL	26k	40	✓
AIST++ [12]	SMPL	1k	N/A	✓*
BEHAVE [1]	SMPL	0.3k	N/A	✓*
HuMMan [2]	SMPL	0.3k	339	✓
GTAHuman [3]	SMPL	20k	N/A	×
Motion-X [13]	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.

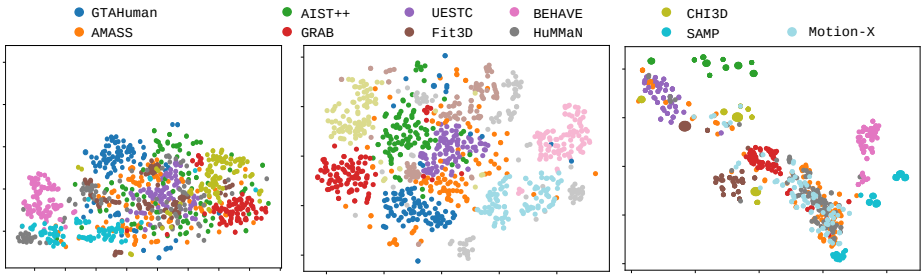

Fig. 1: Motion, KP, and text distribution of Kinematic Phrase Base.

1.6 Global Velocity Phrase

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

2 Kinematic Phrase Base Details

As shown in Tab. 1, 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. 1. Besides, a word cloud visualization of the texts in the database is illustrated in Fig. 2.

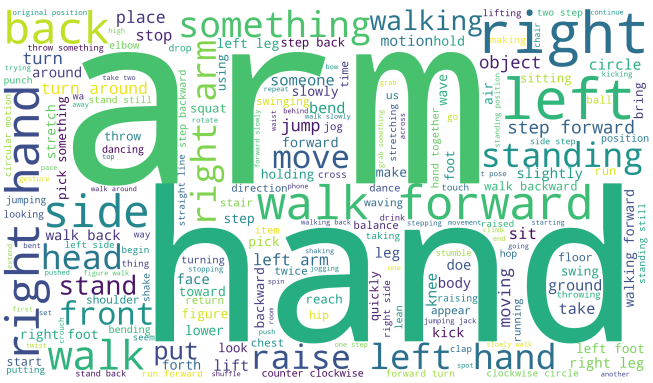

Fig. 2: Word cloud visualization of the texts in Kinematic Phrase Base.

3 Method Details

3.1 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 $A, \hat{A}$:

$$\mathcal{L}_{rec} = \sum_{\cdot \in \{m,p,mp\}} ||M. - \hat{M}.||_1 + ||C. - \hat{C}.||_1 + ||J. - \hat{J}.||_1 + ||V. - \hat{V}.||_1 + ||A. - \hat{A}.||_1. \tag{1}$$

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

$$\mathcal{L}_{KL} = KL(\phi_m, \pi) + KL(\phi_p, \pi). \tag{2}$$

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

$$\mathcal{L}_{da} = KL(\phi_m, \phi_p) + KL(\phi_p, \phi_m). \tag{3}$$

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

$$\mathcal{L}_{emb} = ||z_m - z_p||_1. \tag{4}$$

The overall loss is calculated as

$$\mathcal{L} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{KL} + \lambda_3 \mathcal{L}_{da} + \lambda_4 \mathcal{L}_{emb}, \tag{5}$$

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

4 Kinematic Prompt Generation Details

4.1 Prompts

We provide the 7,776 text prompts converted from KP in the file `kpg.txt`.

5 Experiment Details

5.1 Implementation Details

Sequences are sampled to 15 FPS and randomly clipped into short clips with lengths between 30 frames and 150 frames. 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.

5.2 Motion Generation Settings

For HumanML3D [8], 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 similarly to [8]. 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.

5.3 Model Size comparison.

We compare the number of parameters in our model and previous SOTAs in Tab. 2. As shown, with a model size comparable to MLD [4] and significantly lower than T2M-GPT [19], we achieve competitive performance on conventional benchmarks and even better performance with the newly proposed KPG.

Method	MDM [17]	MLD [4]	T2M-GPT [19]	Ours
#params	23M	42.7M	228M	45.1M

Table 2: Model Size Comparison.

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 [17].

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 [4].

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 [19].

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 3: Detailed user study results on HumanML3D.

5.4 User Study Details

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.

User Study on Conventional Text-to-Motion Detailed results of the HumanML3D user study are demonstrated in Tab. 3. As shown, both FID and R-P@1 are not consistent with the user reviews, indicating these black-box-based metrics might be sub-optimal for motion generation evaluation. Meanwhile, the

Accuracy = 50%		Semantic consistency			Sum
		Yes	Partially	No	
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 [19].

Accuracy = **54%**		Semantic consistency			Sum
		Yes	Partially	No	
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 4: Detailed user study results on 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 5: Detailed consistency statistics between KP-inferred Accuracy and user-reviewed semantic consistency.

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.

User Study on KPG Detailed user study results on KPG are demonstrated in Tab. 4. 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.

Furthermore, we provide detailed consistency statistics between KP-inferred Accuracy and user-reviewed semantic consistency in Tab. 5. 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. Concerning 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. 3, 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. 3, extra right-hand waving motion), while our method manages to provide more concise responses.

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, extending KP to amplitude might 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

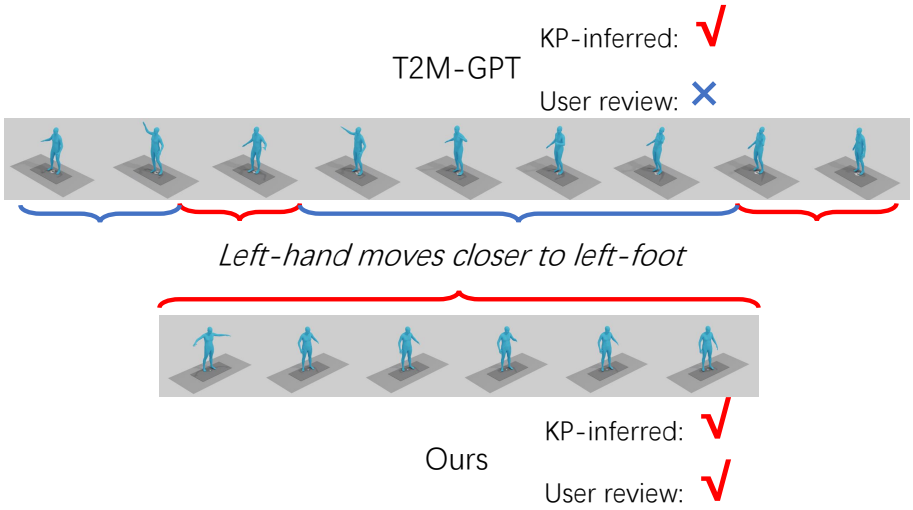

Fig. 3: For KPG, we generate more concise motion than T2M-GPT [19].

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.

5.5 Failure Analysis on KPG

With the attached video `1286.mp4`, we further demonstrate visualization results on KPG. An interesting finding is that different methods show different failure patterns. Limited motion amplitude is usually observed for MDM [17] and MLD [4]. Also, MLD [4] could misunderstand commands on certain body parts. T2M-GPT [19] tends to generate over-active motion sequences with redundancy, which could explain its low accuracy for repetitive prompts. ReMoDiffuse [20] produces jerky motion. MoMask [7] could surprisingly mystify left with right. Our model also shows redundancy. Moreover, most models tend to execute the prompts indirectly, which could indicate the potential over-fitting of motion style. KPG prompts are simple body-joint relations like exercising instructions; however, they are not usually explicitly described in general datasets. Thus, the models tend to produce everyday activity motion, which contains the required relations, instead of directly fulfilling the requirements. This reveals that current models could be sub-optimal in real understanding of the human body structure.

5.6 More Visualizations

More visualizations are included in the video `1286.mp4`.

6 Extensive Discussion

Relation with phase-based methods. Some previous efforts [10,15,18] adopted phase-based motion representation, which is similar to Kinematic Phrase in movement representation. However, the term "Phrase" emphasizes the seamless conversion between our phrases and semantic descriptions, which is not explored in previous efforts.

Representing complex motions with KP. Currently, the KP-based complex motion semantics representation could be conducted demonstration-based. That is, given a motion demonstrating Tai-Chi, we could convert it into KPs for a *basic* KP representation of Tai-Chi. Then minor modifications could be made to the KP representation for diversity to produce diverse Tai-Chi motions. Further exploration of KP-based semantics representation, *e.g.*, the introduction of LLMs might be promising given the symbolic nature of KP. We believe future works on this would be promising.

Further exploration on KP-based evaluation. Thanks for your constructive comment. A current limitation of KP-aided evaluation is the trade-off between reliability and generality. Initially, we considered comparing the KP similarity of the generated and GT motions for general prompts. However, as the GT might not fully cover expected semantics, this design sacrifices reliability, which is a common issue of previous metrics. Therefore, we limit the current KPG to atomic/two-gram prompts to guarantee reliability and obtain helpful insights. Enhancing KPG with more generality would be promising in future works. Also, KP distribution analysis would be a helpful interpretative analysis tool.

FineMoGen Comparison. We evaluate it on KPG, with a 46.79% Acc for simultaneous prompts (ours 43.08%) and 36.52% overall Acc (ours 57.86%). FineMoGen is trained with LLM-extended descriptions similar to the simultaneous prompts. However, it is also biased toward them, resulting in a degenerated overall performance. More details will be updated in the revision.

References

1. Bhatnagar, B.L., Xie, X., Petrov, I., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2022)
2. Cai, Z., Ren, D., Zeng, A., Lin, Z., Yu, T., Wang, W., Fan, X., Gao, Y., Yu, Y., Pan, L., Hong, F., Zhang, M., Loy, C.C., Yang, L., Liu, Z.: Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 557–577. Springer Nature Switzerland, Cham (2022)
3. Cai, Z., Zhang, M., Ren, J., Wei, C., Ren, D., Lin, Z., Zhao, H., Yang, L., Liu, Z.: Playing for 3d human recovery. arXiv preprint arXiv:2110.07588 (2021)
4. 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)

5. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Three-dimensional reconstruction of human interactions. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
6. Fieraru, M., Zanfir, M., Pirlea, S.C., Olaru, V., Sminchisescu, C.: Aifit: Automatic 3d human-interpretable feedback models for fitness training. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2021)
7. 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)
8. Guo, C., Zou, 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)
9. 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)
10. Holden, D., Komura, T., Saito, J.: Phase-functioned neural networks for character control. ACM Transactions on Graphics (TOG) **36**(4), 1–13 (2017)
11. 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)
12. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Learn to dance with aist++: Music conditioned 3d dance generation (2021)
13. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. arXiv preprint arXiv:2307.00818 (2023)
14. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019)
15. Starke, S., Zhao, Y., Komura, T., Zaman, K.: Local motion phases for learning multi-contact character movements. ACM Transactions on Graphics (TOG) **39**(4), 54–1 (2020)
16. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: European Conference on Computer Vision (ECCV) (2020), `https://grab.is.tue.mpg.de`
17. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
18. Zhang, H., Starke, S., Komura, T., Saito, J.: Mode-adaptive neural networks for quadruped motion control. ACM Transactions on Graphics (TOG) **37**(4), 1–11 (2018)
19. 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)
20. 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)