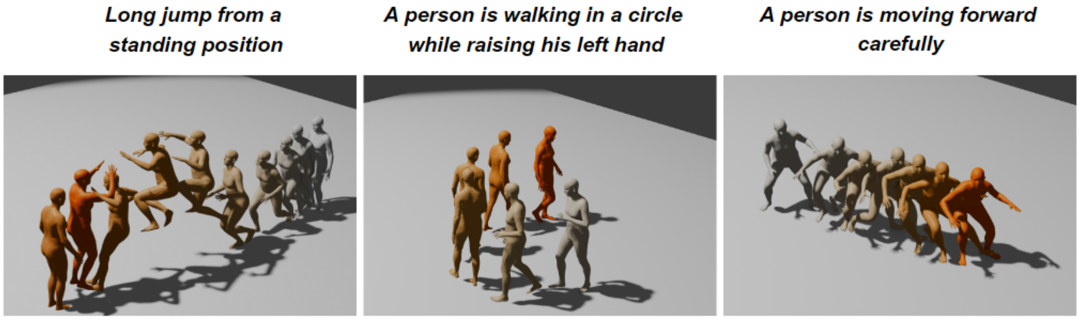
Kinematic Motion Diffusion: Towards Semantic-adaptive Motion Synthesis via Kinematic Guidance

LEIZHI LI*, JINCHENG YANG*, YICHAO ZHONG*, and JUNXIAN GUO*, Shanghai Jiao Tong University, China



Long jump from a standing position

A person is walking in a circle while raising his left hand

A person is moving forward carefully

Fig. 1. Our proposed **Kinematic Motion Diffusion (KMD)** generates high-quality and feasible motion with the following text prompts given: a) Long Jump from a standing position. b) A person is walking in a circle while raising his left hand. c) A person is moving forward carefully. The brighter the colors, the later in time.

3D human motion synthesis is a crucial task with broad applications across various industries. Although recent advancements in data-driven models like GANs, VAEs, and denoising diffusion models have improved the quality and diversity of generated motions, the integration of natural language as a descriptor for motion synthesis presents unique challenges. Current methods often fail to ensure coherence between text semantics and the resulting motion sequences, a gap that is particularly evident in applications requiring subtle semantic interpretation. In this paper, we introduce Kinematic Motion Diffusion (KMD), a novel model that leverages kinematic phrases (KP) as an intermediate representation to bridge the domain gap between language and motion. By establishing multiple classifiers based on KPs, KMD utilizes a high level of motion abstraction grounded in human knowledge to generate more semantically coherent motions. We propose a semantic adaptive activation method that dynamically adjusts classifier compounds, enabling the model to accommodate diverse motion generation scenarios. KMD achieves a considerable performance in many metrics on HumanML3D with a much simpler architecture compared with other state-of-the-art works.

CCS Concepts: • **Computing methodologies** → **Activity recognition and understanding**.

Additional Key Words and Phrases: motion synthesis, diffusion probabilistic models, computer vision

1 INTRODUCTION

3D human motion synthesis represents a crucial task in numerous domains such as gaming, human-computer interaction, film production, and robotic assistance. With the recent advancements in data-driven generative models such as Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE) [?] [?], solutions of increasing satisfaction

[*] Both authors contributed equally to this research.

Authors' address: Leizhi Li, lileizhi@sjtu.edu.cn; Jincheng Yang, modem514@sjtu.edu.cn; Yichao Zhong, zhongyichao0618@sjtu.edu.cn; Junxian Guo, guojunxian@sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China.

have been proposed, with the denoising diffusion model surpassing its predecessors in terms of both quality and diversity [?] [?] [?] [?] [?].

Language, being the primary means through which humans convey concepts, indisputably serves as an accurate and convenient medium for describing objective motion, especially in real-world applications. The prevalent method of incorporating language information into diffusion-based motion generation models involves utilizing text descriptions as conditional inputs alongside denoised motion sequences. Existing research predominantly focuses on ensuring the model's output aligns with specific external signals, such as trajectory and key-frame inpainting. However, there is a lack of studies that carefully consider the coherence between text semantics and the motions, which argue that this type of coherence is crucial in real-world applications, particularly for motion as a modality that contains more subtle semantic information. A clear illustration of this can be seen in Figure 3, where a recent motion generation model, GMD, generates a sequence of a figure raising its right arm, despite the input prompts indicating the left arm should be raised. We believe that this discrepancy arises primarily for two reasons.

First, the inherent high degrees of freedom in human motions present a significant challenge to current methods. While the abundance of data in the image generation field somewhat mitigates this issue, the motion generation field lacks such extensive datasets. Another challenge is the absence of universal objectives to evaluate the synthesized motions, with mainstream practice relying on human feedback for accurate assessment.

Second, the use of language as a cross-modal auxiliary is primarily limited to combining text embedding with motion sequence as input to the diffusion model in the reverse denoising process. This method does infuse semantic information into synthesized motions but struggles to bridge the domain gap between natural language and human motions.

Given these observations, we believe it's essential to introduce a quantitative intermediate representation of motion sequences to extract appropriate and interpretable abstractions. Previous studies proposed kinematic human motion representations, which recent works have extended into Kinematic Phrases (KP) [?]. KP has been applied in multiple tasks related to human motion, but its use in motion generation is very limited. The challenge here is that, even when using KP as an intermediate language, a domain gap still exists between KP and motion.

Recent advancements in classifier guidance offer an indirect approach to enhance generation quality by intervening in the diffusion sampling process. By developing classifiers based on KP, the motion sampled by diffusion can be evaluated using interpretable and concrete classifiers. This approach guides diffusion towards converging motions that align more closely with the input text semantics.

Another challenge arises from the fact that different motion sequences necessitate distinct KPs for description. For instance, the action of walking primarily focuses on the relative position of the hands, while sitting does not. To adaptively trigger different KP classifiers for various language descriptions, we propose a semantic adaptive activation method. This approach assigns a trainable embedding to each classifier, and in each iteration, we first encode the given language description and map it to the k-nearest classifiers. Only the activated classifiers contribute to the sampling process.

In awareness of such observations, we propose our model Kinematic Motion Diffusion(KMD). The evaluations we conduct on HumanML3D [?] show that KMD attains a considerable performance with a much simpler architecture compared with other state-of-the-art works. The examples displayed in Fig 1 shows that KMD acquires semantic abilities in multiple levels. Our contributions can be summarized as follows:

- We introduce kinematic prior into traditional motion diffusion by establishing multiple classifiers based on Kinematic Phrases. By utilizing its high level of abstraction and characterization of human motions, we manage to establish motion generation on the ground of human knowledge and understanding.
- We propose semantic adaptive activation methods that dynamically adjust our classifier compound to better accommodate diverse motion generation scenarios. This new way of semantic infusion is the first to be used in the human motion generation field, bypassing the domain gap between motion and language.

2 RELATED WORK

2.1 Diffusion Probabilistic Models (DPM)

DPMs have attracted much attention recently for their incredible performance and versatility. As generative models, DPMs have been widely used on different tasks like image generation [?], video generation [?], drug discovery [?], and reinforcement learning [?]. Apart from that, there are some guided sampling methods for DPMs to handle conditional generation problems, such as classifier guided sampling [?], classifier-free guided sampling [?]. Specifically, the classifier guidance sampling method needs an extra classifier and can generate samples with label attributes. Moreover, this extra classifier is independent of the diffusion model, enabling a core diffusion model to be deployed with different classifiers without retraining. In addition, the classifier can even be a deterministic function rather than a trained network.

2.2 Human Motion Generation

Human motion generation has recently been a hot yet challenging research area. Previous approaches mostly implemented human motion generation via generative models like VAE, GAN, and diffusion models. Initially, some works address human motion prediction as a basic version of generation. ?] predicts 3D body movement based on history trajectories based on VAE. The work emphasized some non-joint DoFs and achieved progress. ?] aims to enable virtual humans to navigate cluttered indoor scenes and interact naturally with objects. The approach predicts more diverse human motions with a scheduled sampling strategy on training, which reuses the current-step prediction as the input of the next step with a certain probability. As for works in human motion generation, Motion Latent Diffusion [?] combines VAE with the latent diffusion model by plugging the latent diffusion model into the Encoder sampling process and achieving competitive performance. There are also dozens of generative methods under the guidance of specific signals like textual prompts, trajectory keyframe conditions, etc. ?] proposed a GAN-based transformer named actFormer to address single-to-multi-personal human motion generation. The GAN is conditioned on labels of the source of the data.

TEMOS [?] employed two transformer-based VAE encoders, one for motion and another for text, and one VAE decoder, to achieve text-to-motion generation. Also aiming to address text-to-motion tasks, Guided Motion Diffusion [?] utilizes the classifier guidance sampling method to solve the textual guided motion generation. Its diffusion sample input is a text embedding concatenated to a motion-trajectory-shaped noise, and the conditional signal function is built by the keyframe trajectory conditions z. However, the condition signal excludes the semantic information, which would be considered in our work. Above all, most previous works on motion synthesis hold a large and complex model, while ours is simpler without loss of performance.

2.3 Kinematic Phrases (KP)

KP is a recent work proposed by **?**] as a phrase system representing the kinematic semantics of a motion sequence to bridge the gap between motion and action semantics. The KP consists of 420 objective-level kinematic indicators that score a motion frame on its relativity to their objective kinematic fact on semantics, covering six types of kinematic hierarchies: joint positions, joint pairwise relative positions, joint pairwise distances, limb angular movements, limb oriental movements, and whole-body movements. Specifically, KP receives a frame of human motion sequence as input and then outputs 420 KP scores within $[0, 1]$, which represents how strong each KP's kinematic semantic is stimulated by the input motions. Triggered by KP, they also proposed a text-to-motion generation benchmark called Kinematic Prompts Generation (KPG). In one track of our work, the classifiers for classifier guidance sampling are built based on KPs.

3 METHODS

3.1 Preliminaries

Denoising Diffusion Probabilistic Model (DDPM). DDPM [**?**] utilizes a parameterized diffusion process, represented as $p_\theta(x_0) = \int p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) x_{1:T}$, to model how the pure noise, denoted as $x_T = \mathcal{N}(0, I)$, is denoised into real data x_0. The sequence $x_{T:0}$ is a Markov chain with learned Gaussian transitions. Consider the forwarding process $x_{0:T}$. Each step is denoted as adding Gaussian noise to the data according to a variance schedule $\beta_{1:T}$: $x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1 - \alpha_t} \epsilon_t$, and therefore $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon(x_t, t)$, where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_1^t \alpha_i$ and $\epsilon(x_t, t) \sim \mathcal{N}(0, \mathbf{I})$. As for the reverse direction, it is parameterized with $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta^2(x_t, t))$, where:

$$\Sigma_\theta^2(x_t, t) = \beta_t \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}$$

and

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} (x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t)) \ . \tag{1}$$

However, the noise variable ϵ is still unknown. To address this issue, a network ϵ_θ is employed to learn the noise generation process. The training session is performed by maximizing the evidence lower bound(ELBO):

$$\mathcal{L}_{t-1} \coloneqq \mathbb{E}_{x_0, \epsilon} [\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + (\sqrt{1 - \bar{\alpha}_t}) \epsilon, t) \|^2] \ . \tag{2}$$

Here, ϵ is sampled from the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ and $\mathcal{L}_{t-1}$ represents the loss function at diffusion step $t - 1$ for $t > 1$. During deployment, we sample iteratively according to the reverse process.

Classifier Guidance Sampling Method. Classifier guidance needs an extra classifier p_ϕ that outputs the probabilities for a sample to be estimated as different labels. Compared to DDPM, experimentally, the only change is adding a classifier term at the denoising process:

$$p(x_t | x_{t+1}, y) = \mathcal{N}(\mu(x_t, t) + s \Sigma(x_t, t) g, \Sigma(x_t, t)) \ , \tag{3}$$

where $g = \nabla_{x_t} \log p_\phi(y|x_t)|_{x_t = \mu}$ and s serves as the gradient scale.

Moreover, a typical implementation is to set the labels as $\{0, 1\}$, which refers to infeasible and feasible, and some metric functions that map samples to $\{0, 1\}$ are employed as classifiers, then input label=1 while deploying to achieve feasible sample generation. Specifically, Guided motion diffusion [**?**] utilizes the classifier guidance sampling method

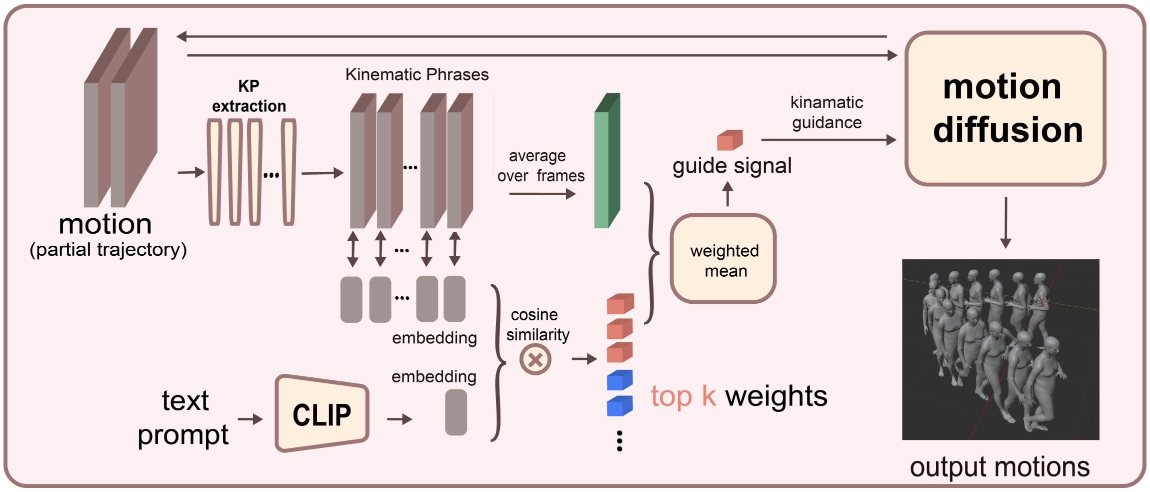


Fig. 2. Overall framework of **KMD**: KMD contains a CLIP, a motion diffusion, and kinematic classifiers as a guidance signal into motion diffusion. KMD is designed to accomplish a one-stage pipeline to generate human motions

and a mathematical trick:

$$\nabla_{x_t} \log p_\phi(G(x_t) = 0|x_t)|_{x_t} \simeq \nabla_{x_t} G(x_t)\,, \tag{4}$$

This trick enables us to give a scalar signal to the kinematic guidance function as Figure 2 shows, as long as samples with label 0 are our target to generate.

3.2 Overview

This work aims to generate human motion based on a natural language description. Our framework consists of a KP-guided motion diffusion model and a set of learnable kinematic classifiers. We adopt a one-stage pipeline to generate motion as shown in Figure 2, i.e., we directly generate motion sequences without first generating the corresponding trajectory, nor do we use any keyframes in our pipeline. KMD contains a frozen CLIP text encoder $\mathcal{E}$ of dimension D. The kinematic classifiers consists of m KP extract functions $\mathbf{F} = [f_1.f_2, \dots, f_m]$ and individual phrase is assigned with a semantic embedding $\mathbf{E} = [e_1, e_2, \dots, e_m]$ of the same dimension as T.

With each incoming text description of desired motion that is first decoded by CLIP [?] into text token t, KMD calculates the affinity with kinematic classifier embeddings:

$$\mathbf{A} = [\frac{t \cdot e_i}{|t| \cdot |e_i|}]_{i=1}^{m} \tag{5}$$

KMD uses the affinity scores to filter the kinematic classifiers and as the weight to assemble individual KP classifiers for diffusion sampling. Because guidance only functions in the sampling procedure of the diffusion model without intervening in the reverse process network, we adopt a pre-trained motion diffusion model provided by Guided Motion Diffusion [?].

3.3 Kinematic Classifier

As we explain in the introduction, recent works of human motion generation can not firmly adhere to the semantic instructions, especially in critical details, which severely limits their applications. We believe such behavior is partially the consequence of the data-driven approach to motion generation. The current data sets cannot strongly support arbitrary motion generation in text. Human motions' innate high degree of freedom can not fit in abstract semantic space. To alleviate this gap between the two modalities, KP is proposed, which takes the objective kinematic facts of human motion and depicts qualitative categorical representations by capturing sign changes with minimal pre-defined standards [**?**].

KP extraction is based on a skeleton sequence $X = \{x_i | x_i \in \mathbb{R}^{n_k \times 3t}\}_{j=i=1}$, 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. With the incoming skeleton sequence X, at each specific time t, each phrase P_j calculates a scalar indicator $I^t \in \{0, 1\}$. The indicator sequence forms the kinematic signal. We extract the mean over frames as the final label to form the scalar classifier to be passed to the diffusion model.

$$P_i(\mathbf{X}) = \frac{1}{t} \sum_{j=1}^{t} \mathrm{sign}(f_i(\mathbf{X})) \tag{6}$$

The original complete set of KP contains 400 phrases and will introduce too much noise in practice. So, as we show in the overview, we adopt a top-k selection of the kinematic classifiers based on the similarity with text token $\mathbf{A}$ to block out signals irrelevant to our specific scenario. Besides, we also use the similarity to weigh the strength of each kinematic classifier.

$$G(\mathbf{X}) = \sum_{i \in \mathrm{argtopk}(a_i)} a_i \cdot P_i(\mathbf{X}) \tag{7}$$

$G(\mathbf{X})$ is eventually transmitted to the diffusion model as kinematic guidance.

3.4 Semantic Adaptive Activation

Concerning diversity in human motions, the representations of different kinematic phrases are diverse in their compatibility with specific circumstances. Consequently, it is necessary to establish an alignment between kinematic classifiers and language latent space. A vanilla attempt grants each kinematic classifier an embedding generated by directly processing the description on the physical truth of each KP with CLIP. However, considering CLIP is trained on general text-image pares, the latent space of texts regarding human motions is only a subspace for the whole CLIP. In our case, this leads to a severe collapse of the space spanned by the embeddings of all kinematic classifiers. Consequently, KMD is unable to adapt to text description effectively. This tendency is reflected in our experiments as a noticeable concentration of activated kinematic classifiers.

We introduce flexible parameters on both ends to align the classifier and scenario semantic space better while maintaining low extra costs. On the encoder end, we introduce a multi-layer perceptron to conduct transformation on the CLIP output instead of fine-tuning CLIP. We also replace fixed classifier embedding with trainable parameters $\hat{\mathbf{E}} = [\hat{e}_i]_{i=1}^{m}$.

$$\hat{T} = \mathrm{MLP}(\mathcal{E}(T))$$

The affinity is calculated by

$$\hat{\mathbf{A}} = [\cos(\hat{T}, \hat{e}_i)]_{i=1}^{m} \tag{8}$$

The expressibility brought by learnable parameters requires extra training to build semantic representation and projection on CLIP output. To train the classifier embeddings, we utilize the ground truth of the dataset. Mathematically, on a given ground truth motion sequence of F frames $\mathbf{X} = \{x^t\}_{t=1}^{F}$ and correspondent text description T, we follow the pipeline shown in figure 2 and calculate the corresponding text-classifier similarity and the average kinematic signal of each classifier. The goal is to minimize the Cosine Ensemble Loss on these two scalars for all kinematic classifiers:

$$\mathcal{L} = 1 - \cos(\hat{\mathbf{A}}, \mathbf{P}(\mathbf{X})) \tag{9}$$

Notice that $\mathbf{P}(\mathbf{X})$ is the kinematic classifier output mean over frames.

4 EXPERIMENTS

4.1 Settings

Datasets. We adopt the HumanML3D [?] dataset to evaluate our methods for text-to-motion generation. As a collection of text-annotate motion sequences from multiple datasets including AMASS [?], KIT-ML [?] and HumanAct12 [?], HumanML3D [?] contains 14,646 motions and 44,970 motion annotations and covers a broad range of human actions.

Evaluation Metrics. We evaluate generative text-to-motion models using some standard metrics introduced by Guo et al [?]. These include Frechet Inception Distance (FID), R-Precision, and diversity. **FID** measures the distance between the distributions of ground truth and generated motion using a pretrained motion encoder. **R-Precision** evaluates the relevance of the generated motion and its text prompt. **Diversity** measures the variability within the generated motion.

Implementation Details. For text feature extraction, we employ a standard CLIP (Vit-B/32) [?], fine-tuned with a three-layer MLP. Embeddings of each kinematic phrases are initialized as the text embedding obtained from their corresponding KPG prompts [?]. Our motion DPM is based on UNET with AdaGN. We utilized DDPM [?] with $T = 1,000$ denoising steps for training and inference.

4.2 Training-free Approach

As introduced by **?**], KP also proposed KPG that each KP indicator has its text prompts and its embeddings, which can be further employed in text-guided generation tasks. With this idea, we initially implemented a training-free approach as a baseline and observed some progress in our results. This basic pipeline demonstrated promising performance, as shown in Table 2, but suffered from a significant issue. Since the majority of our target text prompts consist of kinematic expressions, which means their semantics information are similar in a general view, all the text prompts are mapped to a small subspace of the whole latent space for embeddings, leading to a phenomenon that the mutual similarities among our CLIP embeddings are generally high (e.g., ≥ 0.7 for all). As a result, our top-k selection strategy would be susceptible to noise and perturbations due to this embedding concentration.

Table 1. **Observation on CLIP embeddings of KPG prompts.** In this experiment, we fix the text description t ="a man walks backward," and then we list some of the **KPG prompts** $\{s_i\}_{i=1}^{5}$ together with the cosine similarity between $CLIP(s_i)$ and $CLIP(t)$.

KPG prompts	cosine similarity
left hand moves forwards	0.919
the left hand is below head then above head	0.890
left hand moves away from head	0.899
left arm bends	0.903
the person moves backwards	0.931

Table 2. **Quantitative evaluation on the HumanML3D** [?] **test set**. For each metric, we repeat the evaluation 20 times and report the average with a 95% confidence interval. The right arrow → means closer to real data is better. **Bold** face indices the best result among all selected work.

	FID↓	R-Precision ↑ (Top3)	Diversity→	
Real motion	$0.002^{\pm.000}$	$0.797^{\pm.002}$	$9.503^{\pm.065}$	
TM2T [?]	$1.501^{\pm.017}$	$0.729^{\pm.002}$	$8.589^{\pm0.76}$	-
T2M [?]	$1.067^{\pm.002}$	$0.740^{\pm.003}$	$9.188^{\pm.002}$	
MDM [?]	$0.544^{\pm.044}$	$0.611^{\pm.007}$	$\mathbf{9.559}^{\pm.086}$	
MLD [?]	$0.473^{\pm.013}$	$0.772^{\pm.002}$	$9.724^{\pm.082}$	
GMD [?]	$\mathbf{0.212}^{\pm.021}$	$0.670^{\pm.006}$	$9.440^{\pm.057}$	
Ours (training-free)	$0.457^{\pm.010}$	$0.701^{\pm.008}$	$9.235^{\pm.032}$	
Ours (trainable)	$0.320^{\pm.008}$	$\mathbf{0.749}^{\pm.009}$	$9.423^{\pm.034}$	

4.3 Trainable Approach

To overcome the issue of embedding concentration and further improve our results, we employed fine-tuning by introducing a trainable 3-layer MLP at the end of CLIP. The objective was to project the embeddings into a latent space that better separates the textual representations related to kinematics, thereby reducing the concentration effect.

4.3.1 Compared to Training-free Approach. We evaluated the training and training-free approach on their cosine similarities. As shown in Table 1, the trainable approach's embedding similarities show less concentration than the training-free approach, and the model with trained 3-layer MLP reaches better evaluation scores.

4.3.2 Ablation studies on Top-k strategies. Our motivation to apply top-k strategies is diminishing the effects of the unrelated indicators, which may negatively affect our final guide signal. We implement this idea by setting these indicators' weights to 0. To distinguish whether a KP indicator is 'activated,' we rank their weights and regard the KP indicators with at least k-th highest weights as activated.

We did ablation studies over hyper-parameter k to prove this strategy works. We evaluated the model's performances with $k = 10$ and $k \to \infty$(all selected). As Table 3 displays, we noted a distinct improvement when employing the top-50 strategy. The model achieved notably better alignment than selecting all KP indicators, thereby validating the effectiveness of our top-k approach. When k is too low, the abundance of 400 kinematic classifiers of KP is not fully utilized to adapt to different scenario requirements. When k is too low, we may introduce too much noise in the diffusion sampling process and impair model performance.

Moreover, we evaluate KMD and some previous state-of-the-art works with FID, R-Precision, and Diversity metrics. As shown in Table 1, our work achieves a considerable performance with a much simpler architecture, especially on R-Precision metrics. Our extensive adjustments, including fine-tuning procedures and applying top-k strategies, effectively addressed the challenges posed by embedding concentration and yielded favorable outcomes across multiple evaluation criteria.

Finally, We did some visualizations, and some of our experimental observations are as 3 shows. We select some text prompts as input and visualize KMD and GMD's generations with them. KMD outperforms its base model, GMD, especially in the cases we display. a) As the yellow boxes indicate, GMD's generation mistakes which hand to raise when walking, while our KMD generates more reasonable motion sequences. b) Given the prompt 'Long jump from a

Table 3. **Ablation study on top-k strategies.** In this ablation study, we evaluate the performance of our method for $k = 10, 50, 100, 200$ and $k = \infty$ (i.e., no KP indicator is ignored when calculating G). **Bold** face indices the best result, while underscore refers to the second best one.

	FID↓	R-Precision ↑ (Top 3)	Diversity →
Real motion	0.002	0.797	9.503
$k = 10$	0.341	0.710	9.233
$k = 50$	**0.320**	0.749	**9.423**
$k = 100$	0.323	**0.751**	9.419
$k = 200$	0.330	0.729	9.239
$k = \infty$	0.379	0.713	9.102

standing position', the GMD-generated sample fails to represent a 'long' jump and even jump backward, while KMD shows great diversity as the two samples differ in the approach-run distance, but both achieved a 'long' jump.

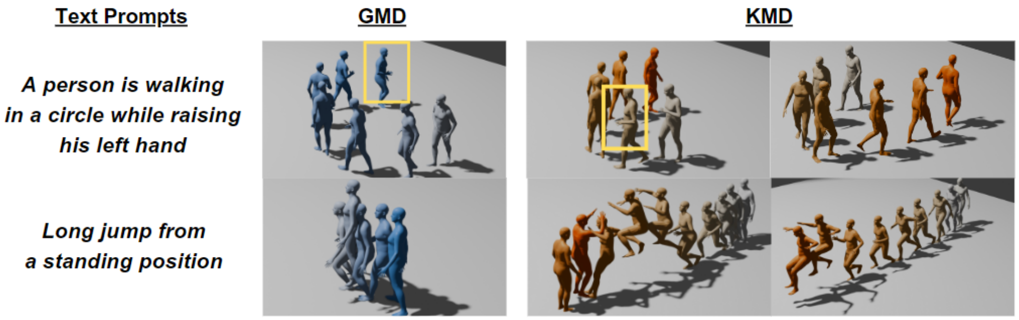

Text Prompts
GMD
KMD
A person is walking in a circle while raising his left hand
Long jump from a standing position

Fig. 3. **Comparison examples with GMD on some generations.** KMD's two columns refer to two experiments with different seeds. The brighter the colors, the later in time.

5 CONCLUSION

In conclusion, we have presented KMD, a novel model that significantly improves the performance of 3D human motion synthesis in various scenarios, with a simpler architecture compared to other state-of-the-art models. Our approach introduces a kinematic prior to traditional motion diffusion by establishing multiple classifiers based on Kinematic Phrases (KP), effectively grounding motion generation in human knowledge and understanding. Furthermore, we have proposed a semantic adaptive activation method that dynamically adjusts our classifier compound to accommodate diverse motion generation scenarios better. This novel method of semantic infusion, the first of its kind in human motion generation, effectively bridges the domain gap between motion and language. The evaluations we conducted on HumanML3D demonstrate the considerable performance achieved by KMD.

ACKNOWLEDGEMENT

We would like to convey our warmest gratitude to Prof. Cewu Lu, Prof. Yong-lu Li and Dr. Lixing Yang for their guidance during our research on this topic. We also appreciate Xinpeng Liu for his instruction on the usage of Kinematic Phrases (KP) and for his advice on our work.

REFERENCES

[] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. 2023. Executing your Commands via Motion Diffusion in Latent Space. arXiv:2212.04048 [cs.CV]

[] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

[] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv:1406.2661 [stat.ML]

[] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions from Text. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5142–5151. https://doi.org/10.1109/CVPR52688.2022.00509

[] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions from Text. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 5142–5151. https://api.semanticscholar.org/CorpusID:250602257

[] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022. TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts. arXiv:2207.01696 [cs.CV]

[] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2021–2029.

[] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. 2021. Stochastic Scene-Aware Motion Prediction. arXiv:2108.08284 [cs.CV]

[] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239 [cs.LG]

[] Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. arXiv:2207.12598 [cs.LG]

[] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. 2022. Video Diffusion Models. arXiv:2204.03458 [cs.CV]

[] Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. 2022. Planning with Diffusion for Flexible Behavior Synthesis. arXiv:2205.09991 [cs.LG]

[] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. 2023. Guided Motion Diffusion for Controllable Human Motion Synthesis. arXiv:2305.12577 [cs.CV]

[] Diederik P Kingma and Max Welling. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML]

[] Xinpeng Liu, Yong-Lu Li, Ailing Zeng, Zizheng Zhou, Yang You, and Cewu Lu. 2023. Bridging the Gap between Human Motion and Action Semantics via Kinematic Phrases. arXiv:2310.04189 [cs.CV]

[] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision*. 5442–5451.

[] Xuran Pan, Tianzhu Ye, Dongchen Han, Shiji Song, and Gao Huang. 2022. Contrastive Language-Image Pre-Training with Knowledge Graphs. arXiv:2210.08901 [cs.CV]

[] Mathis Petrovich, Michael J. Black, and Gül Varol. 2022. TEMOS: Generating diverse human motions from textual descriptions. arXiv:2204.14109 [cs.CV]

[] Matthias Plappert, Christian Mandery, and Tamim Asfour. 2016. The KIT Motion-Language Dataset. *Big Data* 4, 4 (dec 2016), 236–252. https://doi.org/10.1089/big.2016.0028

[] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. 2022. Human Motion Diffusion Model. arXiv:2209.14916 [cs.CV]

[] 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. 2022. ActFormer: A GAN-based Transformer towards General Action-Conditioned 3D Human Motion Generation. arXiv:2203.07706 [cs.CV]

[] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. 2022. GeoDiff: a Geometric Diffusion Model for Molecular Conformation Generation. arXiv:2203.02923 [cs.LG]

[] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. 2023. PhysDiff: Physics-Guided Human Motion Diffusion Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[] Yan Zhang, Michael J. Black, and Siyu Tang. 2021. We are More than Our Joints: Predicting how 3D Bodies Move. arXiv:2012.00619 [cs.CV]

[] Zixiang Zhou and Baoyuan Wang. 2022. UDE: A Unified Driving Engine for Human Motion Generation. arXiv:2211.16016 [cs.CV]