ENERGYMOGEN: Compositional Human Motion Generation with Energy-Based Diffusion Model in Latent Space
Supplementary Material

Jianrong Zhang[1], Hehe Fan[2,†], Yi Yang[2]
[†]Corresponding author
[1]ReLER, AAII, University of Technology Sydney [2]CCAI, Zhejiang University
https://jiro-zhang.github.io/EnergyMoGen/

In this supplementary material, we present:

- Section 1: Training details of ENERGYMOGEN.
- Section 2: Additional results of skeleton-based diffusion models.
- Section 3: Ablations on λ_s, λ_l, and λ_m for Synergistic Energy Fusion.
- Section 4: Ablations on the number of latent vectors in motion VAE.
- Section 5: Ablation study of hyper-parameters in energy-based cross-attention.
- Section 6: Results of inference time.
- Section 7: Evaluation of foot sliding.
- Section 8: More visual results of energy distributions.
- Section 9: More details on datasets and evaluation metrics.
- Section 10: Limitation and failure case

1. Implementation Details

We first provide training details of ENERGYMOGEN. For Motion VAE, both the encoder $\mathcal{E}$ and decoder $\mathcal{D}$ comprise 9 layers of transformer blocks with a dimension d =256. We use 10 additional tokens (mean and various tokens) to sample N =5 latent vectors representing the motion. We use an AdamW optimizer with a batch size of 1024. We train 300K iterations in total, and the learning rate changes from 0.0001 to 0.00001 after 200K iterations. The weights of reconstruction loss and KL loss are set to 1 and 0.0001. As for the latent diffusion, we apply a frozen CLIP ViT-L/14 to encode the textual descriptions. Regarding the denoising autoencoder, we use a 9-layer transformer with a dimension of 256. To acquire an accurate mapping from textual data to latent vectors during training, γ is initialized with 0. We utilize the AdamW optimizer to train the model with a batch size of 512, with an initial learning rate of 0.0001 for 200K iterations and decayed to 0.00001 for another 100K iterations. The diffusion model is learned using classifier-free guidance [3] with an unconditional score estimation rate of 10%. For experiments on CompML, we only finetune the latent diffusion model for 100K iterations in total with a learning rate of 0.00005.

For the skeleton-based approach, we use an 8-layer transformer with a dimension of 512. We follow [12] to train the model using Adam optimizer with a batch size of 1024. We train 8000 epochs in total and employ the CosineAnnealing learning policy with the learning rate from 0.0002 to 0.00002.

2. Additional Results of Skeleton-Based Diffusion Models

2.1. Text-to-Motion Generation

We conduct experiments on HumanML3D [2] to evaluate the performance of text-to-motion generation. We use evaluation models from Guo *et al*. [2] and use the same metrics. The training details of skeleton-based ENERGYMOGEN are provided in Section 1. Experimental results are shown in Table 1. Our approach outperforms current state-of-the-art skeleton-based methods, *i.e*., ReMoDiffuse [13] and FineMoGen [14] on R-Precision, Diversity, and MM-Dist, while achieves comparable results on FID and MModality.

2.2. Motion Temporal Composition

Following FineMoGen [14] and PriorMDM [9], we use the motion temporal composition task to measure the compositional capacity of our approach. We perform latent-aware

Methods	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity →	MModality ↑
	Top-1	Top-2	Top-3				
Real motion	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
MDM [10]	$0.418^{\pm.005}$	$0.604^{\pm.001}$	$0.707^{\pm.004}$	$0.489^{\pm.025}$	$3.630^{\pm.023}$	$\mathbf{9.450}^{\pm.066}$	$\mathbf{2.870}^{\pm1.11}$
MotionDiffuse [12]	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$\underline{9.410}^{\pm.049}$	$1.553^{\pm.042}$
ReMoDiffusion [13]	$\underline{0.510}^{\pm.005}$	$\underline{0.698}^{\pm.006}$	$\underline{0.795}^{\pm.004}$	$\mathbf{0.103}^{\pm.004}$	$\underline{2.974}^{\pm.016}$	$9.018^{\pm.075}$	$1.795^{\pm.043}$
FineMoGen [14]	$0.504^{\pm.002}$	$0.690^{\pm.002}$	$0.784^{\pm.004}$	$0.151^{\pm.008}$	$2.998^{\pm.008}$	$9.263^{\pm.094}$	$\underline{2.696}^{\pm.079}$
EnergyMoGen (skeleton)	$\mathbf{0.528}^{\pm.003}$	$\mathbf{0.718}^{\pm.003}$	$\mathbf{0.810}^{\pm.002}$	$\underline{0.139}^{\pm.007}$	$\mathbf{2.902}^{\pm.010}$	$9.386^{\pm.078}$	$2.549^{\pm0.104}$

Table 1. **Comparison with the state-of-the-art diffusion models on the HumanML3D [2] test set.** We repeat the evaluation 20 times for each metric and report the average with a 95% confidence interval. Bold and underlined indicate the best and second-best results.

Methods	R-Precision ↑	FID ↓	Diversity →	MM-Dist ↓
Ground Truth	0.80	$1.6 \times 10{-}3$	9.62	2.96
PriorMDM [9] (Double take)	0.59	0.60	<u>9.50</u>	5.61
PriorMDM [9] (First take)	0.59	1.00	9.46	5.63
MotionDiffuse [15]	0.62	1.76	8.55	5.40
ReMoDiffuse [13]	<u>0.64</u>	**0.40**	9.35	<u>5.24</u>
FineMoGen [9]	<u>0.64</u>	0.45	9.23	5.27
Ours	**0.67**	<u>0.43</u>	**9.52**	**5.22**

Table 2. **Quantitative results on the HumanML3D [2] test set.** R-Presicion denotes Top-3 accuracy. Bold and underlined indicate the best and second-best results.

composition to tackle this task.

Specifically, denoting $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ as two concepts. $M^t_{\boldsymbol{c}_1} \in \mathbb{R}^{N_1 \times d_m}$, $M^t_{\boldsymbol{c}_2} \in \mathbb{R}^{N_2 \times d_m}$ denote predicted scores corresponding to two concepts at t-th step, N_i is the motion length, and d_m is the dimension of motions. $M^t_3 \in \mathbb{R}^{N' \times d_m}$ indicates the overlapping part, where N' is the number of interval frame. Each reverse process can be formulated as:

$$\begin{aligned} M^t_{\boldsymbol{c}_1,\boldsymbol{c}_2} =& M^t_{\boldsymbol{c}_1}[:N_1 - N'] \oplus (M^t_{\boldsymbol{c}_1}[N_1 - N':] + \\ & M^t_{\boldsymbol{c}_2}[:N'] - M^t_3) \oplus M^t_{\boldsymbol{c}_2}[N':], \end{aligned} \tag{1}$$

where $M^t_{\boldsymbol{c}_1,\boldsymbol{c}_2}$ is the final score at t-th step, $\oplus$ is the concatenate operation. We conduct experiments on the HumanML3D dataset, and the results are shown in Table 2. We implement MotionDiffuse [12], ReMoDiffuse [13], and FineMoGen [14] using the "first take" from PriorMDM [9]. Our approach is implemented based on Equation 1, and exhibits performance advantages compared with previous methods. We provide visual comparisons with PriorMDM, which can be found on the project page.

2.3. Multi-Concept Motion Generation

In Table 3, we show quantitative results on the MTT [7] dataset. Our approach without Adaptive Gradient Descent (AGD) yields results that are competitive with existing state-of-the-art methods. By combining AGD, our approach achieves superior performance on R-Precision, TMR-Score, and Transition distance.

3. Ablations on Synergistic Energy Fusion

We show the effect of hyper-parameters λ_l, λ_s, and λ_m in Table 4. The results in the first three rows correspond to "Ours (latent only)", "Ours (semantic only)", and "Ours" in Table 3 of the main paper, respectively.

Then we conduct ablative experiments on the weights of the two spectra of the energy-based model (latent-aware and semantic-aware), as shown in the middle five rows (Table 4). We find that as the weight of λ_s increases, the results align more closely with the text (R-Precision and TMR-Score), while larger weights for λ_l produce smoother motions (Transition distance).

Meanwhile, we also demonstrate that combining multi-concept motion generation can further improve the performance, as shown in the last 4 rows. It can be seen that Synergistic Energy Fusion with $\lambda_l = 0.1$, $\lambda_s = 0.7$, and $\lambda_m = 0.2$ achieves best performance.

Note that a more intuitive comparison can be found in Figure 1.

Methods	R-Presicion R@1 ↑	R@3 ↑	TMR-Score ↑ M2T	M2M	FID ↓	Transition distance ↓
MotionDiffuse [12]	10.9	21.3	0.558	0.546	**0.621**	**1.9**
MDM [10]	9.5	19.7	0.556	0.549	0.666	2.5
ReModiffuse [13]	7.4	18.3	0.531	0.534	0.699	3.3
FineMoGen [14]	5.4	11.7	0.504	0.533	0.948	9.4
EnergyMoGen (skeleton)	**11.5**	22.6	0.550	0.549	0.670	2.2
EnergyMoGen (skeleton) + AGD	**11.5**	**24.4**	**0.560**	**0.552**	0.643	**1.9**

Table 3. **Quantitative comparison of skeleton-based diffusion on MTT [7].** We compute metrics following STMC [7]. 'AGD' denotes the adaptive gradient descent

λ_l	λ_s	λ_m	Per-crop semantic correctness R@1 ↑	R@3 ↑	TMR-Score ↑ M2T	M2M	Realism FID ↓	Transition distance ↓
1.0	0.0	0.0	9.7	19.6	0.547	0.521	0.917	1.6
0.0	1.0	0.0	15.1	27.5	0.585	**0.567**	**0.569**	2.2
0.0	0.0	1.0	12.7	25.4	0.570	0.562	0.592	2.7
0.5	0.5	0.0	13.3	25.5	0.584	0.551	0.740	**1.4**
0.4	0.6	0.0	12.7	25.0	0.584	0.555	0.730	**1.4**
0.3	0.7	0.0	13.4	26.4	0.589	0.559	0.694	**1.4**
0.2	0.8	0.0	13.6	26.9	0.590	0.558	0.668	1.5
0.1	0.9	0.0	14.2	27.5	0.587	0.563	0.613	1.7
0.3	0.4	0.3	14.5	27.1	0.588	0.560	0.669	1.6
0.2	0.5	0.3	14.4	26.9	0.590	0.563	0.628	1.6
0.1	0.7	0.2	**15.7**	**28.0**	**0.591**	**0.567**	0.604	1.6
0.1	0.8	0.1	14.9	26.7	0.587	0.563	0.615	1.6

Table 4. **Ablation of hyper-parameters in Synergistic Energy Fusion on MTT [7].** We find that as the weight of λ_s increases, the results align more closely with the text (R-Precision and TMR-Score), while larger weights for λ_l produce smoother motions (Transition distance).

4. Ablations on the Number of Latent Vectors *N* in Motion VAE

The results are provided in Table 5. For reconstruction, 7 latent vectors achieve the best results. However, it increases the difficulty of latent diffusion models in training. Using 5 latent vectors to represent the motion obtains the best text-to-motion generation performance.

5. Ablation Study of Hyper-parameters in Cross-Attention

We investigate the impact of γ_{attn} and γ_{reg} We follow [5] to split γ into attention step size γ_{attn} and regularization step size γ_{reg} for compositional and multi-concept motion generation) in energy-based cross-attention, and the results are presented in Table 6. We notice that $\gamma_{attn}, \gamma_{reg} >= 0.1$ significantly degrades the performance, and $\gamma_{attn}, \gamma_{reg} = [0.001, 0.002]$ achieves best results on MTT.

6. Inference Time

Since our method, like most others, is based on Transformer, we compare its inference time with SOTA Transformer-based diffusion models in Table 7.

7. Evaluation of Foot Sliding

Physical Foot Contact score (PFC), proposed in EDGE [11], is used to evaluate the foot sliding problem. We provide a PFC comparison on MTT in Table 8, demonstrating the effectiveness of the proposed Synergistic Energy Fusion.

8. Energy Distribution Visualization

We show additional contour maps of energy distributions in Figure 2. We provide two examples of the concept conjunc-

N	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity →
	Top-1	Top-2	Top-3			
			Reconstruction			
1	$0.493^{\pm.002}$	$0.681^{\pm.002}$	$0.787^{\pm.003}$	$0.170^{\pm.001}$	$3.160^{\pm.015}$	$\mathbf{9.589}^{\pm.074}$
3	$0.501^{\pm.002}$	$0.696^{\pm.002}$	$0.792^{\pm.004}$	$0.117^{\pm.000}$	$3.037^{\pm.007}$	$9.621^{\pm.091}$
5	$0.508^{\pm.003}$	$0.700^{\pm.003}$	$0.795^{\pm.002}$	$0.080^{\pm.000}$	$3.004^{\pm.008}$	$9.620^{\pm.098}$
7	$\mathbf{0.513}^{\pm.002}$	$\mathbf{0.704}^{\pm.003}$	$\mathbf{0.797}^{\pm.002}$	$\mathbf{0.022}^{\pm.000}$	$\mathbf{2.984}^{\pm.009}$	$9.603^{\pm.085}$
			Generation			
1	$0.498^{\pm.003}$	$0.686^{\pm.004}$	$0.791^{\pm.004}$	$0.424^{\pm.009}$	$3.085^{\pm.009}$	$9.705^{\pm.097}$
3	$0.523^{\pm.004}$	$0.712^{\pm.002}$	$0.814^{\pm.002}$	$0.418^{\pm.025}$	$2.946^{\pm.009}$	$9.443^{\pm.136}$
5	$\mathbf{0.523}^{\pm.003}$	$\mathbf{0.715}^{\pm.002}$	$\mathbf{0.815}^{\pm.002}$	$\mathbf{0.188}^{\pm.006}$	$\mathbf{2.915}^{\pm.007}$	$\mathbf{9.488}^{\pm.099}$
7	$0.514^{\pm.004}$	$0.713^{\pm.005}$	$0.813^{\pm.003}$	$0.291^{\pm.006}$	$2.938^{\pm.012}$	$9.456^{\pm.130}$

Table 5. **Study on the number of latent vectors in motion VAE on the HumanML3D [2] test set.**

γ_{attn}	γ_{reg}	Per-crop semantic correctness				Realism	
		R@1 ↑	R@3 ↑	TMR-Score ↑ M2T	TMR-Score ↑ M2M	FID ↓	Transition distance ↓
0.0	0.0	12.7	25.4	**0.570**	**0.562**	0.592	2.7
0.1	0.2	1.4	4.2	0.498	0.495	1.083	6.1
0.01	0.02	9.1	20.1	0.551	0.547	0.623	2.9
0.005	0.01	6.7	15.3	0.531	0.523	0.806	**1.9**
0.005	0.005	13.8	25.6	**0.570**	0.558	0.591	2.7
0.001	0.002	**14.0**	**26.3**	**0.570**	0.560	**0.587**	2.7

Table 6. **Ablation of step size in Adaptive Gradient Descent on MTT [7].**

Methods	FineMoGen	MLD	GUESS	EnergyMoGen
AIT (s)	2.54	**0.21**	1.79	0.66

Table 7. **Inference time**. AIT (s) denotes the Average inference time per sentence in seconds.

Methods	multi-concept	latent-only	semantic-only	Ours SEF
PFC ↓	0.61	0.54	1.05	**0.51**

Table 8. **Evaluation of Foot Sliding**. 'PFC' denotes the Physical Foot Contact score.

tion. We visualize the energy distributions of motion latent representations generated from the denoising autoencoder. Multi-concept generation combines concepts from both (a) and (b). Compared with multi-concept generation (*i.e.*, (d) in Figure 2), energy distributions of composed motion (*i.e.*, (c) in Figure 2) show consistent high- and low-energy regions. This demonstrates that complex motion latent distributions can be composed of simple distributions, indicating that our method is explainable. Such results further explain the effectiveness of EnergyMoGen.

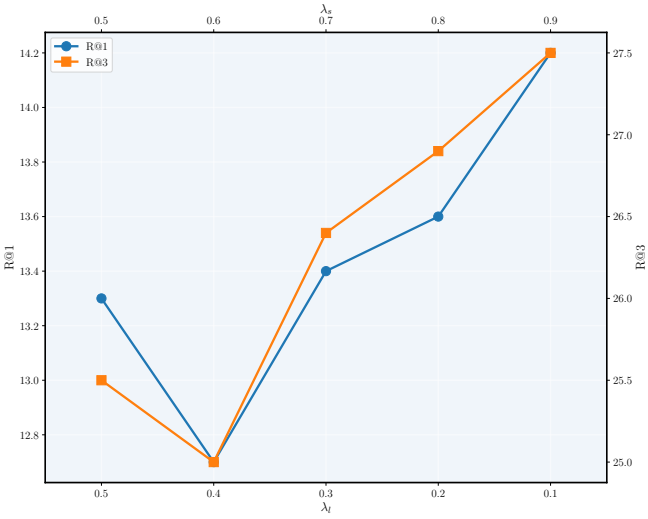
9. More Details on Datasets and Evaluation Metrics

9.1. Datasets

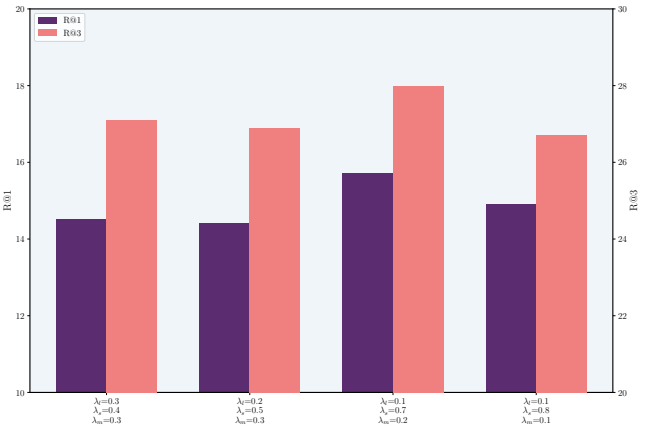
Our experiments are conducted on three datasets: HumanML3D [2], KIT-ML [8], and MTT [7].

- **HumanML3D** [2] is a large-scale text-to-motion generation benchmark that contains 14,616 human motions with 44,970 textual descriptions. The dataset is split with proportions of 80%, 5%, and 15% for training, validation, and testing, respectively.
- **KIT-ML** [8] is another leading benchmark for motion generation from relatively short text. It has 3,911 finely annotated human motions, with 4888/300/830 for the training, validation, and test sets.
- **MTT** [7] has 60 textual descriptions with body part annotations. The corresponding motions are collected from the AMASS dataset [4]. Three texts are randomly composed based on body parts and motion duration through some conjunction words (*e.g.*, "and", "while"), resulting in a test set with 500 samples.

The three datasets use the same motion representation pro-



(a)



(b)

Figure 1. **Ablation of hyper-parameters in Synergistic Energy Fusion on MTT [7].**

posed in [2].

9.2. Evaluation Metrics

We use evaluation models from Guo *et al*. [2] to measure the performance of text-driven human motion generation. We adopt the same metrics as previous works, including Frechet Inception Distance (FID) for motion quality, Retrieval Precision (R-Precision) and Multi-Modal Distance (MM-Dist) for text-motion consistency, and Diversity and MultiModality (MModality) for the diversity of generated motions. We denote two sets of features from the ground truth and generated motion as m and $\hat{m}$, respectively.

FID. The Fréchet Inception Distance (FID) measures the quality of generated motions by comparing their feature distributions to ground truth motions. It evaluates both the mean and covariance. Lower FID scores indicate better quality and closer resemblance to real data. FID can be calculated as:

$$\text{FID} = ||\mu_m - \mu_{\hat{m}}||^2 - \texttt{TR}(\Sigma_m + \Sigma_{\hat{m}} - 2(\Sigma_m \Sigma_{\hat{m}})^{\frac{1}{2}}) \quad (2)$$

where μ_m and $\mu_{\hat{m}}$ are mean of two sets of features. Σ is the covariance matrix.

MM-Dist. MM-Dist is used to measure the distance between the generated motion and text directly:

$$\text{MM-Dist} = \frac{1}{N} \sum_{i=1}^{N} ||m_i - \hat{m}_i|| \quad (3)$$

where N is the total number of the motion.

Diversity. To assess the diversity among motions generated by different textual descriptions in the test set, we randomly select 300 pairs of motions and compute this metric as follows:

$$\text{Diversity} = \frac{1}{300} \sum_{i=1}^{300} ||\hat{m}_1 - \hat{m}_2|| \quad (4)$$

MModality. Similar to **Diversity**, MModality is used to measure the diversity among motions generated by the same text. We follow Guo *et al*. [2] to generate 30 motion samples from one text and randomly select two subsets, each containing 10 motions. The formulation of MModality is similar to the **Diversity** described above.

For compositional motion generation, we follow STMC [7] to use R-Precision, TMR-Score, FID, and transition distance to evaluate the performance. Similar to MM-Dist, TMR-Score computes the cosine similarity between the generated motion embedding and text embedding with TMR model [6]. TEACH [1] calculates Euclidean distance between two consecutive frames as transition distance.

10. Limitation and Failure Case

Existing latent diffusion models encode motion into a single (or a fixed number of) latent vector(s), which limits the use of per-frame composition algorithms. We propose using an energy function to directly model the latent vector(s) encapsulating the overall features, *e.g*., temporal and skeletal features. The energy function (or energy) is additive. This property enables motion composition by composing energy functions (generated latent vectors) from different concepts together via conjunction and negation. However, our method still struggles with completely novel concepts. We provide a failure case in Figure 3.

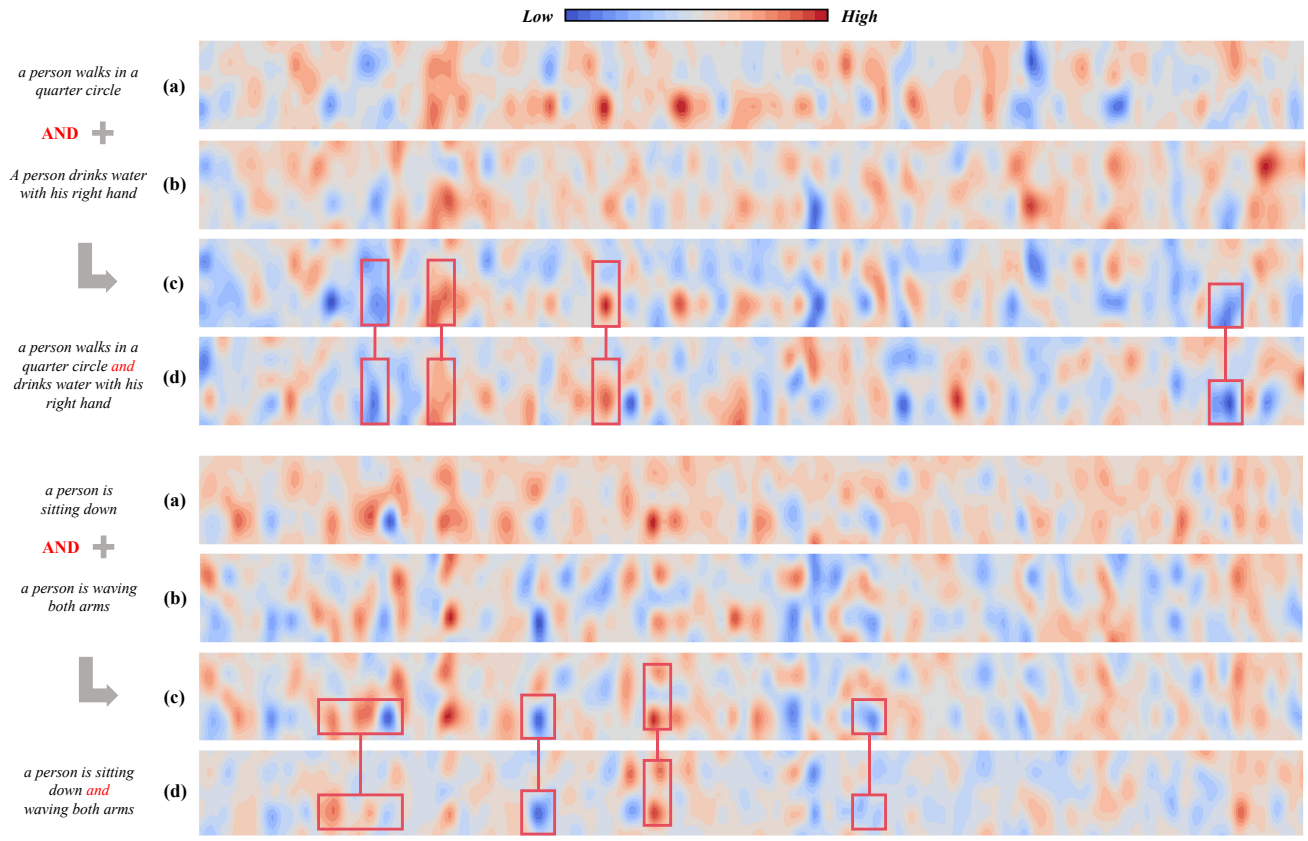


Figure 2. **Visual results of energy distributions.** For a clear illustration, energy distributions are calculated with interpolation and Gaussian smoothing and then visualized as contour maps. (a) Concept 1, (b) Concept 2, (c) Compositional motion generation, (d) Multi-concept motion generation. Similar regions are highlighted in red

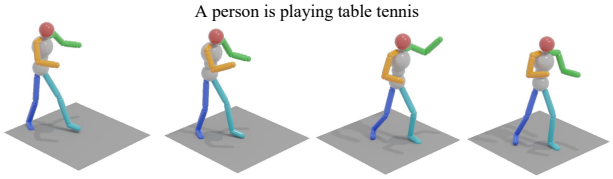


Figure 3. **Failure case.**

References

[1] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. TEACH: Temporal Action Compositions for 3D Humans. In *International Conference on 3D Vision (3DV)*, 2022. 5

[2] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 4, 5

[3] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[4] 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 International Conference on Computer Vision (ICCV)*, 2019. 4

[5] Geon Yeong Park, Jeongsol Kim, Beomsu Kim, Sang Wan Lee, and Jong Chul Ye. Energy-based cross attention for bayesian context update in text-to-image diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[6] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 5

[7] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. 2, 3, 4, 5

[8] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 2016. 4

[9] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior.

In *International Conference on Learning Representations (ICLR)*, 2024. 1, 2

[10] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *arXiv*, 2022. 2, 3

[11] Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. Edge: Editable dance generation from music. *arXiv preprint arXiv:2211.10658*, 2022. 3

[12] 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*, 2022. 1, 2, 3

[13] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. *arXiv*, 2023. 1, 2, 3

[14] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 3

[15] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation. In *European Conference on Computer Vision*, pages 265–282. Springer, 2025. 2