# Towards Decompositional Human Motion Generation with Energy-Based Diffusion Models
## *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/DeMoGen/

In this supplementary material, we present:

- Section 1: Training and inference details of DEMOGEN.
- Section 2: Ablations on the hyper-parameters of DEMOGEN-EXP.
- Section 3: Ablations on orthogonalization loss $\mathcal{L}_{\text{Ortho}}$ weight $\alpha_o$ for DEMOGEN-OSS.
- Section 4: Ablations on orthogonalization loss $\mathcal{L}_{\text{Ortho}}$ weight $\alpha_o$ and semantic consistency loss $\mathcal{L}_{\text{SC}}$ weight $\alpha_{sc}$ for DEMOGEN-SC.
- Section 5: Additional results of multi-concept motion generation.
- Section 6: Ablation studies on the number of decomposed motion concepts $K$.
- Section 7: Additional results on motion temporal composition.
- Section 8: Inference time comparison.
- Section 9: Text-to-motion evaluation on the extended HumanML3D, *i.e.*, DeCompML.
- Section 10: Additional results of our compositional training paradigm on MLD and MotionDiffuse.
- Section 11: Ablations on text embeddings.
- Section 12: Comparison with discrete representation-based text-to-motion models.
- Section 13: More details on datasets and evaluation metrics.

## 1. Implementation Details

### 1.1. Training

For motion VAE, we apply downsampling in both the temporal and spatial dimensions to obtain a latent feature $\boldsymbol{z} \in \mathbb{R}^{L' \times N_j \times d'_m}$, where $L'$ is the latent feature length with a downsampling rate of 4. $N_j$ and $d'_m$ denote the number of joints and latent feature dimension, which are set to 7 and 32, respectively. The VAE is initialized with pretrained weights from [4] for fair comparison. For the diffusion model, we use a 5-layer transformer with a dimension of 256. The learning rate is initialized at 0.0002 and subsequently reduced to 0.00002 following 50,000 training iterations.

On the extended HumanML3D dataset, *i.e.*, DeCompML, we finetune the latent diffusion models of two recent state-of-the-art approaches, EnergyMoGen [10] and SALAD [4], for 100K iterations and 300 epochs. The learning rates are set to 0.00001 and 0.00002, respectively. For the experiments in Section 10, we apply our compositional training paradigm to two classic and representative methods: MLD [8] and MotionDiffuse [11]. We follow their original training configurations.

### 1.2. Inference

During inference, we sample 50 diffusion steps to generate motion from texts. The pseudocode for text-to-motion generation (DEMOGEN-EXP, -OSS, -SC) and compositional motion generation (DEMOGEN-EXP) inference is illustrated in Algorithm 1. Note that DEMOGEN-EXP adopts a text replacement rate $\tau$ during training, enabling high-quality text-to-motion generation by duplicating the single textual description. The ablation study for $\tau$ is provided in Section 2. Our model is trained with $K = 2$, but it can still support a larger number of concept compositions

**Algorithm 1** Text-to-Motion Generation and Compositional Motion Generation

**Require**: Frozen VAE decoder $\mathcal{D}$, denoising network $\epsilon_\theta$, diffusion timestep $T$, text embeddings $\boldsymbol{C} = \{\boldsymbol{c}_k\}_{k=1}^K$, a latent feature $\boldsymbol{z}_T \sim \mathcal{N}(0, 1)$
**for** $t = T, \ldots, 1$ **do**
  **if** mod is **latent-aware** **then**
    $\epsilon_{pred} \leftarrow \frac{1}{K} \sum_{k=1}^K \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{c}_k, t)$
  **end if**
  **if** mod is **semantic-aware** **then**
    $\epsilon_{pred} \leftarrow \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{C}, t)$ *// Replace the cross-attention with DCA*
  **end if**
  *// Run denoising step*
  $\boldsymbol{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{z}_t - \eta_t \epsilon_{pred}) + \mathcal{N}(0, \hat{\beta}_t I)$.
    where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, and $\eta_t = \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}$.
**end for**

**Algorithm 2** Decompositional Motion Generation

**Require**: Frozen VAE decoder $\mathcal{D}$, denoising network $\epsilon_\theta$, diffusion timestep $T$, text embeddings $\boldsymbol{C} = \{\boldsymbol{c}_k\}_{k=1}^K$, a set of latent features $\{\boldsymbol{z}_T^k\}_{k=1}^K$ with $\boldsymbol{z}_T^k \sim \mathcal{N}(0, 1)$
**for** $t = T, \ldots, 1$ **do**
  *// For both latent-aware and semantic-aware settings*
  **for** $k = 1, \ldots, K$ **do**
    $\epsilon_{pred}^k \leftarrow \epsilon_\theta(\boldsymbol{z}_t^k, \boldsymbol{c}_k, t)$
    *// Run denoising step*
    $\boldsymbol{z}_{t-1}^k = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{z}_t^k - \eta_t \epsilon_{pred}^k) + \mathcal{N}(0, \hat{\beta}_t I)$.
      where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, and $\eta_t = \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}$.
  **end for**
**end for**

by averaging energy scores based on Algorithm 1. For example, we directly average the conditional distribution of 3 concepts on the MTT dataset. Furthermore, Algorithm 2 presents the pseudocode for decompositional motion generation, where decomposed concepts can be recombined using Algorithm 1.

## 2. Ablations on DEMOGEN-EXP

We conduct ablative experiments for DEMOGEN-EXP on HumanML3D, DeCompML, and MTT. Text-to-motion results are provided in Table 1. As $\tau$ increases from 0.0 to 0.7, we observe that the model consistently shows improved R-Precision, FID, and MM-Dist. The results for compositional motion generation on the DeCompML dataset (in Table 2) demonstrate similar findings. $\tau$ =0.7 achieves the best results under both latent-aware and semantic-aware settings. We also investigate the impact of $\tau$ on a more complicated benchmark, *i.e.*, MTT. As shown in Table 3, we find that the latent-aware DEMOGEN-EXP significantly outperforms the current state-of-the-art model, *i.e.*, EnergyMoGen [10] across all settings of $\tau$. As for the semantic-aware model, we achieve improvements in most metrics, such as R-Precision, TMR-Score, and Transition distance, except for a decline in FID.

It is worth noting that the semantic-aware model performs better when the input conditions remain within the training distribution (HumanML3D and DeCompML). In contrast, the latent-aware model demonstrates stronger generalization in out-of-domain scenarios (MTT). Please note that all experimental results in Table 1–3 are obtained using models trained on the HumanML3D dataset (with decomposed textual descriptions from DeCompML).

## 3. Ablations on DEMOGEN-OSS

We study the impact of orthogonalization loss weight $\alpha_o$ in DEMOGEN-OSS, and the results are reported in Table 4. All tested values of $\alpha_o$ deliver competitive performance. Notably, $\alpha_o = 2.0$ yields the best performance for the latent-aware configuration, while $\alpha_o = 1.0$ performs best under the semantic-aware setting.

| $\tau$ | R-Precision ↑ | | | FID ↓ | MM-Dist ↓ | Diversity → |
|---|---|---|---|---|---|---|
| | Top-1 | Top-2 | Top-3 | | | |
| ***latent-aware*** | | | | | | |
| 0.0 | $0.532^{\pm.005}$ | $0.726^{\pm.004}$ | $0.820^{\pm.003}$ | $0.255^{\pm.012}$ | $2.926^{\pm.012}$ | $\mathbf{9.707}^{\pm.111}$ |
| 0.3 | $0.562^{\pm.004}$ | $0.756^{\pm.004}$ | $0.844^{\pm.003}$ | $0.108^{\pm.004}$ | $\underline{2.753}^{\pm.017}$ | $\underline{9.766}^{\pm.104}$ |
| 0.5 | $\underline{0.565}^{\pm.004}$ | $\underline{0.757}^{\pm.003}$ | $\underline{0.845}^{\pm.003}$ | $\underline{0.095}^{\pm.004}$ | $2.757^{\pm.011}$ | $9.849^{\pm.102}$ |
| 0.7 | $\mathbf{0.569}^{\pm.004}$ | $\mathbf{0.760}^{\pm.005}$ | $\mathbf{0.850}^{\pm.004}$ | $\mathbf{0.078}^{\pm.003}$ | $\mathbf{2.708}^{\pm.012}$ | $9.774^{\pm.099}$ |
| ***semantic-aware*** | | | | | | |
| 0.0 | $0.564^{\pm.003}$ | $0.756^{\pm.004}$ | $0.845^{\pm.002}$ | $0.297^{\pm.016}$ | $2.764^{\pm.012}$ | $\mathbf{9.784}^{\pm.057}$ |
| 0.3 | $0.579^{\pm.004}$ | $0.769^{\pm.003}$ | $0.858^{\pm.003}$ | $0.118^{\pm.007}$ | $2.694^{\pm.009}$ | $9.858^{\pm.054}$ |
| 0.5 | $\underline{0.583}^{\pm.002}$ | $\underline{0.774}^{\pm.004}$ | $\mathbf{0.865}^{\pm.002}$ | $\mathbf{0.100}^{\pm.006}$ | $\underline{2.632}^{\pm.006}$ | $\underline{9.807}^{\pm.063}$ |
| 0.7 | $\mathbf{0.586}^{\pm.005}$ | $\mathbf{0.776}^{\pm.003}$ | $\underline{0.863}^{\pm.002}$ | $\underline{0.116}^{\pm.008}$ | $\mathbf{2.623}^{\pm.008}$ | $9.873^{\pm057}$ |

Table 1. **Ablation of text replacement rate $\tau$ in DEMOGEN-EXP for text-to-motion on the HumanML3D test set. Bold** and underlined denote the best and second-best results, respectively.

| $\tau$ | R-Precision ↑ | | | FID ↓ | MM-Dist ↓ | Diversity → |
|---|---|---|---|---|---|---|
| | Top-1 | Top-2 | Top-3 | | | |
| ***latent-aware*** | | | | | | |
| 0.0 | $0.544^{\pm.005}$ | $0.738^{\pm.005}$ | $0.830^{\pm.003}$ | $\mathbf{0.089}^{\pm.006}$ | $2.850^{\pm.010}$ | $\mathbf{9.703}^{\pm.090}$ |
| 0.3 | $\underline{0.551}^{\pm.005}$ | $\underline{0.743}^{\pm.006}$ | $\underline{0.834}^{\pm.004}$ | $0.109^{\pm.007}$ | $\underline{2.817}^{\pm.023}$ | $\underline{9.735}^{\pm.104}$ |
| 0.5 | $0.550^{\pm.004}$ | $\underline{0.743}^{\pm.002}$ | $\underline{0.834}^{\pm.003}$ | $\underline{0.093}^{\pm.007}$ | $\underline{2.817}^{\pm.015}$ | $9.803^{\pm.105}$ |
| 0.7 | $\mathbf{0.559}^{\pm.004}$ | $\mathbf{0.754}^{\pm.004}$ | $\mathbf{0.842}^{\pm.004}$ | $\mathbf{0.089}^{\pm.005}$ | $\mathbf{2.758}^{\pm.014}$ | $9.756^{\pm.115}$ |
| ***semantic-aware*** | | | | | | |
| 0.0 | $0.547^{\pm.004}$ | $0.737^{\pm.002}$ | $0.827^{\pm.004}$ | $0.163^{\pm.010}$ | $2.826^{\pm.007}$ | $9.864^{\pm.056}$ |
| 0.3 | $0.561^{\pm.005}$ | $0.750^{\pm.003}$ | $0.841^{\pm.003}$ | $0.111^{\pm.006}$ | $2.765^{\pm.010}$ | $9.809^{\pm.062}$ |
| 0.5 | $\underline{0.565}^{\pm.004}$ | $\underline{0.757}^{\pm.004}$ | $\underline{0.845}^{\pm.003}$ | $\mathbf{0.095}^{\pm.006}$ | $\underline{2.720}^{\pm.012}$ | $\mathbf{9.771}^{\pm.056}$ |
| 0.7 | $\mathbf{0.567}^{\pm.004}$ | $\mathbf{0.760}^{\pm.003}$ | $\mathbf{0.846}^{\pm.002}$ | $\underline{0.102}^{\pm.007}$ | $\mathbf{2.719}^{\pm.009}$ | $\underline{9.786}^{\pm.058}$ |

Table 2. **Ablation of text replacement rate $\tau$ in DEMOGEN-EXP for motion composition on DeCompML.**

## 4. Ablations on Loss Weights $\alpha_{sc}$ and $\alpha_o$ for DEMOGEN-SC

The results are provided in Table 5. We first conduct an ablation study on the scaling loss weight $\alpha_{sc}$, building on the findings presented in Section 3. We then further examine the performance of the model without orthogonalization loss. For the latent-aware DEMOGEN-SC, although setting $\alpha_{\text{sc}}$ to 1 slightly degrades performance, we retain this value to encourage the model to learn from the decomposed text. Conversely, for the semantic-aware model, setting $\alpha_{\text{sc}}$ to 1 achieves the best performance. Meanwhile, the performance decreases when the orthogonalization loss is removed, further demonstrating its effectiveness.

## 5. Additional Results of Multi-concept Motion Generation

The results are shown in Table 6. We observe that the semantic-aware setting surpasses the latent-aware setting on both DEMOGEN-OSS and DEMOGEN-SC, indicating its stronger ability to generate motions from a single complex textual description. Please refer to Table 3 for the results and analysis of motion composition.

## 6. Ablation Studies on the Number of Decomposed Motion Concepts $K$

We explore the effect of the number of decomposed motion concepts $K$ on HumanML3D. GPT-4.1 is used to generate decomposed texts with three or four components (see Sec-

| Methods | R-Precision R@1 ↑ | R@3 ↑ | TMR-Score ↑ M2T | M2M | FID ↓ | Transition distance ↓ |
|---|---|---|---|---|---|---|
| ***latent-aware*** | | | | | | |
| EnergyMoGen (latent) | 9.7 | 19.6 | 0.547 | 0.521 | 0.917 | 1.6 |
| DEMOGEN-EXP ($\tau$=0.0) | **16.5** | **31.9** | 0.594 | 0.566 | 0.630 | 1.7 |
| DEMOGEN-EXP ($\tau$=0.3) | 16.3 | 31.7 | 0.594 | **0.570** | **0.607** | 1.6 |
| DEMOGEN-EXP ($\tau$=0.5) | 15.3 | 31.1 | 0.596 | 0.563 | 0.648 | **1.5** |
| DEMOGEN-EXP ($\tau$=0.7) | 16.2 | **31.9** | **0.597** | **0.570** | 0.621 | 1.6 |
| ***semantic-aware*** | | | | | | |
| EnergyMoGen (semantic) | 15.1 | 27.5 | 0.585 | 0.567 | **0.569** | 2.2 |
| DEMOGEN-EXP ($\tau$=0.0) | 13.5 | 27.1 | 0.577 | 0.560 | 0.606 | 2.1 |
| DEMOGEN-EXP ($\tau$=0.3) | 14.1 | 28.2 | 0.581 | 0.563 | 0.648 | **1.8** |
| DEMOGEN-EXP ($\tau$=0.5) | **15.2** | **30.1** | **0.594** | 0.564 | 0.643 | **1.8** |
| DEMOGEN-EXP ($\tau$=0.7) | 14.8 | 29.3 | 0.584 | **0.568** | 0.631 | **1.8** |

Table 3. **Quantitative comparison on the MTT [6] dataset.** We compare our approach with EnergyMoGen and analyze the impact of the text replacement rate $\tau$.

| $\alpha_o$ | R-Precision ↑ Top-1 | Top-2 | Top-3 | FID ↓ | MM-Dist ↓ | Diversity → |
|---|---|---|---|---|---|---|
| ***latent-aware*** | | | | | | |
| 0.0 | $0.582^{\pm.005}$ | $0.771^{\pm.003}$ | $0.855^{\pm.003}$ | $0.138^{\pm.005}$ | $2.642^{\pm.008}$ | $9.851^{\pm.147}$ |
| 0.1 | $0.584^{\pm.007}$ | $0.771^{\pm.005}$ | $0.857^{\pm.001}$ | $0.097^{\pm.004}$ | $2.641^{\pm.016}$ | $9.781^{\pm.096}$ |
| 0.5 | $0.584^{\pm.003}$ | $0.775^{\pm.004}$ | $0.860^{\pm.003}$ | $0.092^{\pm.005}$ | $2.629^{\pm.009}$ | $\mathbf{9.729}^{\pm.145}$ |
| 1.0 | $0.583^{\pm.003}$ | $0.772^{\pm.005}$ | $0.857^{\pm.003}$ | $\mathbf{0.074}^{\pm.004}$ | $2.635^{\pm.009}$ | $9.799^{\pm.107}$ |
| 2.0 | $\mathbf{0.588}^{\pm.004}$ | $\mathbf{0.778}^{\pm.002}$ | $\mathbf{0.861}^{\pm.003}$ | $0.092^{\pm.003}$ | $\mathbf{2.625}^{\pm.007}$ | $9.779^{\pm.120}$ |
| ***semantic-aware*** | | | | | | |
| 0.0 | $0.583^{\pm.004}$ | $0.771^{\pm.003}$ | $0.856^{\pm.003}$ | $\mathbf{0.093}^{\pm.005}$ | $2.649^{\pm.009}$ | $\mathbf{9.797}^{\pm.123}$ |
| 0.1 | $0.583^{\pm.005}$ | $\mathbf{0.776}^{\pm.002}$ | $\mathbf{0.861}^{\pm.003}$ | $0.113^{\pm.005}$ | $2.648^{\pm.011}$ | $9.928^{\pm.112}$ |
| 0.5 | $0.583^{\pm.003}$ | $\mathbf{0.776}^{\pm.003}$ | $\mathbf{0.861}^{\pm.002}$ | $0.112^{\pm.005}$ | $2.662^{\pm.010}$ | $9.929^{\pm.126}$ |
| 1.0 | $\mathbf{0.584}^{\pm.002}$ | $0.774^{\pm.003}$ | $0.858^{\pm.001}$ | $0.104^{\pm.005}$ | $\mathbf{2.637}^{\pm.014}$ | $9.877^{\pm.127}$ |
| 2.0 | $0.578^{\pm.004}$ | $0.769^{\pm.003}$ | $0.855^{\pm.004}$ | $0.107^{\pm.004}$ | $2.665^{\pm.010}$ | $9.905^{\pm.114}$ |

Table 4. **Ablation of orthogonalization loss weight $\alpha_o$ for DEMOGEN-OSS on the HumanML3D test set.** We find that $\alpha_o$ =2 and $\alpha_o$ =1 yield the best performance for latent-aware and semantic-aware DEMOGEN-OSS, respectively.

tion 13 for more details), and a latent-aware DEMOGEN-EXP is trained accordingly. As shown in Table 7, with an increasing number of decomposed motion concepts $K$, several key metrics, including FID and R-Precision, consistently decline on DEMOGEN-EXP. We believe that the decrease is mainly due to the limited information in HumanML3D texts, which constrains the extraction of high-quality components under three- or four-part decompositions. To verify this, we further conduct experiments on DEMOGEN-OSS, which does not require decomposed textual instructions. We find that $K = 4$ achieves performance comparable to that of $K = 2$. Furthermore, we conduct experiments on energy aggregation methods, *i.e.*, summation and averaging over $K$ concepts. Table 8 shows that, compared with mean aggregation, summation leads to worse FID.

| $\alpha_o$ | $\alpha_{sc}$ | R-Precision ↑ Top-1 | Top-2 | Top-3 | FID ↓ | MM-Dist ↓ | Diversity → |
|---|---|---|---|---|---|---|---|
| | | | ***latent-aware*** | | | | |
| 2.0 | 0.1 | $\mathbf{0.569}^{\pm.002}$ | $\mathbf{0.761}^{\pm.004}$ | $\mathbf{0.847}^{\pm.006}$ | $\mathbf{0.112}^{\pm.006}$ | $\mathbf{2.712}^{\pm.016}$ | $\mathbf{9.763}^{\pm.125}$ |
| 2.0 | 0.5 | $0.560^{\pm.004}$ | $0.756^{\pm.003}$ | $\underline{0.846}^{\pm.002}$ | $0.152^{\pm.004}$ | $2.754^{\pm.010}$ | $9.935^{\pm.116}$ |
| 2.0 | 1.0 | $\underline{0.565}^{\pm.003}$ | $\underline{0.757}^{\pm.004}$ | $\underline{0.846}^{\pm.003}$ | $\underline{0.121}^{\pm.005}$ | $\underline{2.739}^{\pm.009}$ | $\underline{9.933}^{\pm.131}$ |
| 0.0 | 1.0 | $0.560^{\pm.004}$ | $0.752^{\pm.004}$ | $0.845^{\pm.003}$ | $0.185^{\pm.008}$ | $2.742^{\pm.006}$ | $9.976^{\pm.151}$ |
| | | | ***semantic-aware*** | | | | |
| 1.0 | 0.1 | $\underline{0.564}^{\pm.003}$ | $\underline{0.755}^{\pm.006}$ | $\underline{0.845}^{\pm.005}$ | $\underline{0.132}^{\pm.007}$ | $2.741^{\pm.016}$ | $9.954^{\pm.182}$ |
| 1.0 | 0.5 | $0.557^{\pm.003}$ | $0.751^{\pm.002}$ | $0.841^{\pm.003}$ | $0.174^{\pm.006}$ | $2.774^{\pm.012}$ | $9.924^{\pm.154}$ |
| 1.0 | 1.0 | $\mathbf{0.565}^{\pm.002}$ | $\mathbf{0.758}^{\pm.003}$ | $\mathbf{0.846}^{\pm.003}$ | $0.138^{\pm.005}$ | $\mathbf{2.727}^{\pm.010}$ | $\underline{9.769}^{\pm.158}$ |
| 0.0 | 1.0 | $0.561^{\pm.003}$ | $\underline{0.755}^{\pm.004}$ | $0.844^{\pm.003}$ | $\mathbf{0.080}^{\pm.004}$ | $\underline{2.734}^{\pm.009}$ | $\mathbf{9.712}^{\pm.145}$ |

Table 5. **Ablation of loss hyper-parameters in DEMOGEN-SC on the HumanML3D test set.**

| Methods | R-Precision R@1 ↑ | R@3 ↑ | TMR-Score ↑ M2T | M2M | FID ↓ | Transition distance ↓ |
|---|---|---|---|---|---|---|
| Multi-concept motion generation (single text) | | | | | | |
| DEMOGEN-OSS(latent) | <u>14.9</u> | <u>29.5</u> | <u>0.584</u> | **0.57** | <u>0.580</u> | <u>2.6</u> |
| DEMOGEN-OSS (semantic) | **15.4** | 29.2 | **0.585** | <u>0.571</u> | 0.598 | **2.4** |
| DEMOGEN-SC (latent) | 14.3 | **29.7** | 0.578 | 0.568 | 0.585 | 2.7 |
| DEMOGEN-SC (semantic) | 14.8 | 29.1 | 0.578 | 0.569 | **0.575** | <u>2.6</u> |

Table 6. **Quantitative results on MTT [6].** We present the multi-concept generation results of the semantic-aware DEMOGEN. The metrics are computed following STMC [6].

## 7. Additional Results on Motion Temporal Composition

We also present results for long-motion generation on the HumanML3D dataset. The results in Table 9 show that our model outperforms existing state-of-the-art approaches. Our approach is implemented using the "first take" from PriorMDM [7]. Please note that InfiniDreamer [12] is specifically designed for long-motion generation, while our approach offers a general training paradigm that improves multiple tasks.

## 8. Inference Time Comparison

We compare the inference time of our approach with existing state-of-the-art methods in Table 10. All experiments are conducted using a batch size of 1 on a single V100 GPU.

## 9. Text-to-motion Evaluation on the DeCompML Dataset

In Section 4.5 of the main paper, we demonstrate that leveraging the decomposed motions as additional training data can improve text-to-motion performance. In Table 11, we provide complementary evaluation metrics and further validate our dataset by presenting results on a classical VQ-VAE–based approach, *i.e.*, T2M-GPT [9]. The improvement for SALAD is mainly reflected in the FID metric, which we attribute to its already strong performance in terms of R-precision and MM-Dist. Both T2M-GPT and EnergyMoGen show improvements in motion smoothness (FID) and text–motion consistency (R-precision and MM-Dist).

## 10. Additional Results of Our Compositional Training Paradigm on MLD and MotionDiffuse

To comprehensively validate our approach, we apply the compositional training paradigm to MLD [8] and MotionDiffuse [11], the results are presented in Table 12. We train both models based on orthogonal self-supervision (OSS) under a latent-aware setting. Notably, we find that MotionDiffuse achieves better performance when predicting $\boldsymbol{x}_0$. The experimental results demonstrate that our approach serves as a general training paradigm, which can be seam-

| $K$ | R-Precision ↑ | | | FID ↓ | MM-Dist ↓ | Diversity → |
|---|---|---|---|---|---|---|
| | Top-1 | Top-2 | Top-3 | | | |
| DEMOGEN-EXP | | | | | | |
| 2 | $\mathbf{0.569}^{\pm.004}$ | $\mathbf{0.760}^{\pm.005}$ | $\mathbf{0.850}^{\pm.004}$ | $\mathbf{0.078}^{\pm.003}$ | $\mathbf{2.708}^{\pm.012}$ | $\underline{9.774}^{\pm.099}$ |
| 3 | $\underline{0.556}^{\pm.005}$ | $\underline{0.749}^{\pm.005}$ | $\underline{0.840}^{\pm.003}$ | $\underline{0.110}^{\pm.005}$ | $\underline{2.786}^{\pm.016}$ | $9.822^{\pm.133}$ |
| 4 | $0.553^{\pm.005}$ | $0.748^{\pm.004}$ | $0.836^{\pm.002}$ | $0.157^{\pm.005}$ | $2.795^{\pm.011}$ | $\mathbf{9.743}^{\pm.116}$ |
| DEMOGEN-OSS | | | | | | |
| 2 | $\mathbf{0.588}^{\pm.004}$ | $\mathbf{0.778}^{\pm.002}$ | $\underline{0.861}^{\pm.003}$ | $\mathbf{0.092}^{\pm.003}$ | $\mathbf{2.625}^{\pm.007}$ | $\mathbf{9.779}^{\pm.120}$ |
| 4 | $\underline{0.584}^{\pm.007}$ | $\underline{0.776}^{\pm.002}$ | $\mathbf{0.862}^{\pm.002}$ | $\underline{0.100}^{\pm.006}$ | $\underline{2.636}^{\pm.016}$ | $\underline{9.801}^{\pm.247}$ |

Table 7. **Ablation on the number of decomposed concepts $K$ on the HumanML3D test set.**

| $K$ | R-Precision ↑ | | | FID ↓ | MM-Dist ↓ | Diversity → |
|---|---|---|---|---|---|---|
| | Top-1 | Top-2 | Top-3 | | | |
| mean value | $0.569^{\pm.004}$ | $0.760^{\pm.005}$ | $0.850^{\pm.004}$ | $0.078^{\pm.003}$ | $2.708^{\pm.012}$ | $9.774^{\pm.099}$ |
| summation | $0.568^{\pm.004}$ | $0.760^{\pm.003}$ | $0.850^{\pm.002}$ | $0.119^{\pm.011}$ | $2.713^{\pm.007}$ | $9.831^{\pm.149}$ |

Table 8. **Ablation study of different energy aggregation methods.**

| Methods | R-Precision ↑ | FID ↓ (motion) | FID ↓ (transition) |
|---|---|---|---|
| PriorMDM* | 0.603 ± 0.009 | 1.36 ± 0.03 | 3.19 ± 0.29 |
| InfiniDreamer* | 0.679 ± 0.007 | 0.47 ± 0.12 | 2.04 ± 0.28 |
| latent-Exp (Ours) | 0.679 ± 0.008 | 0.34 ± 0.05 | 2.01 ± 0.17 |

Table 9. **Quantitative results on HumanML3D.** $*$ results are from InfiniDreamer [12]. R-Precision denotes the Top-3 accuracy.

| Methods | FineMoGen | EMG | SALAD | semantic-Exp | latent-Exp |
|---|---|---|---|---|---|
| AIT (s) | 2.54 | 0.66 | **0.61** | 0.80 | <u>0.65</u> |

Table 10. **Inference time.** AIT (s) indicates Average Inference Time per sentence in seconds.

lessly integrated with existing diffusion models.

## 11. Ablations on Text Embedding

For DEMOGEN-OSS and DEMOGEN-SC, we split word-level text features along feature dimensions and introduce additional layers (*e.g.*, Transformers and MLPs) to learn the decomposition patterns. Results of latent-aware DEMOGEN-OSS in Table 13 show that global embeddings largely degrade performance, whereas using two separate networks on non-split embeddings (with $\mathcal{L}_{\text{Ortho}}$) results in a minor drop.

## 12. Comparison with Discrete Representation-based Text-to-Motion Models

Since our method is diffusion-based, the main paper only includes comparisons with diffusion-based methods. Table 14 shows that our approach also largely outperforms discrete representation-based methods, *i.e.*, M2D2M [1] and MoMask [3], in Top-1,2,3 accuracy and MM-Dist, while achieving comparable FID.

## 13. More Details on Datasets and Evaluation Metrics

### 13.1. DeCompML

This section describes how decomposed text in DeCompML is generated. We utilize a large language model and design prompts to automatically perform the decomposition of the given text. The text prompt is shown in Table 15. We experiment with multiple large language models, including GPT-4o, GPT-4.1, GPT-5, and GPT-OSS-120B (OpenAI's 120B open-source model). GPT-OSS-120B and GPT-4o often fail to effectively decompose the text, frequently producing empty outputs. GPT-5 requires longer inference time and sometimes generates additional irrelevant content. Therefore, we ultimately select GPT-4.1 for our decomposition tasks.

### 13.2. Evaluation Metrics

We evaluate text-driven human motion generation using the models and metrics from Guo *et al.* [2]. Motion quality

| Method | R-Precision ↑ | | | FID ↓ | MM-Dist ↓ | Diversity → |
|---|---|---|---|---|---|---|
| | Top-1 | Top-2 | Top-3 | | | |
| T2M-GPT [9] | $0.491^{\pm.003}$ | $0.680^{\pm.003}$ | $0.775^{\pm.002}$ | $0.116^{\pm.004}$ | $3.118^{\pm.011}$ | $9.761^{\pm.081}$ |
| T2M-GPT [9]* | $0.497^{\pm.004}$ | $0.685^{\pm.002}$ | $0.780^{\pm.003}$ | $0.102^{\pm.008}$ | $3.017^{\pm.007}$ | $9.748^{\pm.074}$ |
| EnergyMoGen [10] | $0.523^{\pm.003}$ | $0.715^{\pm.002}$ | $0.815^{\pm.002}$ | $0.188^{\pm.006}$ | $2.915^{\pm.007}$ | $9.488^{\pm.099}$ |
| EnergyMoGen [10]* | $0.526^{\pm.004}$ | $0.718^{\pm.002}$ | $0.818^{\pm.003}$ | $0.147^{\pm.004}$ | $2.884^{\pm.009}$ | $9.392^{\pm.084}$ |
| SALAD [4] | $0.581^{\pm.003}$ | $0.769^{\pm.003}$ | $0.857^{\pm.002}$ | $0.076^{\pm.002}$ | $2.649^{\pm.009}$ | $9.696^{\pm.096}$ |
| SALAD [4]* | $0.580^{\pm.003}$ | $0.769^{\pm.003}$ | $0.857^{\pm.003}$ | $0.060^{\pm.005}$ | $2.651^{\pm.009}$ | $9.379^{\pm.149}$ |

Table 11. **Quantitative comparison on the HumanML3D test set.** $*$ denotes that the model is finetuned on DeCompML.

| Method | R-Precision ↑ | | | FID ↓ | MM-Dist ↓ | Diversity → |
|---|---|---|---|---|---|---|
| | Top-1 | Top-2 | Top-3 | | | |
| MLD [8] | $0.481^{\pm.003}$ | $0.673^{\pm.003}$ | $0.772^{\pm.002}$ | $0.473^{\pm.013}$ | $3.196^{\pm.010}$ | $9.724^{\pm.082}$ |
| DEMOGEN-MLD [8] | $0.490^{\pm.003}$ | $0.680^{\pm.004}$ | $0.776^{\pm.003}$ | $0.288^{\pm.005}$ | $3.124^{\pm.011}$ | $9.785^{\pm.115}$ |
| MotionDiffuse [11] | $0.491^{\pm.001}$ | $0.681^{\pm.001}$ | $0.782^{\pm.001}$ | $0.630^{\pm.001}$ | $3.113^{\pm.001}$ | $9.410^{\pm.049}$ |
| MotionDiffuse [11]§ | $0.523^{\pm.002}$ | $0.714^{\pm.004}$ | $0.807^{\pm.002}$ | $0.287^{\pm.005}$ | $2.912^{\pm.010}$ | $9.456^{\pm.107}$ |
| DEMOGEN-MotionDiffuse [11]§ | $0.527^{\pm.002}$ | $0.719^{\pm.003}$ | $0.820^{\pm.004}$ | $0.141^{\pm.006}$ | $2.908^{\pm.009}$ | $9.628^{\pm.096}$ |

Table 12. **Quantitative results on the HumanML3D test set.** § indicates that we train MotionDiffuse using an $\boldsymbol{x}_0$-prediction objective and apply DDIM for inference.

is measured by FID, text-motion alignment by R-Precision and MM-Dist, and diversity by Diversity and Multimodality metrics. The feature sets of ground-truth and generated motions are denoted as $\boldsymbol{m}$ and $\hat{\boldsymbol{m}}$, respectively.

**R-Precision.** For each motion sequence, we use 32 text descriptions (one ground-truth and 31 randomly selected mismatched) and rank them based on the Euclidean distance between motion and text embeddings. We report Top-1, Top-2, and Top-3 accuracy for motion-to-text retrieval.

**FID.** The Fréchet Inception Distance (FID) serves as a key indicator of the realism and quality of synthetic motions. It calculates the statistical difference between the feature space of the generated samples ($\hat{\boldsymbol{m}}$) and the real data ($\boldsymbol{m}$):

$$\text{FID} = ||\mu_{\hat{\boldsymbol{m}}} - \mu_{\boldsymbol{m}}||^2 + \text{TR}(\Sigma_{\hat{\boldsymbol{m}}} + \Sigma_{\boldsymbol{m}} - 2\sqrt{\Sigma_{\hat{\boldsymbol{m}}}\Sigma_{\boldsymbol{m}}}). \quad (1)$$

Here, $\mu$ represents the feature set mean, and $\Sigma$ is the covariance matrix. A key property is that lower values of FID imply higher fidelity to the real data distribution.

**MM-Dist.** We compute the average Euclidean distance between each text embedding and the corresponding motion embedding generated from that text, providing a measure of alignment between text and motion features.

**Diversity.** We quantify the variation of motions generated from different text descriptions in the test set by randomly sampling 300 motion pairs. For each pair, we compute the Euclidean distance between their feature representations, and define the Diversity metric as:

$$\text{Diversity} = \frac{1}{300}\sum_{i=1}^{300}\left\|\hat{\boldsymbol{m}}_1^{(i)} - \hat{\boldsymbol{m}}_2^{(i)}\right\|, \quad (2)$$

where $\hat{\boldsymbol{m}}_1^{(i)}$ and $\hat{\boldsymbol{m}}_2^{(i)}$ denote the motion features of the $i$-th pair.

**Multimodality.** To evaluate the variation among motions generated from the same textual description, we adopt a procedure similar to Diversity. Following Guo *et al*. [2], we generate multiple samples per text and partition them into two random subsets. The Multimodality metric is calculated as the average Euclidean distance between paired features from these subsets, following the same formulation as the Diversity metric.

Following STMC [6] and EnergyMoGen [10], we evaluate compositional motion generation using R-Precision, TMR-Score, FID, and Transition Distance. The TMR-Score quantifies motion-text alignment via the cosine similarity of embeddings derived from the TMR model [5], similar to MM-Dist. Transition Distance is computed as the Euclidean distance between consecutive frames.

| Methods | Top1↑ | Top2↑ | Top3↑ | FID↓ | MM Dist↓ | Diversity → |
|---|---|---|---|---|---|---|
| global embed | $0.517^{\pm.004}$ | $0.712^{\pm.003}$ | $0.806^{\pm.001}$ | $0.187^{\pm.004}$ | $3.050^{\pm.021}$ | $9.845^{\pm.035}$ |
| w/o embed split | $0.585^{\pm.002}$ | $0.774^{\pm.003}$ | $0.857^{\pm.001}$ | $0.112^{\pm.005}$ | $2.641^{\pm.012}$ | $9.878^{\pm.108}$ |
| DeMoGen-OSS (latent) | $\mathbf{0.588}^{\pm.004}$ | $\mathbf{0.778}^{\pm.002}$ | $\mathbf{0.861}^{\pm.003}$ | $\mathbf{0.092}^{\pm.003}$ | $\mathbf{2.625}^{\pm.007}$ | $\mathbf{9.779}^{\pm.120}$ |

Table 13. **Ablations on different text embeddings.**

| Methods | Top1↑ | Top2↑ | Top3↑ | FID↓ | MM Dist↓ | Diversity→ |
|---|---|---|---|---|---|---|
| MoMask [3] | $0.521^{\pm.002}$ | $0.713^{\pm.002}$ | $0.807^{\pm.002}$ | $0.045^{\pm.002}$ | $2.958^{\pm.008}$ | - |
| M2D2M [1] | - | - | $0.799^{\pm.002}$ | $0.087^{\pm.004}$ | $3.018^{\pm.008}$ | $9.672^{\pm.086}$ |
| DeMoGen-Exp (latent) | $0.569^{\pm.004}$ | $0.760^{\pm.005}$ | $0.850^{\pm.004}$ | $0.078^{\pm.003}$ | $2.708^{\pm.012}$ | $9.774^{\pm.099}$ |

Table 14. **Comparison with discrete representation-based models on the HumanML3D [2] test set.**

## References


[1] Seunggeun Chi, Hyung-gun Chi, Hengbo Ma, Nakul Agarwal, Faizan Siddiqui, Karthik Ramani, and Kwonjoon Lee. M2d2m: Multi-motion generation from text with discrete diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 6, 8

[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. 6, 7, 8

[3] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6, 8

[4] Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 7

[5] 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. 7

[6] 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. 4, 5, 7

[7] 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. 5

[8] Chen Xin, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. *arXiv*, 2022. 1, 5, 7

[9] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5, 7

[10] Jianrong Zhang, Hehe Fan, and Yi Yang. Energymogen: Compositional human motion generation with energy-based diffusion model in latent space. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 7

[11] 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, 5, 7

[12] Wenjie Zhuo, Fan Ma, and Hehe Fan. Infinidreamer: Arbitrarily long human motion generation via segment score distillation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025. 5, 6

**Prompt:**
I have a text description of a human motion. Your task is to split this description into exactly two separate sentences.

Each sentence should describe one distinct human motion. If the original text contains style, emotion, speed, or environment information, you must preserve it in the corresponding sentences. No additional information should be added if it is absent in the original text.

Focus on accurately capturing the original meaning. The two sentences, when combined as sequential or simultaneous motions, should reproduce the meaning of the original description.

STRICT OUTPUT FORMAT (must follow exactly):
- Return ONLY the two sentences joined by a single "#".
- Both `<sentence1>` and `<sentence2>` must be non-empty and contain meaningful words (not just spaces or punctuation).
- No labels, no explanations, no numbering, no quotes, no code fences, no extra spaces, and do not repeat the input.
- The output must match the pattern: `<sentence1>#<sentence2>`

Example:
Input: `a person is walking forward while waving his left hand.`
Output: `a person is walking forward#a person is waving his left hand.`
Now split the following sentence:

Table 15. **Prompt for splitting human motion descriptions into two distinct sentences.**