

Supplementary Materials

Anonymous WACV Algorithms Track submission

Paper ID 2310

1. Experiment Derails

In this chapter, we will provide more experimental details, including the detailed introduction of the evaluation metrics and datasets used in the experiments.

1.1. Evaluation Metric

1.1.1 Unconditional Text Generation

When assessing the motion-to-text task, we mainly focus on metrics such as Length, Bleu [2], Rouge [1], Cider [3], and BertScore [4]. Length: This metric calculates the average length of the generated text, providing insights into text length consistency. Bleu: BLEU measures the similarity between the generated text and reference text based on n-gram overlap. Rouge: ROUGE evaluates the quality of the generated text by comparing it with reference text in terms of recall of n-grams. Cider: CIDEr assesses the quality of generated text by considering consensus among human annotators. BertScore: BERTScore evaluates the quality of generated text by comparing it with reference text using contextual embeddings from pre-trained BERT models.

1.1.2 Motion Completion

For the motion completion task, we also utilize metrics like Average Displacement Error (ADE) and Final Displacement Error (FDE). ADE: ADE measures the average deviation between the predicted positions of key points or joints in a motion sequence and their actual positions. It calculates the average Euclidean distance of corresponding points across all frames in the sequence. FDE: FDE quantifies the disparity between the predicted position of the last frame in a motion sequence and its actual position. It quantifies the accuracy of the model in predicting the final state of the motion.

1.2. Dataset

1.2.1 ROCStories

The ROCStories dataset serves as a benchmark dataset for evaluating natural language understanding and generation

models. It comprises 98,162 stories, each consisting of 5 sentences. These stories aim to capture common sense, emotions, and temporal relationships expressed in everyday life.

1.2.2 AG News Topic Classification

The AG News Topic Classification dataset is a collection of news articles designed for text classification tasks. It encompasses four main topic categories: World, Sports, Business, and Technology and comprises a total of 120,000 training instances, each containing article titles and descriptions.

1.2.3 Motion-X

The inclusion of Motion-X, 52.4k motions annotated by more detailed textual descriptions, aims to enhance the tokenizer’s representation of human motion, facilitating integration with textual information and thereby benefiting motion generation. We employed the following methods for data normalization and augmentation:

1. **Uniform Skeleton Normalization:** To ensure consistency across various data sources, motion data undergoes normalization to a uniform skeleton model. This involves scaling the skeleton based on leg lengths to adjust for differences in skeletal structure and proportions.
2. **Floor Normalization:** This step involves aligning the motion data vertically by setting the lowest point to the floor level. By doing so, all motions are standardized to a common vertical starting point, ensuring that the motions are grounded and realistic when rendered or analyzed.
3. **Rotation and Position Recovery:** Cumulative summation of rotation and linear velocities is employed to yield the global orientation and position, enabling the translation of local joint movements into a cohesive global pose.

Table 1. Qualitative Results of Unconditional Text Generation Tasks on ROCStories and AG News Dataset

"ROCStories dataset"	
Megan was excitedly putting the finishing touches on her intricate LEGO creation when suddenly,	it collapsed into pieces. Devastated, she burst into tears. Megan's father rushed to her side, offering a warm embrace.
Todd and Lulu took their young daughter on a trip to South America. They all enjoyed swimming	in the Caribbean and tasting local food. By the end of their journey, they felt tired but happy.
Sophie eagerly planned her birthday party and wanted a specific cake design.	She visited several bakeries but found none that matched her vision. Finally, she discovered the perfect cake design at a small bakery tucked away in a quiet corner.
"AG News dataset"	
Last Wednesday, the German government announced it would authorize an Australian scientist to conduct cloning research to advance medical research and innovation, bringing new hope to those seeking widespread tuberculosis vaccine coverage.	
At the recent sports event, a previously unknown athlete captured the gold medal with a remarkable performance, becoming the center of global attention.	
A young artist showcased a captivating oil painting at a contemporary art exhibition, earning enthusiastic applause from the audience and acclaim from the art community.	

4. **Foot Contact Detection:** Determining when feet make contact with the ground is essential for ensuring the realism of the motion. This is achieved by using velocity thresholds to detect significant changes in foot position between frames:

$$contact = \sqrt{(\Delta P_x)^2 + (\Delta P_y)^2 + (\Delta P_z)^2} < threshold, \quad (1)$$

where ΔP indicates the change in foot position, and the threshold is defined based on the expected velocity of foot contacts. This mechanism allows for the identification of key moments in the motion where feet interact with the ground.

5. **Motion Mirroring:** This involves flipping the motion data across the vertical axis, effectively creating a left-right inversion of the original motion. This augmentation step doubles the dataset size and introduces variation that helps improve the robustness and generalizability of motion analysis and synthesis models.

Benefits for Motion Generation: The processing and augmentation of motion datasets are pivotal for training motion tokenizers, facilitating the generation of realistic and contextually coherent human motions. Enhancements in data diversity and motion realism, along with the seamless synthesis of textual and motion data, underscore the comprehensive approach to motion generation.

2. Experimental Results

2.1. Qualitative Results of Motion Generation

For the motion generation task, our approach delivers diverse, smooth, and realistic visualizations. Figure 1 shows

the visual results of our methods.

2.2. Qualitative Results of Editing Tasks

Our method supports mixed modal input as editing condition, such as generation task with motion or text as guidance direction. Figure 2 shows the variety of our methods in editing tasks.

2.3. Qualitative Results of Unconditional Text Generation Tasks

We showcase our method's unconditional generation samples on both the ROCStories and AG News datasets in Table 1, demonstrating coherent and diverse outputs, featuring specialized nouns.

2.4. Quantitative Results on KIT Dataset

In addition to the HumanML3D and Human-X datasets, we also validated our method on the KIT dataset. Table 2 summarizes the results, which indicate that our method achieves superior results in R-Precision, FID, and MM-Dist, and also shows commendable diversity. The outcomes tabulated in Table 2 demonstrate our method's exceptional performance, particularly in terms of R-Precision, FID, and MModality metrics, alongside its commendable diversity.

3. Ablation Studies

In this section, we assess the impact of various elements within our framework under controlled conditions.

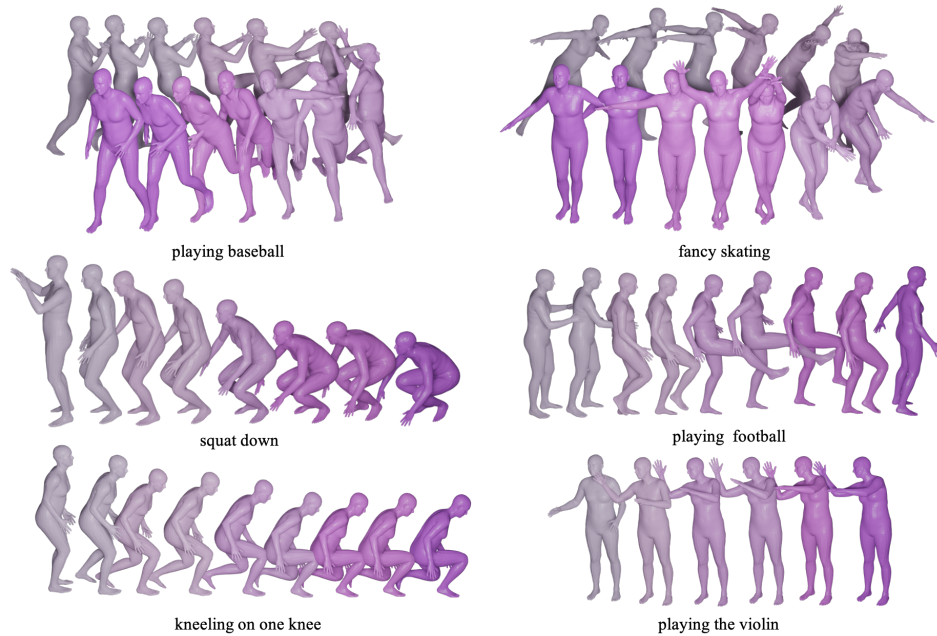


Figure 1. Results of motion generation task.

Table 2. Quantitative results of text-to-motion task on the KIT test set

Method	R-Precision \uparrow			FID \downarrow	MM-Dist \downarrow	Diversity \rightarrow	MModality \uparrow
	Top1	Top2	Top3				
Real	0.424 \pm .005	0.649 \pm .006	0.779 \pm .006	0.031 \pm .004	2.788 \pm .012	11.08 \pm .097	-
MDM	0.164 \pm .004	0.291 \pm .004	0.396 \pm .004	0.497 \pm .021	9.191 \pm .022	10.85 \pm .109	1.907 \pm .214
T2M	0.370 \pm .005	0.569 \pm .007	0.693 \pm .007	2.770 \pm .109	3.401 \pm .008	10.91 \pm .119	1.482 \pm .065
MotionDiffuse	0.417 \pm .004	0.621 \pm .004	0.739 \pm .004	1.954 \pm .062	2.958 \pm .005	11.10 \pm .143	0.730 \pm .013
MLD	0.390 \pm .008	0.609 \pm .008	0.734 \pm .007	0.404 \pm .027	3.204 \pm .027	10.800 \pm .117	2.192 \pm .071
Ours	0.422 \pm .003	0.630 \pm .005	0.750 \pm .006	0.392 \pm .023	3.087 \pm .012	11.127 \pm .083	2.300 \pm .055

3.1. Contrastive Text-Motion Variational Autoencoder

3.1.1 Variational Design

To investigate the role of Variational Autoencoder (VAE), we conducted experiments using a Deterministic Autoencoder as a comparative baseline. In the Deterministic encoder setup, the input no longer consists of two learnable tokens representing distributions, but rather a single token representing an embedding. The encoder’s output does not require sampling; instead, it directly represents the latent space vector. During loss computation, all KL losses are removed, retaining only the cosine loss. The experimental results are documented in the Table 3. To investigate the effect of the sampling size in the VAE, we compared experiments using a single random sample versus ten samples from the latent space. The results are documented in the Table 3.

3.1.2 Loss Design

In this module, we investigate the impact of the loss used to train the CTMV on motion generation, with results recorded in the Table 4. When removing the motion encoder and using only a Gaussian prior KL loss, there is a significant drop in R-precision, decreasing to 0.482; removing the cosine loss results in a decrease in R-precision to 0.486; removing the Gaussian prior KL loss leads to a decrease in R-precision to 0.490; removing the cross-modal KL loss results in a smaller decrease in R-precision to 0.494. It can be observed that removing any component of the training loss diminishes the performance of motion generation, indicating the effectiveness of each component.

3.2. Diffusion Model

In this subsection, we delineate the methodical ablation study performed on the diffusion model within our UniT-MGE framework. Our analysis is systematically segmented

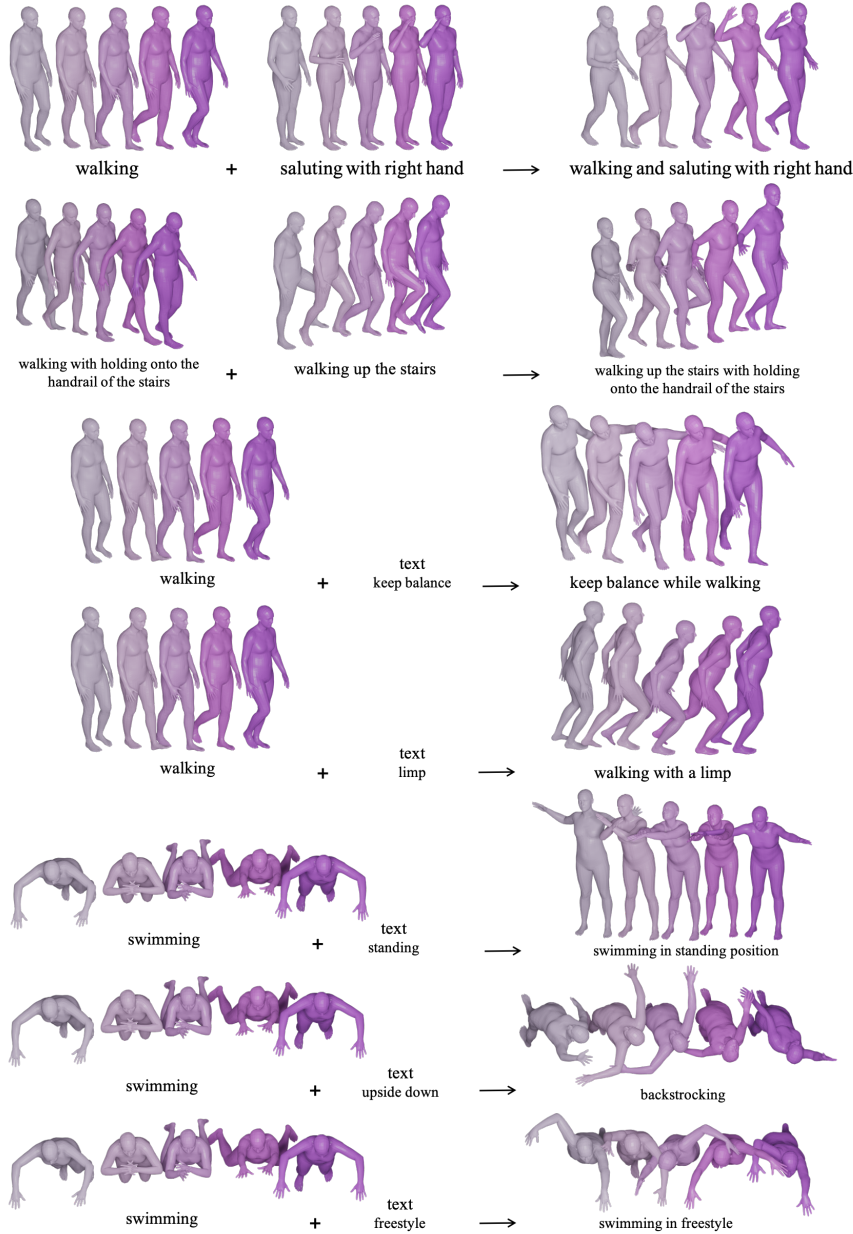


Figure 2. Results of editing tasks. UniTMGE edits the base motion by adding multimodal conditions, including text and motion, to the latent space.

Table 3. Evaluation of variational or deterministic autoencoder on motion generation

Pattern	Sampling	Top1	R-Precision↑ Top2	Top3	FID↓	MM-Dist↓	Diversity→	MModality↑
Real		0.511±.003	0.703±.003	0.797±.002	0.002±.000	2.974±.008	9.503±.065	-
Deterministic	no	0.47±.005	0.664±.005	0.764±.004	0.386±.012	3.489±.011	9.417±.049	2.367±.079
Variational	1 random sample	0.499±.003	0.683±.003	0.780±.002	0.339±.009	3.087±.008	9.527±.053	2.500±.083
Variational	10 random average	0.489±.001	0.675±.002	0.774±.002	0.341±.004	3.067±.005	9.535±.023	2.495±.053

Table 4. Evaluation of the loss fuction on motion generation

L_{KL}	L_{cos}	Top1	R-Precision \uparrow Top2	Top3	FID \downarrow	MM-Dist \downarrow	Diversity \rightarrow	MModality \uparrow
Real		0.511 \pm .003	0.703 \pm .003	0.797 \pm .002	0.002 \pm .000	2.974 \pm .008	9.503 \pm .065	-
$KL(\phi_t, \psi)$	\times	0.482 \pm .004	0.674 \pm .005	0.770 \pm .004	0.406 \pm .012	3.189 \pm .009	9.667 \pm .038	2.417 \pm .084
$KL(\phi_t, \phi_m) + KL(\phi_t, \psi) + KL(\phi_m, \psi)$	\times	0.486 \pm .003	0.677 \pm .003	0.775 \pm .003	0.366 \pm .009	3.123 \pm .008	9.587 \pm .053	2.450 \pm .063
$KL(\phi_t, \phi_m)$	\checkmark	0.490 \pm .003	0.679 \pm .002	0.778 \pm .003	0.358 \pm .006	3.105 \pm .008	9.554 \pm .048	2.478 \pm .084
$KL(\phi_t, \psi) + KL(\phi_m, \psi)$	\checkmark	0.494 \pm .003	0.680 \pm .002	0.779 \pm .003	0.342 \pm .006	3.095 \pm .008	9.534 \pm .048	2.494 \pm .069
$KL(\phi_t, \phi_m) + KL(\phi_t, \psi) + KL(\phi_m, \psi)$	\checkmark	0.499\pm.003	0.683\pm.003	0.780\pm.002	0.339\pm.009	3.087\pm.008	9.527\pm.053	2.500\pm.083

Table 5. Ablation study on the Diffusion Model component of UniTMGE

Models	R Precision Top 1 \uparrow	FID \downarrow	MM Dist. \downarrow	Diversity \rightarrow	MModality \uparrow
Real Data	0.511 \pm .003	0.002 \pm .000	2.974 \pm .008	9.503 \pm .065	-
UniTMGE-1 ($z, R^{1 \times 512}$)	0.499 \pm .003	0.339 \pm .009	3.087 \pm .008	9.527 \pm .053	2.500 \pm .083
UniTMGE-5 ($z, R^{5 \times 512}$)	0.485 \pm .003	0.884 \pm .019	3.311 \pm .008	8.224 \pm .081	2.498 \pm .070
UniTMGE-7 ($z, R^{7 \times 512}$)	0.489 \pm .002	0.901 \pm .019	3.415 \pm .008	8.336 \pm .064	2.483 \pm .079
UniTMGE-10 ($z, R^{10 \times 512}$)	0.479 \pm .005	0.786 \pm .027	3.416 \pm .012	8.506 \pm .067	2.578 \pm .077
UniTMGE-1 ($z, R^{1 \times 256}$)	0.483 \pm .004	0.453 \pm .031	3.160 \pm .012	8.598 \pm .088	2.768 \pm .096
UniTMGE-1 ($z, R^{1 \times 512}$)	0.499 \pm .003	0.339 \pm .009	3.087 \pm .008	9.627 \pm .056	2.440 \pm .074
UniTMGE-1 (ϵ_θ , cross-att)	0.470 \pm .004	0.922 \pm .041	3.980 \pm .015	8.598 \pm .088	2.768 \pm .096
UniTMGE-1 (ϵ_θ , concat)	0.499 \pm .003	0.339 \pm .009	3.087 \pm .008	9.527 \pm .053	2.500 \pm .083
UniTMGE-1 (ϵ_θ , w/o skip)	0.488 \pm .003	0.684 \pm .015	3.343 \pm .010	9.568 \pm .093	2.597 \pm .098
UniTMGE-1 (ϵ_θ , w/ skip)	0.499 \pm .003	0.339 \pm .009	3.087 \pm .008	9.527 \pm .053	2.500 \pm .083
UniTMGE-1 (ϵ_θ , 4 layers)	0.489 \pm .005	0.324 \pm .010	3.159 \pm .009	9.706 \pm .072	2.535 \pm .083
UniTMGE-1 (ϵ_θ , 6 layers)	0.497 \pm .003	0.329 \pm .012	3.119 \pm .012	9.624 \pm .062	2.504 \pm .088
UniTMGE-1 (ϵ_θ , 8 layers)	0.496 \pm .002	0.353 \pm .013	3.096 \pm .010	9.724 \pm .082	2.483 \pm .069
UniTMGE-1 (ϵ_θ , 10 layers)	0.499 \pm .003	0.339 \pm .009	3.087 \pm .008	9.527 \pm .053	2.500 \pm .083

Table 6. Comparative Evaluation of Motion Synthesis under Masked Conditions

Method	FID \downarrow	KID \downarrow	Multimodality \uparrow
ACTOR	48.80	0.53	14.10
MoDi	13.03	0.12	17.57
MDM	31.92	0.36	17.00
UniTMGE-1 (ϵ_θ , w/o skip)	24.40	0.20	17.30
UniTMGE-1 (ϵ_θ , w/ skip)	22.50	0.19	17.85
UniTMGE-1 (Motion Encoder)	28.10	0.27	16.90

to evaluate the influence of various model components on text-to-motion generation quality. We meticulously dissect the impact of the latent space configuration, denoted by UniTMGE-i, where i indexes the dimensionality of the latent vector z in $\mathbb{R}^{i \times 256}$. The choice of latent vector dimension is critical, as it represents the model’s capability to encapsulate motion nuances; a smaller dimensionality may provide computational benefits at the potential cost of expressive power.

We proceed to assess the role of the diffusion component ϵ_θ , particularly focusing on two methods of condition embedding—cross-attention (cross-att) and concatenation (concat). These methods are pivotal in integrating textual information and are expected to exhibit a substantial impact

on the coherence of generated motion sequences with the textual descriptions.

Furthermore, our study incorporates an exploration of the skip connection mechanism, inspired by its success in image processing tasks, hypothesizing its utility in preserving high-fidelity information throughout the diffusion process. We also experiment with varying the number of transformer layers within ϵ_θ to calibrate the model’s depth for effective learning.

The detailed results of these ablation experiments are summarized in the Table 5, Notably, UniTMGE-1, which employs the smallest latent representation, surprisingly outperforms its counterparts in several metrics, hinting at the potential efficiency of a more compact latent space.

As per the architecture involving ϵ_θ , both cross-attention and concatenation are explored for their effectiveness in conditioning the diffusion process. Preliminary results suggest that, akin to the MDM findings, concatenation yields a more beneficial encoder design. Meanwhile, the inclusion of skip connections furnishes a significant performance uplift, corroborating their utility beyond image domains. However, varying the number of transformer layers does not demonstrate a marked difference in performance, suggesting that a plateau may have been reached in terms of learning capacity for this particular dataset.

Finally, we contrast the performance of diffusion-based generation with the latent sampling-based generation from the variational component (V). This distinction is crucial for understanding the unique contributions of each approach within our proposed framework.

3.3. Multimodal Conditional Representation and Editing

The efficacy of MCRE is pivotal for the M2M tasks in our UniTMGE framework. MCRE is the cornerstone that allows for seamless motion synthesis by solely leveraging motion inputs when textual descriptors are unavailable or unnecessary. To distill the contribution of MCRE to the UniTMGE framework, we conduct an ablation study on the HumanML3D test set under a purely motion-driven context.

This ablation investigates three key variants: our model UniTMGE-1 with a latent vector size of $R^{1 \times 512}$, other established M2M capable models, and a configuration that utilizes a motion encoder in isolation. For the motion encoding process, we experiment with a transformer-based motion encoder as an alternative to the standard MCRE module. This encoder mirrors the structure of the motion adapter within the MCRE, featuring an eight-layer transformer network with a comparable latent vector size of $R^{1 \times 512}$. However, it diverges in training methodology; we integrate its training with the diffusion process rather than treating it as a separate component with CLIP text encoder. The training proceeds in two distinct stages: initially, the VAE is trained to establish a robust motion representation. Subsequently, we transition to the joint training of the diffusion process alongside the motion encoder. This training strategy aligns with the protocols established in our original Stage 3.

In our experimental setup, we aim to assess the ability of various models to generate motion sequences when conditioned on real data with an applied Mask. This simulates a scenario where partial information about the desired motion is known, and the model is expected to complete or refine the motion sequence. We test the models' performance using a comprehensive set of metrics to quantify the quality and diversity of the generated motions. Specifically, we employ FID and KID to evaluate the visual authenticity of the motions against the ground truth, gauging the similarity to the original motion data. Diversity and Multimodality are assessed to ensure that the models are not just reproducing variations of a single motion pattern but are capable of generating a wide array of plausible motions.

The following Table 6 presents the outcomes of this ablation, offering a comparative analysis of the model's ability to generate diverse and authentic motion sequences without textual input. Lower FID and KID scores suggest better quality, whereas higher Multimodality scores indicate a greater variety in the synthesized motions, both of which

are desirable attributes for robust M2M synthesis.

References

- [1] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 1
- [2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. 1
- [3] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. 1
- [4] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. 1