Uniform Text-Motion Generation and Editing via Diffusion Model

Ruoyu Wang[*✉]
Tsinghua University
Beijing, 100084, China
wang-ry22@mails.tsinghua.edu.cn

Xiang Li[*]
Li Auto Inc.
Beijing, 101399, China
lixiang960927@gmail.com

Tengjiao Sun[*]
University of Southampton
Southampton, SO17 1BJ, United Kingdom
ts1v23@soton.ac.uk

Tianyu Shi[*]
University of Toronto
Toronto, ON M5S 1A1, Canada
tianyu.shi3@mail.mcgill.ca

Yangfan He[*]
University of Minnesota - Twin Cities
Minneapolis, MN 55455, United States
he000577@umn.edu

Yiting Xie
GenFun
Beijing, 101399, China
xieyiting@genfun.ai

Abstract

Diffusion excels in controllable generation for continuous modalities, ideal for continuous motion generation. However, its flexibility is limited, focusing solely on text-to-motion generation and lacking motion editing capabilities. To address these issues, we introduce UniTMGE, a uniform text-motion generation and editing framework based on diffusion. UniTMGE overcomes single-modality limitations, enabling efficient and effective performance across multiple tasks like text-driven motion generation, motion captioning, motion completion, and multi-modal motion editing. UniTMGE comprises three components: CTMV for mapping text and motion into a shared latent space using contrastive learning, a controllable diffusion model customized for the CTMV space, and MCRE for unifying multimodal conditions into CLIP representations, enabling precise multimodal control and flexible motion editing through simple linear operations. We conducted both closed-world experiments and open-world experiments using the Motion-X dataset with detailed text descriptions, with results demonstrating our model's effectiveness and generalizability across multiple tasks.

1 Introduction

Motion generation tasks are vital in virtual reality, gaming, and robotics, with natural language facilitating intuitive interactions that underscore the significance of text-motion research; however, while diffusion models have advanced to produce smoother animations and greater control than token-based models, existing methods like MDM, MLD, and MotionDiffuse are restricted to single-modal input-output and mainly text-to-motion tasks, revealing an urgent need to broaden diffusion models for diverse text-motion applications. To address this issue, we introduce UniTMGE, a unified text-motion generation and editing framework based on diffusion that employs the Contrastive Text-Motion Variational Autoencoder (CTMV) as Figure 3 shows in the appendix to align text-motion pairs in a shared latent space, along with the Multimodal Conditional Representation and Editing (MCRE) to bridge semantic gaps using CLIP, demonstrating strong generalizability and effectiveness across various tasks, including text-driven motion generation and editing, through extensive experiments on

*Equal Contribution, ✉Corresponding Author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

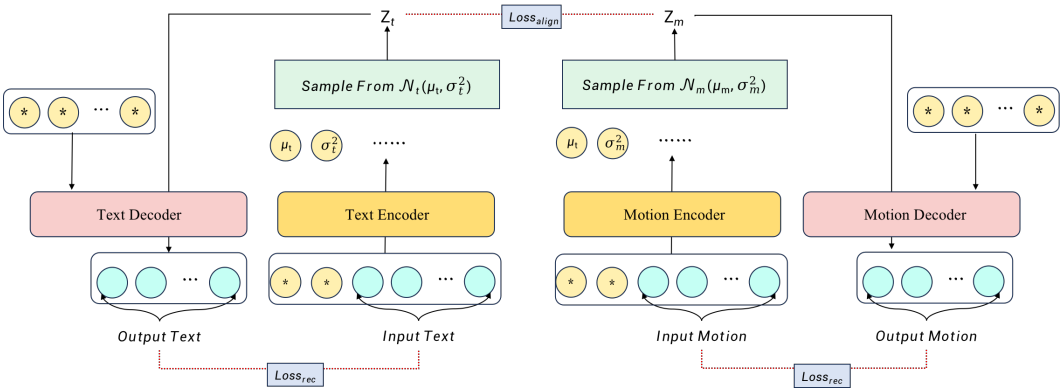

Figure 1: The CTMV module encodes text and motion sequences into latent distributions with learnable tokens, samples latent representations Z_t and Z_m for reconstruction via respective decoders, while applying alignment loss between encoders and reconstruction loss on both VAEs during training.

curated datasets such as Motion-X, while addressing the limitations of existing models that focus primarily on single-modal input-output. To summarize, our contributions can be summarized into three aspects:

- We propose UniTMGE, the first diffusion-based model to handle multiple text-motion tasks simultaneously, demonstrating strong effectiveness and generalization across each task.
- Furthermore, we also propose CTMA module, which maps text and motion into a shared latent representation for multimodal generation, and introduce MCRE, which unifies multi-modal conditions into CLIP representations to facilitate control and editing.
- Closed-world and open-world experimental results demonstrate the advanced effectiveness and generalizability of our approach across multiple tasks.

2 Method

To tackle text-motion generation tasks in a continuous space, we introduce the UniTMGE diffusion model as Figure 1 shows, comprising three key components: the Contrastive Text-Motion Variational Autoencoder (CTMV shown in Figure 3 in appendix), a controllable diffusion model, and Multimodal Conditional Representation and Editing (MCRE) shown in Figure 2. CTMV uses contrastive learning with two Variational AutoEncoders (VAEs) to create a unified representation of text and motion. Text is encoded into representations $I_t = E(t)$ using a pre-trained language encoder, while motion is represented as $I_m = \{I_m^i\}_{i=1}^{l_m}$. The encoders convert these inputs into a shared latent space Z_t and Z_m with length l_s and dimension d_s such that $l_s < \min[l_t, l_m]$ and $d_s < \min[d_t, d_m]$. The decoder reconstructs these representations back to their respective spaces. Training involves separate pre-training of the VAEs, where contrastive learning aligns similar text-motion pairs in latent space. The loss function consists of reconstruction losses L_{rec_t} and L_{rec_m}, KL divergence L_{KL}, and a cosine loss L_{cos}, ensuring effective mapping and reconstruction. The controllable diffusion model operates in the CTMV latent space, generating representations conditioned on input. It simulates a Markov process, progressively adding noise and training a Transformer-based denoising network. The training objective minimizes the difference between model predictions and ground truths, with a classifier-free approach incorporating random masking. During sampling, latent representations are iteratively refined. MCRE leverages CLIP representations for multimodal control, aligning text conditions using the CLIP text encoder and adapting motion conditions through a transformer. Training optimizes a cosine similarity loss to ensure close alignment between modalities. The approach allows for motion editing via linear operations on CLIP embeddings, enhancing the model's versatility in generating and modifying motion based on user inputs. Due to page limit, please refer to appendix for more details about the methods.

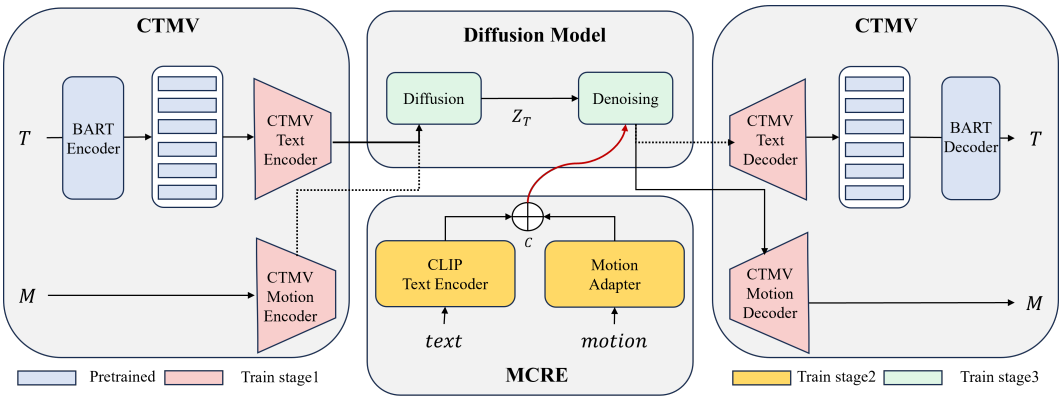

Figure 2: Our model integrates CTMV, a diffusion model, and MCRE, where CTMV encoders unify multimodal data into a common representation, the diffusion model performs generation, and MCRE converts conditional inputs to CLIP representations for precise control, with final outputs decoded back to their original spaces via CTMV decoders.

3 Experiments

3.1 Experiment Setup

3.1.1 Datasets

For motion generation tasks, we train our model on two standard 3D human motion-language datasets: HumanML3D[1] and KIT[16]. HumanML3D is derived from the HumanAct12 and AMASS datasets [12], comprising 14,616 motions and 44,970 descriptions. These motions span various domains such as daily activities, exercise, and artistic performances, with an average duration of 7.1 seconds per action. The descriptions contain 5,371 unique words, with an average description length of 12 words. Similarly, KIT is a dataset with 3,911 motions and 6,278 descriptions. Moreover, we process a subset of Motion-X, which include 2.4k motions annotated by more detailed textual descriptions, for the open-world experiment, to evaluate the model's generalization. For text generation tasks, we utilized two natural language datasets with rich corpora: ROCStories[13] and AG News Topic Classification[21].

3.1.2 Evaluation Metrics

We evaluate generated motions using metrics like R-Precision, MM-Dist, Diversity, MultiModality, FID, ADE, and FDE [27, 30], while motion-to-text tasks are evaluated with Length, Bleu [14], Rouge [9], Cider [25], and BertScore [28] as in [3], and unconditional text generation tasks use MAUVE [15], Perplexity, Diversity, and Memorization metrics as in [11].

3.1.3 Implementation Details

For the CTMV module, both the motion encoder and decoder adopt a nine-layer transformer architecture with skip connections. Similarly, the text encoder and decoder also employ a nine-layer transformer architecture and utilize the pre-trained BART model for text encoding and decoding. The dimensionality of the projected latent space is $Z \in R^{1\times512}$. For the diffusion module, the denoising network employs a nine-layer transformer network with 1000 denoising steps, and the variance β_t scales linearly from 0.0004 to 0.012. For the MCRE module, the text encoder utilizes the CLIP-ViT-B/32 frozen model, while the motion adapter employs an eight-layer transformer network. The dimensionality of the CLIP representation is $c \in R^{1\times512}$. We opt for the Adam optimizer [7] with a learning rate of 0.0002 for model training and a batch size of 128. During the training process, the VAE is trained for 100K epochs, and the diffusion module is trained for 100K epochs as well.

3.2 Generation Task

Our model represents a pioneering diffusion-based approach in breaking modality constraints. It can simultaneously handle various generation tasks, including text-to-motion, motion-to-text, text generation, and motion completion. We compare our approach with state-of-the-art diffusion-based methods for each task on various datasets and conduct open-world generation evaluation on Motion-X.

3.2.1 Text-to-Motion

This task involves generating corresponding motion sequences based on textual input. Table **??** in appendix summarizes the results, which indicate that our method achieves superior results in R-Precision, FID, and MM-Dist, and also shows commendable diversity. The visual results are illustrated in the Figure **??** in the appendix. Both qualitative and quantitative results confirm the high quality and accuracy of our generated results.

3.2.2 Motion-to-Text

This task involves generating corresponding textual descriptions based on motion sequences. Currently, there are no diffusion-based methods capable of performing similar tasks. Therefore, we compared our approach with MotionGPT [5] and TM2T [3], which have shown advanced performance. The experimental results in Table **??** in appendix indicate that our method achieves state-of-the-art performance in this task. Additionally, qualitative results provided in the appendix demonstrate that our generated text can effectively describe the action sequences.

3.2.3 Text Generation

This task involves unconditional text generation. For ease of comparisons, we evaluate our approach on ROCStories and AG News Topic Classification datasets. Apart from diffusion-based models like Diffusion-LM [8] and LD4LG [11], we also employ a fine-tuned GPT-2-Medium as a robust baseline, which is nearly twice the size of our model. As shown in Table **??** in appendix, our method outperforms diffusion-based methods in most metrics, demonstrating advanced generative capabilities.

3.2.4 Motion Completion

This work addresses motion prediction and inbetweening, using the first 20% of a motion sequence as conditional input for prediction and randomly masking 50% for inbetweening. Our diffusion-based approach achieves advanced performance in generation quality and diversity, as shown in Table **??** in appendix. It can handle various text-motion generation tasks simultaneously and shows strong generalizability on the unseen Motion-X dataset, highlighting the potential of diffusion in multitask scenarios.

3.3 Editing Task

Using the MCRE module, multimodal conditions are unified into the CLIP space, where desired editing directions can be derived through linear operations. Our method supports both editing with separate multimodal inputs and mixed-modal inputs. For unitary input (e.g., motion), body part motions are decomposed and recombined for smooth output, while mixed-modal inputs allow for text-guided editing in the latent space, enabling changes in attributes like mood, identity, and style, as illustrated in Figures **??** and **??** in appendix.

4 Conclusion

Our approach introduces the first unified diffusion-based framework for text-motion generation and editing, capable of handling multiple tasks simultaneously. Experimental results show its superior and versatile performance—an advantage lacking in current diffusion methods. Leveraging pre-trained CLIP models for multimodal semantic understanding, we aim to deepen integration with large-scale language models [19, 18, 17, 24] to enhance continuous generation, context comprehension, task decomposition, motion planning [10], and commonsense reasoning [4, 22, 23, 26]. We ultimately aim to unify tasks using language as a universal interface to improve user-friendliness.

References

[1] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng. Generating diverse and natural 3d human motions from text. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5142–5151, 2022. doi: 10.1109/CVPR52688.2022.00509.

[2] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng. Generating diverse and natural 3d human motions from text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5152–5161, 2022.

[3] C. Guo, X. Zuo, S. Wang, and L. Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In European Conference on Computer Vision, pages 580–597. Springer, 2022.

[4] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In International Conference on Machine Learning, pages 9118–9147. PMLR, 2022.

[5] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen. Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems, 36, 2024.

[6] D. P. Kingma and M. Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[7] B. J. Lei et al. Adam: A method for stochastic optimization. Proceedings of ICLR, 2015.

[8] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. B. Hashimoto. Diffusion-lm improves controllable text generation, 2022.

[9] C.-Y. 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. URL https://aclanthology.org/W04-1013.

[10] B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, J. Biswas, and P. Stone. Llm+ p: Empowering large language models with optimal planning proficiency. arXiv preprint arXiv:2304.11477, 2023.

[11] J. Lovelace, V. Kishore, C. Wan, E. Shekhtman, and K. Q. Weinberger. Latent diffusion for language generation, 2023.

[12] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. Amass: Archive of motion capture as surface shapes, 2019.

[13] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 839–849, 2016.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. 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. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

[15] K. Pillutla, S. Swayamdipta, R. Zellers, J. Thickstun, S. Welleck, Y. Choi, and Z. Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers, 2021.

[16] M. Plappert, C. Mandery, and T. Asfour. The kit motion-language dataset. Big Data, 4(4): 236–252, Dec. 2016. ISSN 2167-647X. doi: 10.1089/big.2016.0028. URL http://dx.doi.org/10.1089/big.2016.0028.

[17] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140):1–67, 2020.

[20] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 35: 36479–36494, 2022.

[21] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642, 2013.

[22] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2998–3009, 2023.

[23] H. Sun, Y. Zhuang, L. Kong, B. Dai, and C. Zhang. Adaplanner: Adaptive planning from feedback with language models. Advances in Neural Information Processing Systems, 36, 2024.

[24] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[25] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation, 2015.

[26] Z. Xiao, T. Wang, J. Wang, J. Cao, W. Zhang, B. Dai, D. Lin, and J. Pang. Unified human-scene interaction via prompted chain-of-contacts. arXiv preprint arXiv:2309.07918, 2023.

[27] Y. Yuan and K. Kitani. Dlow: Diversifying latent flows for diverse human motion prediction, 2020.

[28] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert, 2020.

[29] Y. Zhang, M. J. Black, and S. Tang. We are more than our joints: Predicting how 3d bodies move, 2021.

[30] Y. Zhang, M. J. Black, and S. Tang. We are more than our joints: Predicting how 3d bodies move. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3372–3382, 2021.

[31] Z. Zhou and B. Wang. Ude: A unified driving engine for human motion generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5632–5641, 2023.

A Appendix

A.1 Methods

To address text-motion generation tasks in a continuous space using a unified framework, we propose constructing a diffusion model called UniTMGE. The model consists of three main components, as illustrated in Figure 2. The first component is Contrastive Text-Motion Variational Autoencoder(CTMV). In CTMV, using contrastive learning, two Variational AutoEncoder (VAE) [6] are employed to map text and motion to a unified representation and reconstruct them back into their respective spaces. Secondly, a controllable diffusion model is constructed based on the CTMV representation to perform generation tasks. The third component is Multimodal Conditional Representation and Editing(MCRE). In MCRE, multimodal conditions are aligned with the CLIP representation to achieve precise control over the generation and editing processes.

A.2 Contrastive Text-Motion Variational Autoencoder

For the same semantic information, the text consists of a sequence of tokens $t \in N_t^l$. We first employ a pre-trained language encoder $E()$ to encode the input text tokens, obtaining a sequence of text representations $I_t = E(t) = \{I_t^i\}_{i=1}^{l_t} \in R^{d_t}$, where l_t denotes the length of the text representations, and d_t denotes the dimension of the text features. The motion representation consists of a sequence of frame-wise vectors $I_m = \{I_m^i\}_{i=1}^{l_m} \in R^{d_m}$, where l_m denotes the length of the motion representations, and d_m denotes the dimension of the motion representations[2]. The representations of text and motion exhibit different lengths and dimensions, posing challenges for the diffusion generation process. Therefore, we propose the CTMV, and the overall workflow of the module is illustrated in Figure 3.

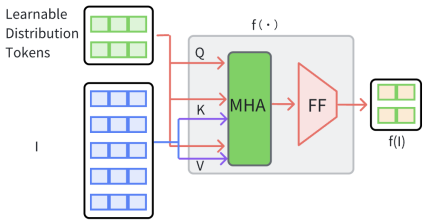

Figure 3: CTMV encoder network using multi-head attention (MHA) and feed-forward (FF) layers.

A.2.1 Encoder

The text encoder aims to unify both text inputs I_t and motion inputs I_m into a shared latent representation. This is achieved through two separate transformer-based encoders, one for text and one for motion, as illustrated in Figure 3. Each encoder is composed of a multi-head attention (MHA) module and a feed-forward network (FFN). These modules take as input variable-length sequences (whether text or motion) alongside two learnable distribution tokens that represent the mean and variance of a Gaussian distribution in the latent space. Utilizing the reparameterization trick [6], a latent representation Z is sampled from this Gaussian distribution. Both text and motion inputs are projected through this transformation, ensuring they are mapped into the same latent space with equal lengths and dimensions. The resulting representations are denoted as $Z_t = f(I_t) = \{Z_t^i\}_{i=1}^{l_s} \in \mathbb{R}^{d_s}$ and $Z_m = f(I_m) = \{Z_m^i\}_{i=1}^{l_s} \in \mathbb{R}^{d_s}$, where l_s and d_s refer to the length and dimension of the latent representation, respectively, ensuring $l_s < \min[l_t, l_m]$ and $d_s < \min[d_t, d_m]$. These constraints guarantee that both the text and motion representations share a common space. The text encoder serves as a compression network, mapping both text and motion inputs into a unified latent space via transformer-based modules, allowing for consistent representations across modalities.

A.2.2 Decoder

The decoder $g()$ also adopts a transformer-based architecture. Its input consists of the generated latent space Z and a sequence of learnable tokens, while the output is the reconstructed text or motion sequence $\widehat{I}_t = g(f(I_t))$, $\widehat{I}_m = g(f(I_m))$. Then, $\widehat{I}_t$ is further mapped back to the text tokens through a pre-trained language decoder $\widehat{t} = D(\widehat{I}_t)$.

A.2.3 Training Strategy

Training the VAE and the diffusion model jointly often leads to unstable training and is prone to collapse. Therefore, we opt to pre-train the VAE networks separately. To unify text and motion into the same representation, we employ a contrastive learning strategy, where similar text-motion pairs are embedded into nearby latent representations, while dissimilar pairs are embedded into distant latent representations. The training data consists of text-motion pairs and the training loss is composed of reconstruction and alignment losses. For text, the reconstruction loss is the cross-entropy loss between the input and the generated text pairs. For motion, the reconstruction loss is the $L2$ loss between the input and the generated motion pairs.

$$L_{rec_t}(I_t, \widehat{I}_t) = \frac{1}{l_t} \sum_{i=1}^{l_t} CELoss(I_t^i, \widehat{I}_t^i) \tag{1}$$

$$L_{rec_m}(I_m, \widehat{I}_m) = \frac{1}{l_m} \sum_{i=1}^{l_m} ||I_m^i - \widehat{I}_m^i||^2 \tag{2}$$

The alignment loss consists of the KL loss and the cosine loss. To align the distributions of text and motion latent spaces, we opt to minimize the KL divergence between the two distributions $\phi_t(Z|I_t) = \mathcal{N}(\mu_t, \sigma_t^2)$, $\phi_m(Z|I_m) = \mathcal{N}(\mu_m, \sigma_m^2)$. Additionally, to regularize the latent space, we utilize KL divergence to bring the distribution of each latent space closer to a standard normal distribution $\psi = \mathcal{N}(0, 1)$. Therefore, the overall KL loss can be represented as:

$$L_{KL} = KL(\phi_t, \phi_m) + KL(\phi_t, \psi) + KL(\phi_m, \psi) \tag{3}$$

After sampling from the distribution of the latent space , we use cosine loss to bring obtained latent representations of text and motion Z_t , Z_m closer together.

$$L_{cos}^i = 1 - cos(Z_t^i, Z_m^i) \tag{4}$$

$$L_{cos} = \frac{1}{l_s} \sum_{i=1}^{l_s} L_{cos}^i \tag{5}$$

Overall, the loss for CTMV can be formulated as follows:

$$L = L_{rec_t} + L_{rec_m} + L_{KL} + L_{cos} \tag{6}$$

By minimizing the reconstruction and alignment losses, we ensure that the CTMV maps text and motion to a unified semantic representation and effectively reconstructs them back to their respective spaces.

A.3 Controllable Diffusion Model

Targeting the unified CTMV latent space, we construct a controllable generation model based on the diffusion model. The objective of the model is to generate a set of latent representations based on the given condition. Diffusion model simulates a Markov process $\{Z^t\}_{t=0}^T$, where t denotes each time step. $Z^0 \in R^{l_s \times d_s}$represents data sampled from the latent space, while Z^t is obtained by adding noise to Z^0.

$$q(Z^t|Z^{t-1}) = \mathcal{N}(\sqrt{\alpha^t}Z^{t-1}, (1-\alpha^t)I), \tag{7}$$

where $\alpha^t \in (0, 1)$ is the hyperparameter. We need to train a Transformer-based denoising network ε_θ incorporating long skip connections[31] to gradually recover the original semantic information from the noisy data based on the conditional information. The training objective of the network is:

$$E_{\varepsilon,t,c}[||\varepsilon - \varepsilon_\theta(Z^t, t, c)||_2^2], \tag{8}$$

where c denotes the condition representation, which can be either text or motion. The denoising network is trained in a classifie-free manner, where the network learns the distributions of the CTMV latent space both conditioned and unconditioned by randomly masking 10% of the conditional information in the samples $c = \phi$ [20, 29]. During sampling, we start with random noise sampled $f(Z^T) \in R^{l_s \times d_s}$ $\mathcal{N}(0, 1)$ and then iteratively denoise it using the trained network.

$$\widehat{\varepsilon} = (1 - w)\varepsilon_\theta(Z^t, t, \phi) + w\varepsilon_\theta(Z^t, t, c) \tag{9}$$

After multiple sampling iterations, the desired result is generated. This generated result is then reconstructed back to the text or motion space through their decoders. A new training module iteratively builds an edited dataset to fine-tune the diffusion model based on user feedback, using the editing loss L_{edit} to capture deviations between outputs and desired user-driven improvements, enhancing variations like text-to-motion alignment or motion dynamics.

$$L_{\text{edit}} = E_{(I_t^c, I_m^c)} \left[\|\varepsilon_\theta(Z^t, t, c) - \varepsilon_{\text{target}}(Z^t, t, c)\|_2^2 \right] \tag{10}$$

$\varepsilon_\theta(Z^t, t, c)$ is the model's prediction, and $\varepsilon_{\text{target}}(Z^t, t, c)$ is the target reflecting user improvements. This L_2-norm loss ensures the model aligns with user preferences while maintaining generalization.

A.4 Multimodal Conditional Representation and Editing

To achieve multimodal control over the generation, we need to attain precise understanding of multimodal semantic information in both text and motion spaces. Therefore, we propose MCRE, which opts to unify the multimodal conditions into the CLIP representation, utilizing its strong semantic understanding and generalization capabilities.

A.4.1 Multimodal Conditional Representation in CLIP

For text conditions I_t^c, we use the pre-trained CLIP text encoder to obtain condition representations in the CLIP space, defined as $c = CLIP_t(I_t^c)$. For action conditions I_m^c, we train an adapter $W(\cdot)$ with a transformer encoder to align these representations with the CLIP space, where the first token of the adapter's output is the final result $c = W(I_m^c)$. The adapter is trained on paired text and motion sequences, and the training objective minimizes the following cosine similarity loss:

$$L = 1 - \cos(CLIP_t(I_t^c), W(I_m^c)) \tag{11}$$

We employ a contrastive learning strategy, similar to CLIP, incorporating both positive and negative samples. The final loss is:

$$L = 1 - \cos(c_{\text{pos}}, c_{\text{motion}}) + \max(0, \cos(c_{\text{neg}}, c_{\text{motion}}) - \epsilon) \tag{12}$$

This ensures positive samples are brought closer, while negative samples are pushed apart. The margin ϵ controls the minimum distance between negative and positive samples, improving the model's discrimination ability.

A.4.2 Motion Editing via Multimodal Conditional Editing

The CLIP space exhibits highly disentangled characteristics, allowing us to perform various motion editing tasks by simple linear operations on the conditional CLIP representations. Specifically, to obtain the edited conditional representation c_{edit} , we add the desired editing direction vector Δc_{edit} to the original motion representation.

$$c_{edit} = c + \Delta c_{edit} \tag{13}$$

The edited conditional representation obtained in this way enables us to control the diffusion to achieve the desired editing result. The editing direction is derived from the difference between CLIP embeddings of the original and target conditions, which can be based on either text or motion.