Cross-Axis Feature Fusion with Joint-Wise Motion Difference Prediction for Text-Based 3D Human Motion Editing

Supplementary Material

A. Generated-to-Source Motion Retrieval

Methods	Generated-to-Source (Batch)			FID↓
	R@1↑	R@2↑	R@3↑	
Ground Truth	74.01	84.52	89.91	-
TMED	71.77	84.07	89.52	0.167
SimMotionEdit	72.71	83.54	87.50	0.110
MotionReFit	**83.47**	<u>90.42</u>	<u>92.84</u>	-
Ours	<u>78.96</u>	**91.04**	**93.33**	**0.097**

Table A. We report generated-to-source motion retrieval results.

To rigorously evaluate the capability of our model to maintain the structural and stylistic properties of the source motion S while performing the instructed edit, we conducted generated-to-source motion retrieval experiments. This analysis is crucial for text-based motion editing tasks, as the generated output M must exhibit a high degree of semantic consistency with S. Following the established TMR [4] retrieval protocol, we used the TMR feature space to check how well the generated motion retrieves its corresponding source motion. The results, shown in Table A, demonstrate that our method significantly outperforms the baselines, TMED [1] and SimMotionEdit [3]. Furthermore, our approach achieves performance comparable to MotionReFit [2], showing particularly strong results in R@2 and R@3. This superior performance indicates that our architecture—with its dedicated axis-anchored transformers and joint-aware supervision—ensures robust preservation of the source motion's semantic information while accurately implementing the desired textual modifications.

B. Experiments on Different Design Choices

We conducted experiments on various design choices within our framework.

B.1. Motion Similarity Prediction Before Fusion

We conducted an ablation study to investigate the optimal placement of the auxiliary motion similarity prediction task, originally introduced in SimMotionEdit [3]. In our final proposed architecture, the task is applied to the features after passing through the cross-axis fusion block. For this ablation, we shifted the motion similarity prediction head to operate on the intermediate frame-wise feature h_{time}, specifically, the output of the time-anchored transformer just before it enters the cross-axis fusion block.

As shown in Table B, the performance achieved by applying the motion similarity prediction task to the intermediate h_{time} feature is consistently lower than our proposed method. This outcome validates our architectural design: the joint-anchored transformer is explicitly trained by the Soft-DTW auxiliary objective ($\mathcal{L}_{\text{aux}}$) to understand which joints to modify and which to preserve. By performing the motion similarity prediction task after the fusion block, the prediction head leverages the rich, contextualized representation (h_{fusion}) that the cross-axis fusion block integrates from both joint-aware and temporally-aware conditioning. This results in superior performance, as the fused feature provides a more robust basis for predicting the overall frame-wise motion similarity.

B.2. Larger Cross-Axis Fusion Block

We further analyzed the impact of the design of the cross-axis fusion block on model performance. In our main design, we employ a single multi-head cross-attention block to integrate the time-anchored feature (h_{time} as Query) and the joint-anchored feature (h_{joint} as Key and Value). For this ablation, we constructed a significantly larger fusion block consisting of an alternating sequence of 4 temporal self-attention blocks and 4 cross-attention blocks.

The results presented in Table B show that the performance with the larger, more complex fusion block was consistently inferior compared to our optimized single cross-attention layer design. We hypothesize that this degradation is due to two primary factors. First, the substantial increase in the number of parameters within the fusion pathway may lead to overfitting on the MotionFix dataset, which is moderate in size for complex models. Second, the repetitive self-attention layers within the fusion block may unnecessarily re-aggregate the features along the temporal axis. This repetition potentially disrupts the explicit disentanglement between the joint-wise and time-wise information that our axis-anchored transformers were designed to establish, thereby diluting the precise and targeted conditioning signal needed for fine-grained motion editing. The simpler, single cross-attention block proves sufficient to perform the required feature integration without introducing undue complexity or feature corruption.

Methods	Generated-to-Target (Batch)				Generated-to-Target (Test Set)				FID↓
	R@1↑	R@2↑	R@3↑	AvgR↓	R@1↑	R@2↑	R@3↑	AvgR↓	
Motion Sim. before fusion	70.00	84.79	90.00	2.00	26.09	41.50	48.81	18.54	0.122
Larger fusion block	72.92	86.88	90.42	2.02	29.45	46.05	55.14	16.65	0.103
Ours	**74.38**	**88.54**	**92.08**	**1.92**	**29.45**	**45.26**	**54.55**	**16.42**	**0.097**

Table B. We report experiment results for different design choices of our framework.

Methods	Generated-to-Target			Generated-to-Source		
	R@1↑	R@2↑	R@3↑	R@1↑	R@2↑	R@3↑
Ours w/ global dist.	71.46	87.71	90.62	77.08	90.42	93.54
Ours w/ joint geodesic dist.	72.50	85.42	89.79	78.33	89.58	93.33
Ours w/ classic DTW	72.71	85.62	90.21	79.58	90.62	93.33
Ours	74.38	88.54	92.08	78.96	91.04	93.33

Table C. Ablation study on distance metrics for joint-wise supervision. We compare the impact of different distance objectives for training the joint-anchored transformer. We evaluate global frame-wise distance, joint-wise geodesic distance on the rotation manifold, and classic DTW against our proposed Soft-DTW.

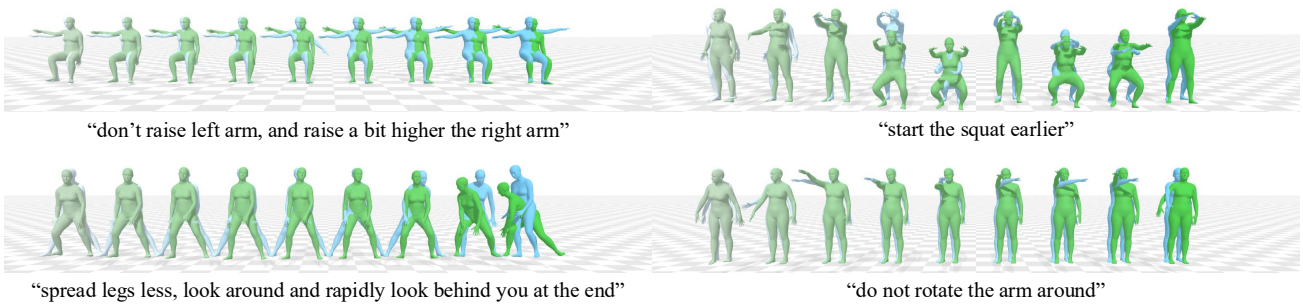

Figure A. We visualized source motions (blue) and motions edited with our method (green) with overlay.

C. Analysis of Distance Metrics for Motion Difference Prediction

We conducted several experiments to justify the design of our auxiliary task in Table C. First, the lower performance of the global distance metric compared to our approach suggests that holistic motion alignment is insufficient for the encoder to reason about joint-specific signals. Regarding the metric formulation, we observed that the joint geodesic distance yields suboptimal results because it collapses multi-dimensional rotation differences into a single scalar, which restricts the information available for supervision. In contrast, our channel-wise prediction provides higher degrees of freedom through finer-grained signals, enabling the encoder to capture the complex dynamics of 6D rotations more effectively. Finally, our method outperforms Classic DTW because the soft-min operator is more robust to high-frequency motion noise than rigid alignment, thereby providing a more stable and differentiable learning signal for the transformer.

D. More Qualitative Results

We provide an overlay visualization of source motions and motions edited with our method in Figure A, superimposing edited motions onto the original source sequences.

E. Future Work

Looking ahead, we aim to extend our framework by leveraging the rich prior knowledge of emerging large-scale text-to-motion foundation models. As motion generation enters the era of large-scale pre-training with massive datasets and significantly increased parameter counts, we intend to investigate how our proposed axis-anchored conditioning and

joint-wise auxiliary task can be effectively utilized to finetune these models for the motion editing task. We believe that integrating our structural and joint-aware supervision with the broad motion priors of foundation models will further enhance the precision and diversity of text-based editing.

Furthermore, we aim to further validate the generalizability of our framework across a wider variety of datasets, such as the STANCE benchmark [2]. While we initially explored an evaluation on the STANCE benchmark, a fair comparison was not feasible at this stage due to the current lack of a publicly available implementation for the full training and evaluation protocol. We look forward to conducting these experiments once the standardized pipeline is released by the authors.

References

[1] Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J. Black, and Gül Varol. Motionfix: Text-driven 3d human motion editing. In *SIGGRAPH Asia 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 1

[2] Nan Jiang, Hongjie Li, Ziye Yuan, Zimo He, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Dynamic motion blending for versatile motion editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22735–22745, 2025. 1, 3

[3] Zhengyuan Li, Kai Cheng, Anindita Ghosh, Uttaran Bhattacharya, Liangyan Gui, and Aniket Bera. Simmotionedit: Text-based human motion editing with motion similarity prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27827–27837, 2025. 1

[4] Mathis Petrovich, Michael J. Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9488–9497, 2023. 1