Supplementary Material

We have introduced a fine-grained text-driven method for generating human motion sequences that conformed with the text prompt in the main paper. This supplementary material presents more details, including:

- Additional fine-grained visual results in section 1;
- More detailed network architectures in section 2.

1. More Fine-Grained Results

We have shown the qualitative results compared with two state-of-the-art methods in the main paper. Our method compares with the Temporal VAE [1] and MotionDif- fuse [2], outperforming these models, especially in terms of grasping the fine-grained details.

In this section, we further apply several subtle modifications to the text prompt, enabling us to assess the quality of the generated motion and demonstrate the robustness of our approach instead of being limited to definite words. Specifically, we implement some changes, such as revising subtle spatial orientation words or temporal prepositions, to the sentences “A person kicks something with his right leg then throws an object with his right hand” and “A person walks forward two steps then backward one step” in the original experiment, as shown in Figure 1 and Figure 2. It can be observed that when replacing some key adjectives or prepositions, our method still yields decent results faithful to the text prompt.

We also show the comparison of some qualitative examples on the ablation study, and present more visual results
in Figure 3 and 4.

2. Network Architectures

As mentioned in the main paper, our method consists of two modules, Linguistics-Structure Assisted Module (LSAM) as the text encoder, and Context-Aware Progressive Reasoning Module (CAPR) as the motion decoder. In this section, we further present more detailed network architectures of these two modules, as listed in Table 1. LSAM comprises node embedding, edge embedding, and graph attention convolution, while CAPR incorporates multi-modal sentence-level feature-fusion and multi-head word-level cross-attention.

References


Figure 3: Qualitative examples on the ablation study and more diverse examples.

Figure 4: **More Visual results**: (a) A person walks to his left side, then bends down and picks something up. (b) A person runs to his right side first, runs to his left, then back to the middle. (c) A person walks forward two steps, pivots 180 degrees then walks two steps back to where he started.
<table>
<thead>
<tr>
<th>Components</th>
<th>Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSAM (Edge)</td>
<td>(Spacy): spacy.load('en_core_web_sm') (Onehot): OneHotEncoder('ignore') (Edge Embedding): Sequential(0): Linear(in_features=45, out_features=512) (1): Silu() (2): Linear(in_features=512, out_features=512))</td>
</tr>
<tr>
<td>CAPR (Sentence-Level)</td>
<td>(Text Embedding): Sequential(0): Linear(in_features=512, out_features=512) (1): LayerNorm(normalized_shape=512)) (Motion Embedding): Sequential(0): Linear(in_features=dim_in, out_features=512) (1): LayerNorm(normalized_shape=512)) (Sent Conv): Conv1d(in_channels=512, out_channels=512, kernel_size=77) (Activation): Sigmoid() (Fusion Ln): LayerNorm(normalized_shape=512) (Query): Linear(in_features=512, out_features=512) (Key): Linear(in_features=512, out_features=512) (Value): Linear(in_features=512, out_features=512)</td>
</tr>
</tbody>
</table>

Table 1: Detail Architecture of our LSAM and CAPR.