LREC-COLING 2024

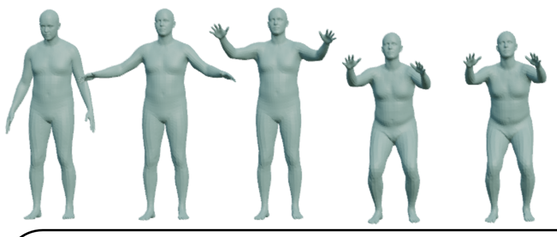
Motion Generation from Fine-grained Textual Descriptions

Kunhang Li [1,2], Yansong Feng [1]

[1] Peking University, [2] The University of Tokyo

Introduction: Fine-grained Text2motion

Descriptive Granularity



coarse-grained description

A man slightly squats.

fine-grained description

<step 1: beginning pose> The man begins standing upright with his feet hip-width apart and his arms relaxed at his sides. </step 1: beginning pose>
<step 2: slight squat> He bends his knees slightly, lowering his hips and shifting his weight slightly towards his heels. His torso remains upright, and his feet remain flat on the ground. </step 2: slight squat>
<step 3: end pose> He holds this slightly squatted position. </step 3: end pose>

Generalizability → Spatial and Temporal Compositionality

A person slightly squats.

A man walks.

A man raises both arms above head.

A man kicks with one leg.

A man slightly squats with both arms raised above head.

A man walks, then kicks with one leg.

Spatial Compositionality ✓

Temporal Compositionality ✓

Our Method

Building the first fine-grained language-motion dataset **FineHumanML3D** using the LLM

No muscle tension:
Muscles cannot be reflected in motion sequences.

Named step marks:
To ensure the chronological order and overall quality

Pseudo-code conversion:
To check the inner consistency of the generated motions

In the first paragraph, please provide a detailed expansion of one coarse-grained motion description. The new description should be in chronological order and step by step. It should specify spatial position changes (including angle changes) of relevant body parts. It should not specify any information related to muscles. In the second paragraph, please convert the description in the first paragraph into a pseudo-code format.
[EXAMPLE1]
[EXAMPLE2]
A man walks, then kicks with one leg.

Our Prompt

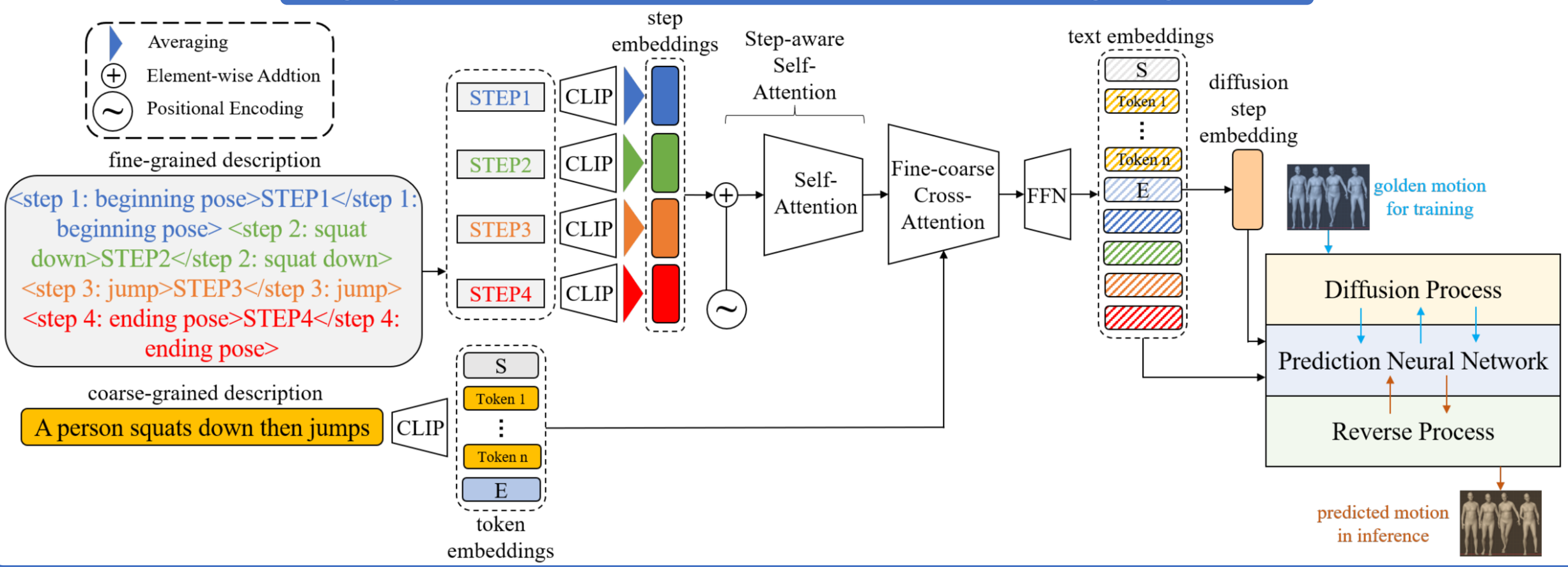
Fine-grained Descriptions

<step 1: beginning pose>The man begins in a standing position with his feet together and his arms at his sides.</step 1: beginning pose>
<step 2: taking a step> ··· </step 2: taking a step>
<step 3: kicking> ··· </step 3: kicking>
<step 4: end pose> ··· </step 4: end pose>

<step 1: beginning pose>stand(feet together, arms at sides)</step 1: beginning pose>
<step 2: taking a step> ··· </step 2:taking a step>
<step 3: kicking> ··· </step 3: kicking>
<step 4: end pose> ··· <step 4: end pose>

Pseudo-code Part

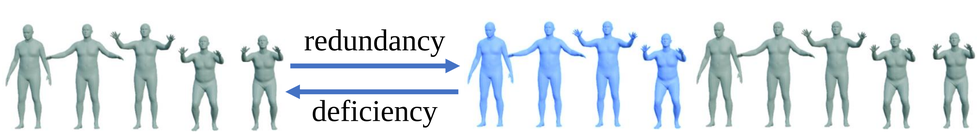
Designing a new text2motion model **FineMotionDiffuse** better modeling fine-grained texts



Human Evaluation of FineHumanML3D

Counts of zero, partial and perfect alignment
2 : 68 : 30

In the majority of partially aligned cases, partial alignment indeed captures the correct time order and relationships among core motions. Issues like redundancy or deficiency are often trivial in nature. Substantive errors rarely occur outside of very complicated motions.

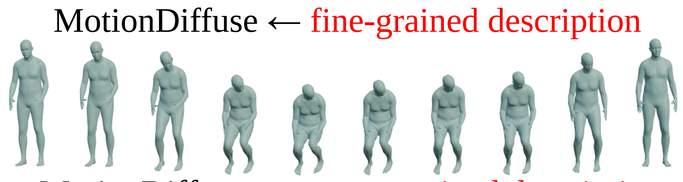


Cases

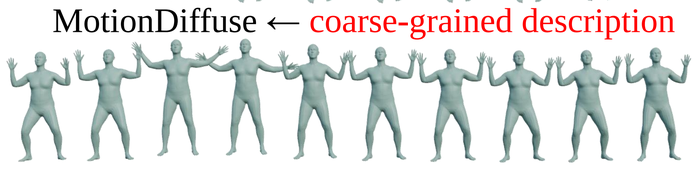
FineMotionDiffuse ← coarse-grained description + fine-grained description: **Refer to the Introduction**

A man slightly squats with both arms raised above head.

MotionDiffuse ← fine-grained description



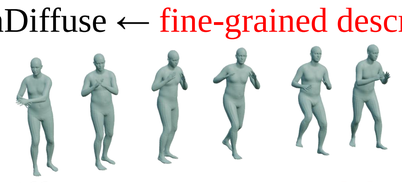
MotionDiffuse ← coarse-grained description



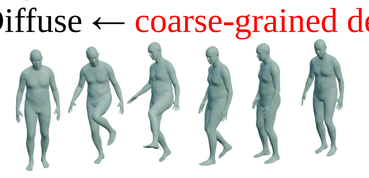
Spatial Compositionality ✗

A man walks, then kicks with one leg.

MotionDiffuse ← fine-grained description



MotionDiffuse ← coarse-grained description



Temporal Compositionality ✗

https://kunhangl.github.io/finemotiondiffuse/