LREC-COLING 2024

Motion Generation from Fine-grained Textual Descriptions

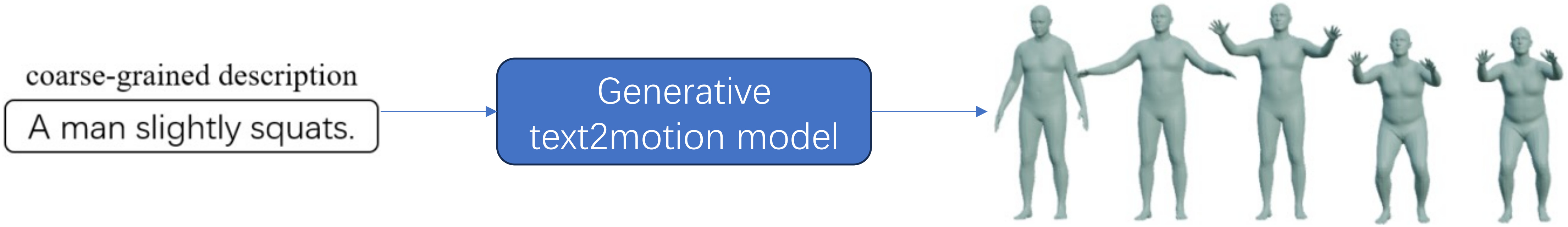
Kunhang Li [1,2], Yansong Feng [1]

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

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

Text2motion

- Generate motion sequences from given textual descriptions



coarse-grained description

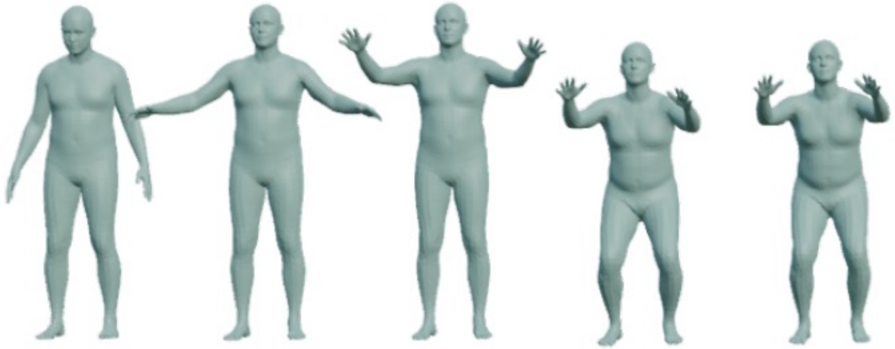
A man slightly squats.

Generative text2motion model

Fine-grained Text2motion

Descriptive Granularity

- Previous works cannot deal with fine-grained text2motion



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>

Expectations of Fine-grained Descriptions

1) Be in time order;

2) Specify spatial movements of relevant body parts;

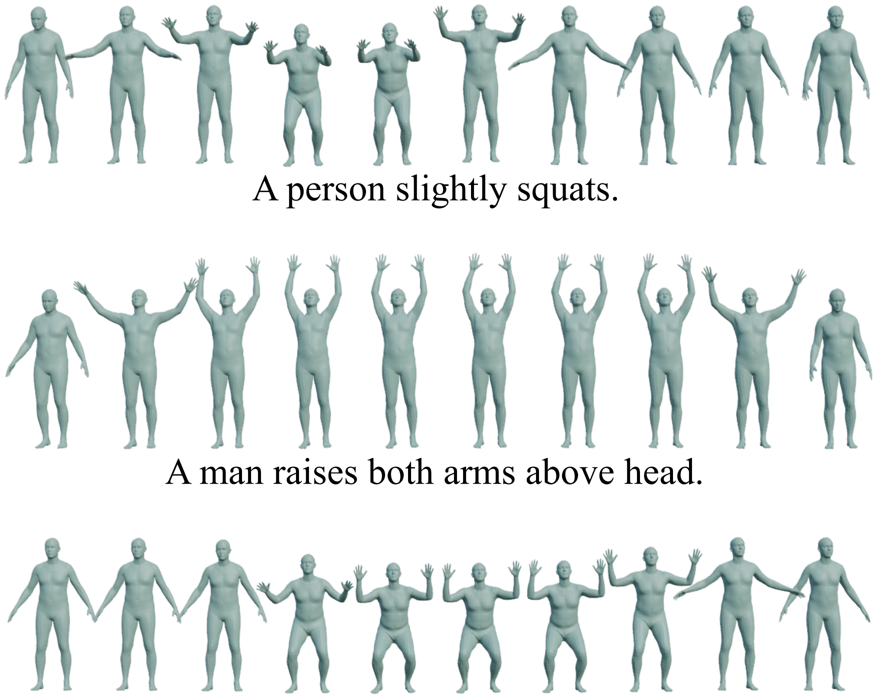
3) Discard unnecessary details regarding muscle tension and human feelings;

4) Conform to human body constraints;

Fine-grained Text2motion

Generalizability → Spatial and Temporal Compositionality

- Previous works cannot deal with fine-grained text2motion

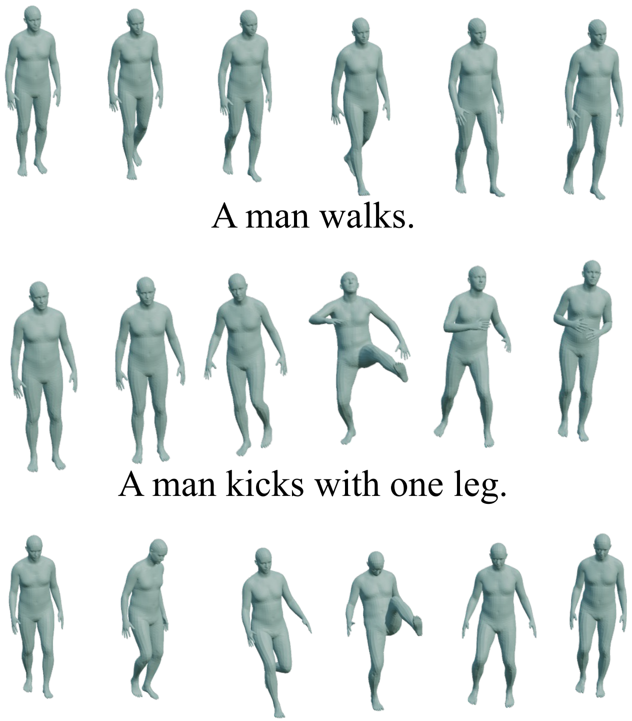


A person slightly squats.

A man raises both arms above head.

A man slightly squats with both arms raised above head.

Spatial Compositionality



A man walks.

A man kicks with one leg.

A man walks, then kicks with one leg.

Temporal Compositionality

Question & Hypothesis

- Question 1: Can we endow the text2motion model with the ability of expected compositionality on textual descriptions of various descriptive granularities?
- Question 2: Do Large Language Models have the appropriate physical knowledge to help us with this problem?

- Hypothesis: Yes! Let's build the first fine-grained language-motion dataset with the LLM and train a new text2motion model!

Our Approach

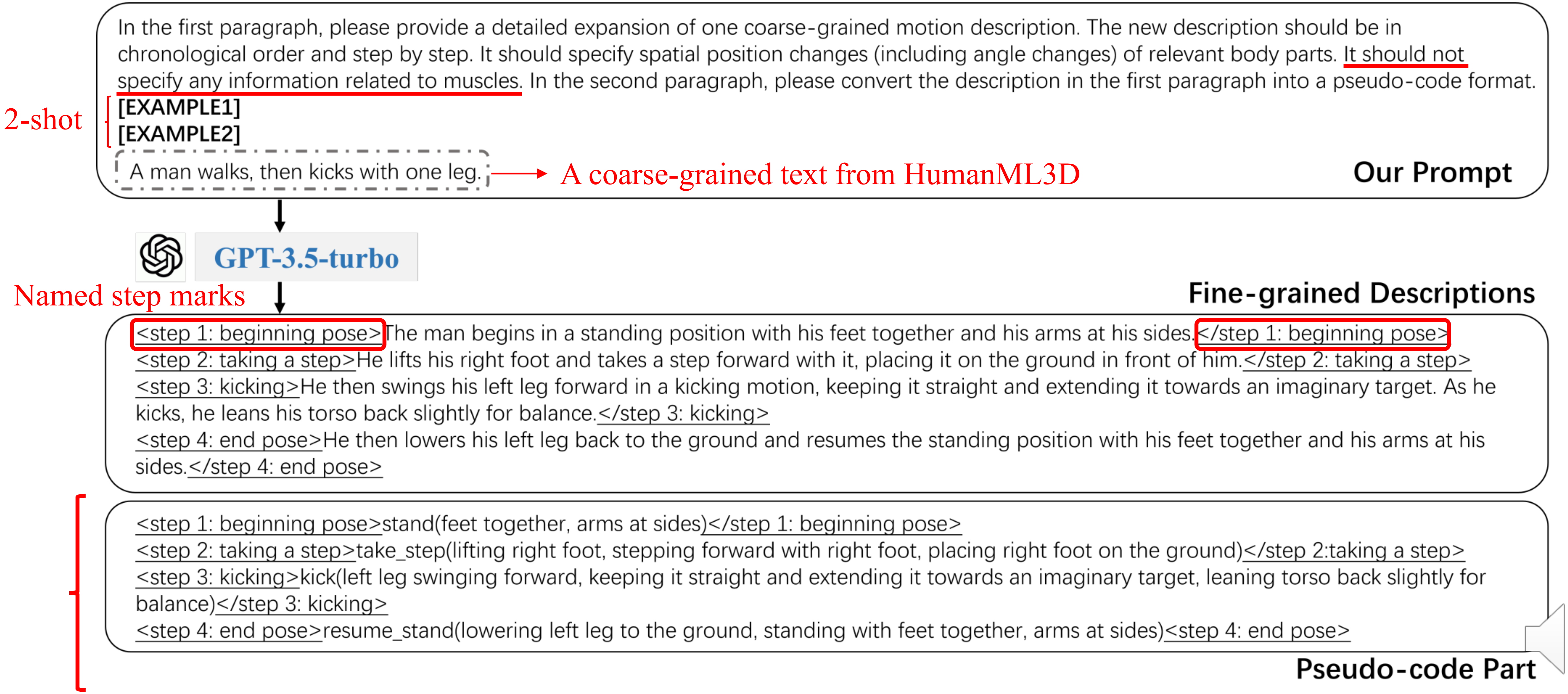
- Automatically building the first fine-grained language-motion dataset FineHumanML3D using GPT-3.5-turbo, from HumanML3D [1]

- Designing a new text2motion model FineMotionDiffuse better modeling fine-grained texts, inspired by MotionDiffuse [2]

[1] Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., & Cheng, L. (2022). Generating Diverse and Natural 3D Human Motions from Text. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 5142-5151.

[2] Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., & Liu, Z. (2022). MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *IEEE transactions on pattern analysis and machine intelligence, PP*.

Prompting Strategies



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.

2-shot

[EXAMPLE1]
[EXAMPLE2]

A man walks, then kicks with one leg. → A coarse-grained text from HumanML3D

Our Prompt

GPT-3.5-turbo

Named step marks

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>He lifts his right foot and takes a step forward with it, placing it on the ground in front of him.</step 2: taking a step>
<step 3: kicking>He then swings his left leg forward in a kicking motion, keeping it straight and extending it towards an imaginary target. As he kicks, he leans his torso back slightly for balance.</step 3: kicking>
<step 4: end pose>He then lowers his left leg back to the ground and resumes the standing position with his feet together and his arms at his sides.</step 4: end pose>

<step 1: beginning pose>stand(feet together, arms at sides)</step 1: beginning pose>
<step 2: taking a step>take_step(lifting right foot, stepping forward with right foot, placing right foot on the ground)</step 2:taking a step>
<step 3: kicking>kick(left leg swinging forward, keeping it straight and extending it towards an imaginary target, leaning torso back slightly for balance)</step 3: kicking>
<step 4: end pose>resume_stand(lowering left leg to the ground, standing with feet together, arms at sides)<step 4: end pose>

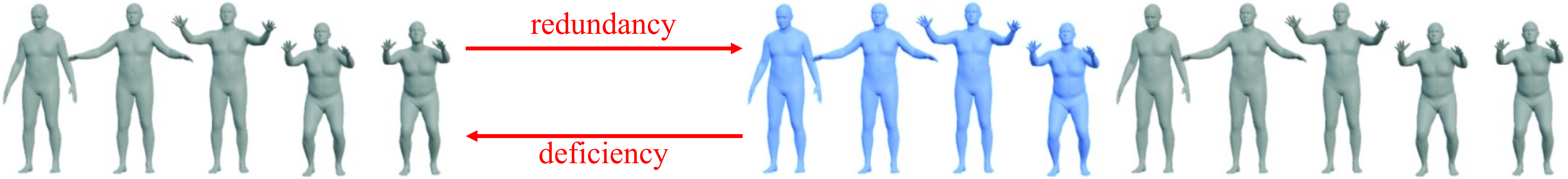
Pseudo-code Part

Human Evaluation of FineHumanML3D

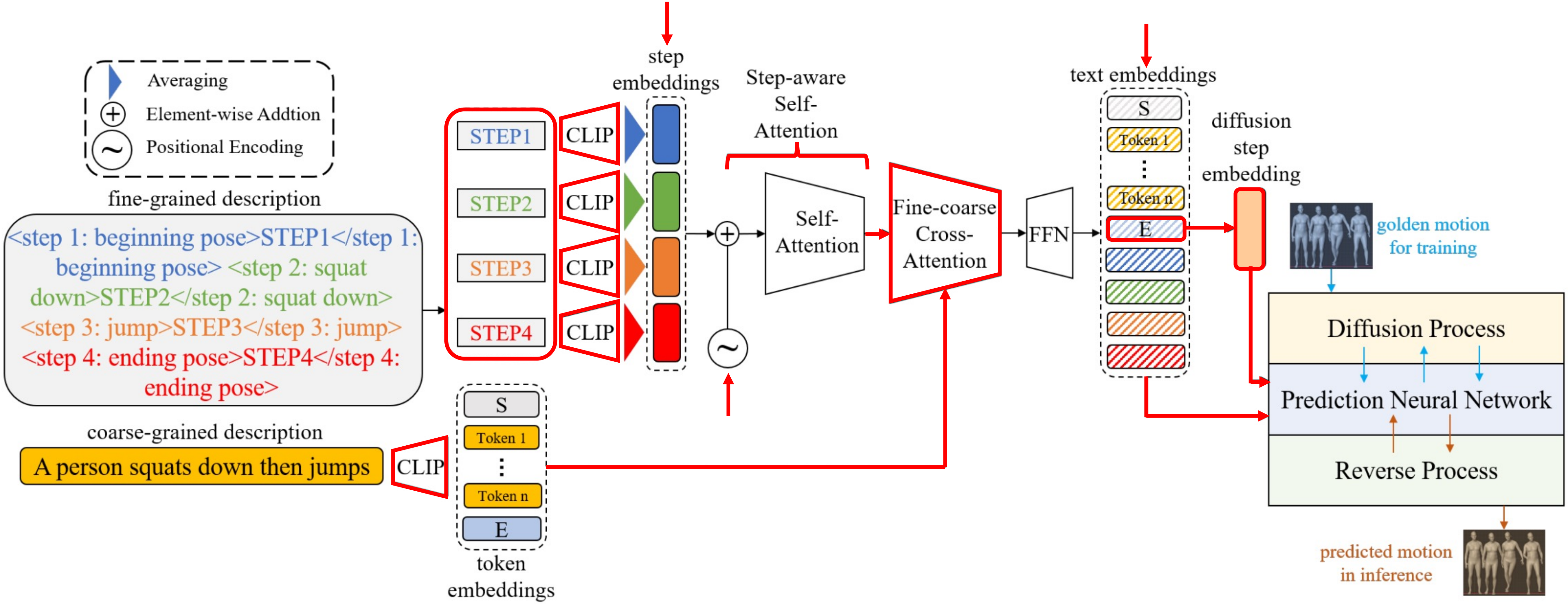
- What is the quality of the automatically built dataset?

- Counts of zero, partial and perfect alignment turn out to be **2 : 68 : 30**

In a large body of the partially aligned cases

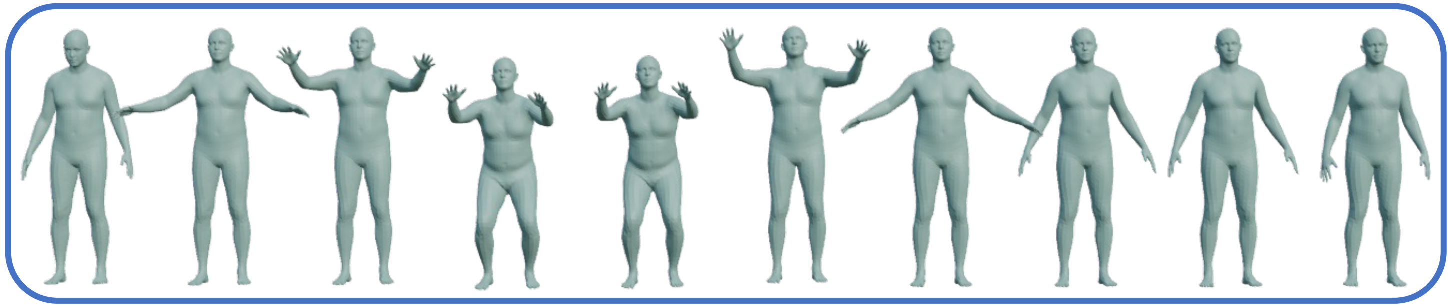


Our FineMotionDiffuse Model

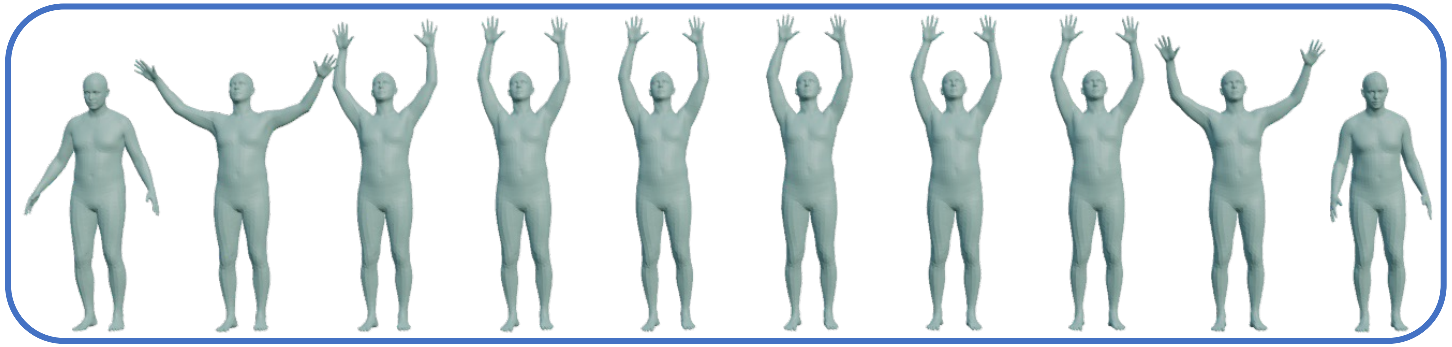

Averaging
Element-wise Addtion
Positional Encoding
fine-grained description
<step 1: beginning pose>STEP1</step 1: beginning pose> <step 2: squat down>STEP2</step 2: squat down> <step 3: jump>STEP3</step 3: jump> <step 4: ending pose>STEP4</step 4: ending pose>
coarse-grained description
A person squats down then jumps
STEP1
STEP2
STEP3
STEP4
CLIP
CLIP
CLIP
CLIP
CLIP
step embeddings
S
Token 1
Token n
E
token embeddings
Step-aware Self-Attention
Self-Attention
Fine-coarse Cross-Attention
FFN
text embeddings
S
Token 1
Token n
E
diffusion step embedding
golden motion for training
Diffusion Process
Prediction Neural Network
Reverse Process
predicted motion in inference

Spatial Compositionality

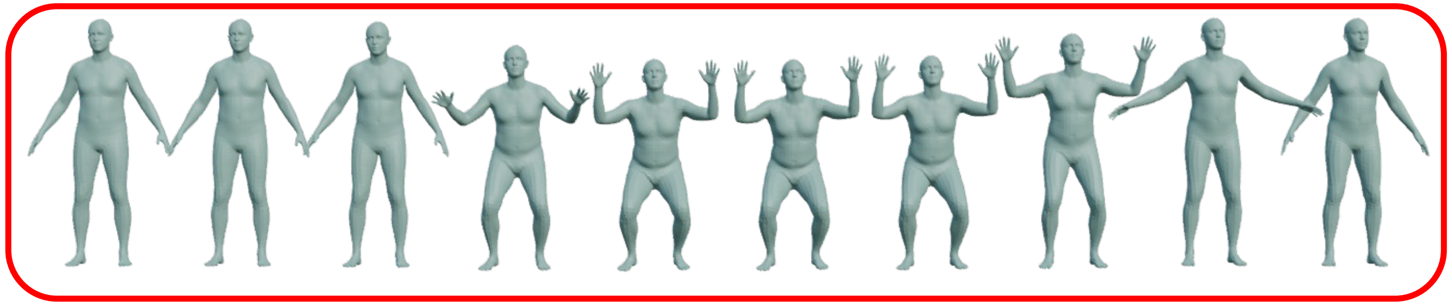
FineMotionDiffuse ← coarse-grained description + fine-grained description



A person slightly squats.



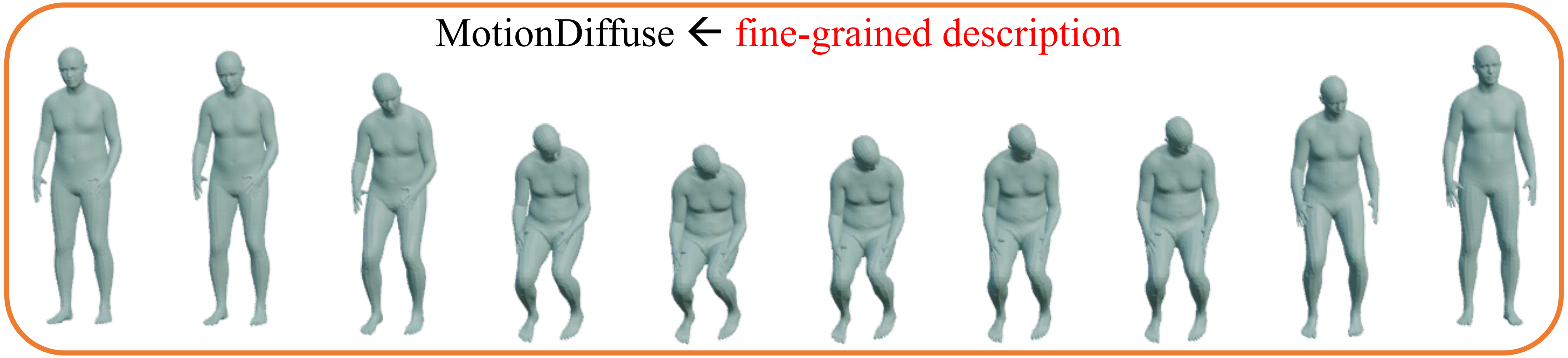
A man raises both arms above head.



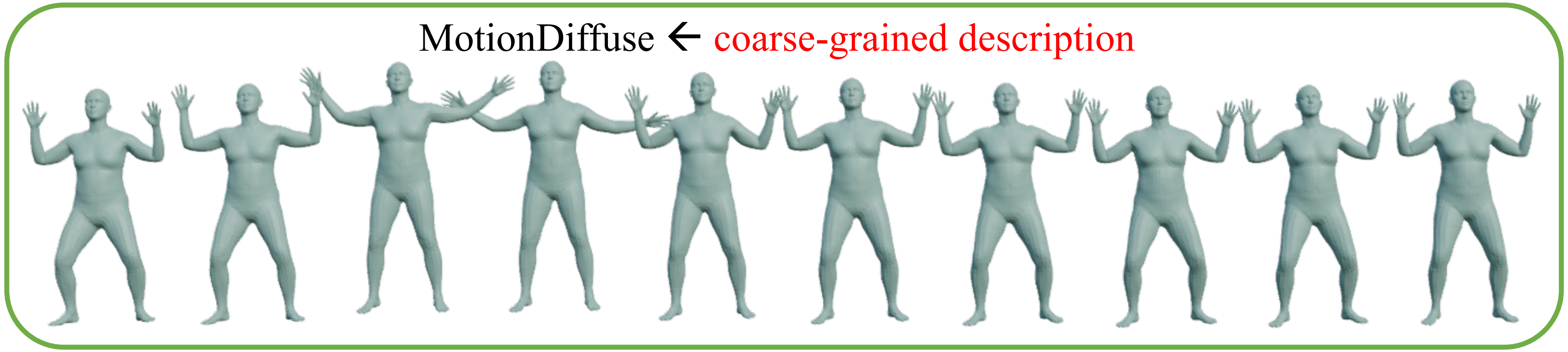
A man slightly squats with both arms raised above head.

Spatial Compositionality

A man slightly squats with both arms raised above head.



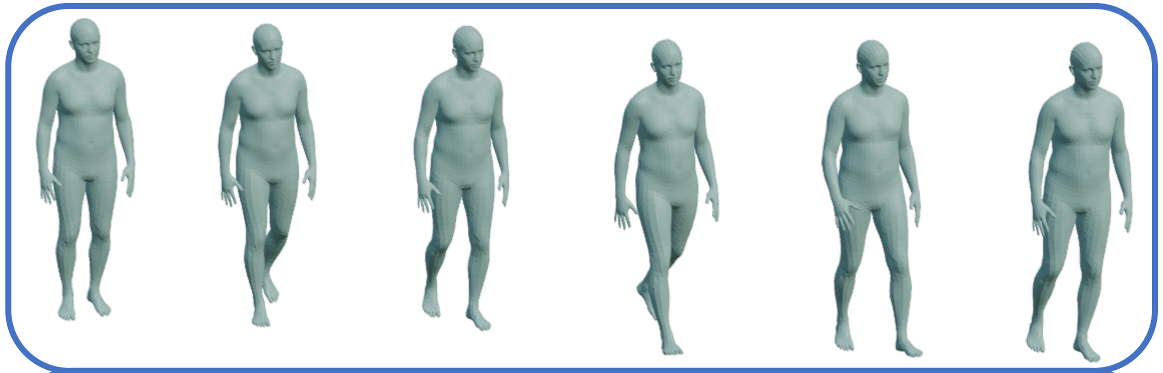
MotionDiffuse ← fine-grained description



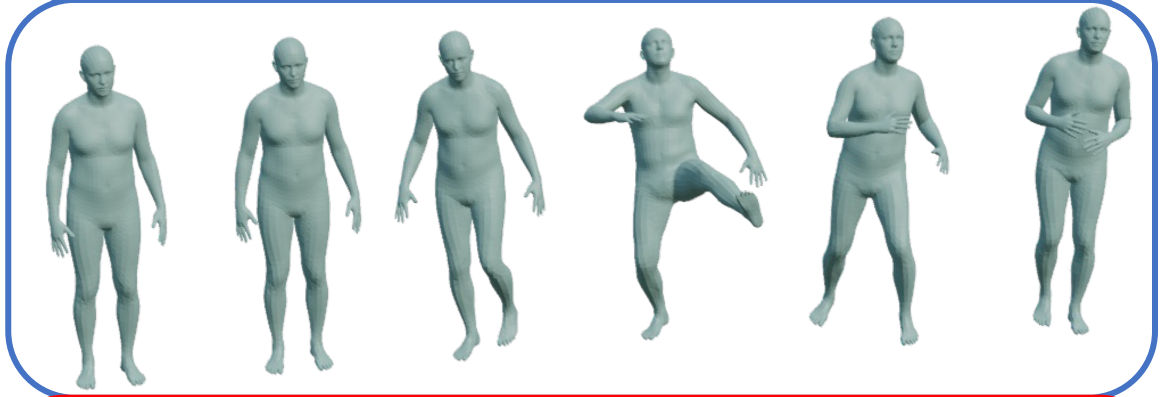
MotionDiffuse ← coarse-grained description

Temporal Compositionality

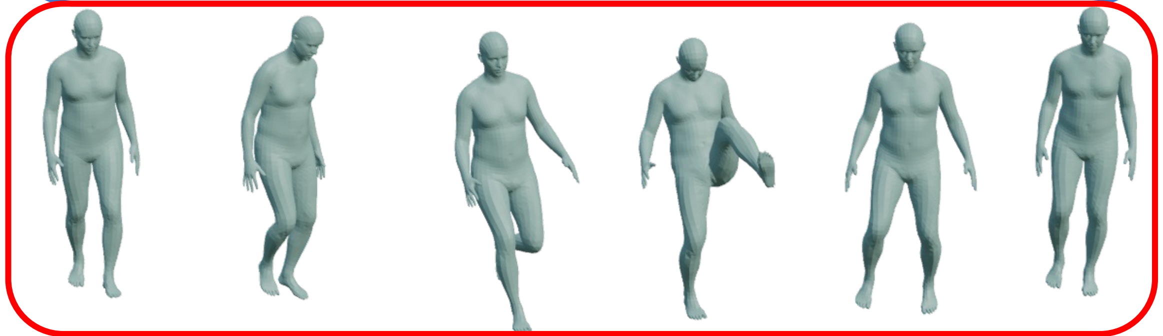
FineMotionDiffuse ← coarse-grained description + fine-grained description



A man walks.



A man kicks with one leg.

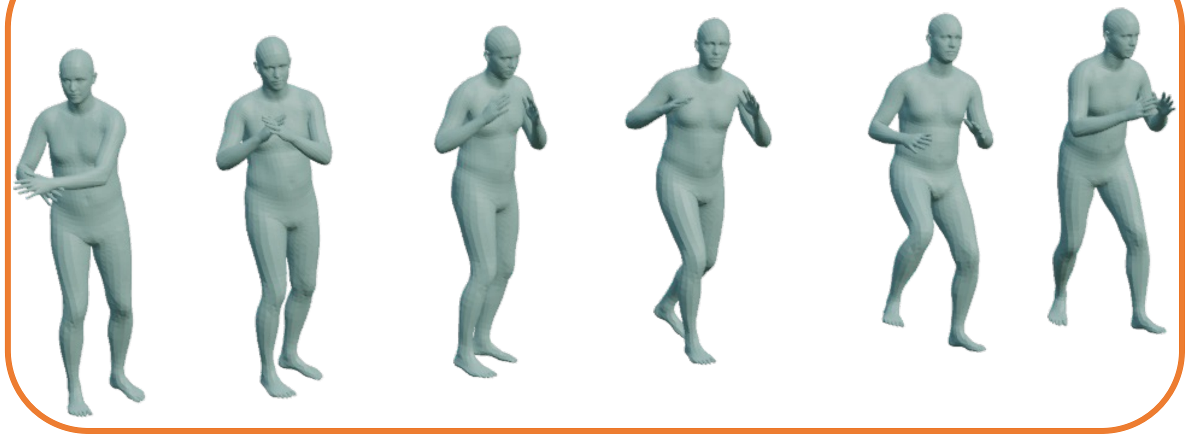


A man walks, then kicks with one leg.

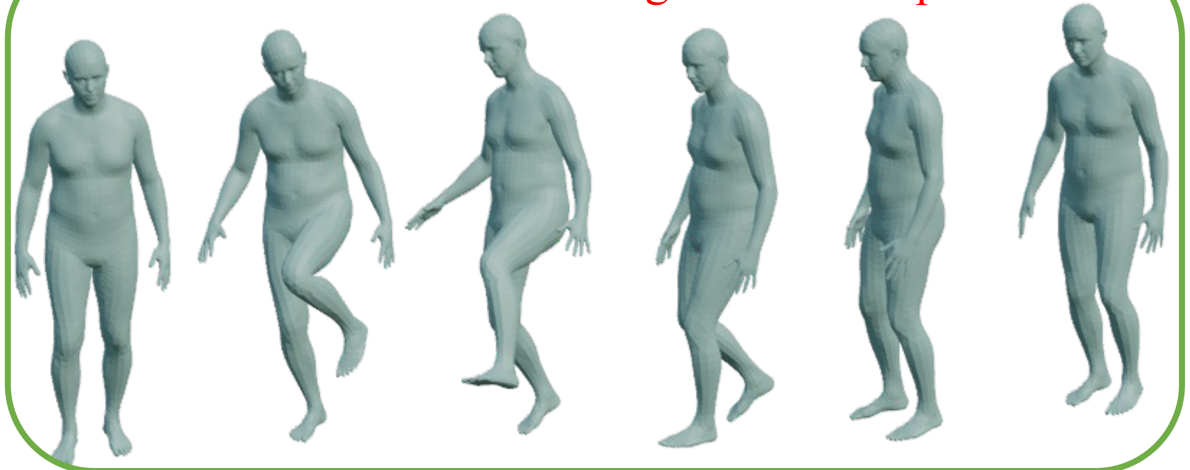
Temporal Compositionality

A man walks, then kicks with one leg.

MotionDiffuse ← fine-grained description



MotionDiffuse ← coarse-grained description



Conclusion

- We find that the LLM has the appropriate knowledge to complement coarse-grained motion descriptions with satisfactory motion details.
- We show that step-aware modeling and fine-coarse cross-attention can make full use of both high-level instruction-like information and fine-grained body-part-related information.
- We find that our method helps the text2motion model learn the mappings from fine-grained descriptions to motion primitives, which leads to the emergence of spatial and temporal compositionality.

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

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