Chronologically Accurate Retrieval for Temporal Grounding of Motion-Language Models

Kent Fujiwara, Mikihiro Tanaka, and Qing Yu

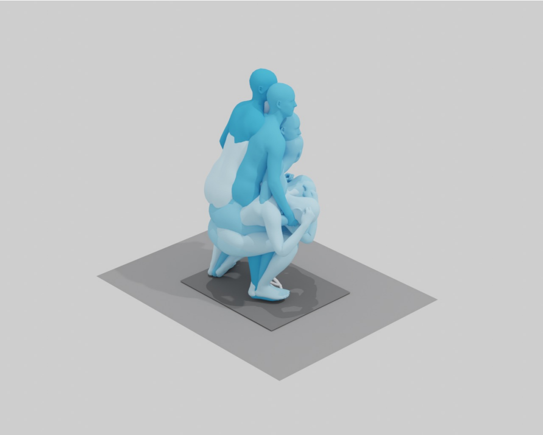
LY Corporation

Research Question

- Can recent motion-language models comprehend chronology of events?

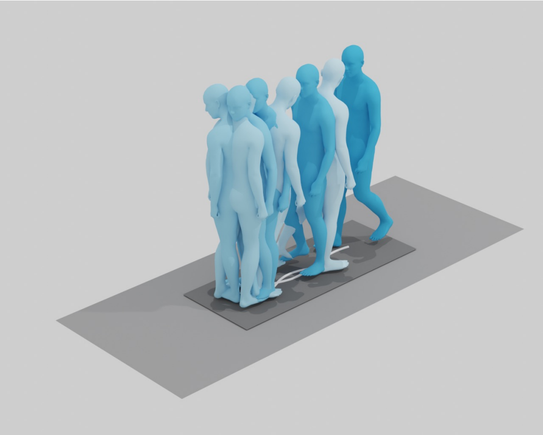
Query

Retrieved text from existing method



1. "a person goes into a ducking position like they are shielding themselves from something."
2. " The person squats down. From 0 seconds to 3 seconds, a person dodges something to his left."
3. "a person squats to lift something up then struggles to carry and put it down."

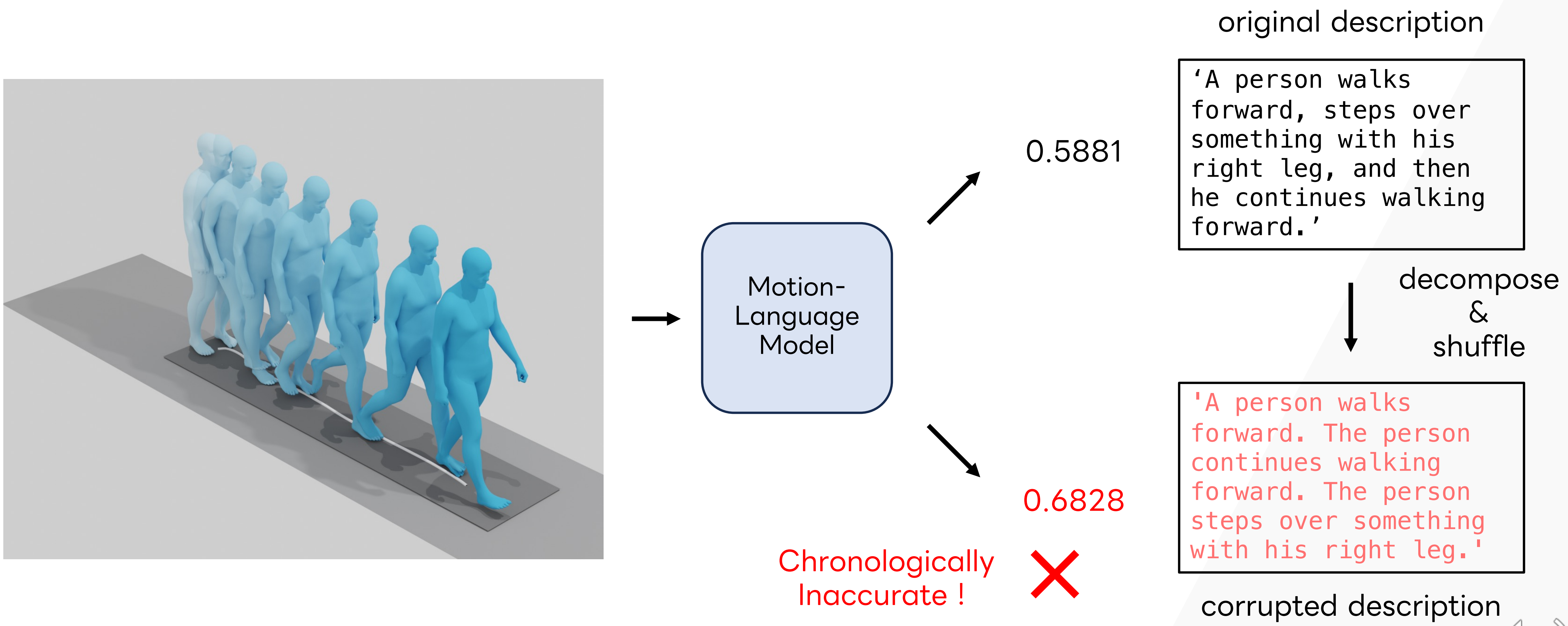
GT: a person goes into a ducking position like they are shielding themselves from something.



1. "The person faces the opposite direction. The person does a full turn. The person moves forward two steps."
2. "The person walks backwards again. A person walks backwards. The person turns around."
3. "The person moves forward two steps. The person faces the opposite direction. The person does a full turn."

GT: a person walks backwards, then turns around then walks backwards again.

CAR: Chronologically Accurate Retrieval



Motion-Language Model

0.5881

original description

```
‘A person walks
forward, steps over
something with his
right leg, and then
he continues walking
forward.’
```

decompose
&
shuffle

```
'A person walks
forward. The person
continues walking
forward. The person
steps over something
with his right leg.'
```

corrupted description

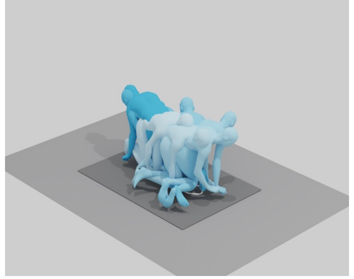
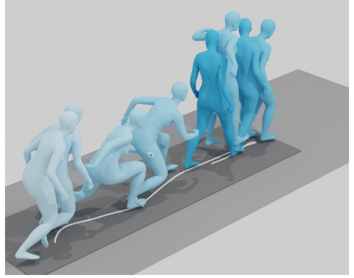
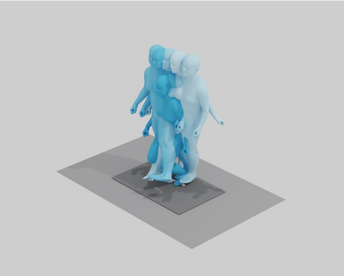
0.6828

Chronologically Inaccurate ! ✗

Decomposing Text into Events

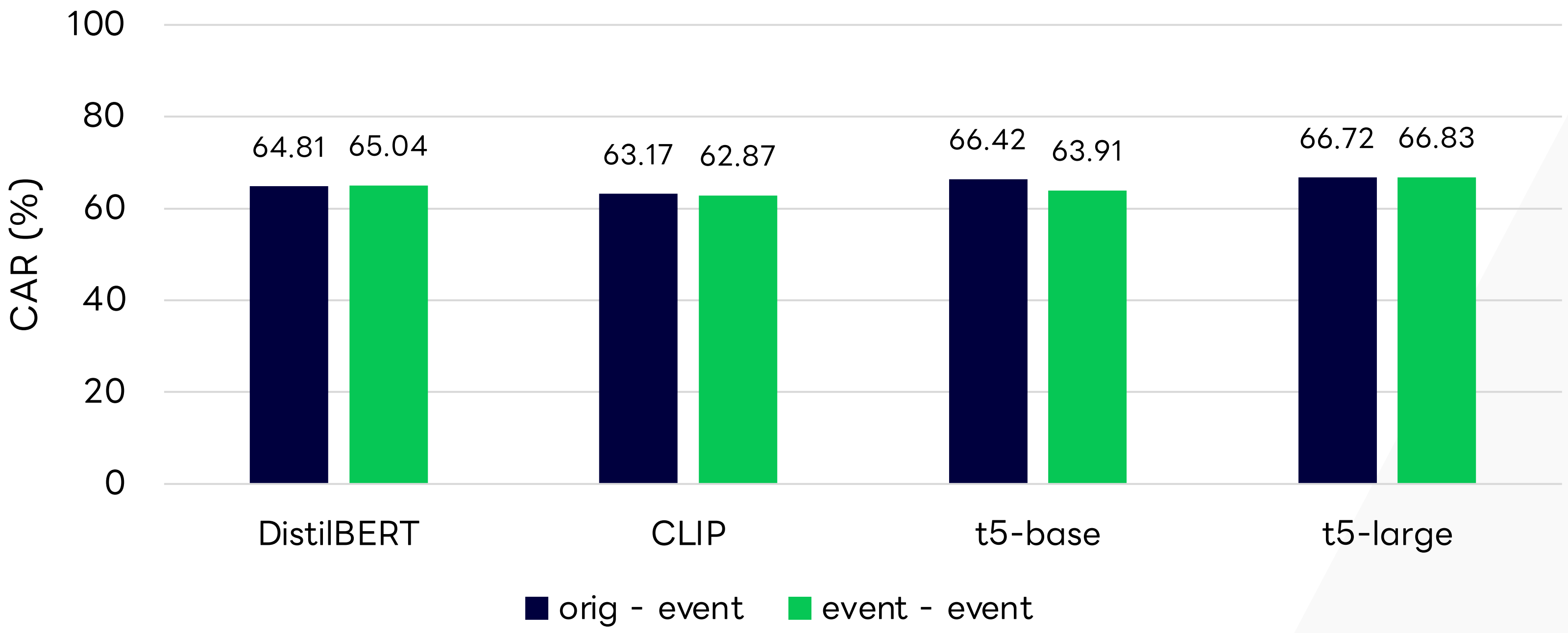
Prompt for decomposition

```
'Please describe the events in the input sentence in the
order in which they occur without omitting any
explanations. Please do not use indicators or pronouns in
sentences. Please include simultaneous actions into one
sentence. Please limit the number of sentences to no more
than the number of verbs. The examples are as follows

Input: A person bends over, using the right leg to bear
weight while kicking back his left leg, and picks
something up with his right hand.

## Examples
Output:
1. A person bends over, using the right leg to bear weight
while kicking back his left leg.
2. The person picks something up with his right hand

Input: A person slowly walked forward.

Output:
1. A person slowly walked forward

Input: Walking forward and then stopping.

Output:
1. The person begins walking forward.
2. The person stops walking.

Input: a person walks slowly while he waves his hands and
then jumps forward.

Output:
1. A person walks slowly while he waves his hands
2. The person jumps forward

________________________

Input: {}'
```

	Original caption	Decomposed Events
	a person gets on his hands and knees and crawls to the left then turns around and crawls back to the right and stands back up on his feet.	"A person gets on his hands and knees." "The person crawls to the left.” "The person turns around.” "The person crawls back to the right." "The person stands back up on his feet."
	a person crouching forward then leaps over something.	"A person crouches forward." "The person leaps over something."
	a person rotates both wrists, wiggles their right foot, wiggles their left foot, bends their knees, then finally sticks their arms out to the side.	"A person rotates both wrists.” "The person wiggles their right foot." "The person wiggles their left foot." "The person bends their knees." "Finally, the person sticks their arms out to the side."

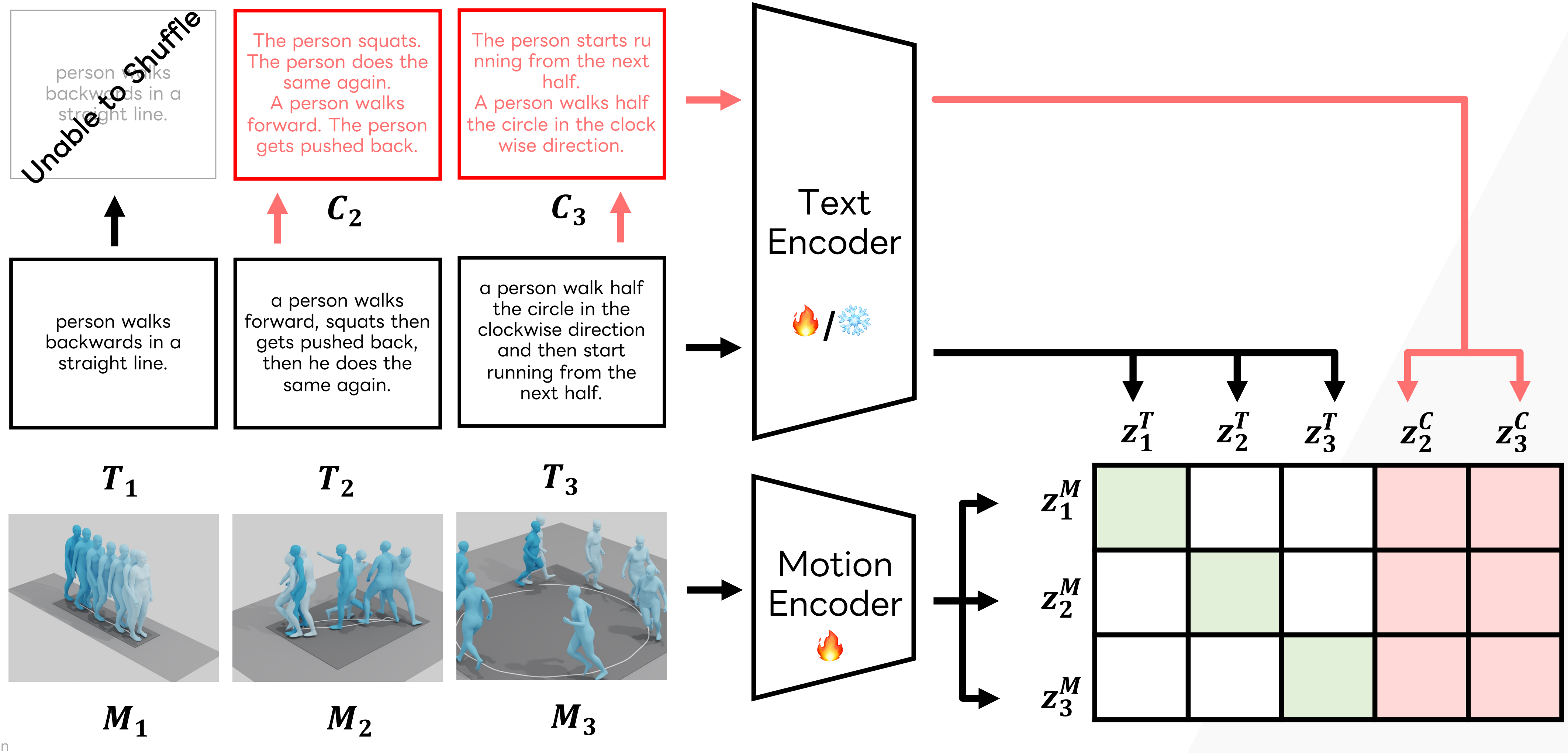
Analysis

- CAR test with original and shuffled texts
  - Base Method: TMR [Petrovich+ ICCV2023]
  - Dataset: HumanML3D [Guo+ CVPR2022]
- Two scenarios
  - "orig – event" : training with original caption, testing with event descriptions
  - "event – event" : training and testing with event descriptions



Proposal

- Reinforcing motion-language model with chronologically inaccurate texts



person walks backwards in a straight line.

Unable to Shuffle

The person squats. The person does the same again. A person walks forward. The person gets pushed back.

The person starts ru nning from the next half. A person walks half the circle in the clock wise direction.

person walks backwards in a straight line.

a person walks forward, squats then gets pushed back, then he does the same again.

a person walk half the circle in the clockwise direction and then start running from the next half.

	z_1^T	z_2^T	z_3^T	z_2^C	z_3^C
z_1^M	🟩			🟥	🟥
z_2^M		🟩		🟥	🟥
z_3^M			🟩	🟥	🟥

Experiment: Motion-to-Text Retrieval

- Motion-to-text retrieval with original and shuffled texts

"orig – event" scenario

Method	R@1↑	R@5↑	R@10↑	MedR↓	CAR↑
TMR	7.89	19.82	28.40	33.75	64.81
Ours	9.38	23.31	34.10	24.00	99.33

”event – event" scenario

Method	R@1↑	R@5↑	R@10↑	MedR↓	CAR↑
TMR	6.93	17.08	26.23	36.00	65.04
Ours	8.90	21.49	31.68	27.50	93.09

Experiment: Retrieval by Fine-tuning

Text-to-motion retrieval

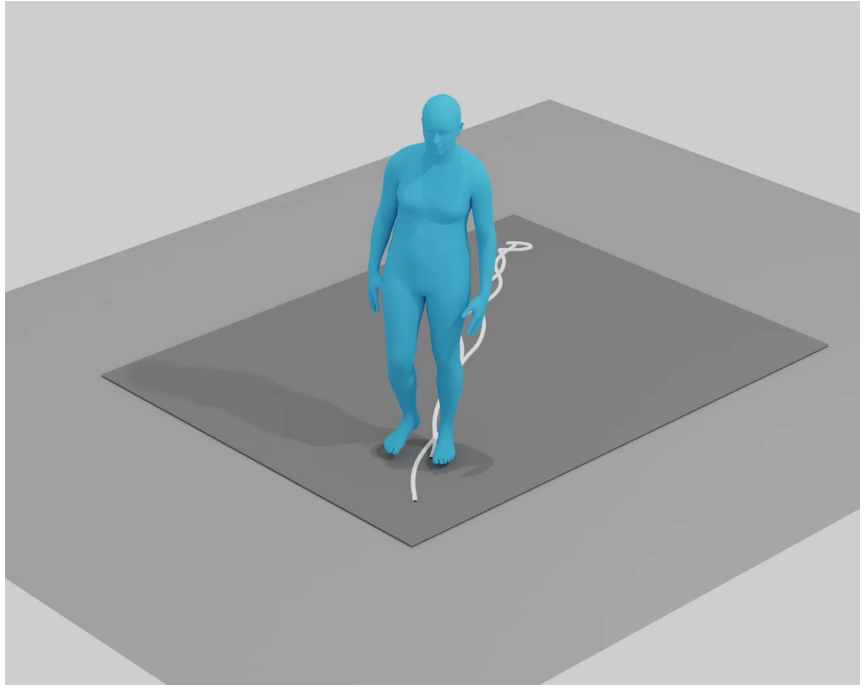
Method	R@1↑	R@5↑	R@10↑	MedR↓
TMR	5.82	21.33	32.76	25.00
Ours+DistilBert	6.55	22.99	34.60	22.00
Ours+CLIP	5.57	20.00	30.06	28.00
Ours+t5-base	6.98	24.98	36.75	19.00
Ours+t5-large	8.03	26.73	38.98	17.00

Motion-to-text retrieval

Method	R@1↑	R@5↑	R@10↑	MedR↓
TMR	9.76	24.13	33.23	23.50
Ours+DistilBert	11.18	25.52	36.38	21.50
Ours+CLIP	8.69	22.06	31.14	27.50
Ours+t5-base	10.90	27.35	38.02	19.50
Ours+t5-large	11.72	28.15	39.23	17.50

Qualitative Results: Retrieval

Query



TMR

1. ”The person faces the opposite direction. The person does a full turn. The person moves forward two steps."
2. "The person walks backwards again. A person walks backwards. The person turns around."
3. "The person moves forward two steps. The person faces the opposite direction. The person does a full turn."

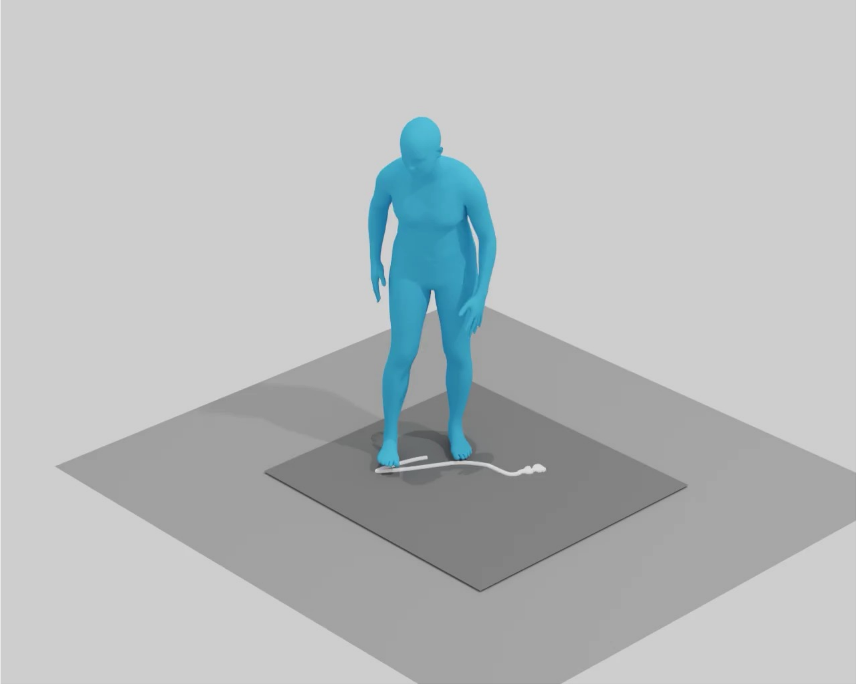
Ours

1. "a person walks backwards, then turns around then walks backwards again.”
2. "a person walks backwards to the right, then turns around and walks backward to the right.”
3. "a person walks backwards to the left, then turns around and walks backward to the left."

GT: a person walks backwards, then turns around then walks backwards again.

Qualitative Results: Retrieval

Query



TMR

1. " The person moves their left hand to the right side in a grabbing motion. The person takes a step lower. A person takes a step."
2. "The person picks up the bottle of shampoo. The person washes the dog.”
3. " The man picks up a brush. The man brushes someone's hair. The man puts the brush back down."

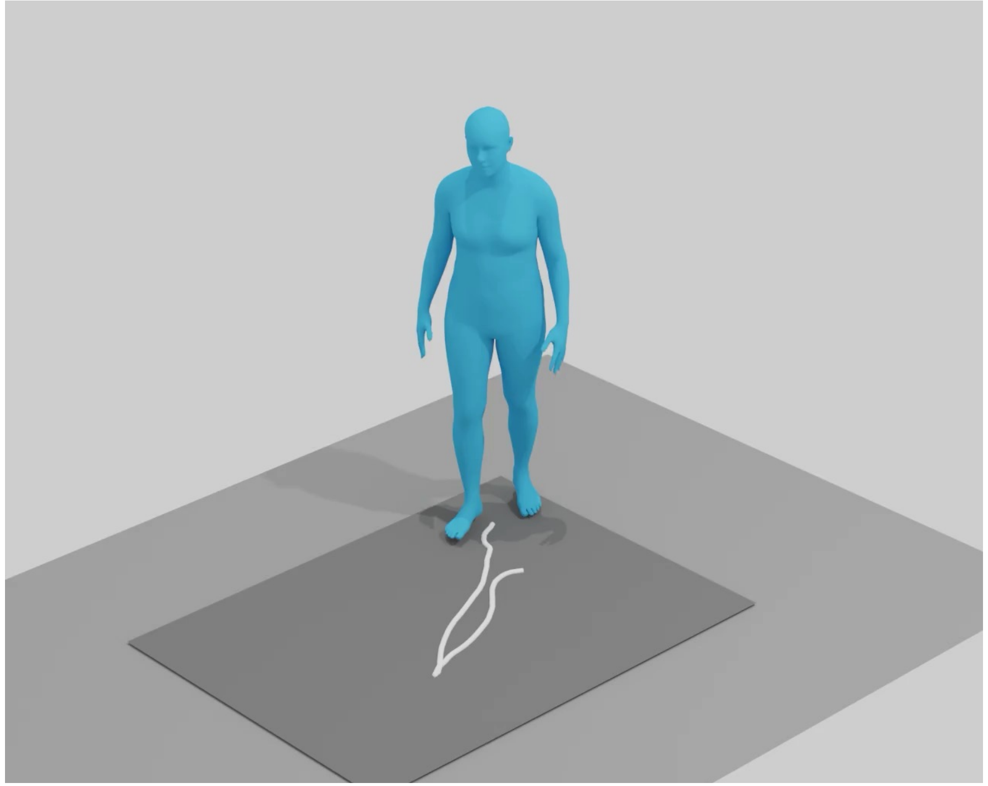
Ours

1. "person bends to pick up something approximately knee high on left side with left hand. he rotates to the right and takes that object and rubs it against something before returning it ."
2. "person bends to pick up something approximately knee high on right side with right hand. he rotates to the left and takes that object and rubs it against something before returning it."
3. "a person bent slightly over and picked something up with left hand and turned to left and shaking item, looks as if is cooking and adding items to a pot of water."

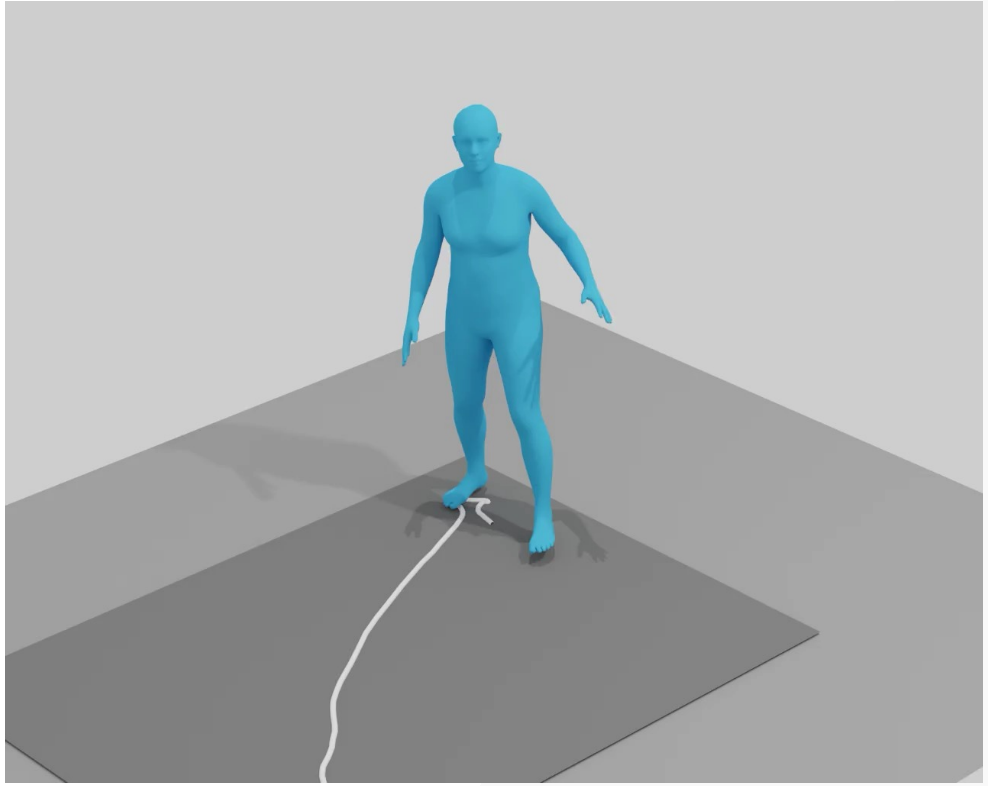
GT: a person bent slightly over and picked something up with left hand and turned to left and shaking item, looks as if is cooking and adding items to a pot of water.

Qualitative Results: Generation

Prompt A person jumps forwards and kicks something.



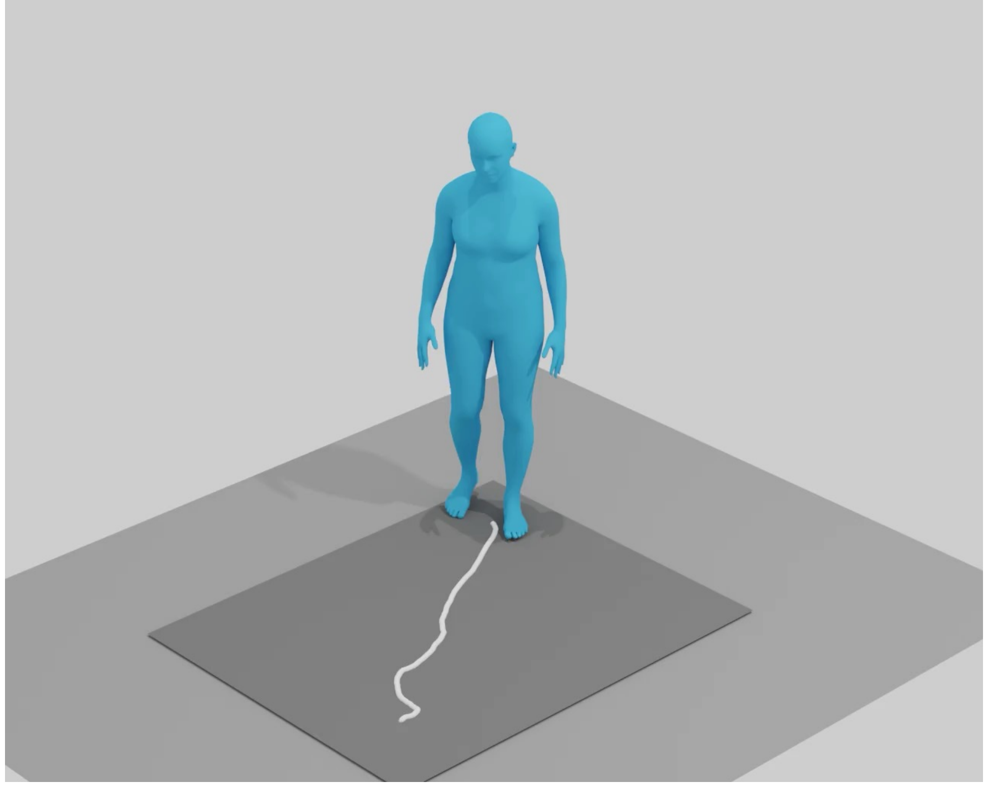
T2M-GPT [Zhang+ CVPR2023]



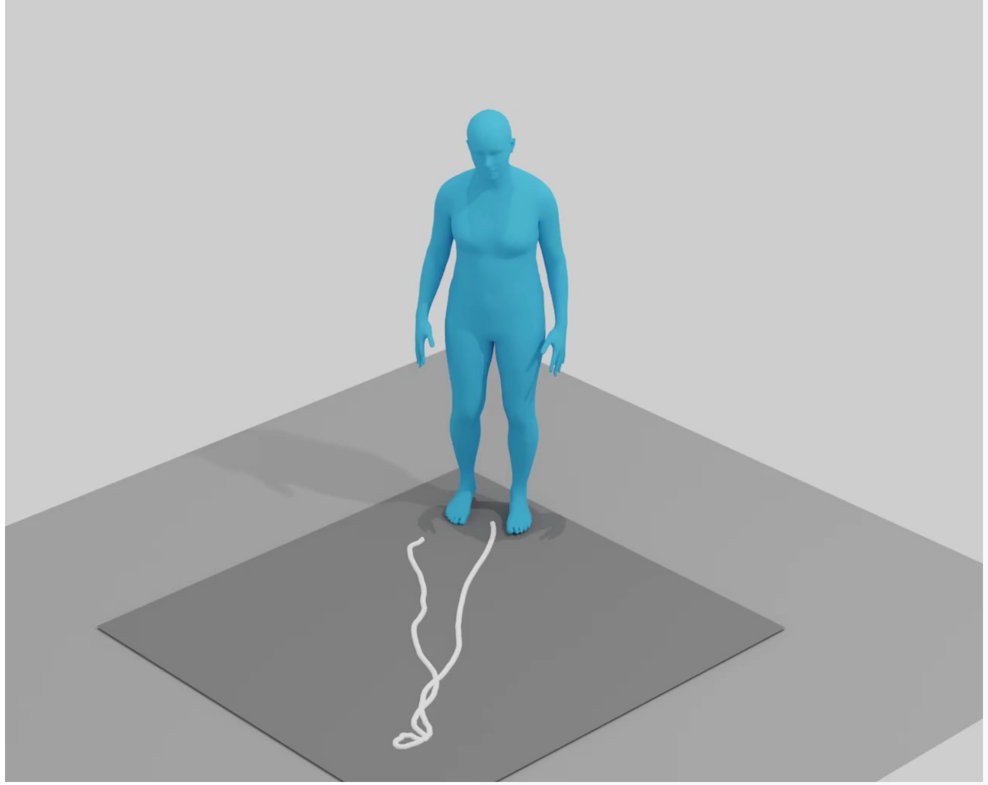
Ours

Qualitative Results: Generation

Prompt A person is ascending a staircase and then descending.



T2M-GPT [Zhang+ CVPR2023]



Ours

Additional Content

- Scan the QR Code for…

- Thorough Analysis

    - Comparison with prior methods
    - Complete retrieval and generation results
    - Effect of pronouns

- Additional Visualization

Paper

Project Page