HumanTOMATO: Text-aligned Whole-body Motion Generation

Page: https://lhchen.top/HumanTOMATO

Ling-Hao CHEN
Tsinghua University, IDEA Research
thu.lhchen@gmail.com

June 5, 2024

Road Map

General Target:
Synthesizing motions we want!

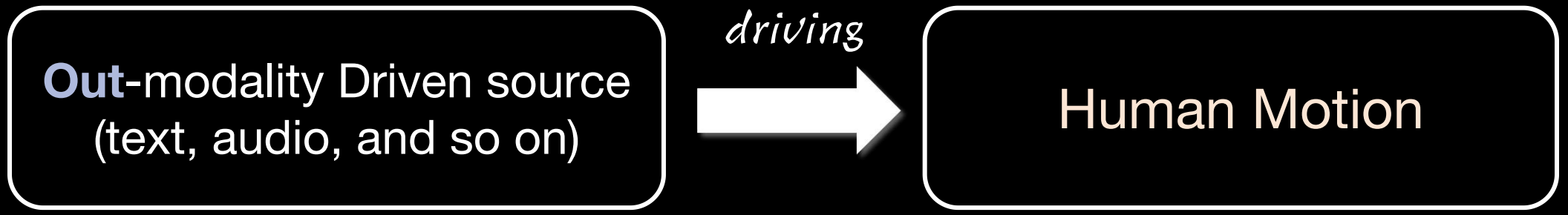
Road Map

General Target:
Synthesizing motions we want!

Human Motion

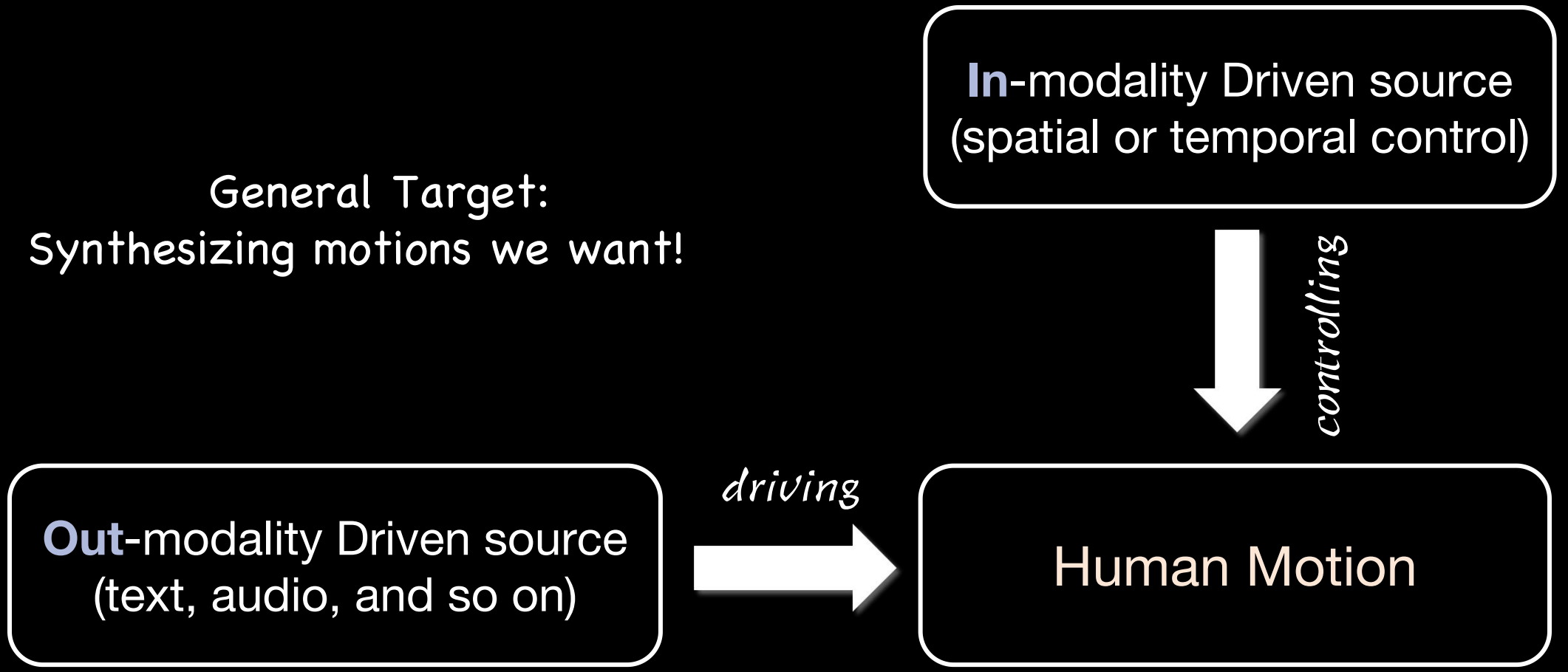
Road Map

General Target:
Synthesizing motions we want!



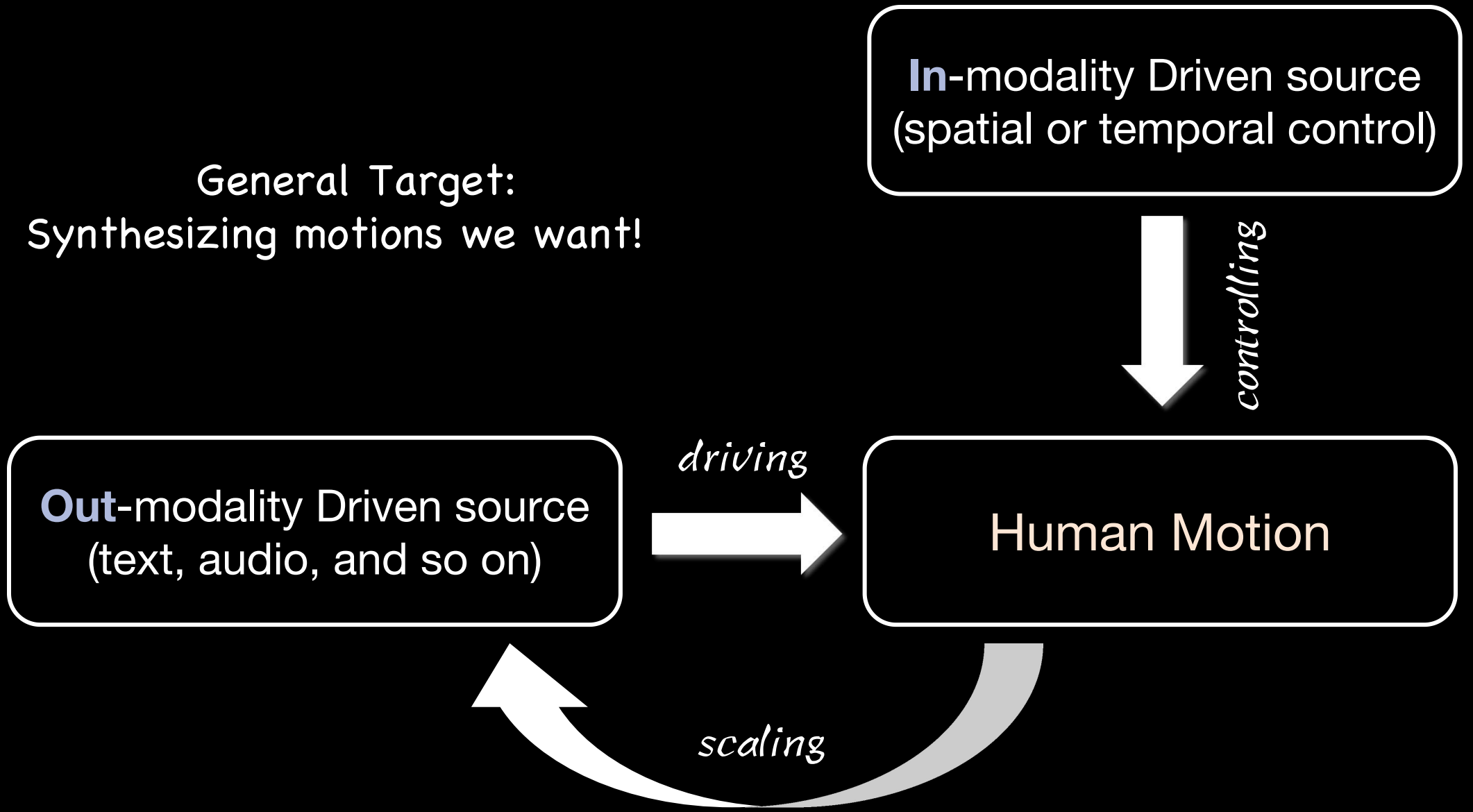
Road Map

General Target:
Synthesizing motions we want!



Road Map

General Target:
Synthesizing motions we want!



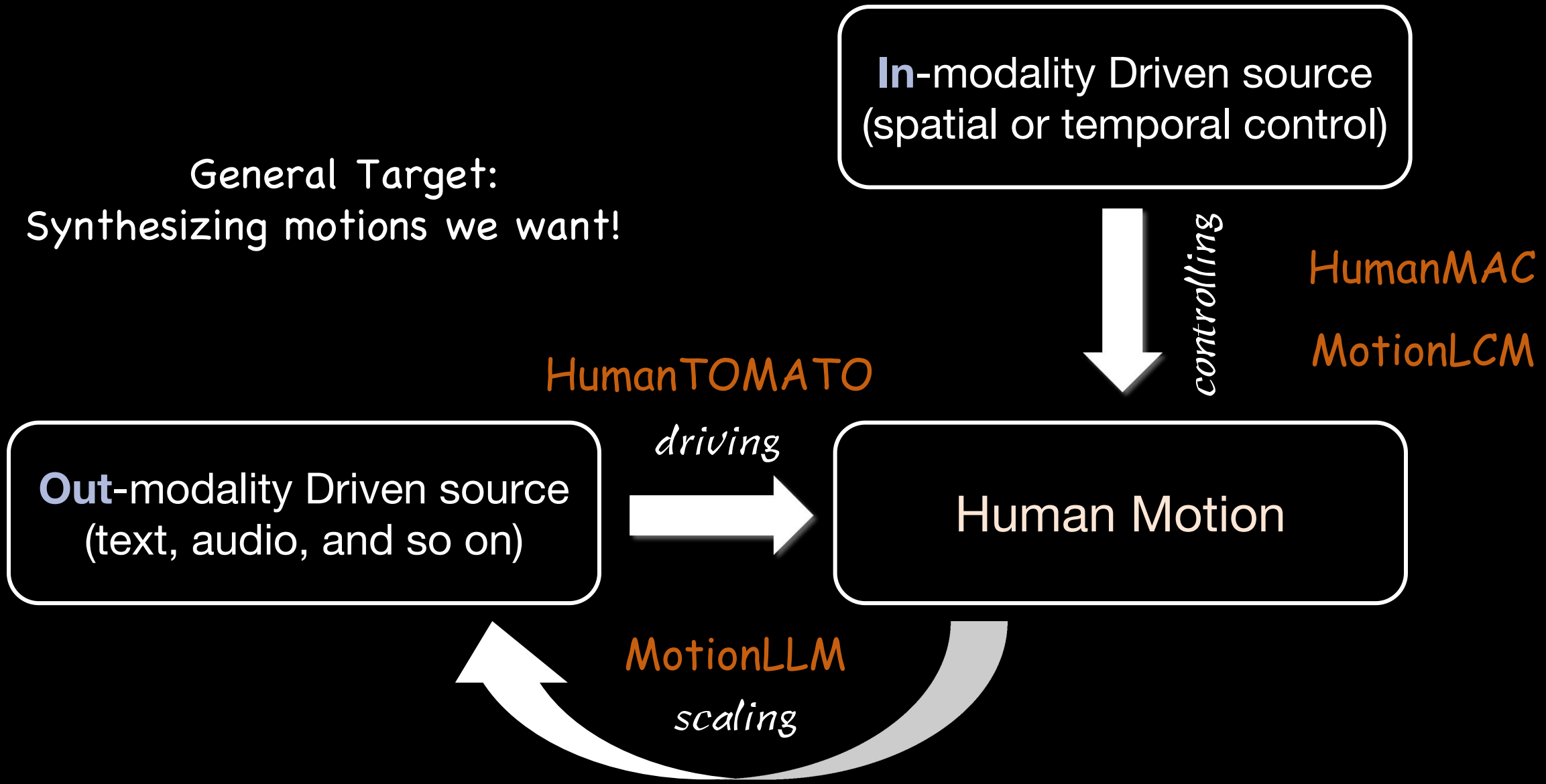
Road Map

General Target:
Synthesizing motions we want!

- ☐ Synthesizing text-aligned human motion (HumanTOMATO)
- ☐ Zero-shot motion controlling (HumanMAC)
- ☐ Tunable, real-time, and multi-modality motion controlling (MotionLCM)
- ☐ Exploring on the scaling law for human motion (MotionLLM)

Road Map

General Target:
Synthesizing motions we want!



Synthesizing text-aligned human motion

Synthesizing text-aligned human motion

The man is playing the ukulele, happily.

Synthesizing text-aligned human motion

- Emphasizing one of my belief:

  The source of intelligence is compression.

- Ilya Sutskever, previously from OpenAI:

  Prediction is compression;

  I find easier to think about compression when talking about unsupervised learning.

- Yi Ma *et al.* :

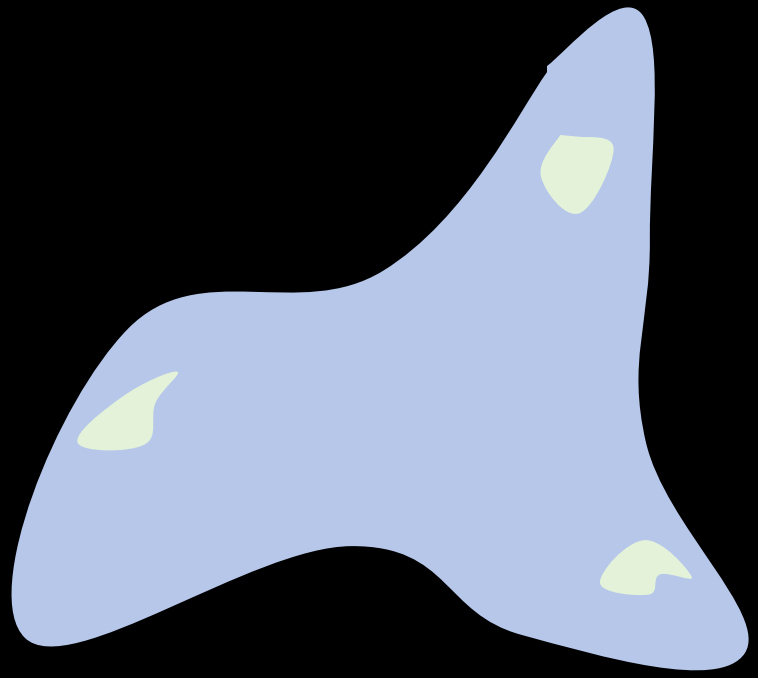
  Closed-loop compressive architectures are ubiquitous for all intelligent beings and at all scales, from the brain to spinal circuits down to DNA.

Synthesizing text-aligned human motion

- Quick question

  Q: Why is the compression so important for human motion?

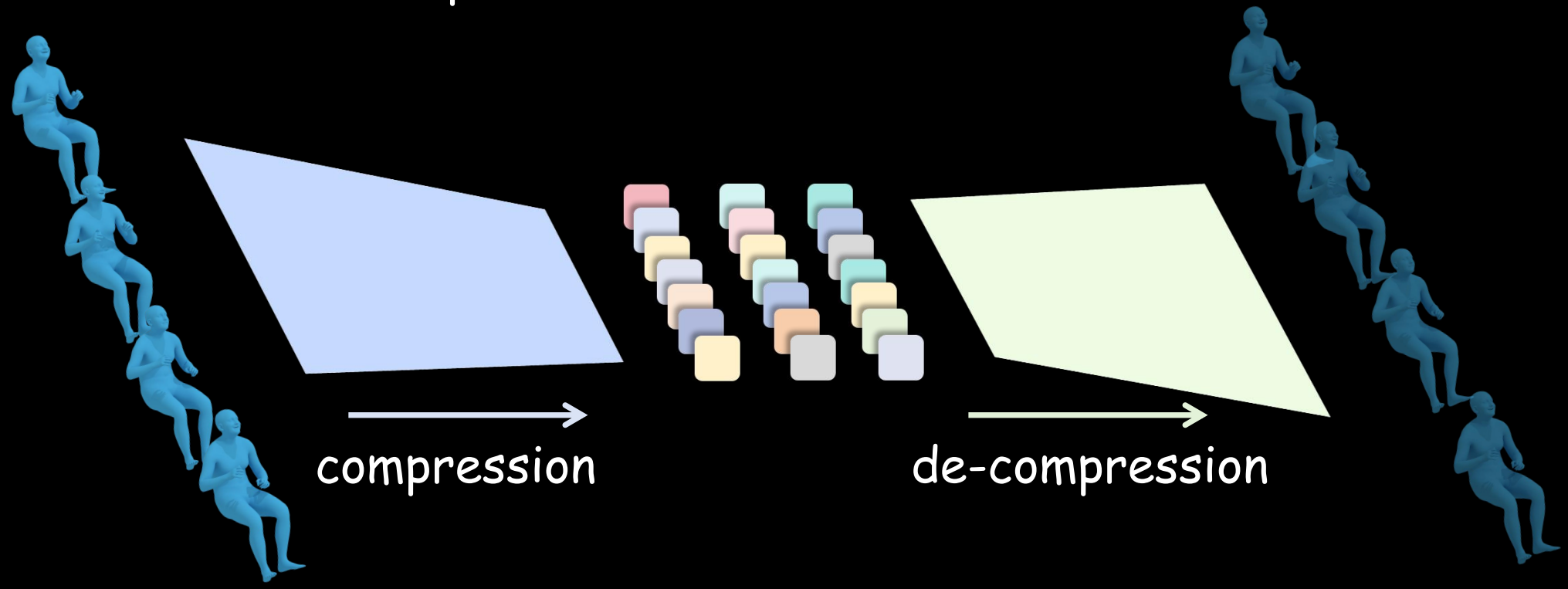
  A: Vinilla motion sapce is too noisy.



$$\frac{\text{vol}(\#\text{motion})}{\text{vol}(\#\text{whole})} \longrightarrow 0$$

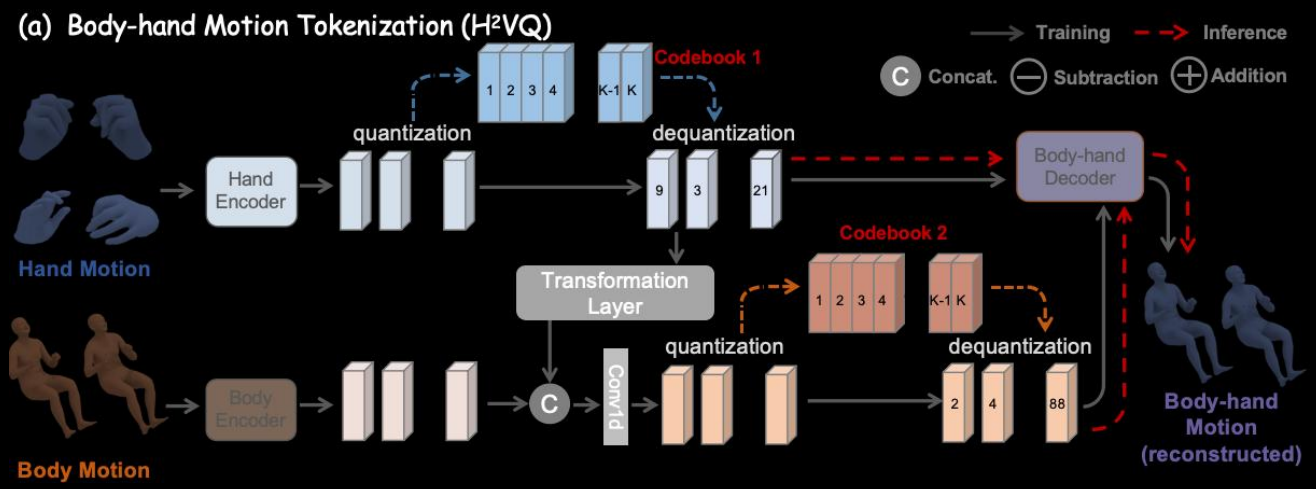
Synthesizing text-aligned human motion

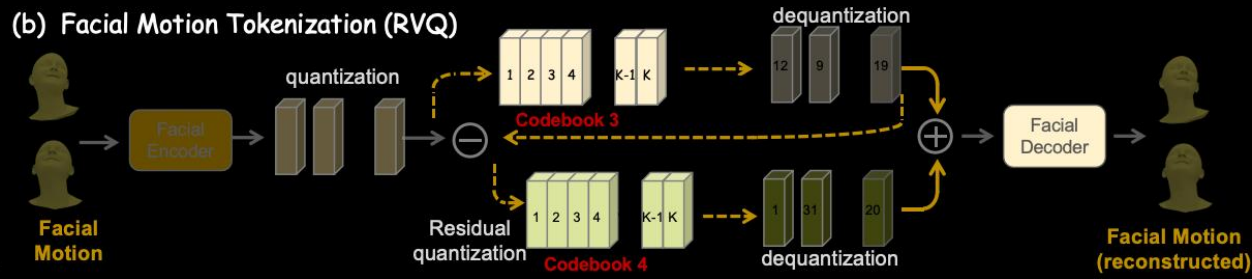
- Quick question
  - Q: Why is the compression so important for human motion?
  - A: Vinilla motion sapce is too noisy.
- Technical solution: Compression motion at first!



Synthesizing text-aligned human motion

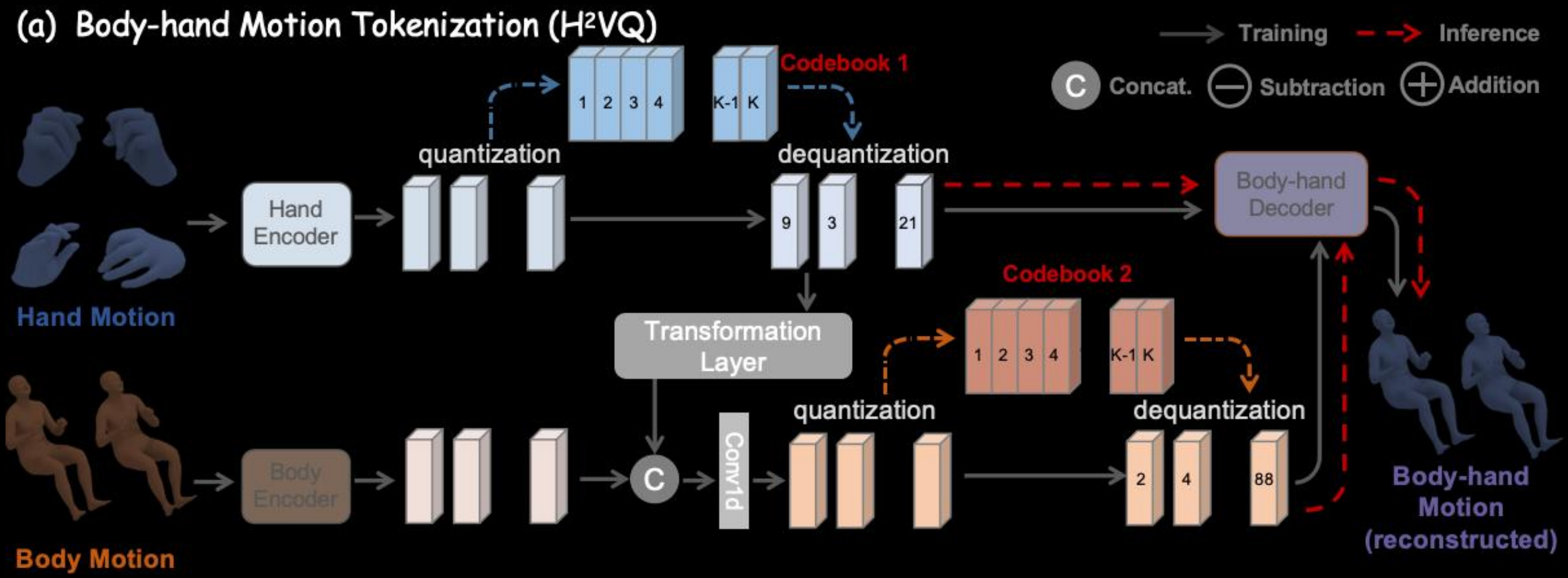
- Quick question
  - Q: Why is the compression so important for human motion?
  - A: Vinilla motion sapce is too noisy.
- Technical solution: Compression motion at first!
- High-quality compression principal: Good reconstruction quality!





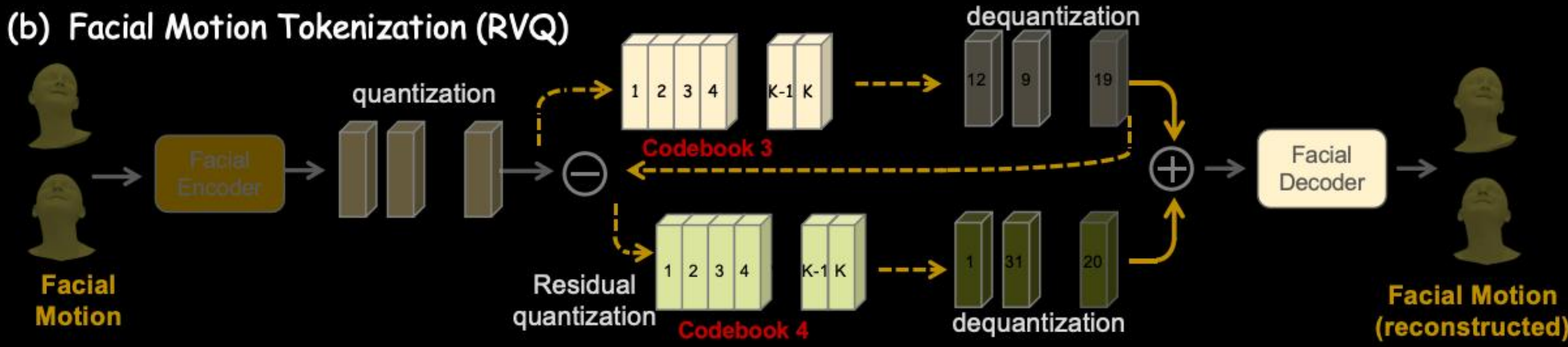
Synthesizing text-aligned human motion

- High-quality compression principal: Good reconstruction quality!



Synthesizing text-aligned human motion

- High-quality compression principal: Good reconstruction quality!

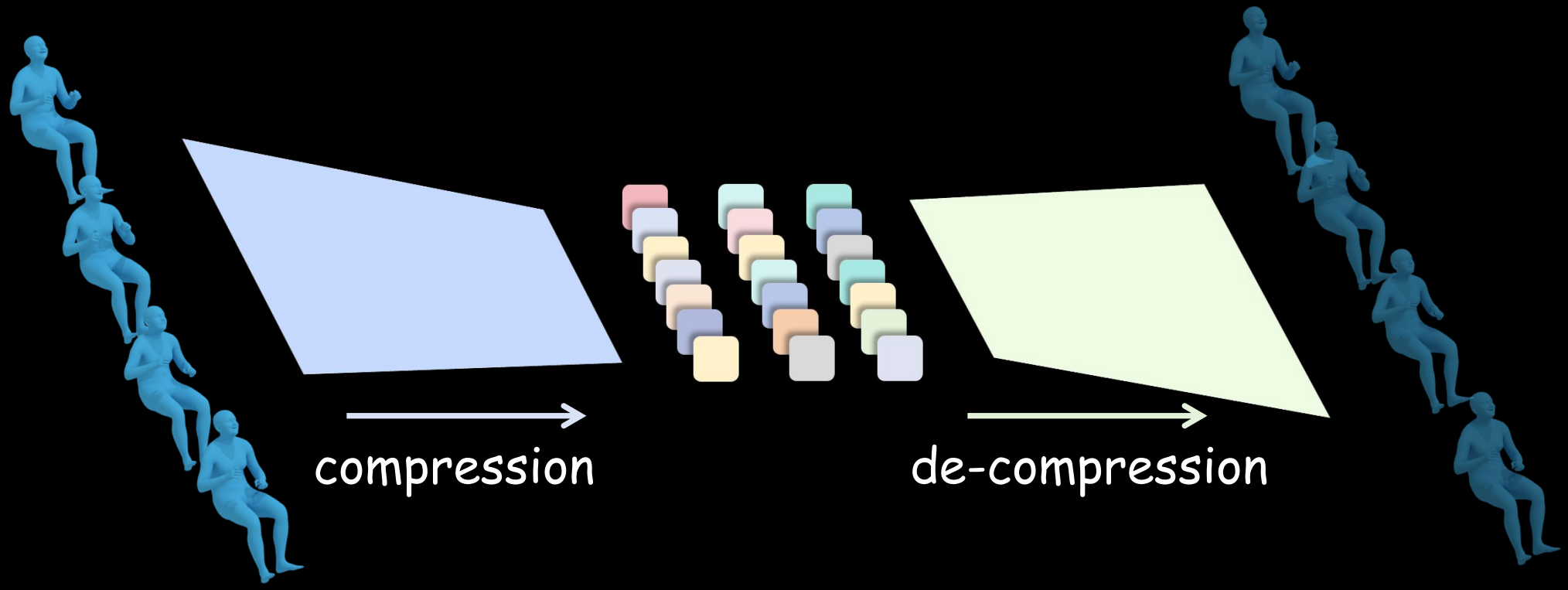


Synthesizing text-aligned human motion

- Text-aligned generation

Synthesizing text-aligned human motion

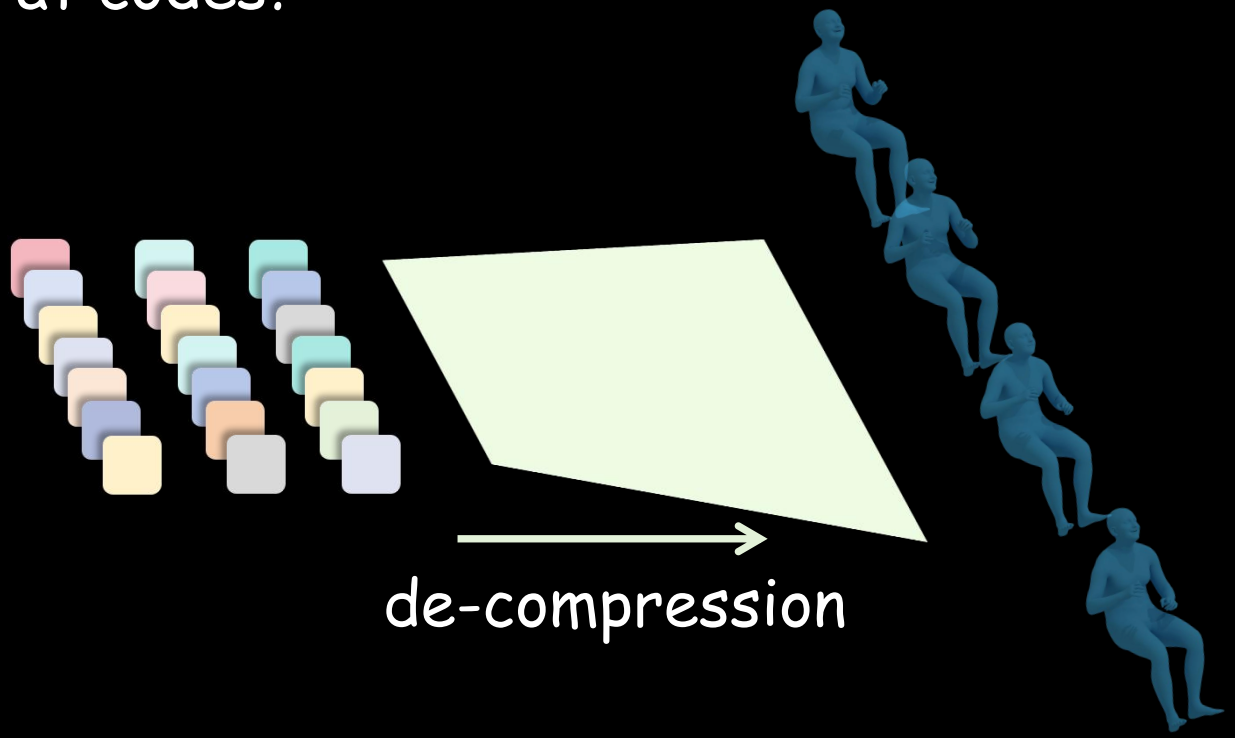
- Text-aligned generation



Synthesizing text-aligned human motion

- Text-aligned generation

  Q1: How generat codes?



Synthesizing text-aligned human motion

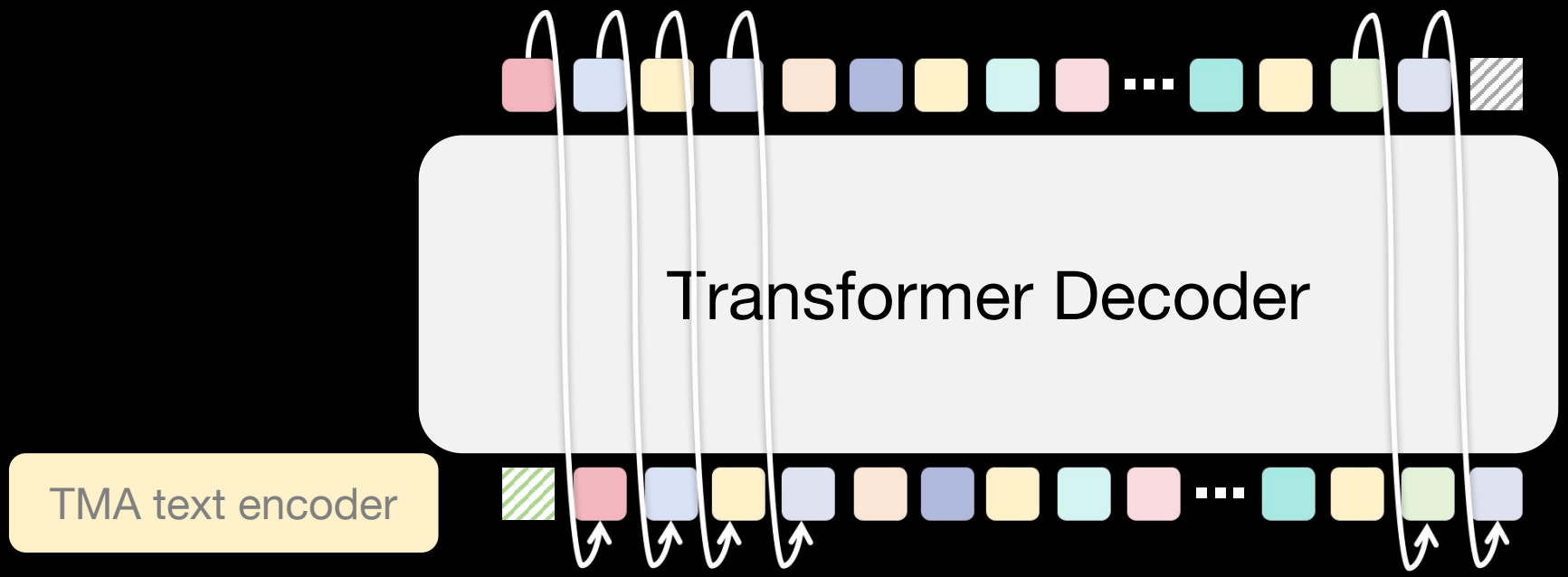
- Text-aligned generation

  Q1: How generat codes?

Synthesizing text-aligned human motion

- Text-aligned generation

  Q1: How generat codes?

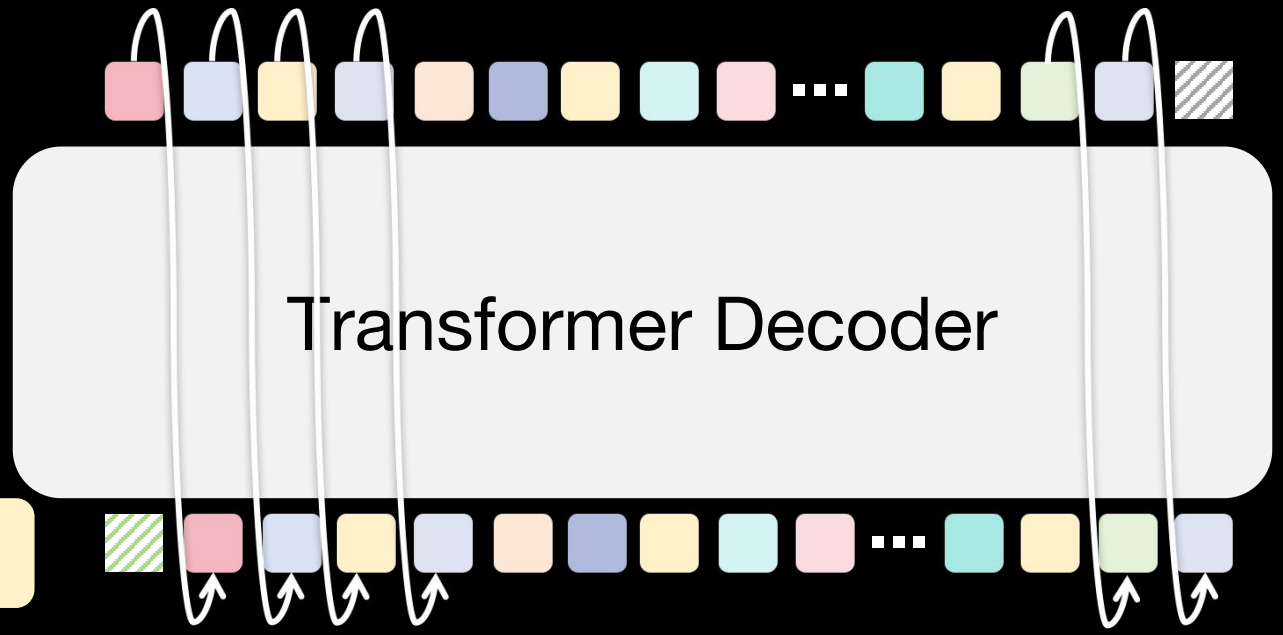


Synthesizing text-aligned human motion

- Text-aligned generation

  Q1: How generat codes?

  Q2: How to achieve text-motion alignment?

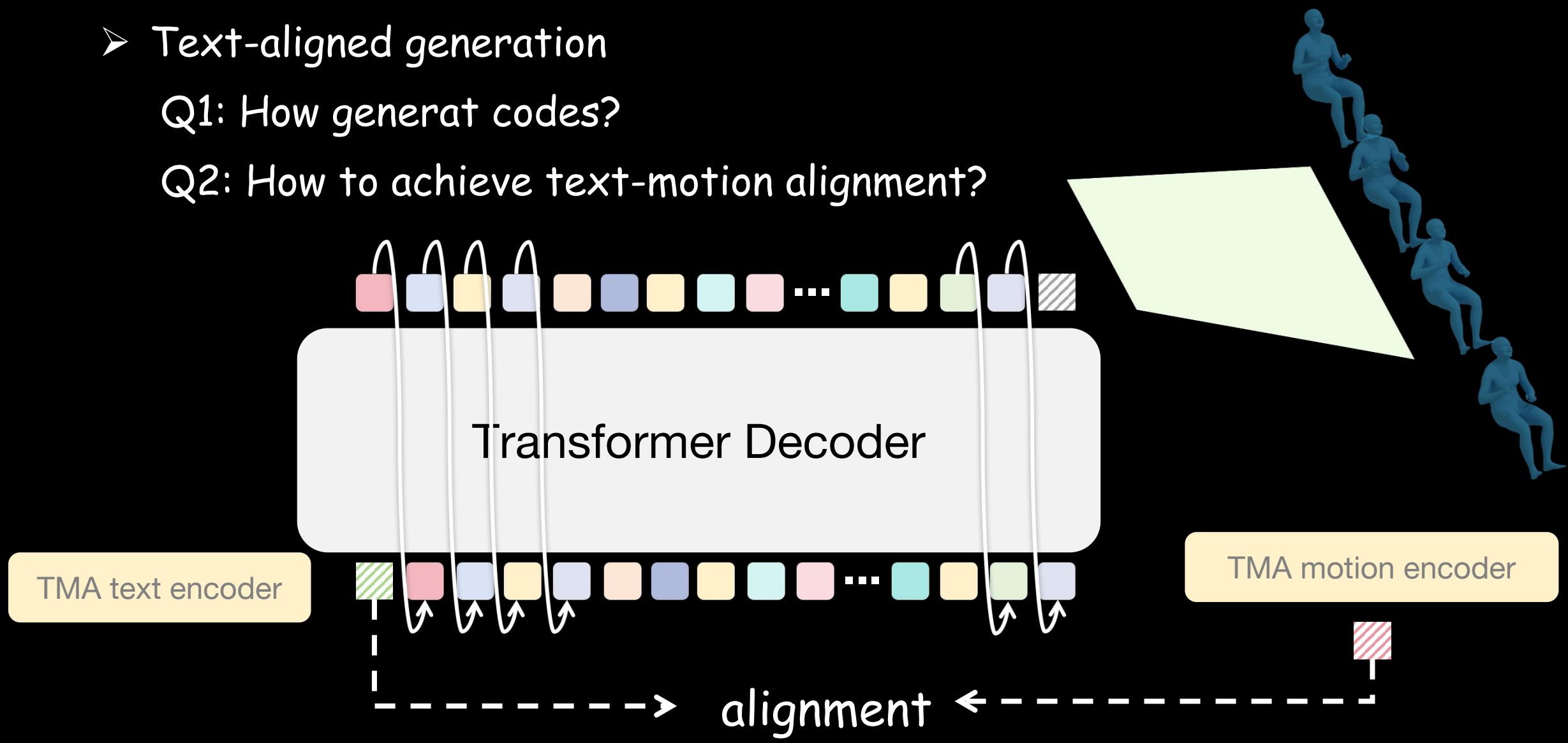


Synthesizing text-aligned human motion

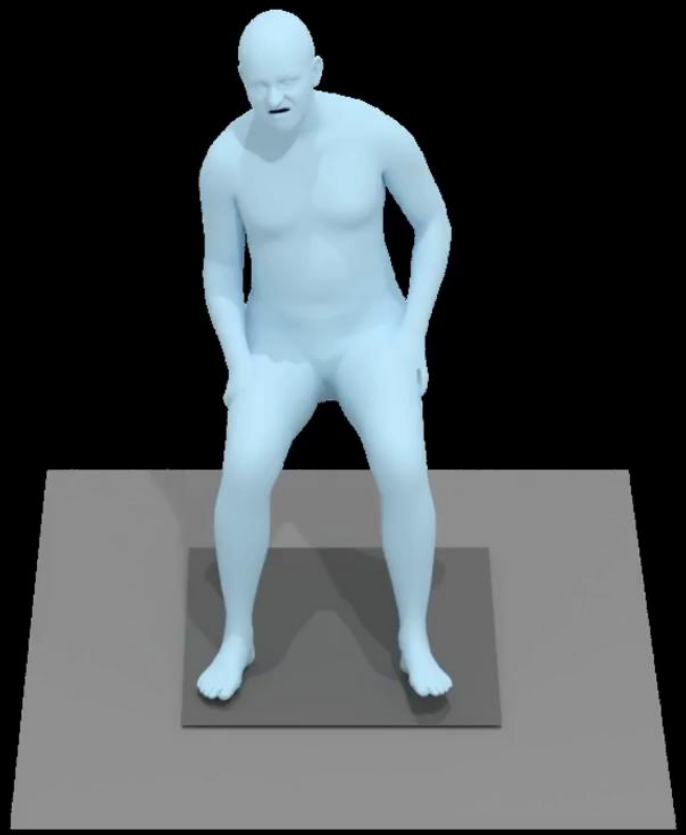
- Text-aligned generation

  Q1: How generat codes?

  Q2: How to achieve text-motion alignment?



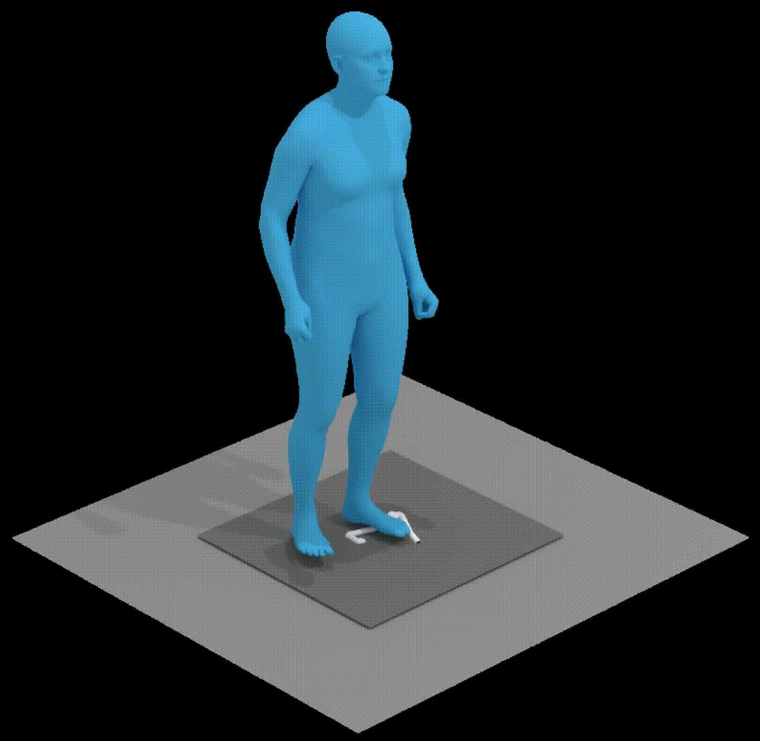
Synthesizing text-aligned human motion



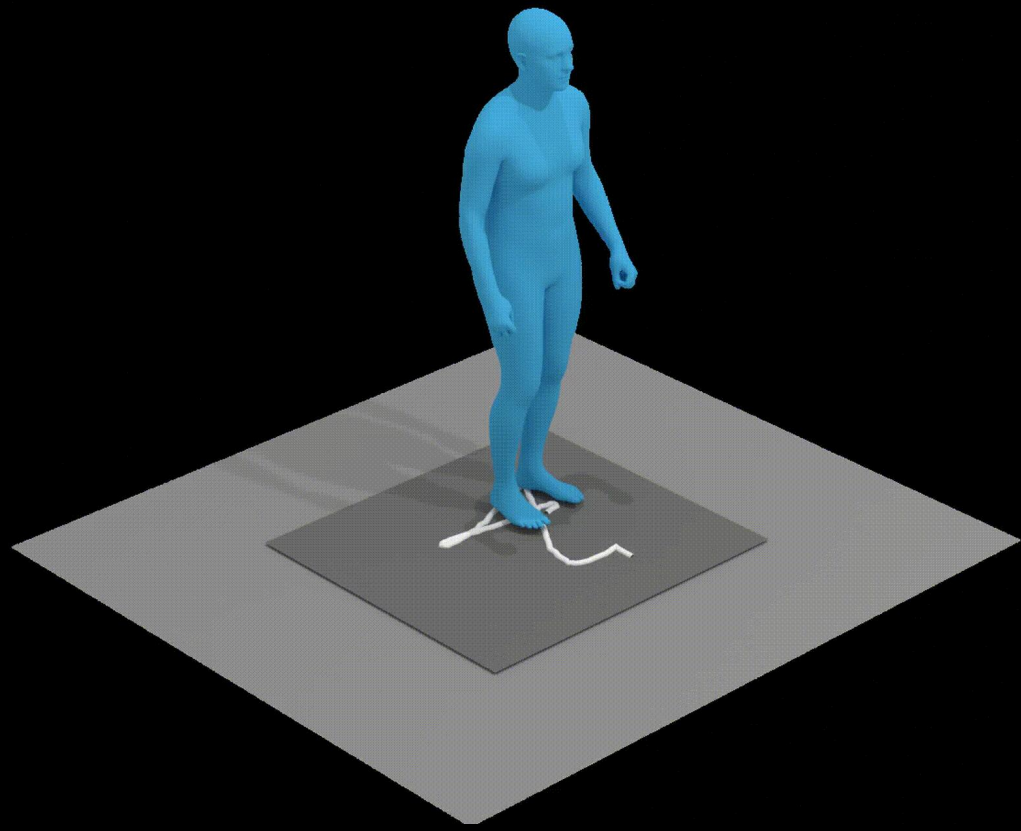
here we show some cases generated by human tomato

Synthesizing text-aligned human motion

man swivels left to right with left hand up towards the face.



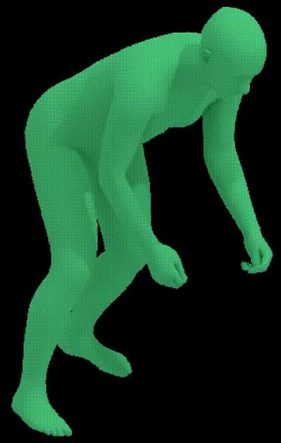
HumanTOMATO



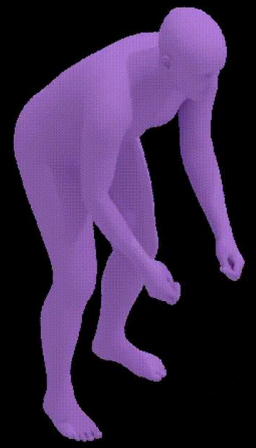
GT

Synthesizing text-aligned human motion

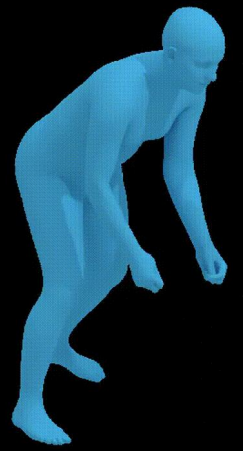
Body-level Comparison



GT



Vanilla VQ

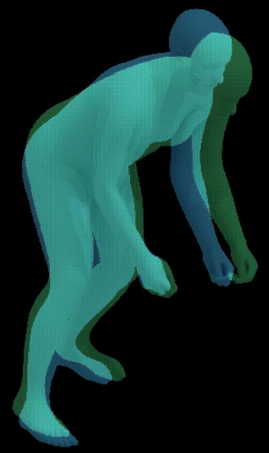


H^2VQ

Synthesizing text-aligned human motion

Body-level

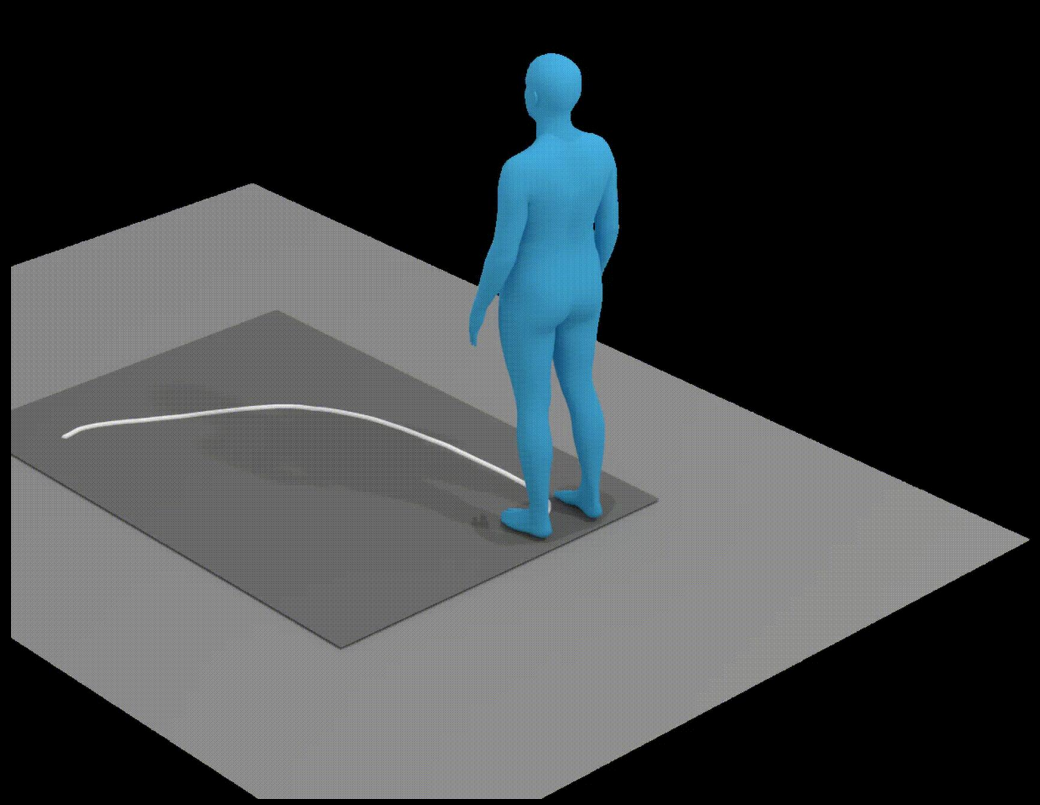
Vanilla VQ



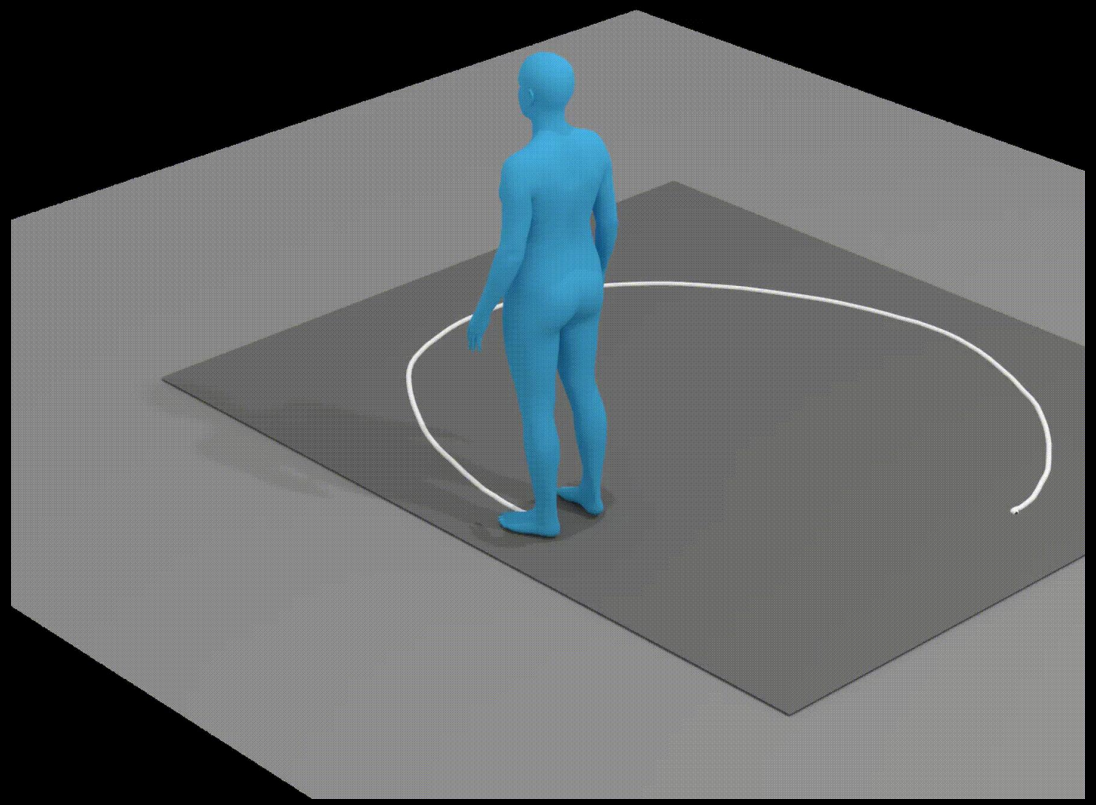
H^2VQ

Synthesizing text-aligned human motion

A person walks clockwise-ly.



w/o TMA prior



w/ TMA prior

Synthesizing text-aligned human motion

	Motion-X			GRAB			HumanML3D
	All↓	Body ↓	Hand↓	All ↓	Body ↓	Hand ↓	Body ↓
Vanilla VQ (512)	140.66	92.20	46.45	78.23	38.29	31.48	77.21
Vanilla VQ (1024)	139.33	91.77	46.40	76.01	37.34	29.89	71.34
RVQ (512×2)	110.94	73.97	40.01	62.94	31.12	27.28	**63.05**
H^2VQ (512×2)	**92.97**	**62.34**	**37.20**	**46.74**	**24.33**	**24.59**	-

Table 7: Comparison of the motion reconstruction errors (MPJPE in mm) of different quantization methods on Motion-X, GRAB, and HumanML3D. Our H^2VQ shows significant improvements.

embedding	supervision	FID ↓	R-Precision$^{(32)}$			TMA-R-Precision$^{(256)}$			Matching-score ↓	TMA-Matching-score ↓
			Top1 ↑	Top2 ↑	Top3 ↑	Top1 ↑	Top2 ↑	Top3 ↑		
GT		0.002	0.500	0.708	0.814	0.407	0.578	0.673	2.888	0.768
CLIP	✗	**1.086**	0.405	0.588	0.695	0.345	0.490	0.573	3.917	0.844
TMA	✗	1.290	0.416	0.596	0.699	0.395	0.550	0.637	3.950	0.815
TMA	✓	1.174	**0.416**	**0.603**	**0.703**	**0.399**	**0.555**	**0.638**	**3.894**	**0.809**

Table 8: Ablation on a pre-trained text-motion-aligned model for motion generation on Motion-X. Both TMA embedding and text-motion alignment supervision help generate text-aligned motions.

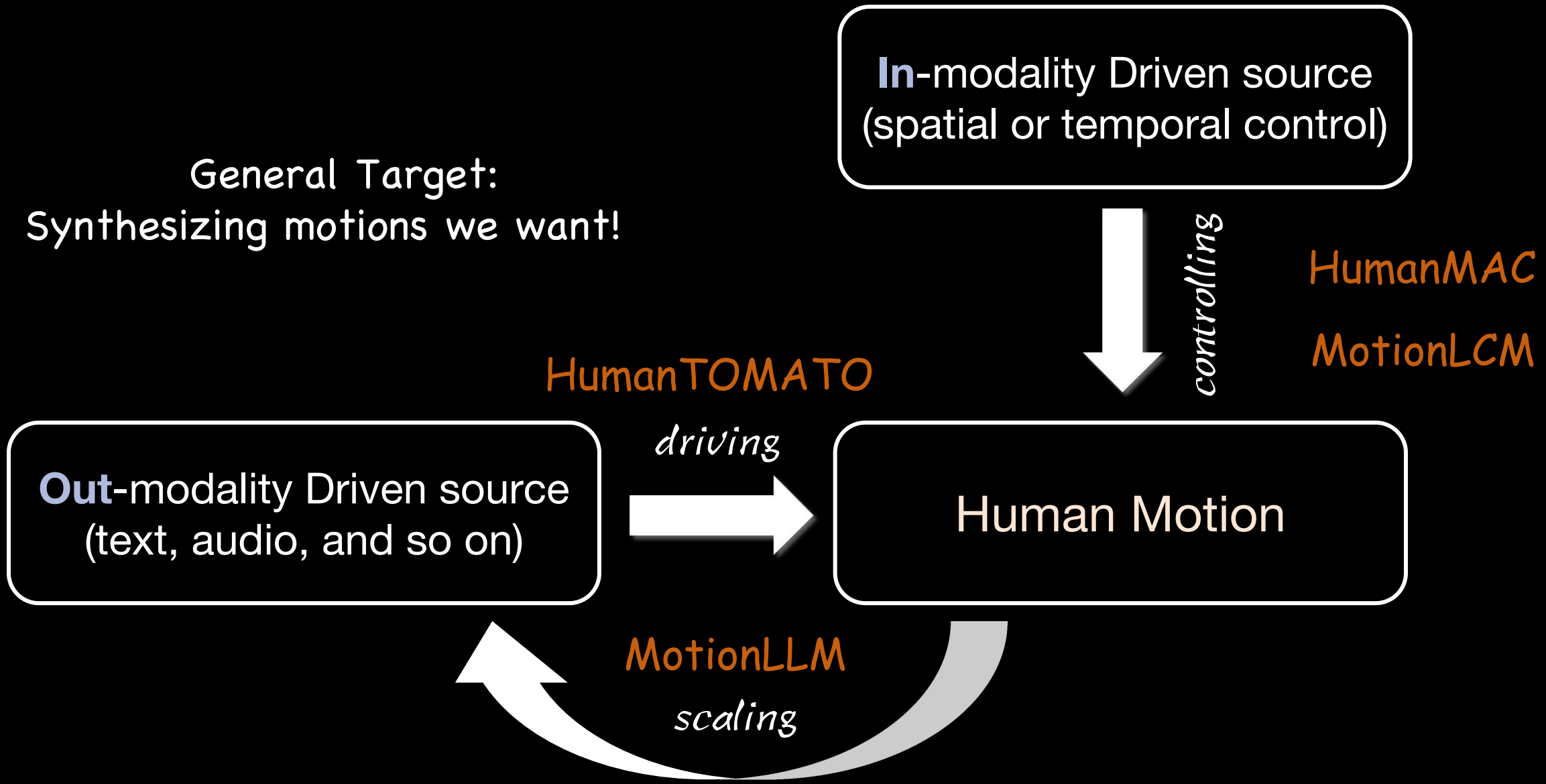
Road Map

General Target:
Synthesizing motions we want!

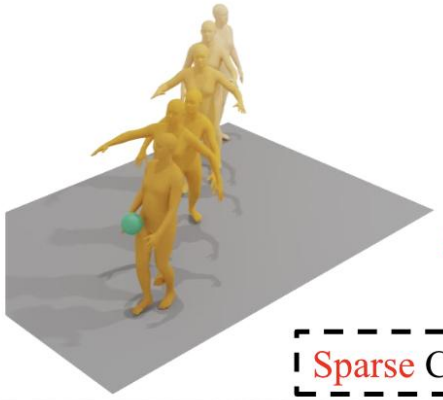
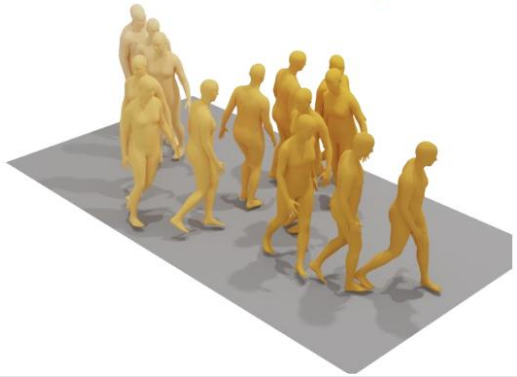
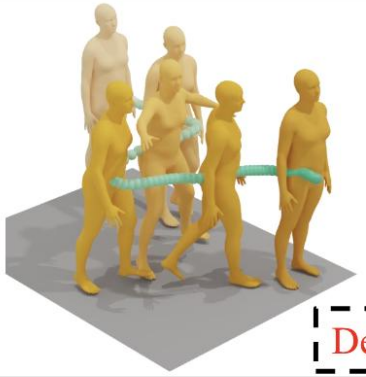
- ☐ Synthesizing text-aligned human motion (HumanTOMATO)
- ☐ Zero-shot motion controlling (HumanMAC)
- ☐ Tunable, real-time, and multi-modality motion controlling (MotionLCM)
- ☐ Exploring on the scaling law for human motion (MotionLLM)

Road Map

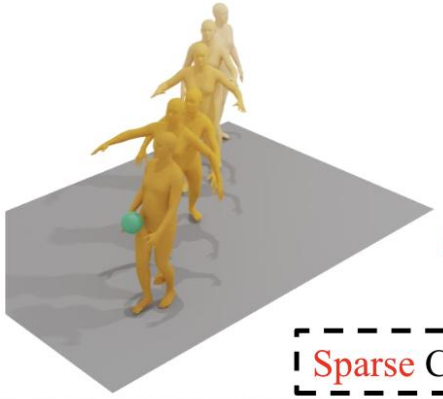
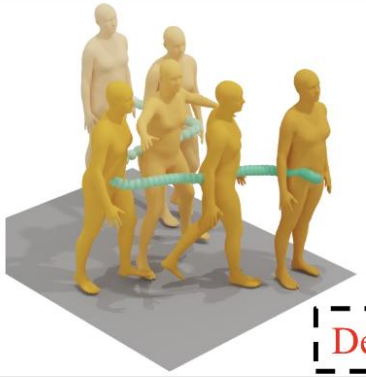
General Target:
Synthesizing motions we want!

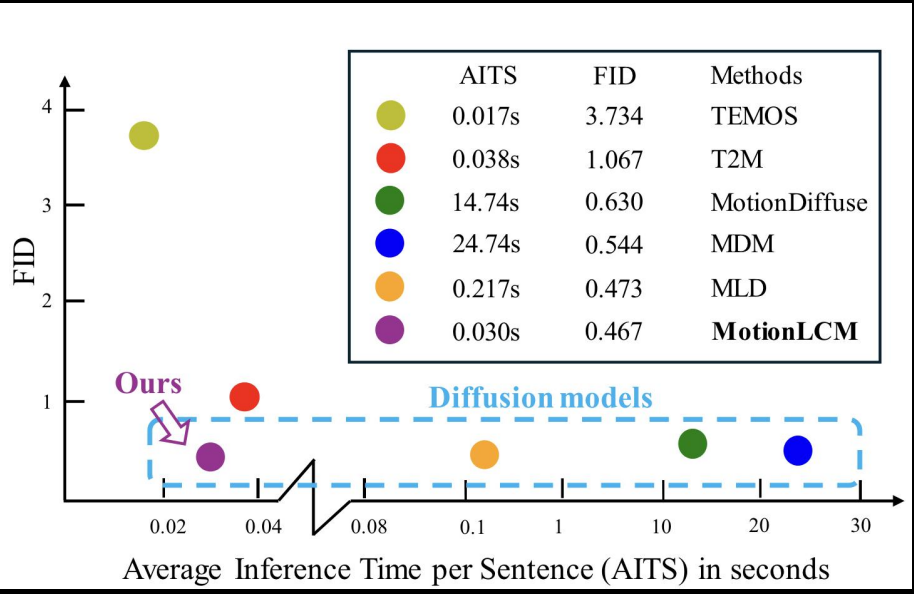


MotionLCM: Real-time Controllable Motion Generation via Latent Consistency Model

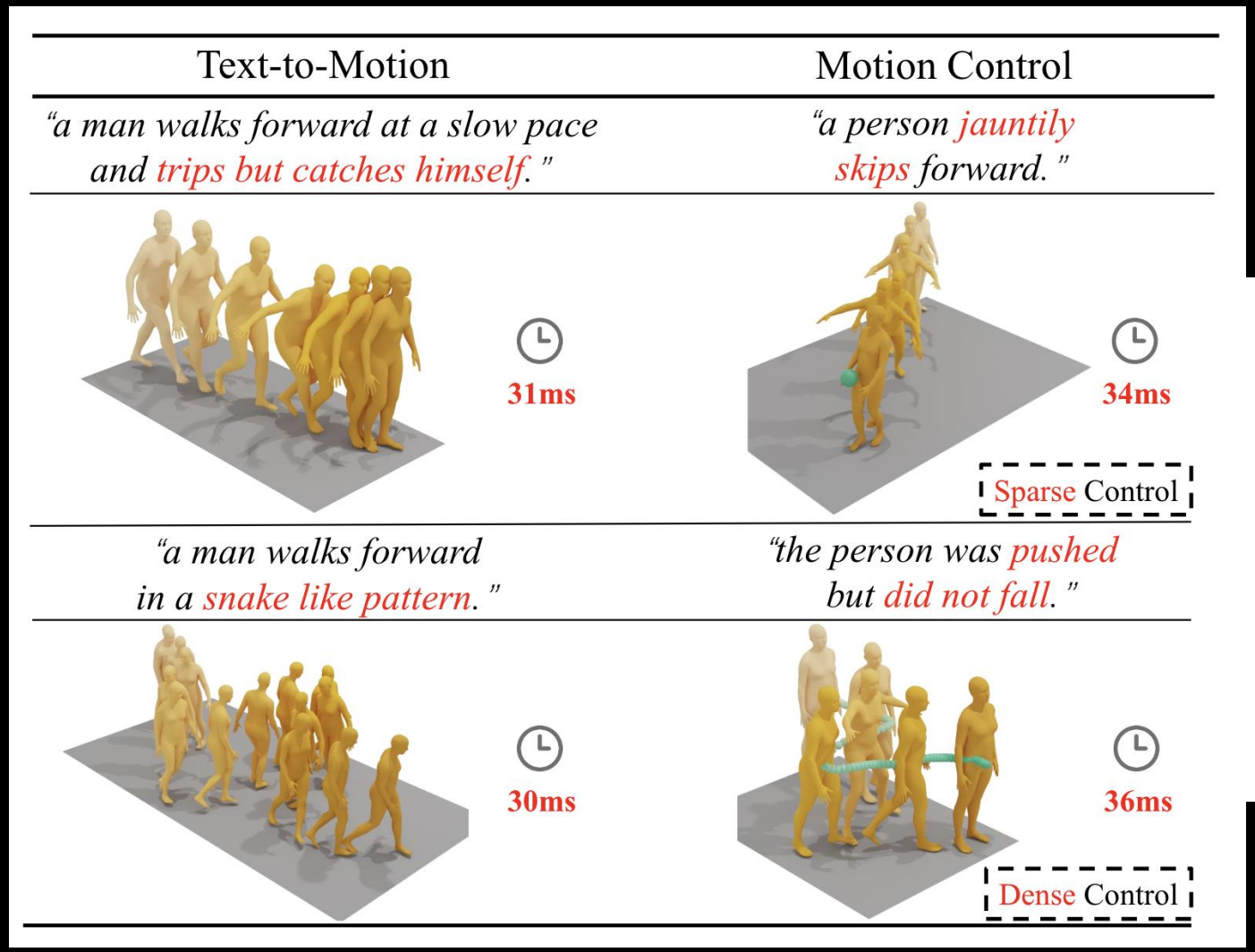
Text-to-Motion	Motion Control
"a man walks forward at a slow pace and trips but catches himself."	*"a person jauntily skips forward."*
 31ms	 34ms Sparse Control
"a man walks forward in a snake like pattern."	*"the person was pushed but did not fall."*
 30ms	 36ms Dense Control

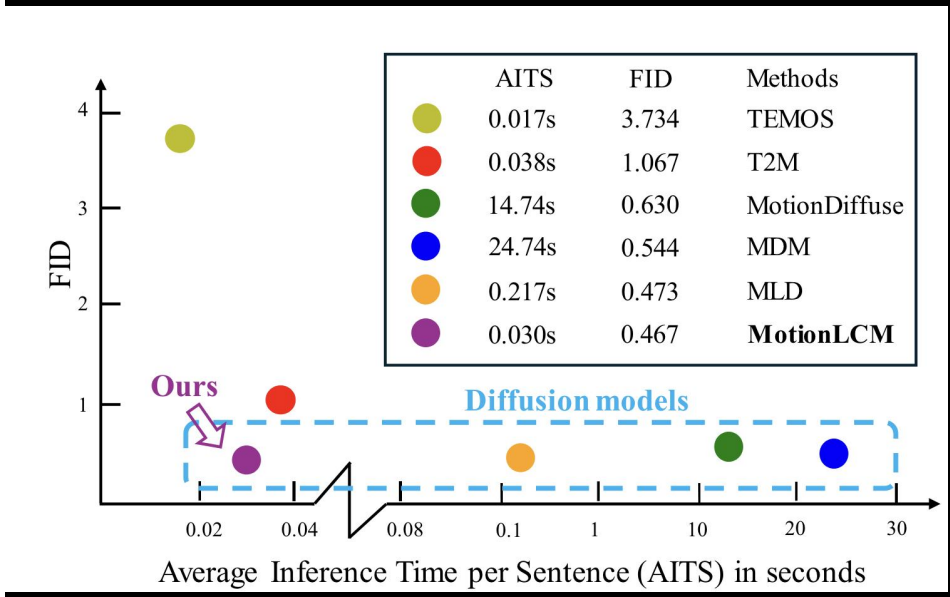
MotionLCM: Real-time Controllable Motion Generation via Latent Consistency Model

Text-to-Motion	Motion Control
"a man walks forward at a slow pace and trips but catches himself."	*"a person jauntily skips forward."*
 31ms	 34ms Sparse Control
"a man walks forward in a snake like pattern."	*"the person was pushed but did not fall."*
 30ms	 36ms Dense Control



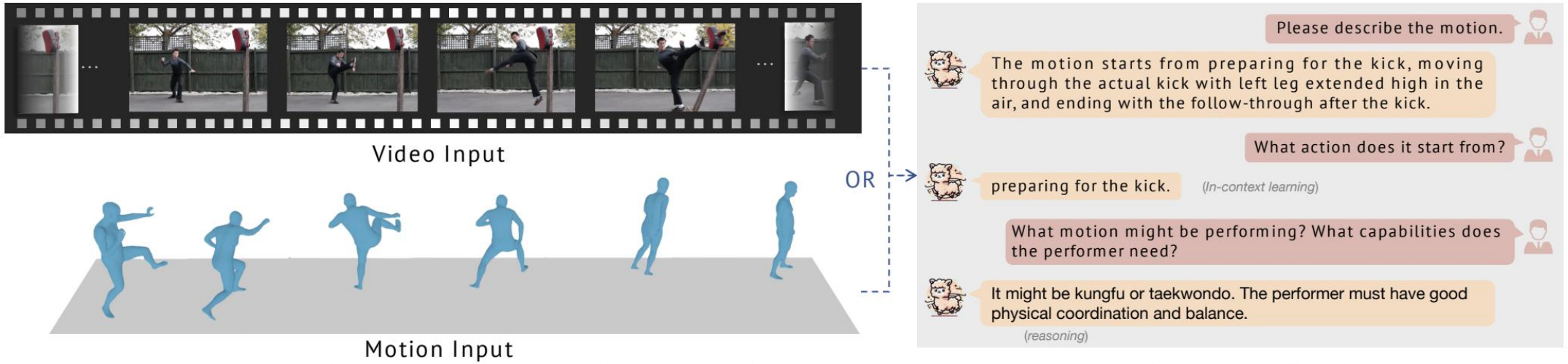
MotionLCM: Real-time Controllable Motion Generation via Latent Consistency Model





Page: https://dai-wenxun.github.io/MotionLCM-page

MotionLLM: Understanding Human Behaviors from Human Motions and Videos



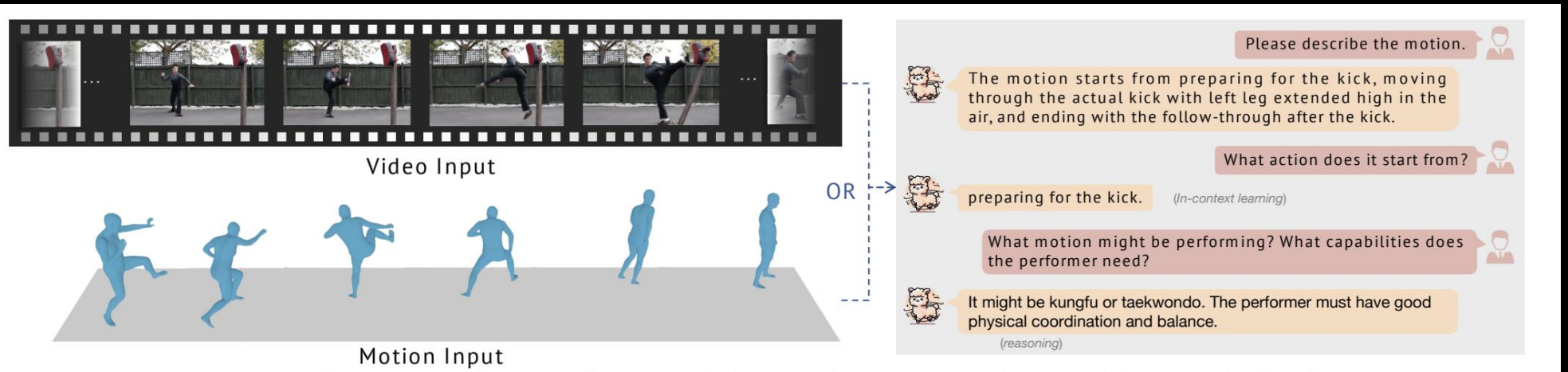
(a) Our MotionLLM takes motions or videos as inputs to understand human behaviors.

MotionLLM as your fitness coach

1. Begin in a supine position with arms at shoulder width. 2. Engage your core as you lift your legs to a 45 degree angle. 3. Bring your feet hip width apart. 4. Slide your hands down your thighs until your hands are on the floor. 5. Put down your feet to return to the supine position.

(b) An application: MotionLLM as your fitness coach based on its caption capability.

MotionLLM: Understanding Human Behaviors from Human Motions and Videos



(a) Our MotionLLM takes motions or videos as inputs to understand human behaviors.

MotionLLM as your fitness coach

1. Begin in a supine position with arms at shoulder width. 2. Engage your core as you lift your legs to a 45 degree angle. 3. Bring your feet hip width apart. 4. Slide your hands down your thighs until your hands are on the floor. 5. Put down your feet to return to the supine position.

(b) An application: MotionLLM as your fitness coach based on its caption capability.

Page: https://lhchen.top/MotionLLM

Thanks!