Human Motion & Interaction Synthesis in the Large Model Era

Libin Liu [http://libliu.info | libin.liu@pku.edu.cn]

School of Intelligence Science and Technology
Peking University

Digital Human and Humanoids

$p(\text{motion}|\text{hidden factors})$

Goals

Personality

Physical status

Emotion

Environment

$p(z|g)$

Behavior Model

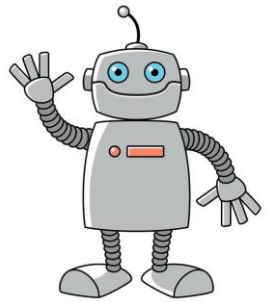
$p(x_{t+1}|x_t; z)$

Motion Model

$p(x_{t+1}, a_t|x_t, z)$

$p(a_t|x_t; z)$ → $p(x_{t+1}|x_t, a_t; z)$

Action Model — World/Simulation



"Brain" / System 2

"Cerebellum" / System 1

Motion synthesis under different conditions

- Generation is often easier with stronger conditions

Basketball playing

$p(\text{ motion} \mid \text{goal})$

difficulty

<

Idling

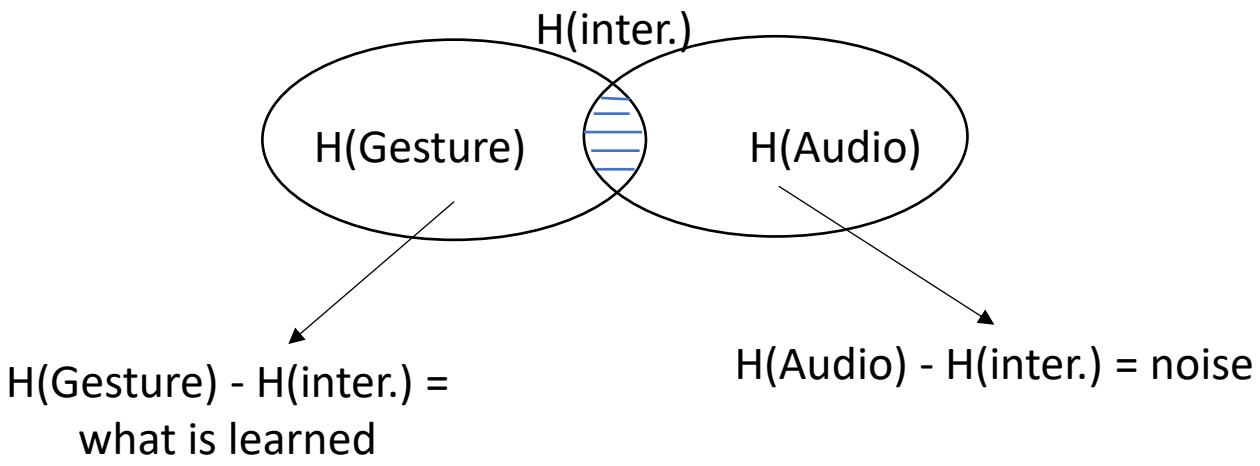
$p(\text{motion}, \text{goal})$

Example: Co-speech Gesture Generation

Example: Co-speech Gesture Generation

Human is sensitive to unnaturalness (Uncanny valley)

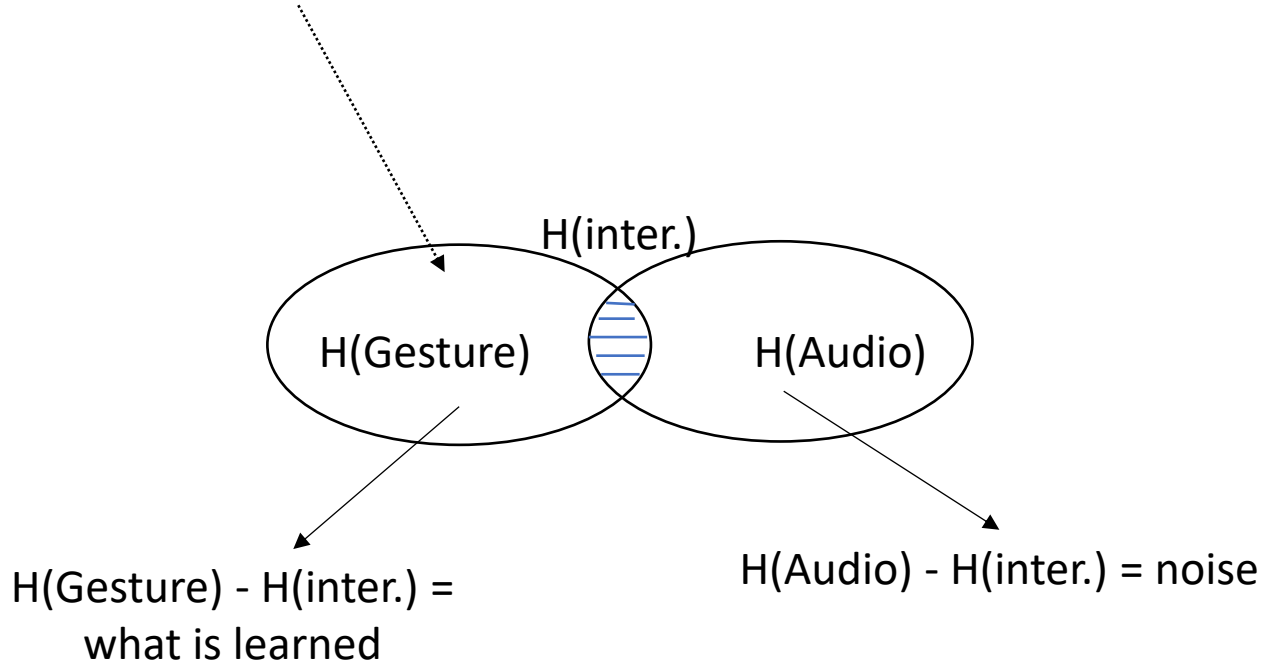
Correspondence among modalities is weak



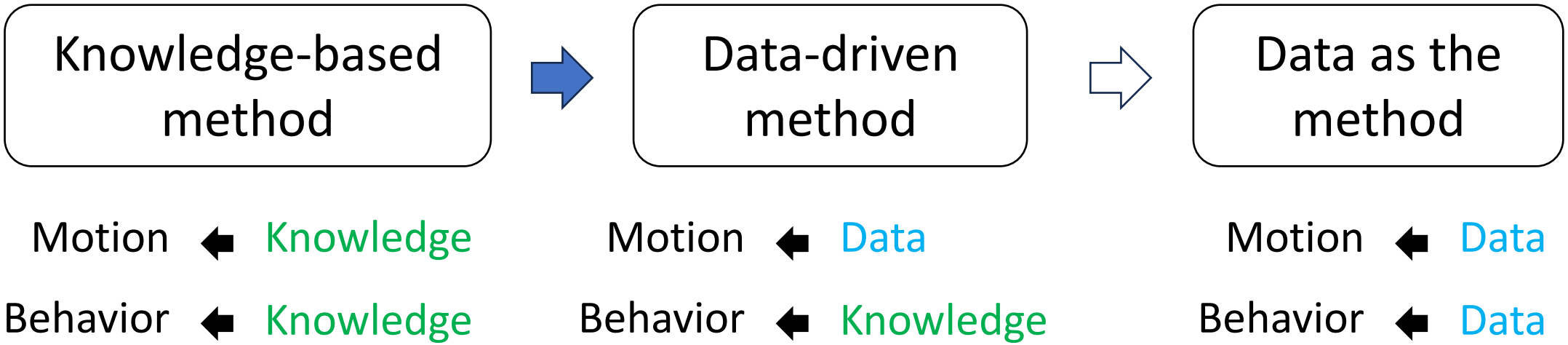
Example: Co-speech Gesture Generation

How to fill this blank?

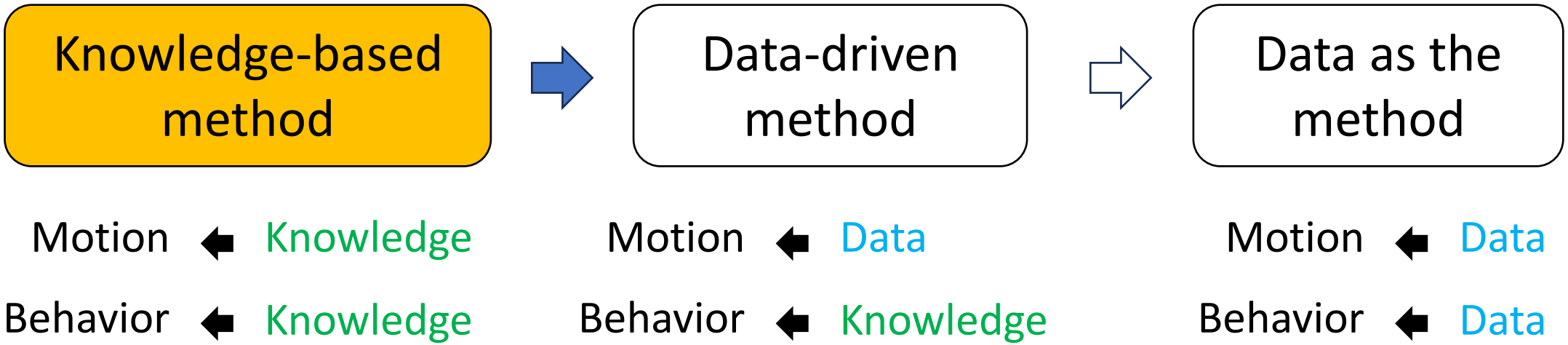
Knowledge ⇆ Data



A path to realistic motion synthesis



A path to realistic motion synthesis



Method based on human knowledge

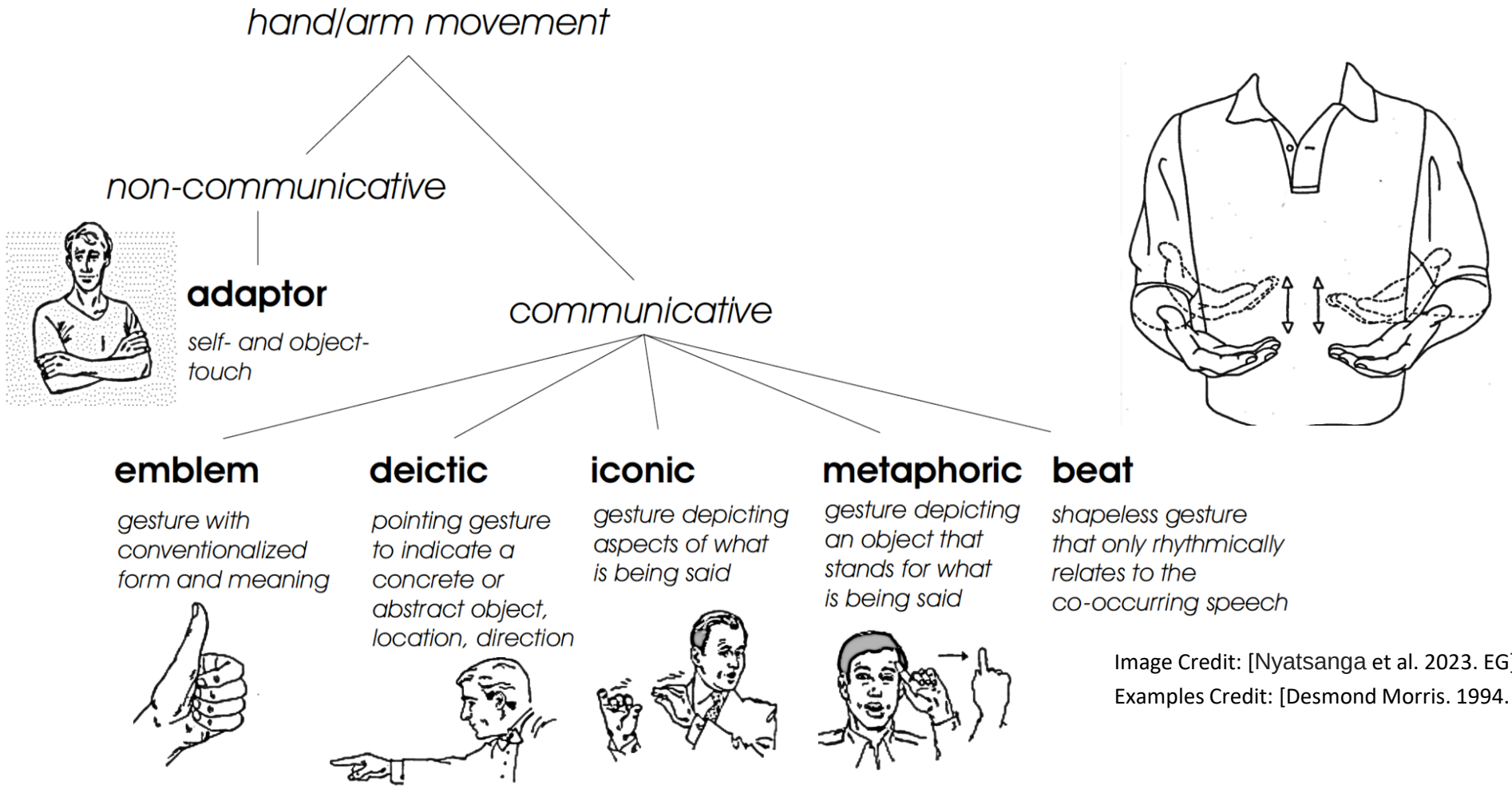
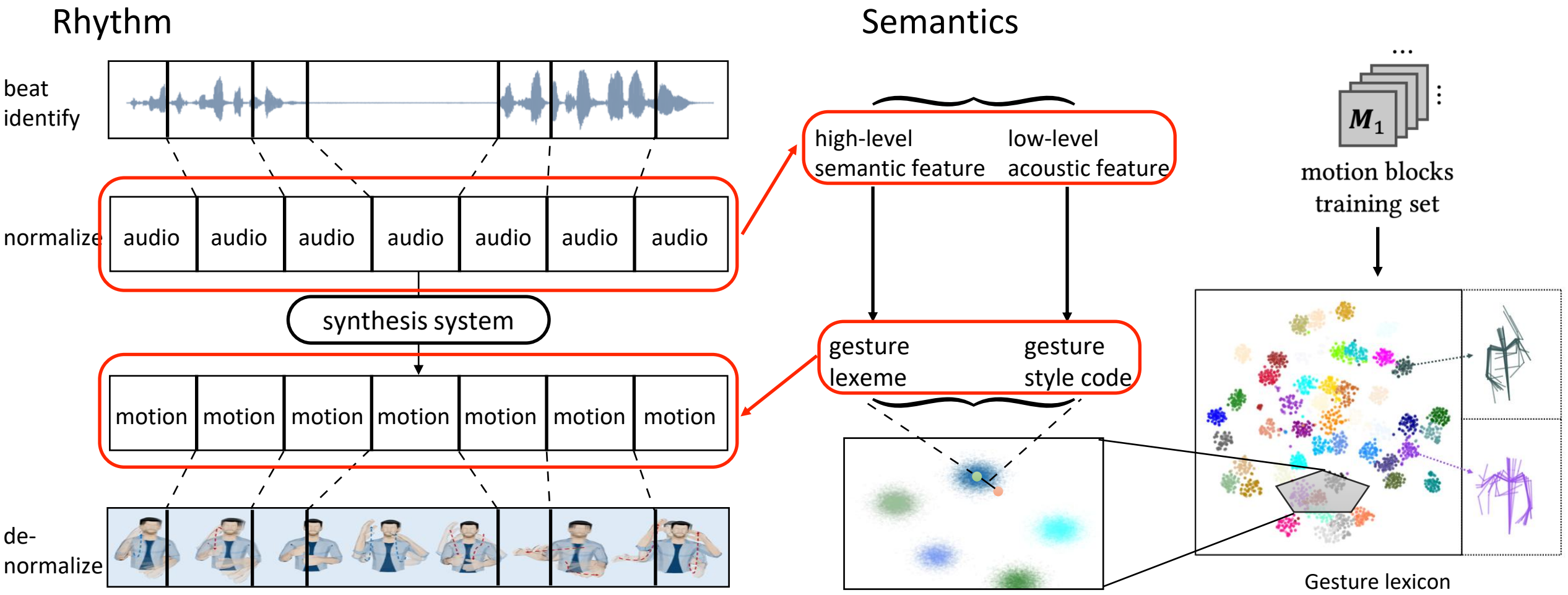


Image Credit: [Nyatsanga et al. 2023. EG]

Examples Credit: [Desmond Morris. 1994. Bodytalk]

Method based on human knowledge

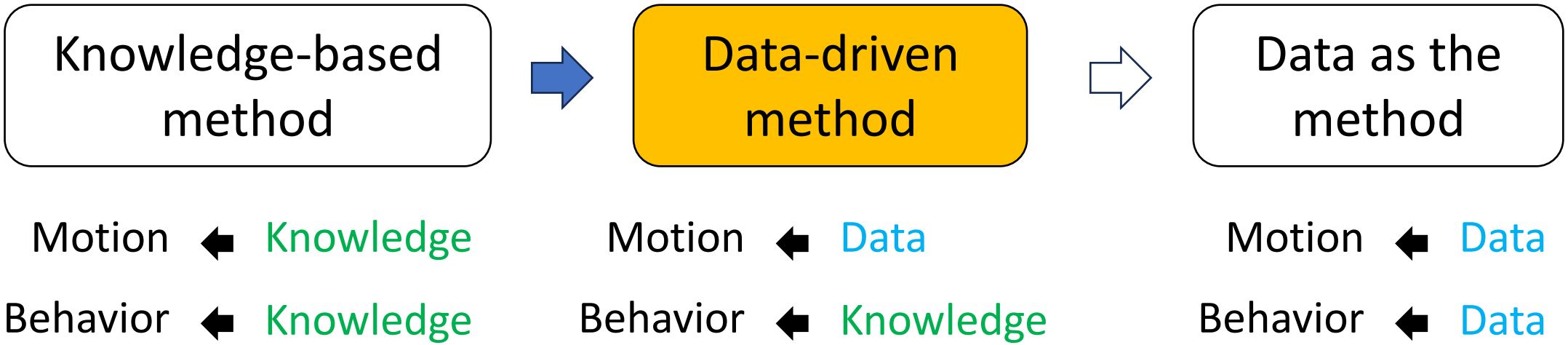


[Rhythmic Gesticulator – Ao et al. 2022, SIGGRAPH Asia 2022 Best Paper Award]

Method based on human knowledge

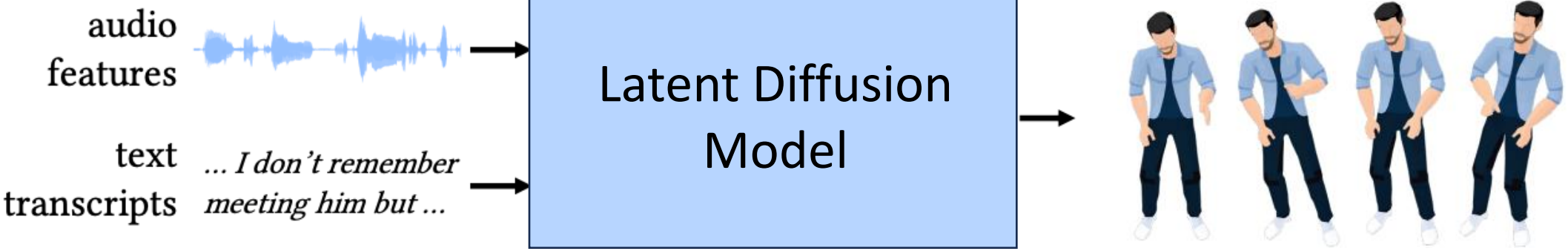
[Rhythmic Gesticulator – Ao et al. 2022, SIGGRAPH Asia 2022 Best Paper Award]

A path to realistic motion synthesis



A task-specific foundation model for motion

Train with ~30 hours carefully curated data



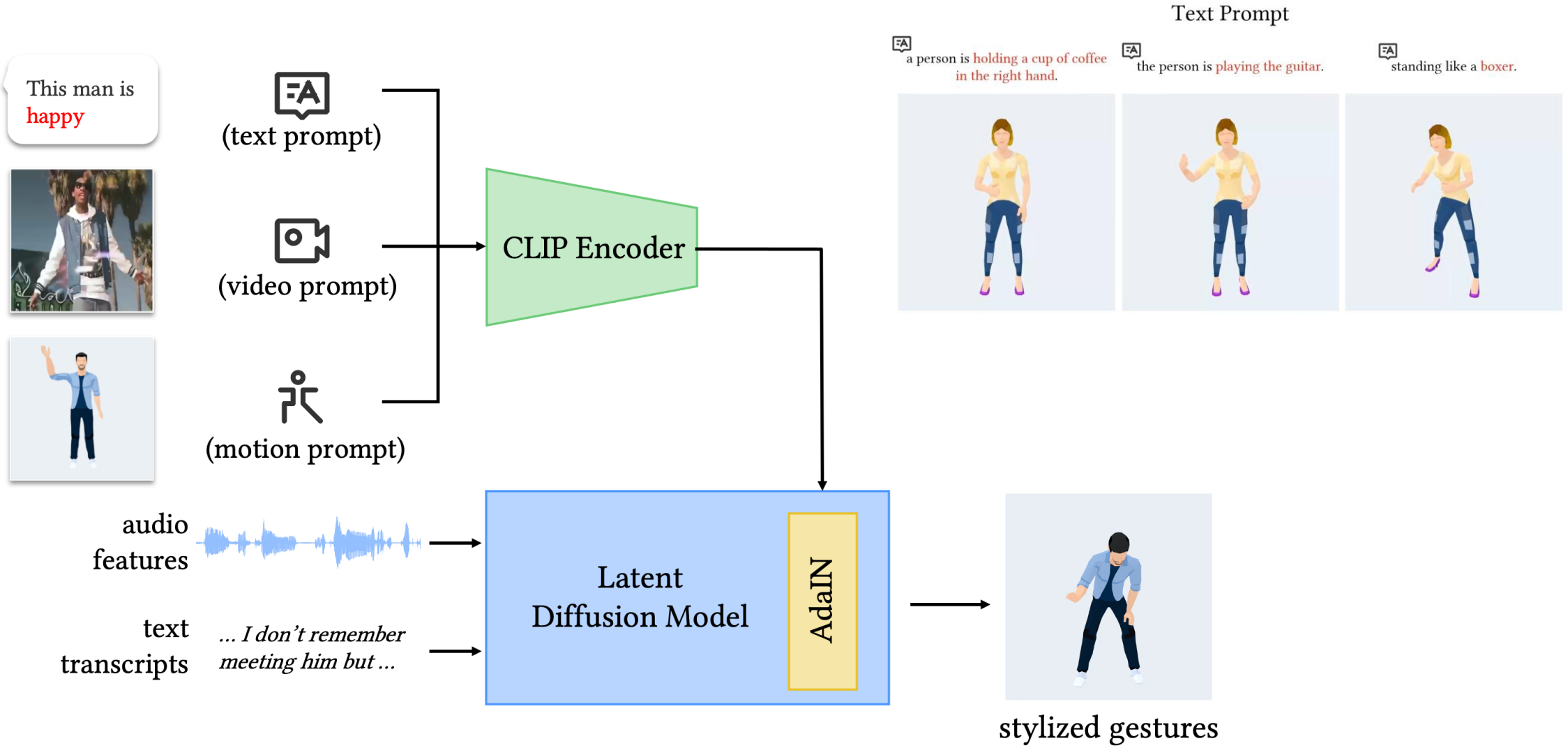
[GestureDiffuClip – Ao et al. 2023, SIGGRAPH 2023 Honorable Mention Award]

A task-specific foundation model for motion

Ground Truth

Ours
(style prompt = ∅)

A task-specific foundation model for motion



[GestureDiffuClip – Ao et al. 2023, SIGGRAPH 2023 Honorable Mention Award]

Knowledge-based behavior control with LLM

Hello, ChatGPT. I want you to act as a public speaking coach.

I will provide you with a speech transcript. Then, you need to provide detailed suggestions about gesture style in a parenthesis after each sentence...

The speech transcript is "The food in my school's cafeteria is awful and I became 3 kg thinner than when I was in high school, so I miss the food my mother cooks for me, especially spaghetti, which is a signature dish."

"The food in my school's cafeteria is awful (disgusted, cross hands) and I became 3 kg thinner than when I was in high school (sad), so I miss the food my mother cooks for me (nostalgic, perhaps hands on heart), especially spaghetti (emphasis), which is a signature dish (proud, perhaps a small chest puff)."

[GestureDiffuClip – Ao et al. 2023, SIGGRAPH 2023 Honorable Mention Award]

Knowledge-based behavior control with LLM

System 2

System 1

I am very happy

when people mentioned this family photo.

But my grandmother just passed away last year.

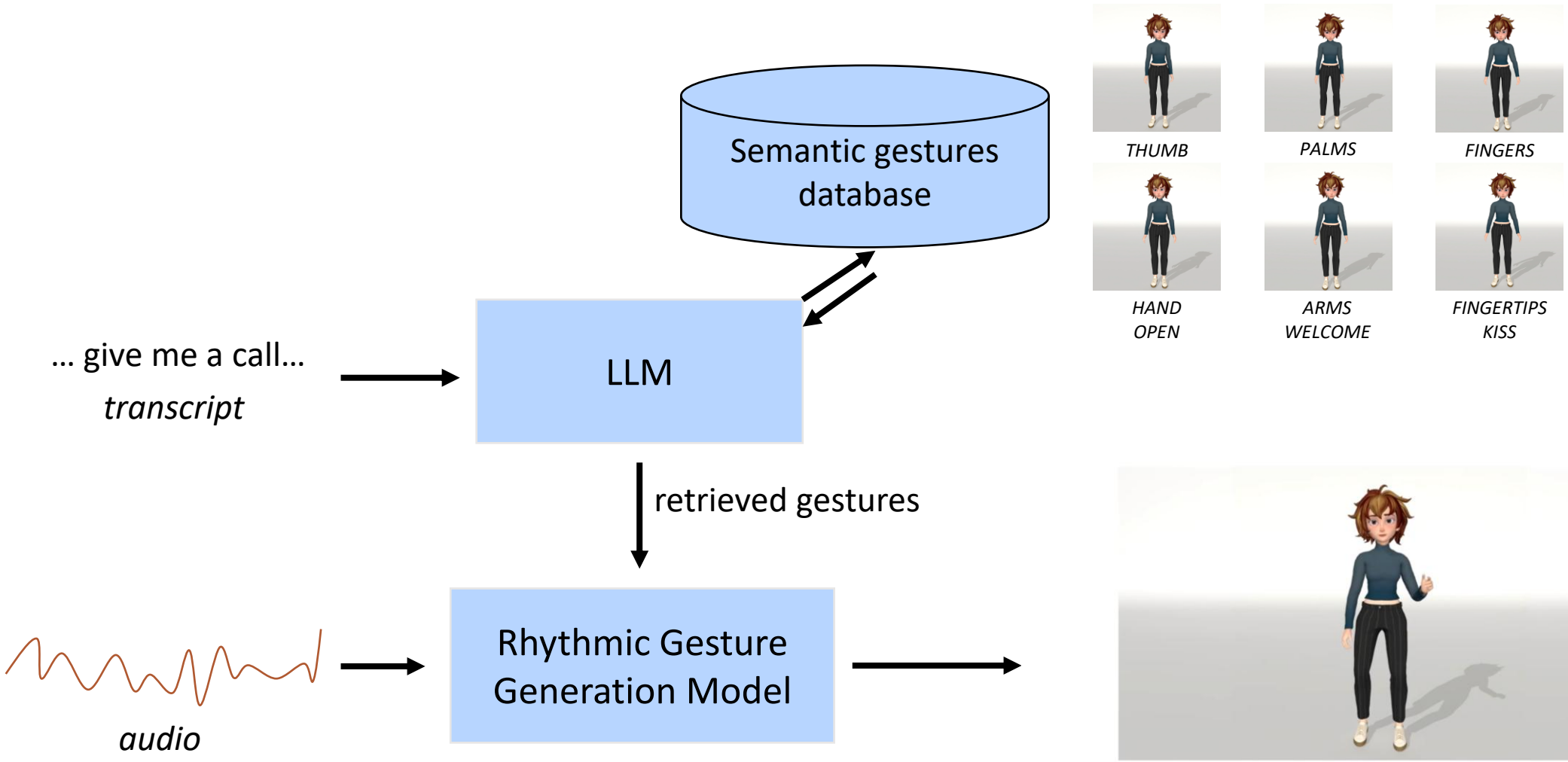
I am very happy
(slightly excited, wave arms)
when people mentioned this family photo.
(gesture towards photo)
But my grandmother just passed away last year.
(sad, lower posture)

Style prompt: disgusted, cross hands

Ours

[GestureDiffuClip – Ao et al. 2023, SIGGRAPH 2023 Honorable Mention Award]

Knowledge-based behavior control with LLM



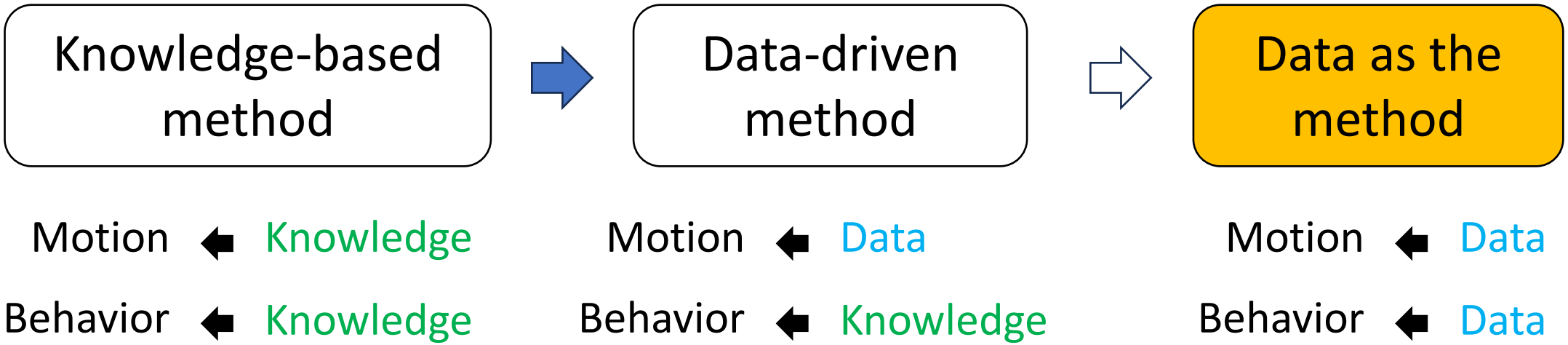
Semantics-enhanced gestures

[Zhang et al 2024. Semantic Gesticulator]

Knowledge-based behavior control with LLM

[Zhang et al. 2025]

A path to realistic motion synthesis



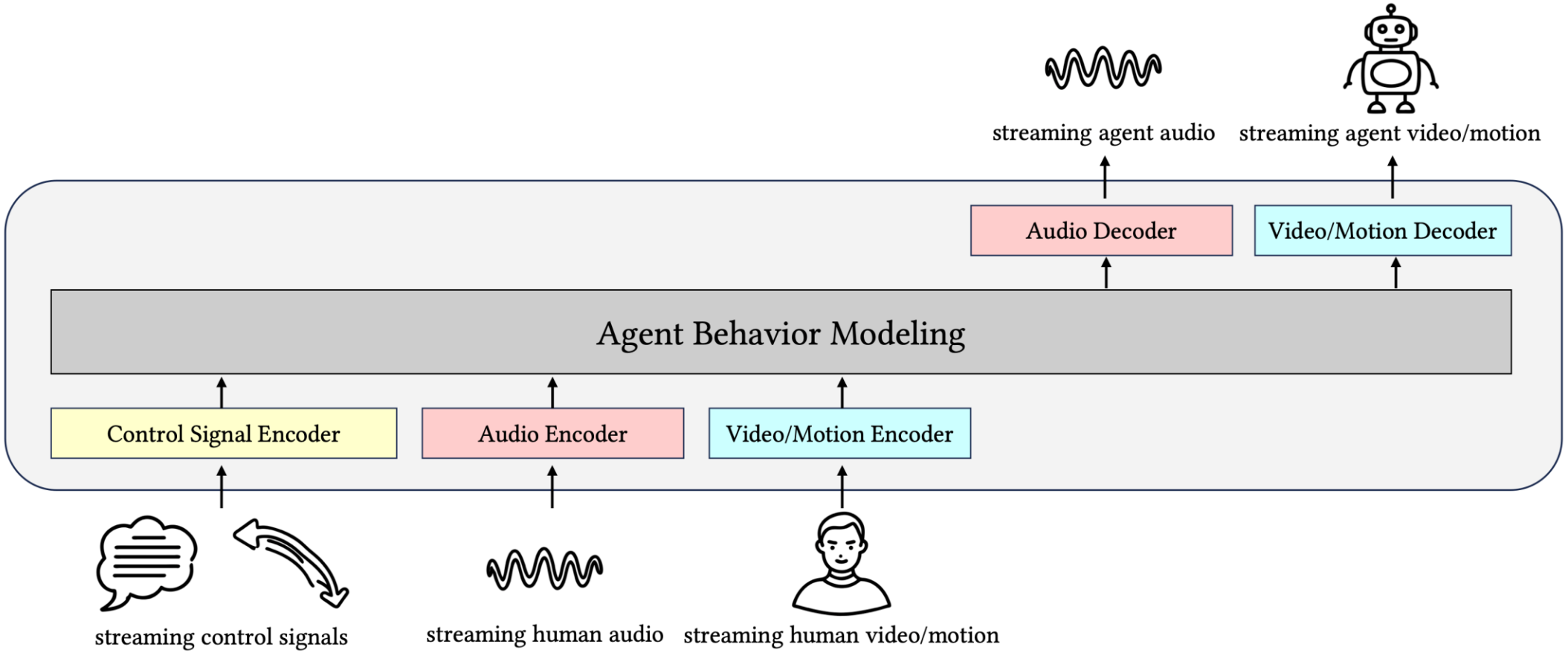
Data as the method

Human

Agent (generated)

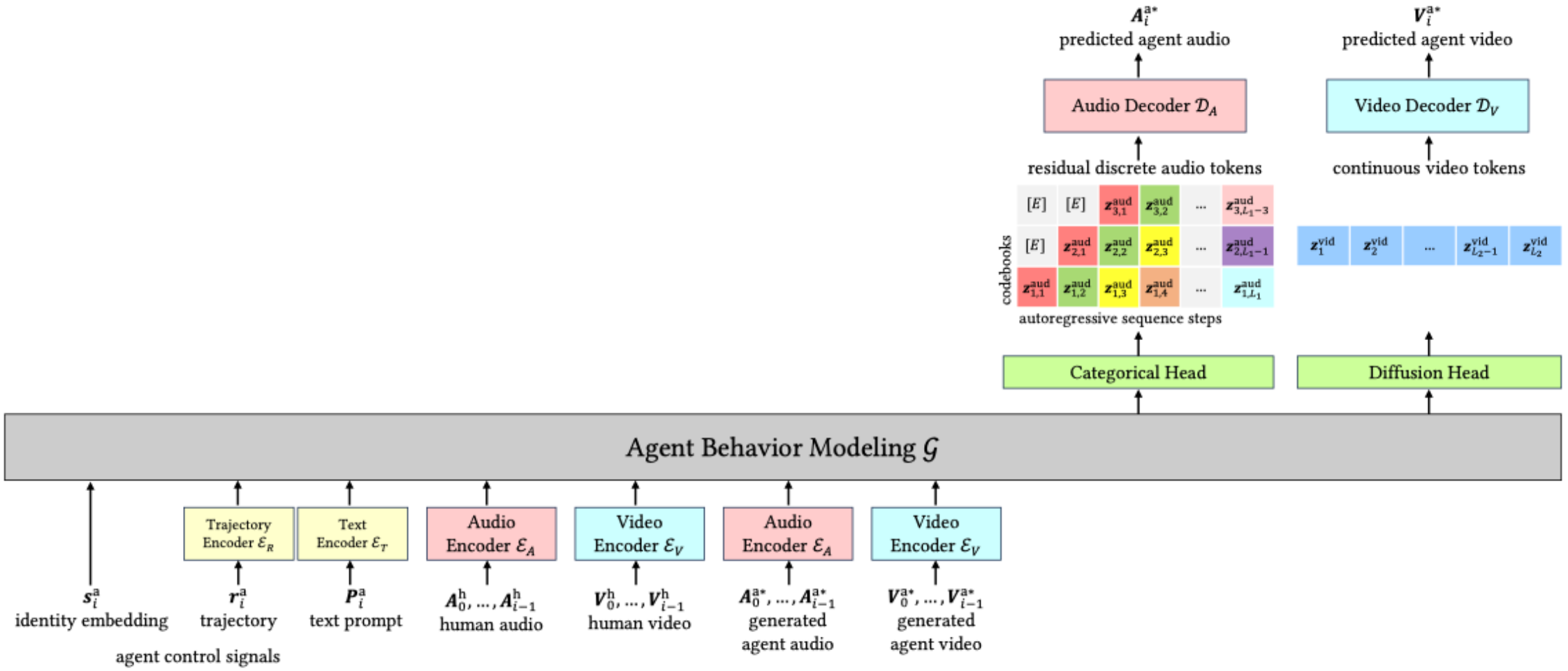
[Body of Her – Ao. 2024]

An autoregressive multimodal model



[Body of Her – Ao. 2024]

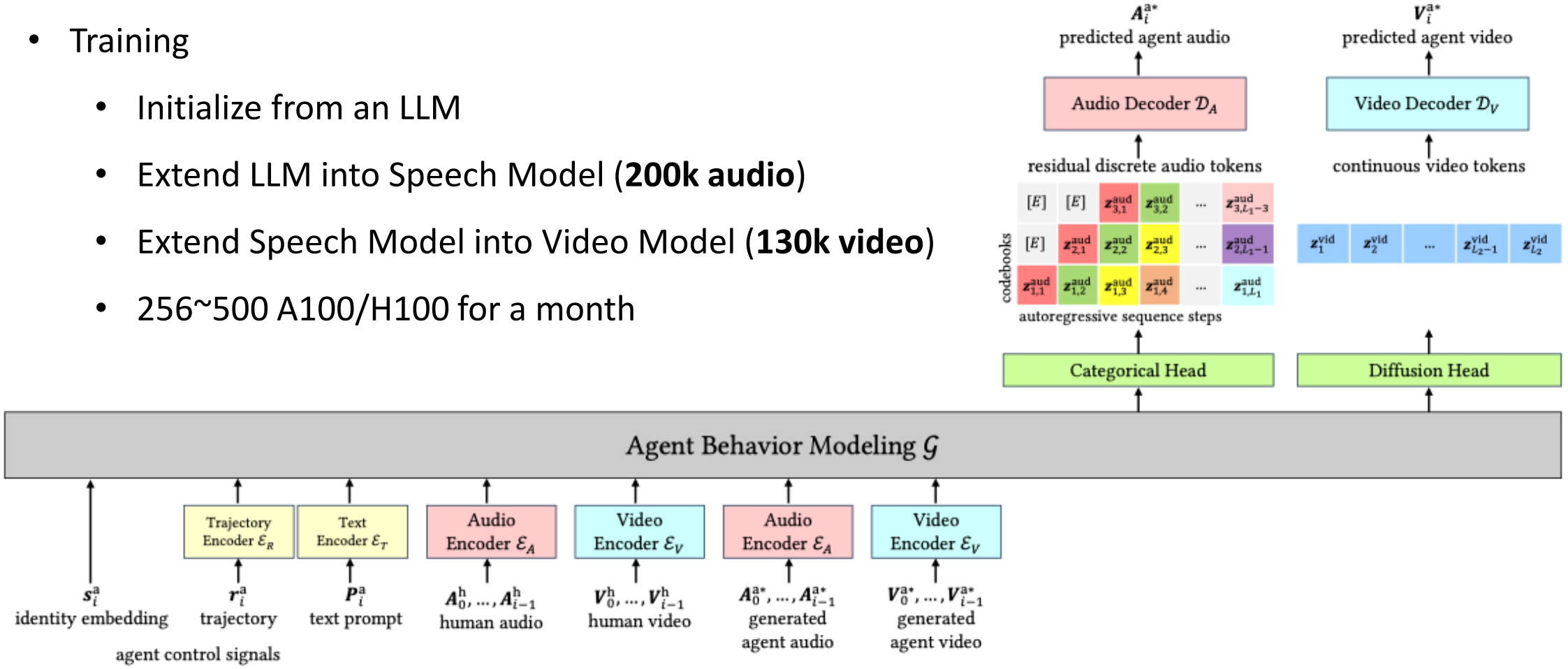
An autoregressive multimodal model



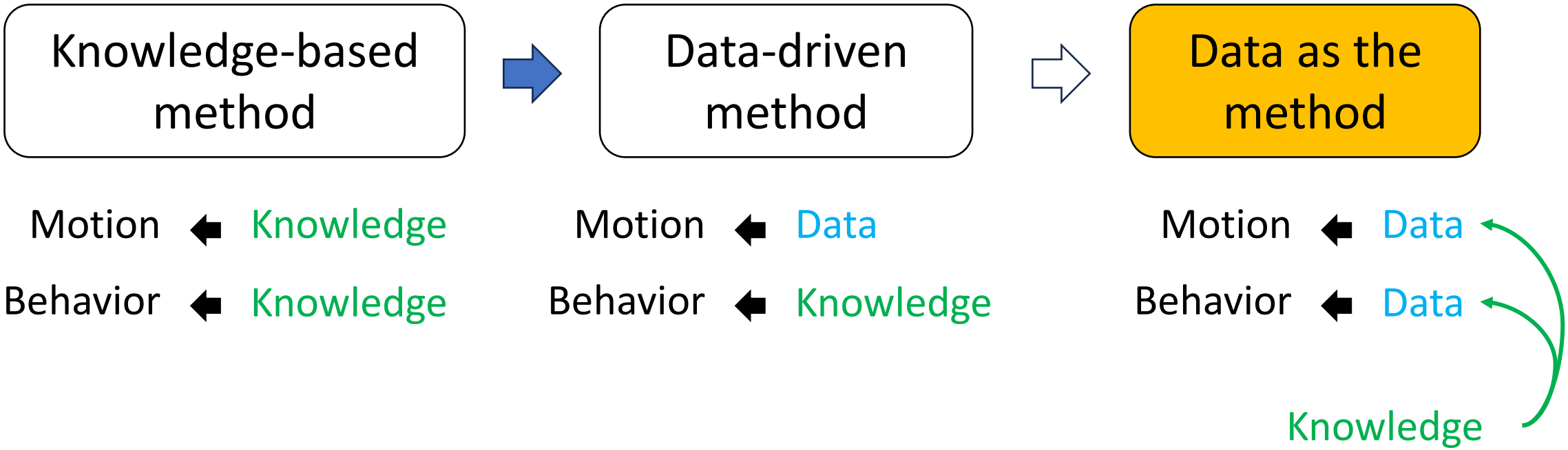
[Body of Her – Ao. 2024]

Data as the Method?

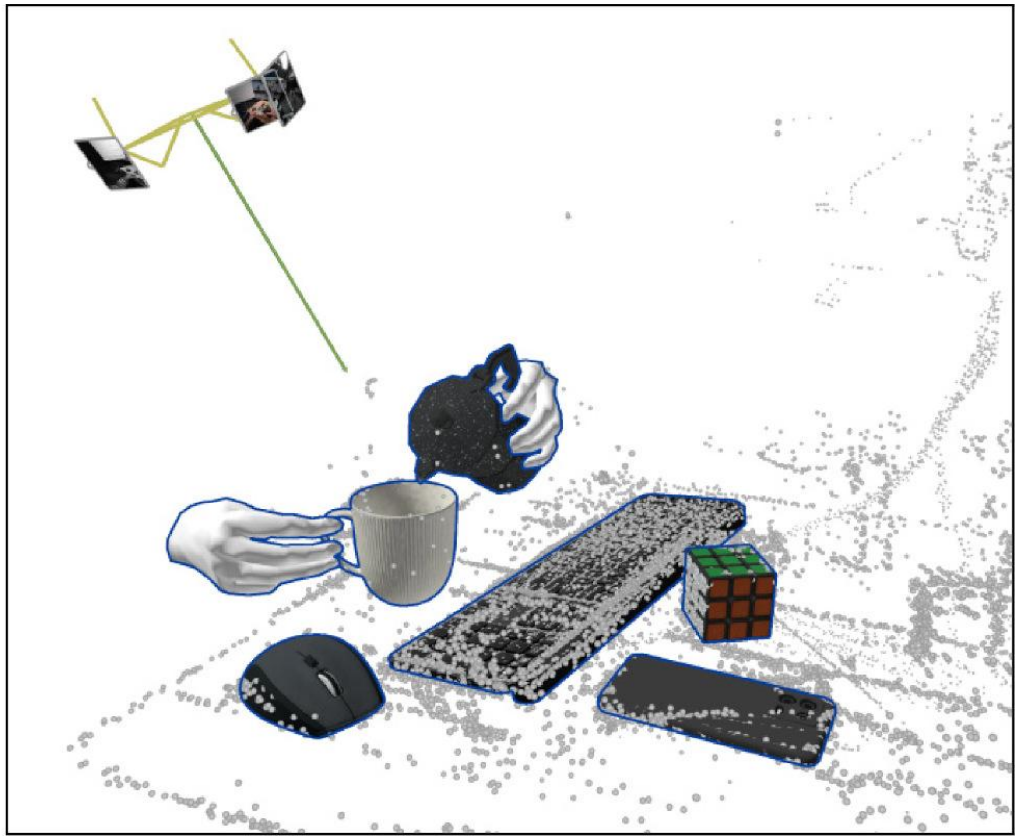
- Training
  - Initialize from an LLM
  - Extend LLM into Speech Model (**200k audio**)
  - Extend Speech Model into Video Model (**130k video**)
  - 256~500 A100/H100 for a month



A path to realistic motion synthesis



Data as the Method



Manipulation: Hot3D dataset (Banerjee et al. 2024)

Realtime chatting and interaction

Human

Agent (generated)

[Body of Her – Ao. 2024]

An idling motion simulator

[Body of Her – Ao. 2024]

Digital Human and Humanoids

$p(\text{motion}|\text{hidden factors})$

Goals

Personality

Physical status

Emotion

Environment

$p(z|g)$

Behavior Model

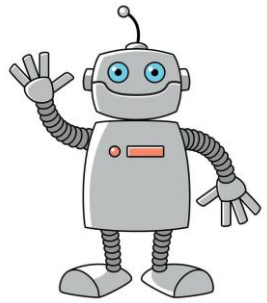
$p(x_{t+1}|x_t; z)$

Motion Model

$p(x_{t+1}, a_t|x_t, z)$

$p(a_t|x_t; z)$ → $p(x_{t+1}|x_t, a_t; z)$

Action Model — World/Simulation



"Brain" / System 2

"Cerebellum" / System 1

Is an end-to-end model necessary/possible?

Latent Behavior Model

Goals

Personality

Physical status

Emotion

Environment

$p(z|g)$

Behavior Model

"Brain" / System 2

Motion Tokenizer

$p(x_{t+1}|x_t; z)$

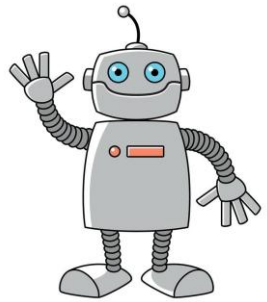
Motion Model

$$\frac{p(x_{t+1}, a_t|x_t, z)}{p(a_t|x_t; z)}$$

Action Model

$p(x_{t+1}|x_t, a_t; z)$

World/Simulation



"Cerebellum" / System 1

Learn large models for humanoids

Motion Tokenizer

$p(x_{t+1}|x_t; z)$

Motion Model

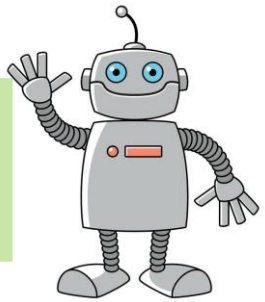
Action Tokenizer

$p(x_{t+1}, a_t|x_t, z)$

$p(a_t|x_t; z)$ → $p(x_{t+1}|x_t, a_t; z)$

Action Model

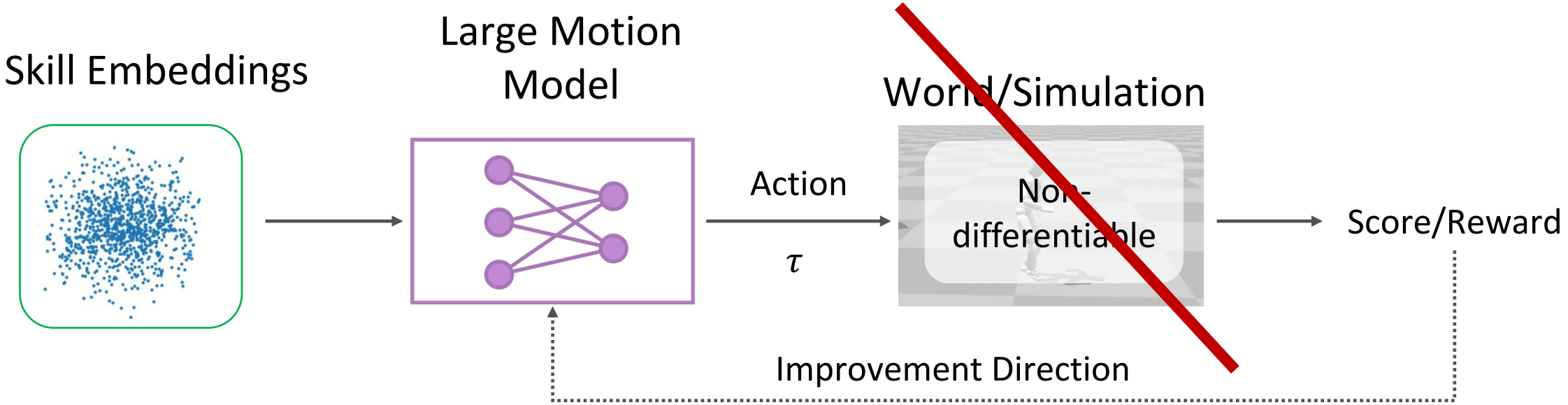
World/Simulation



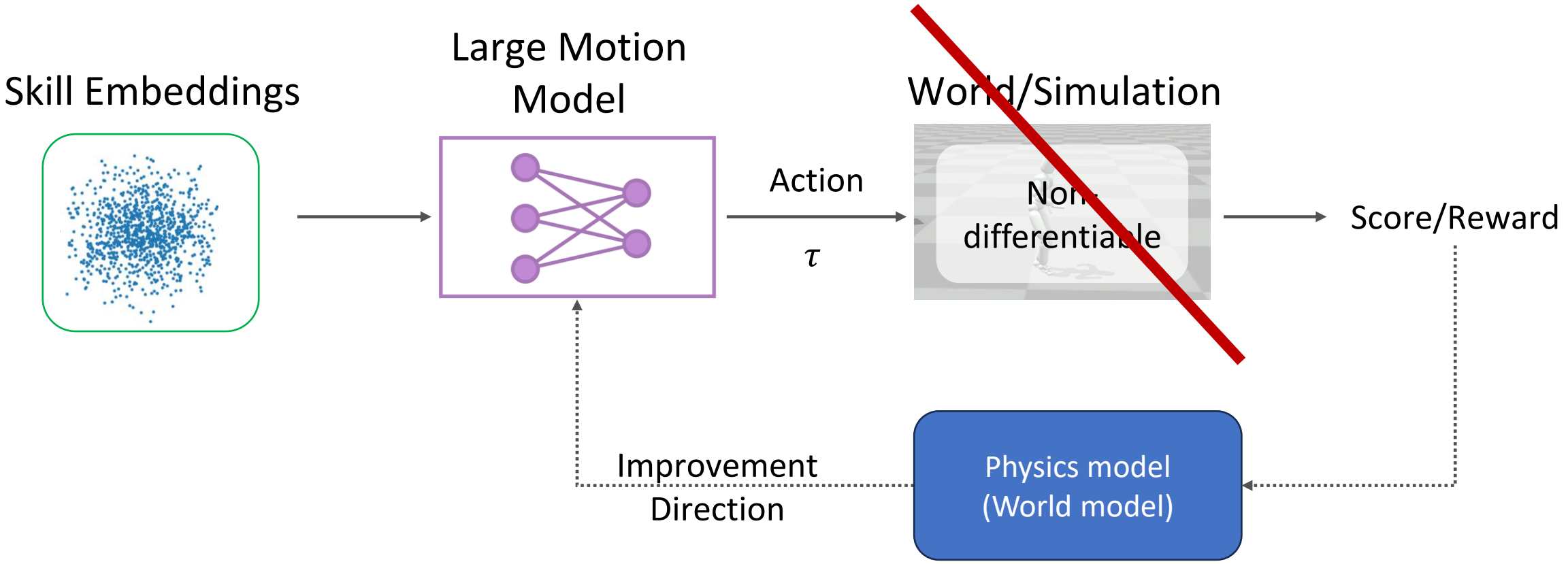
Non-differentiable

Learn large models for humanoids

- Model-free method is not efficient

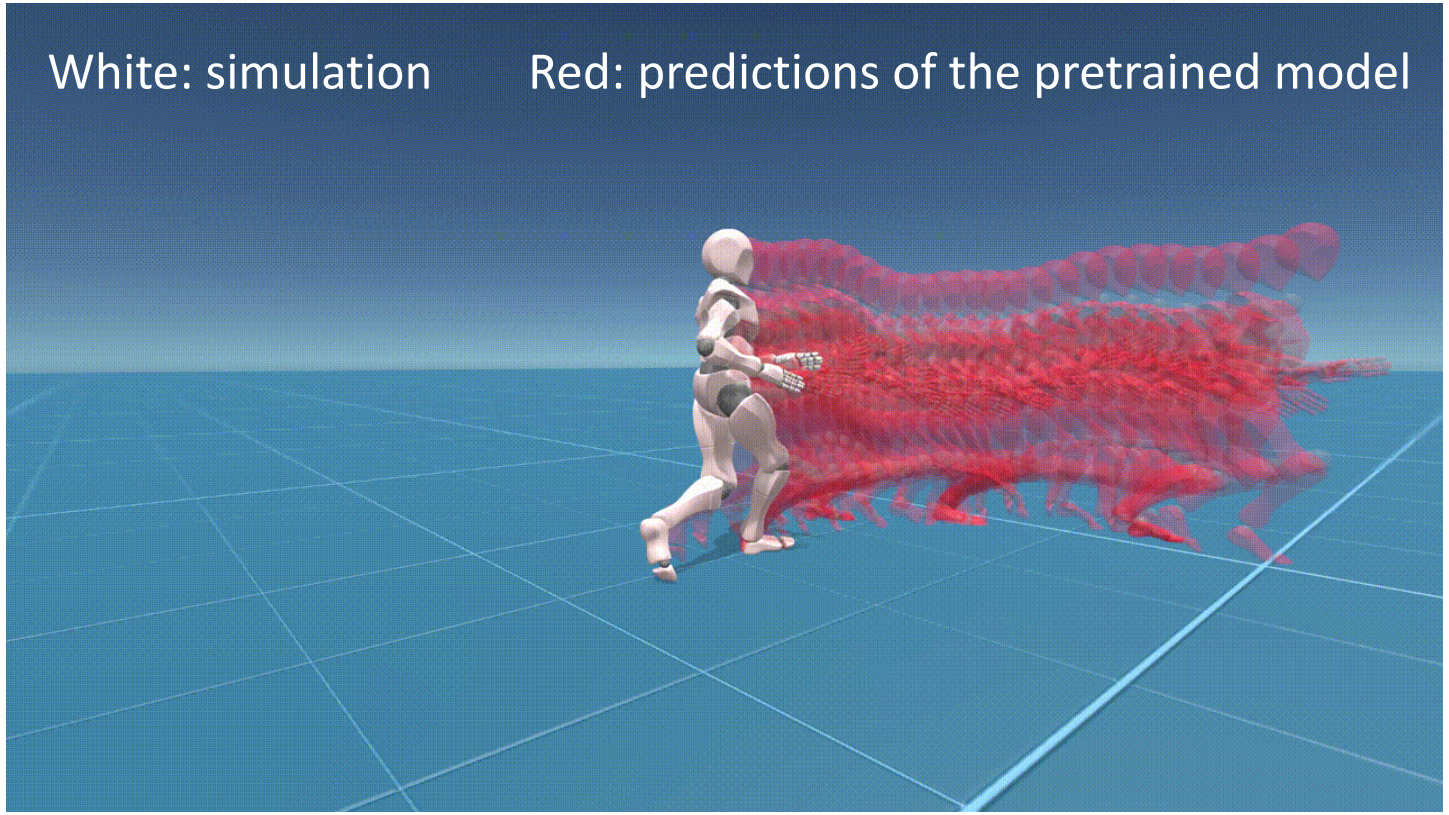


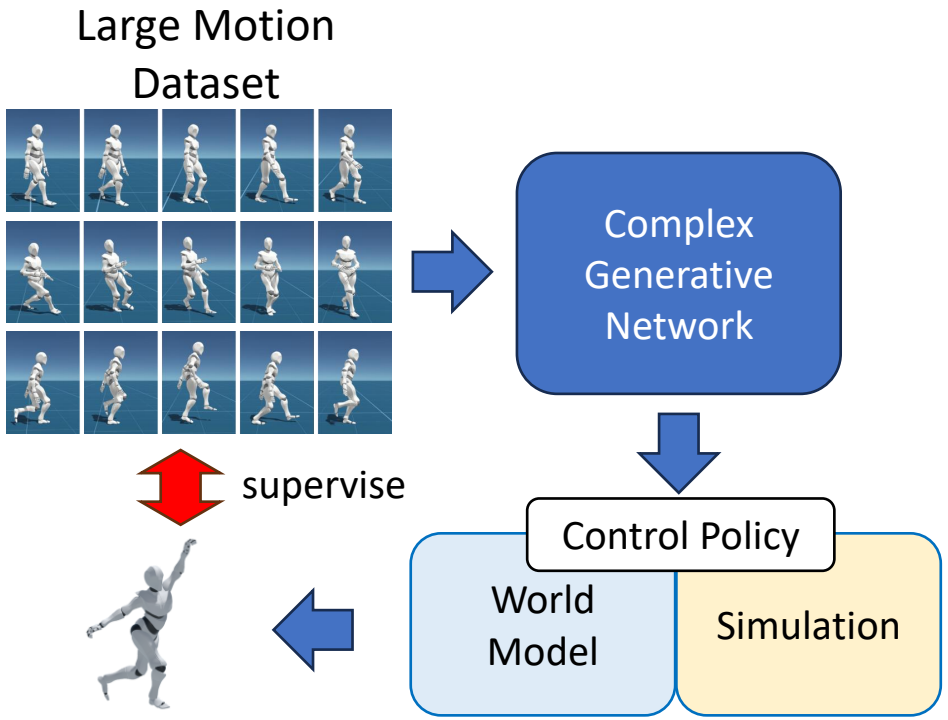
More Efficient Training?



[Yao et al 2022. Control VAE]

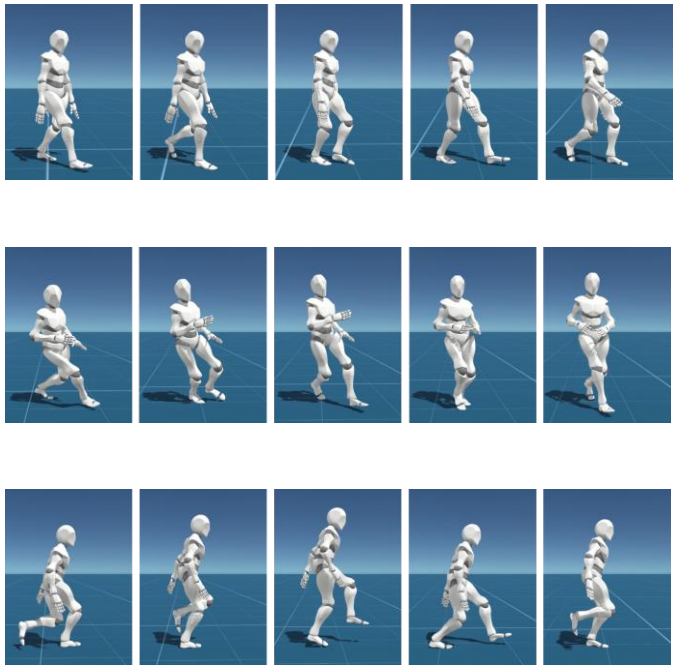
More Efficient Training?



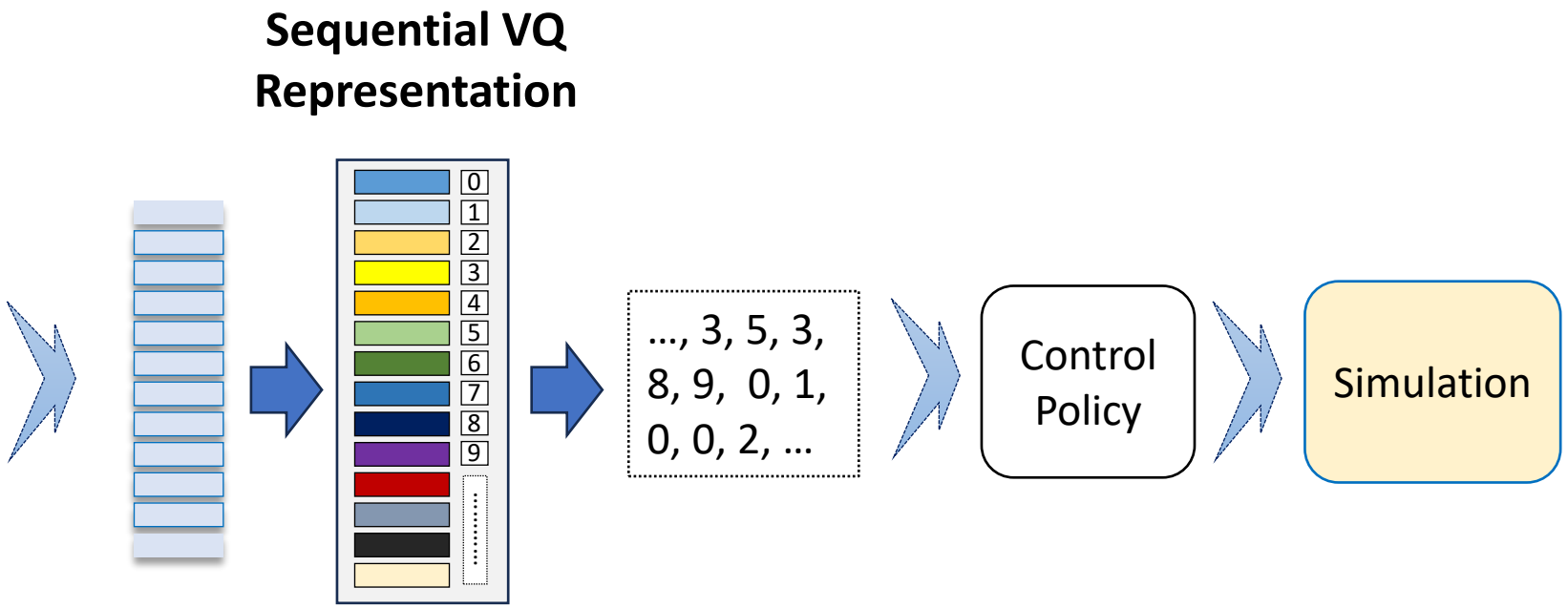


[Yao et al 2022. Control VAE]

Physics-based Tokenizer with Model-based Learning

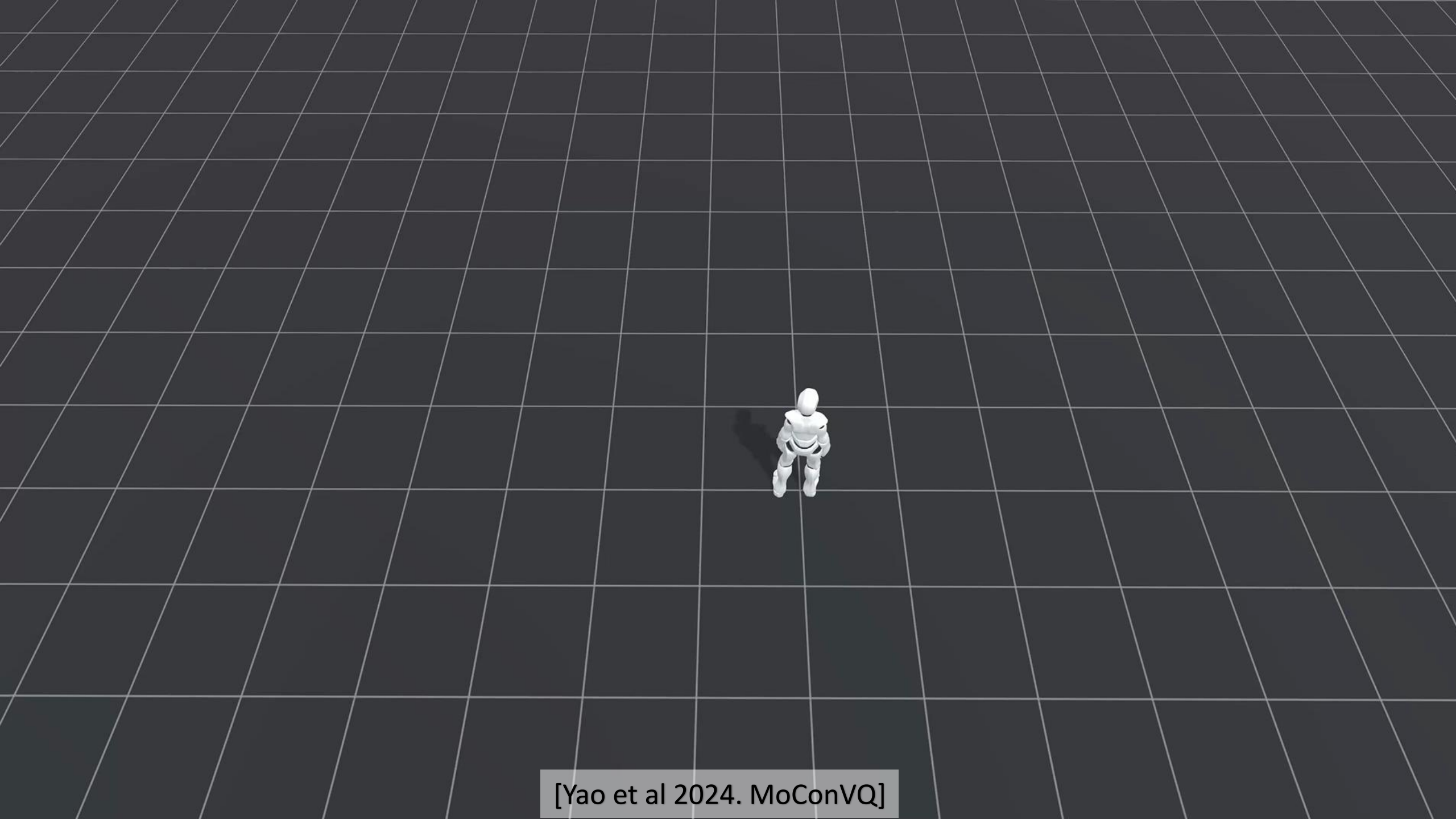


Large Motion Dataset
>25 hours



Sequential VQ Representation

[Yao et al 2024. MoConVQ]



[Yao et al 2024. MoConVQ]

HybrIK
[Li et al 2021]

Simulated (Ours)

[Yao et al 2024. MoConVQ]


Question: “a person walks forward for a long time and kicks, then he begins to dance”
[Yao et al 2024. MoConVQ]

How do you make a character walk in a square trajectory?

I would need the character to take straight walks forward, combined with 90 degree turns at each corner.
Forward walk: [297, 471, 246, 463, 463]
Sharp right turn: [360, 360, 360, 108,...]

[MoConVQ – Yao et al. 2024]

A path to realistic motion synthesis

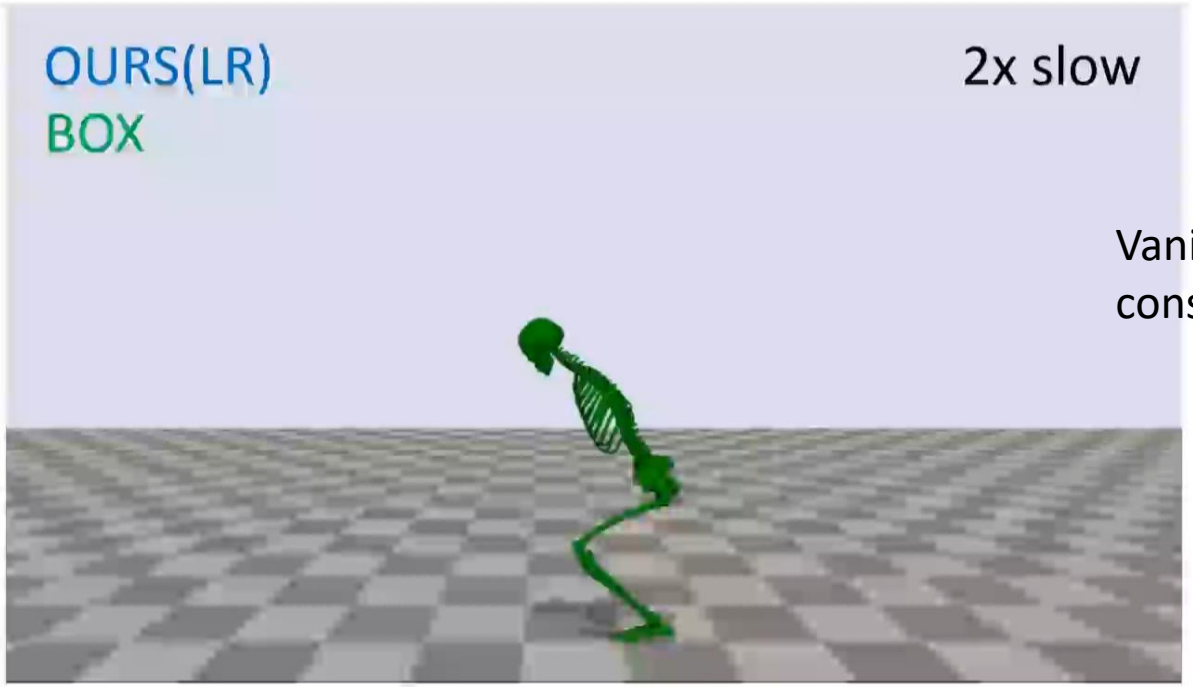
- Data is the key to scaling up
  - Potential extension: VLA
  - Model-based methods are useful for learning large models
- How many data is needed? Just a guess…
  - Hierarchical models
    - Low-level motion: 100 hours, high-quality 3D data
    - High-level behavior: 1000+ hours, the more the better
  - End-to-end models?

Why do we need large data?

- Bad evaluator
  - Models need to learn to evaluate results from data
- Imperfect modeling
  - Agent cannot learn on its own

Learn from the first principles?

- We need an accurate model



Bio-inspired constraints

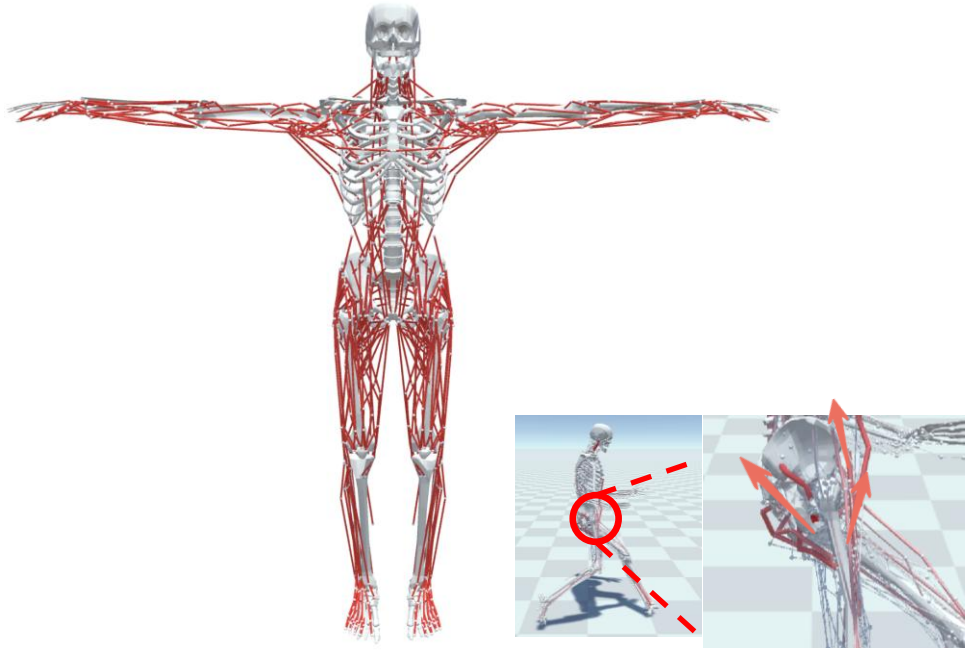
Vanilla constraints

BOX jumps unrealistically high (60cm) by over-bending the leg

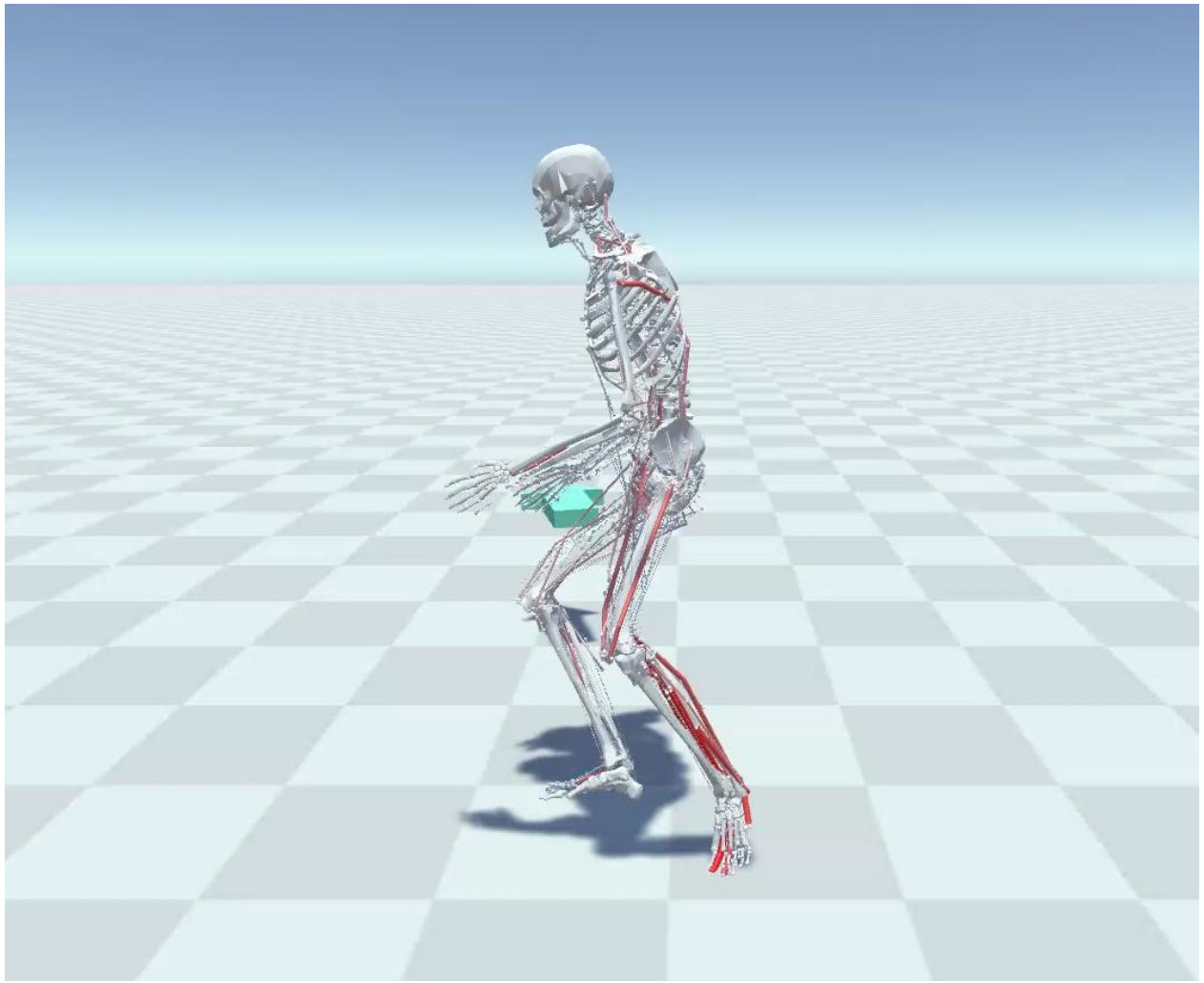
[Jiang et al. 2019 - Synthesis of Biologically Realistic Human Motion Using Joint Torque Actuation]

Learn from the first principles?

- We need an accurate model



[Feng et al 2023. MuscleVAE]

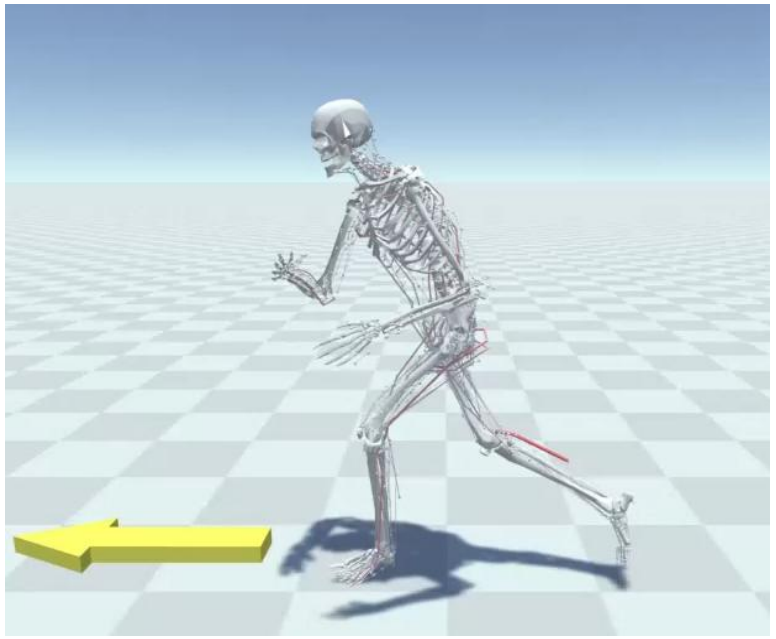


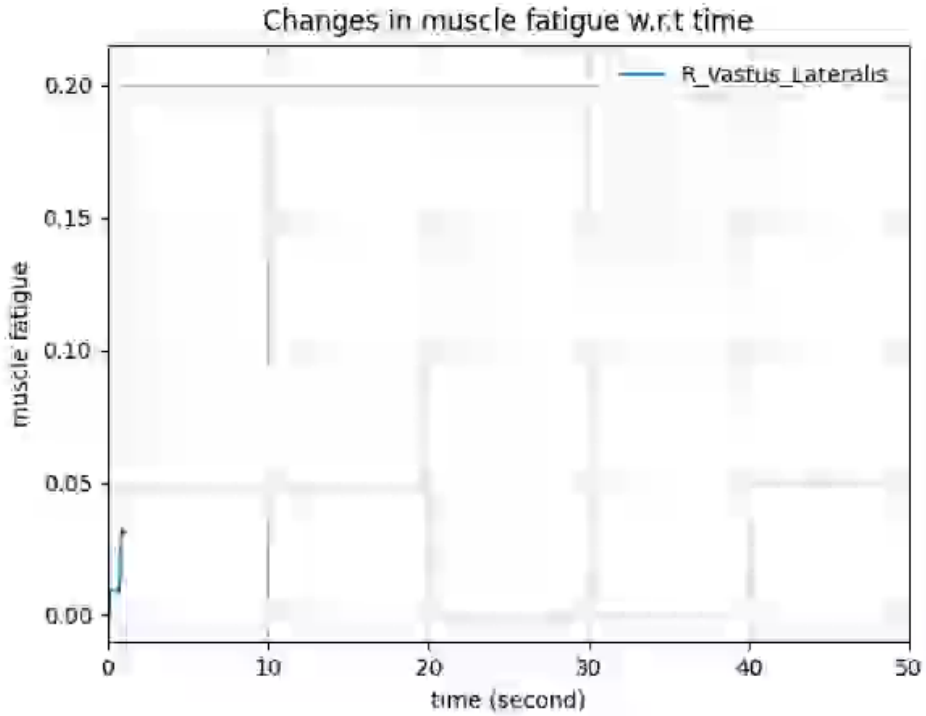
Learn from the first principles?

- We need an accurate model

Marathon's Struggle, from
https://www.youtube.com/watch?v=jrZH3Syx78g

Fatigue Simulation





[Feng et al 2023. MuscleVAE]

Thank you!

[libin.liu@pku.edu.cn | http://libliu.info]