NEURAL INFORMATION PROCESSING SYSTEMS

idea INTERNATIONAL DIGITAL ECONOMY ACADEMY
粤港澳大湾区数字经济研究院

清華大學深圳国际研究生院
Tsinghua Shenzhen International Graduate School

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

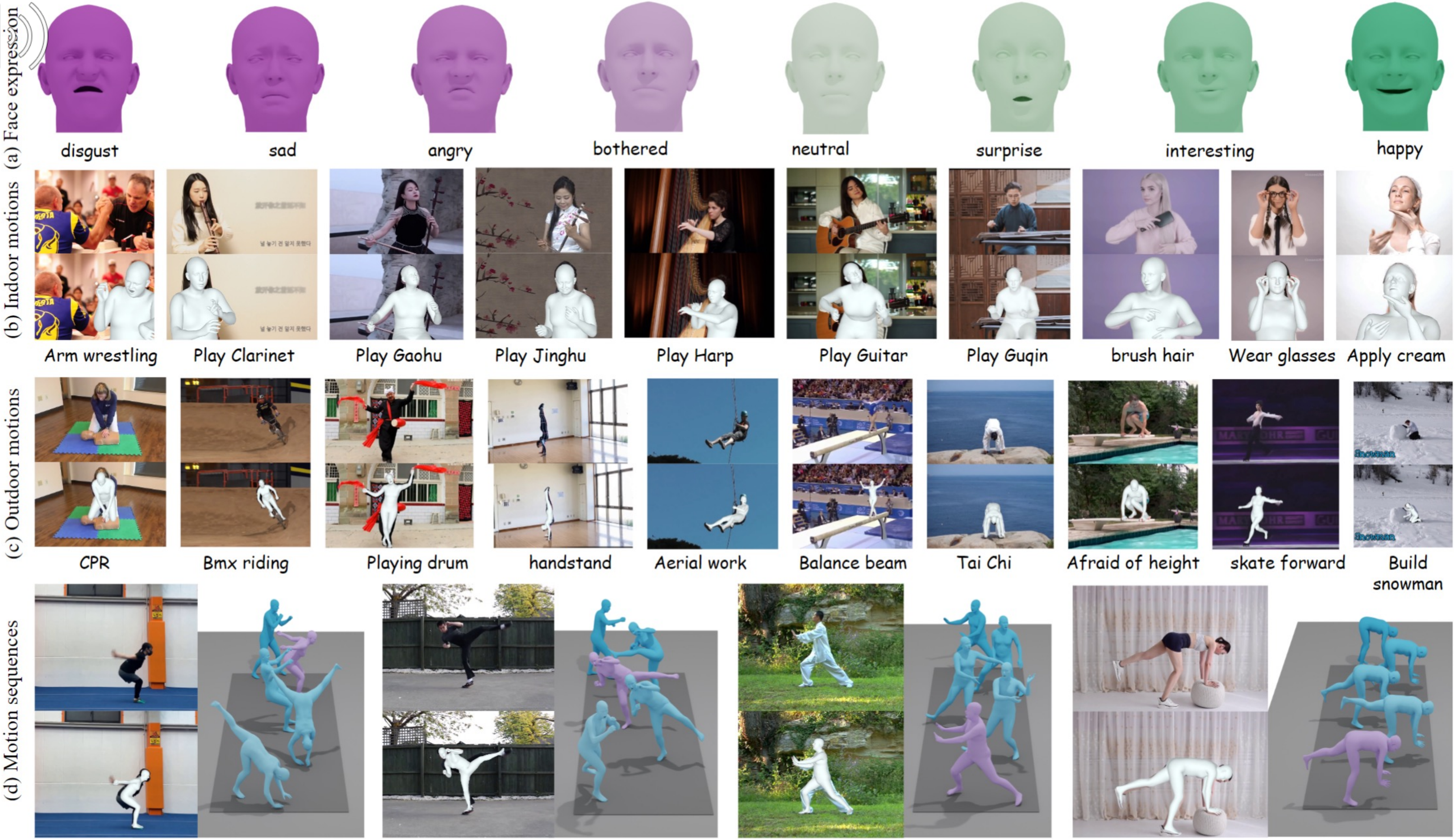
Motion-X:

A Large-scale Expressive Whole-body Human Motion Dataset

Jing Lin [1,2]* Ailing Zeng[1]*,★ Shunlin Lu[3]* Yuanhao Cai[2] Ruimao Zhang[3] Haoqian Wang[2] Lei Zhang[1]

- * Co-first author, ★ Corresponding author
- This work was done when Jing Lin and Shunlin Lu were interns at IDEA.

1. International Digital Economy Academy (IDEA) ,
2. Shenzhen International Graduate School, Tsinghua University
3. The Chinese University of Hong Kong, Shenzhen


(a) Face expression
disgust
sad
angry
bothered
neutral
surprise
interesting
happy
(b) Indoor motions
Arm wrestling
Play Clarinet
Play Gaohu
Play Jinghu
Play Harp
Play Guitar
Play Guqin
brush hair
Wear glasses
Apply cream
(c) Outdoor motions
CPR
Bmx riding
Playing drum
handstand
Aerial work
Balance beam
Tai Chi
Afraid of height
skate forward
Build snowman
(d) Motion sequences

[1] G

[2] A

Whole-body Motion Examples

We annotate text-motion sequences from massive online videos and 7 datasets:

1. Online videos [6.0M]: kungfu, music, performance, … ,and IDEA400[2.6M]

Playing Shaolin Kungfu

Playing the guqin

Blow nose

Whole-body Motion Examples

We annotate text-motion sequences from massive online videos and 7 datasets:

1. Online videos [6.0M]: kungfu, music, performance, … ,and IDEA400[2.6M]
2. Multi-view dance: AIST [0.3M][1]

Break Basic Dance up rock

Break Basic Dance 6 step

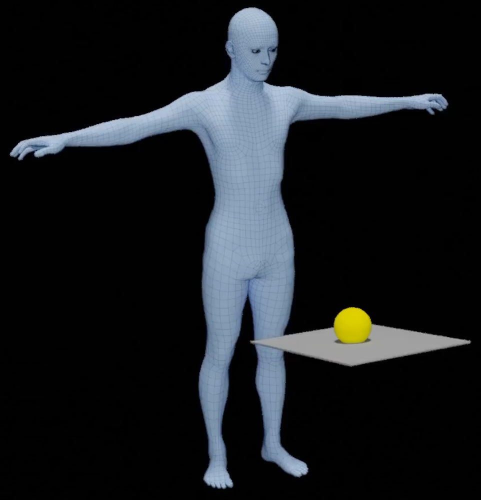
Break Basic Dance indian step

[1] AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing, ISMIR 2019

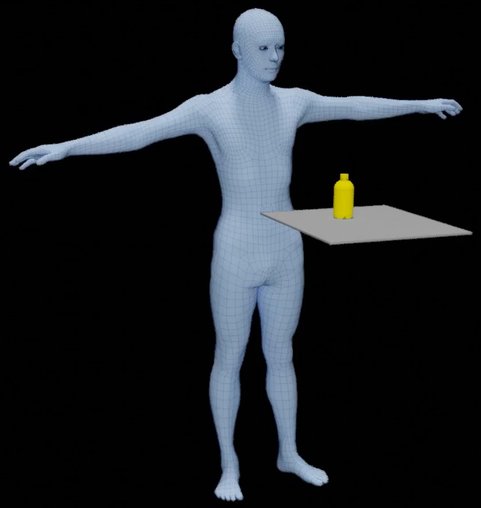
Whole-body Motion Examples

We annotate text-motion sequences from massive online videos and 7 datasets:

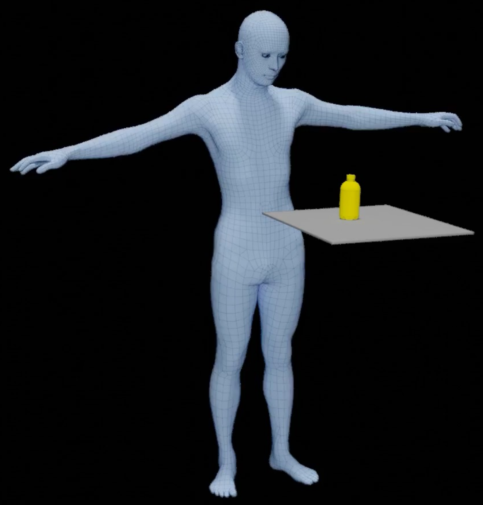
1. Online videos [6.0M]: kungfu, music, performance, … ,and IDEA400[2.6M]
2. Multi-view dance: AIST [0.3M][1]
3. Human-scene-interaction: GRAB [0.4M][2], EgoBody [0.4M][3]



Pick up a sphere



Drink water with a bottle



Pour water with a bottle

[1] AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing, ISMIR 2019
[2] GRAB: A Dataset of Whole-Body Human Grasping of Objects, ECCV 2020
[3] Human Body Shape and Motion of Interacting People from Head-Mounted Devices, ECCV 2022

Whole-body Motion Examples

We annotate text-motion sequences from massive online videos and 7 datasets:

1. Online videos [6.0M]: kungfu, music, performance, … ,and IDEA400[2.6M]
2. Multi-view dance: AIST [0.3M][1]
3. Human-scene-interaction: GRAB [0.4M][2], EgoBody [0.4M][3]
4. Action recognition: HAA500 [0.3M][4], HuMMan [0.1M][5]

Add new car tire

Cardiopulmonary Resuscitation

Baseball pitch

[1] AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing, ISMIR 2019

[2] GRAB: A Dataset of Whole-Body Human Grasping of Objects, ECCV 2020

[3] Human Body Shape and Motion of Interacting People from Head-Mounted Devices, ECCV 2022

[4] HAA500: Human-Centric Atomic Action Dataset with Curated Videos, ICCV 2021

[5] HuMMan: Multi-Modal 4D Human Dataset for Versatile Sensing and Modeling, ECCV 2022

Whole-body Motion Examples

We annotate text-motion sequences from massive online videos and 7 datasets:

1. Online videos [6.0M]: kungfu, music, performance, … ,and IDEA400[2.6M]
2. Multi-view dance: AIST [0.3M]
3. Human-scene-interaction: GRAB [0.4M], EgoBody [0.4M]
4. Action recognition: HAA500 [0.3M], HuMMan [0.1M]
5. Motion Capture-based: AMASS[1]/Babel[2]/HumanML3D [5.4M][3]

Kicks with left leg

Jumps straight to the left

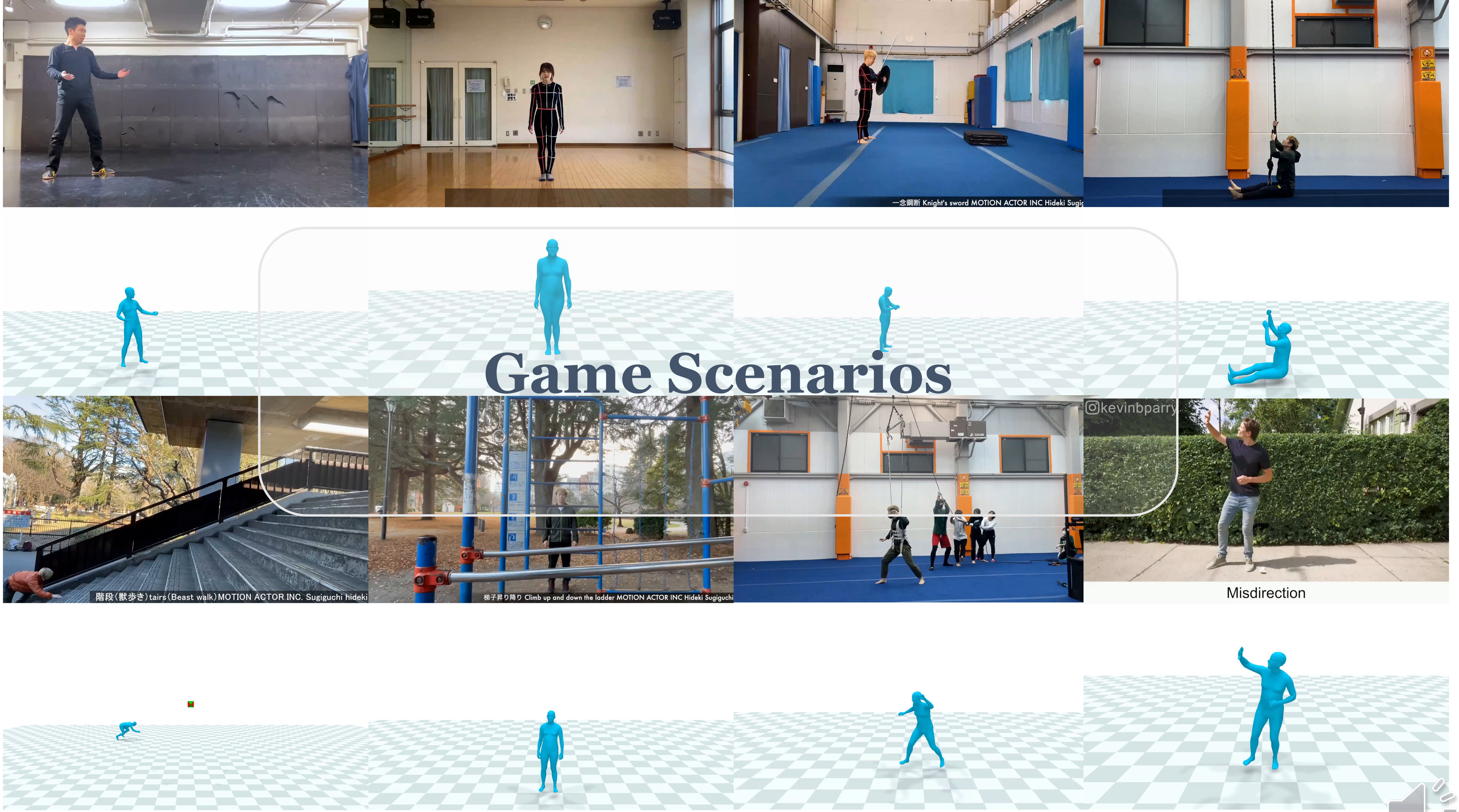
Walking forward

[1] AMASS: Archive of Motion Capture As Surface Shapes, ICCV 2019
[2] BABEL: Bodies, Action and Behavior with English Labels
[3] Generating Diverse and Natural 3D Human Motions From Text, CVPR 2022

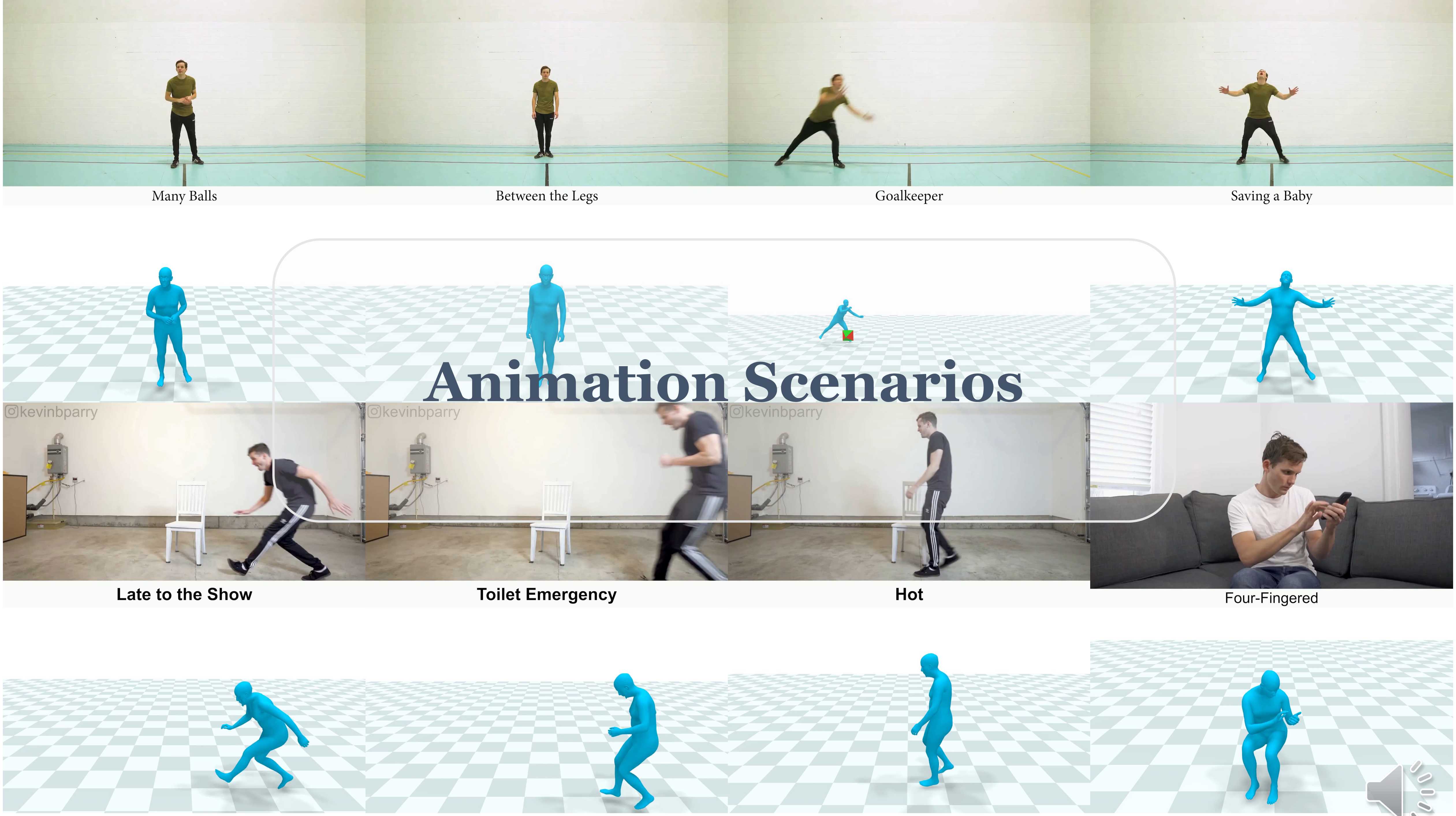
Qualitative SMPL-X Results

Motion-X

Game Scenarios


一念鋼断 Knight's sword MOTION ACTOR INC Hideki Sugig
@kevinbparry
階段(獣歩き)tairs(Beast walk)MOTION ACTOR INC. Sugiguchi hideki
梯子昇り降り Climb up and down the ladder MOTION ACTOR INC Hideki Sugiguchi
Misdirection

Animation Scenarios



Many Balls

Between the Legs

Goalkeeper

Saving a Baby

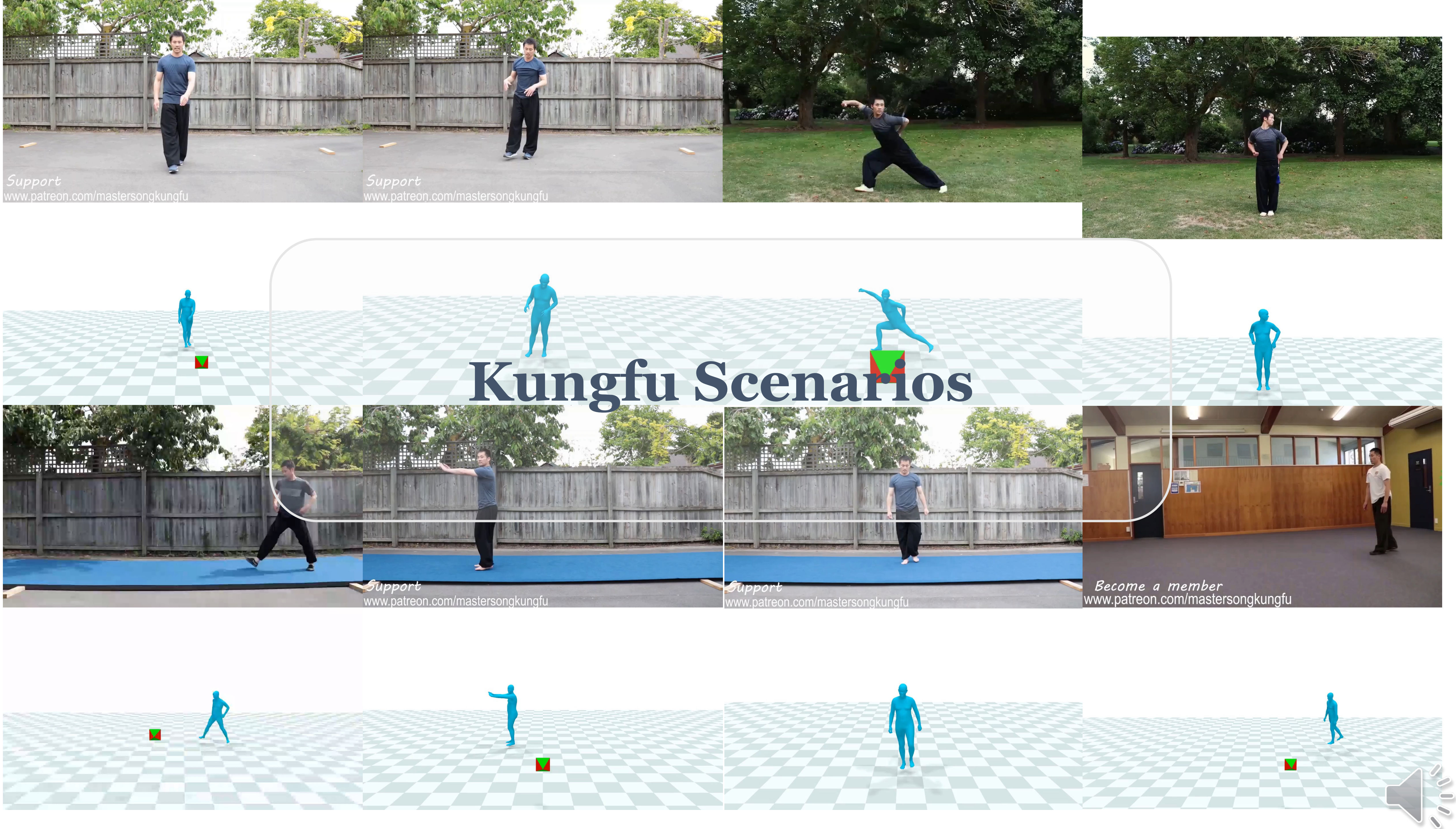
Late to the Show

Toilet Emergency

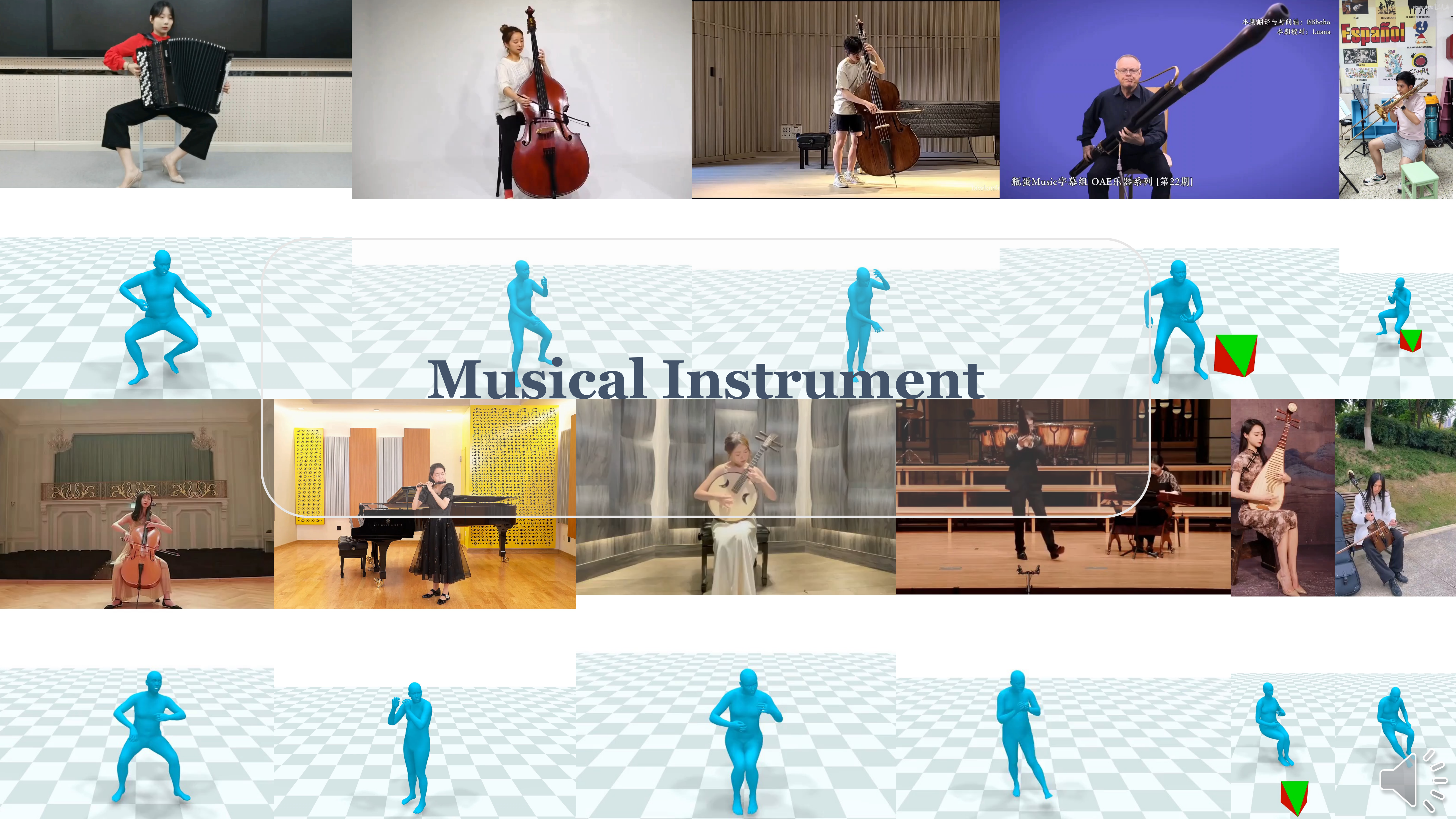
Hot

Four-Fingered

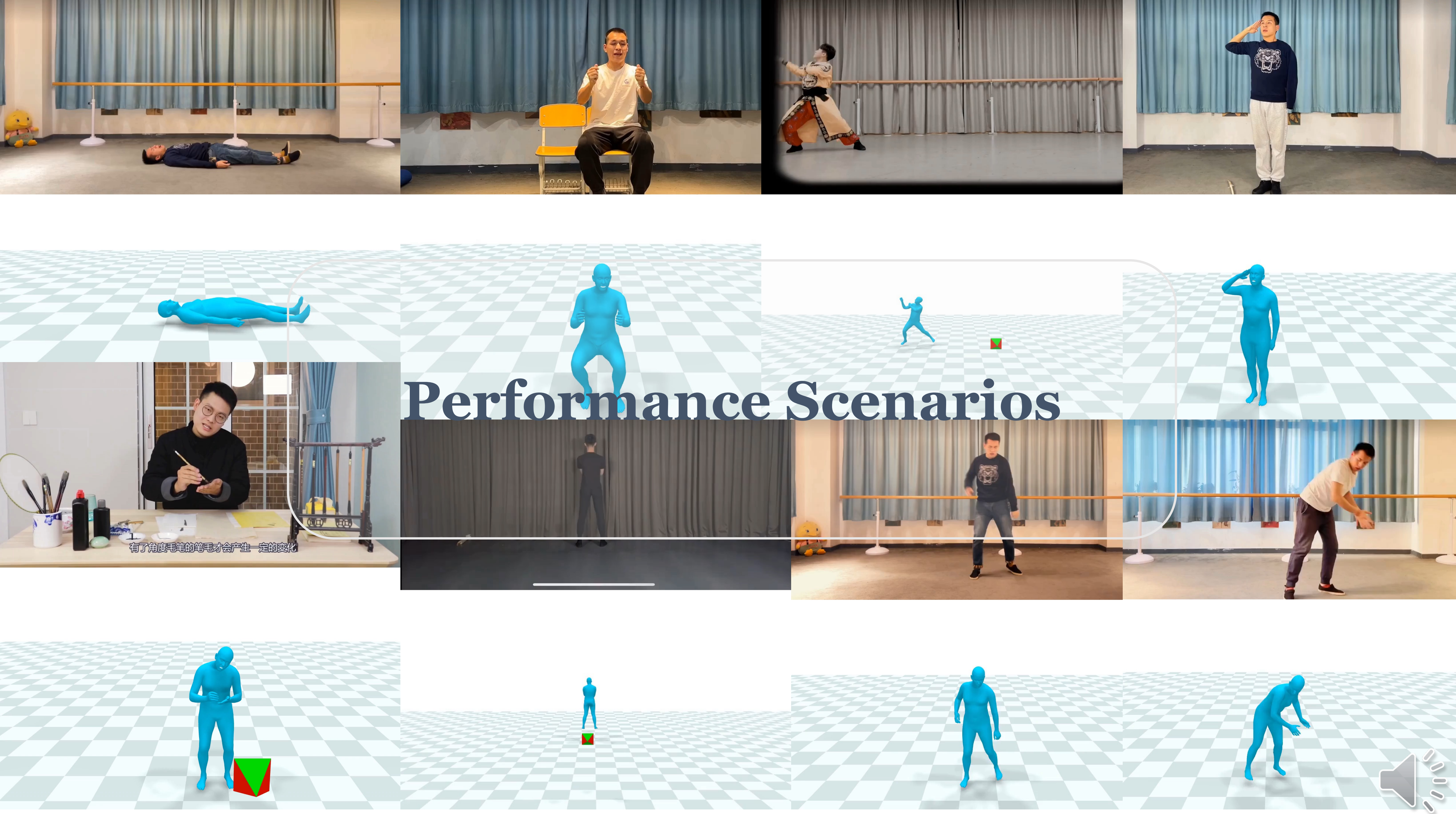
Kungfu Scenarios


Support
www.patreon.com/mastersongkungfu
Support
www.patreon.com/mastersongkungfu
Support
www.patreon.com/mastersongkungfu
Support
www.patreon.com/mastersongkungfu
Become a member
www.patreon.com/mastersongkungfu

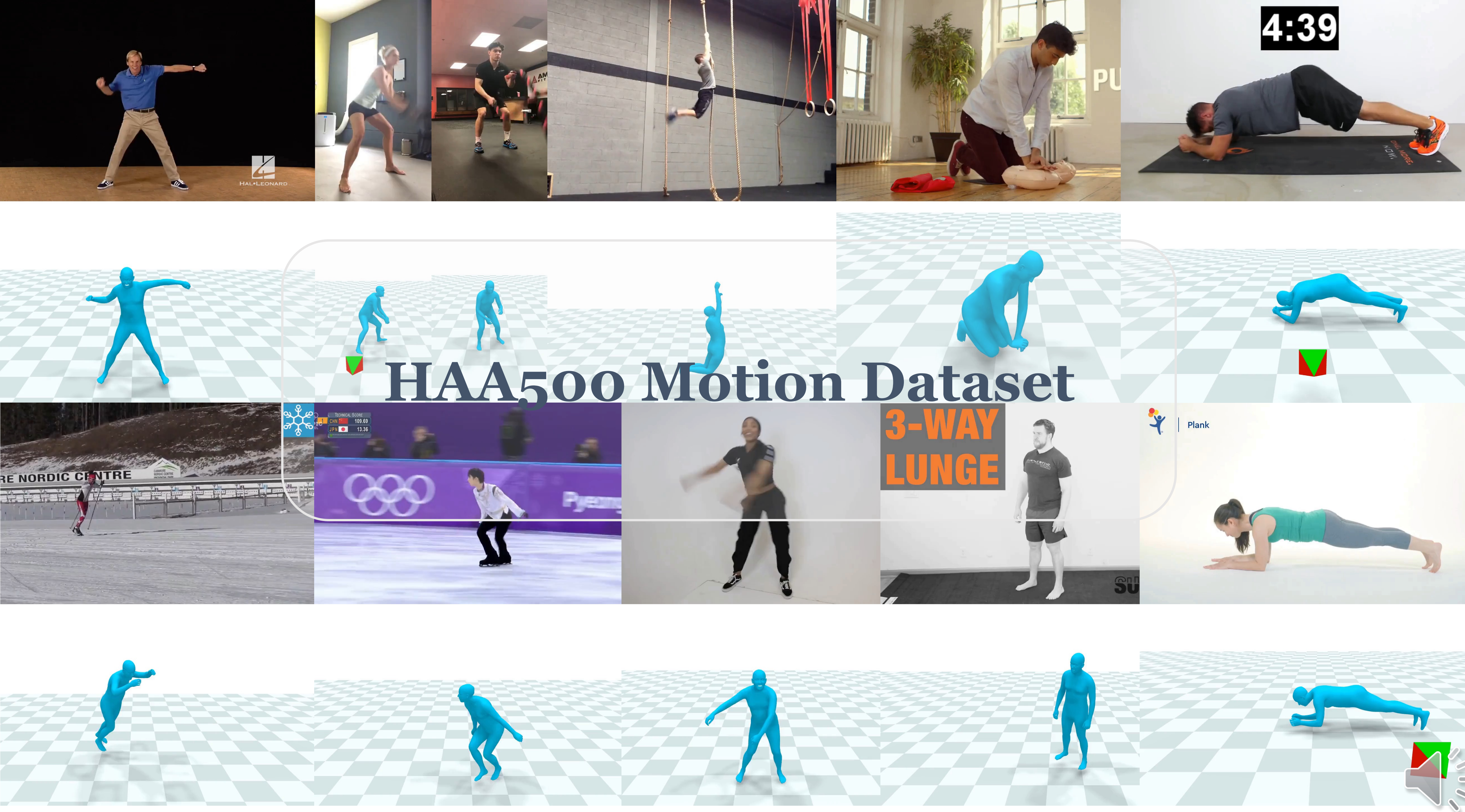
Musical Instrument



Performance Scenarios

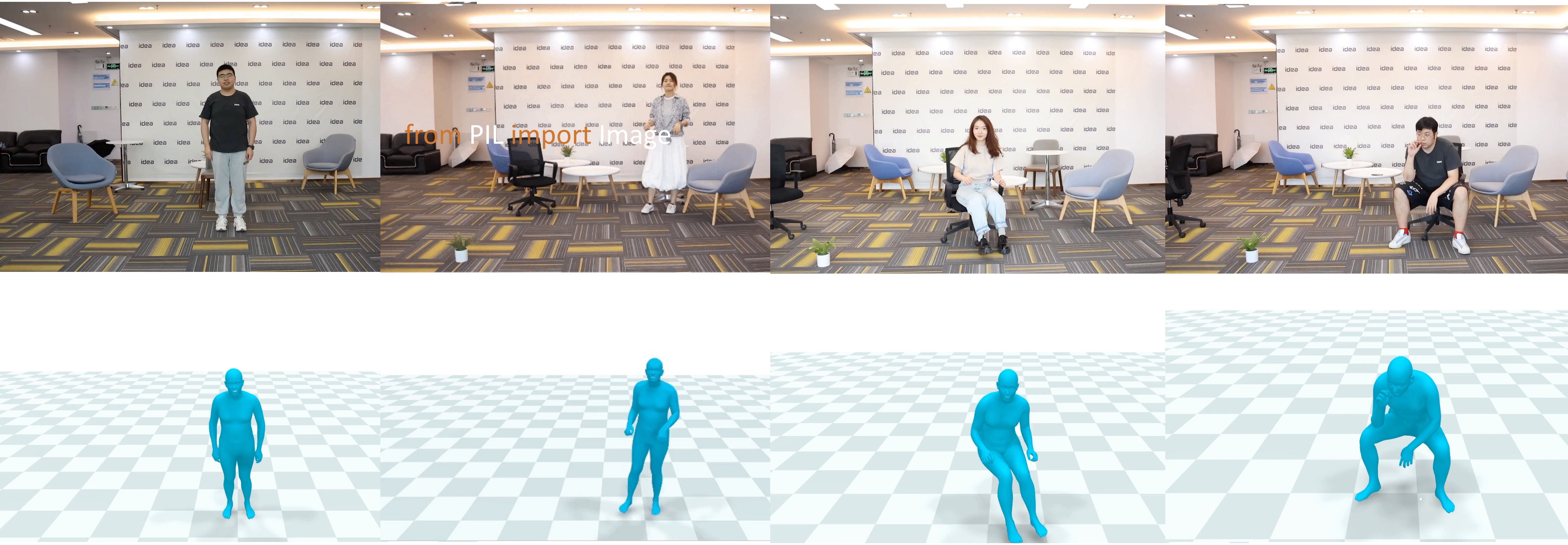


HAA500 Motion Dataset



IDEA400 Motion Dataset

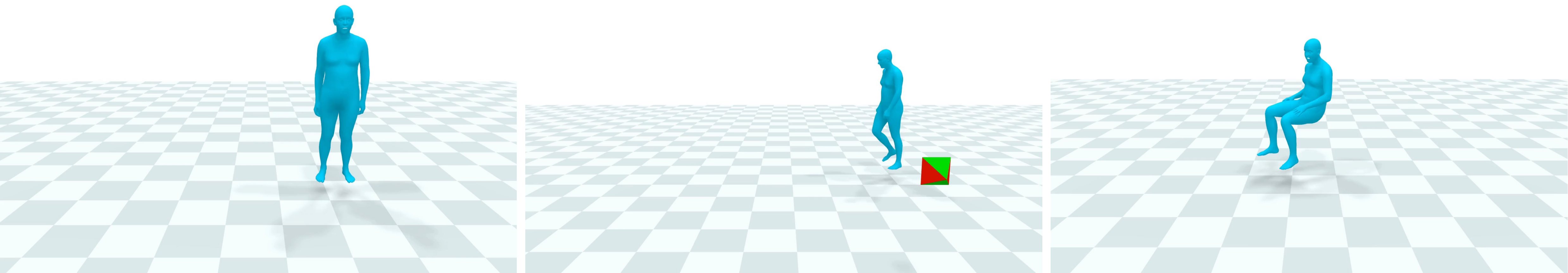
- **400** diverse actions, covering daily, specific motions with various hand gestures and facial expressions. (including 120 actions in NTU RGB+D120[1])



```
from PIL import Image
```

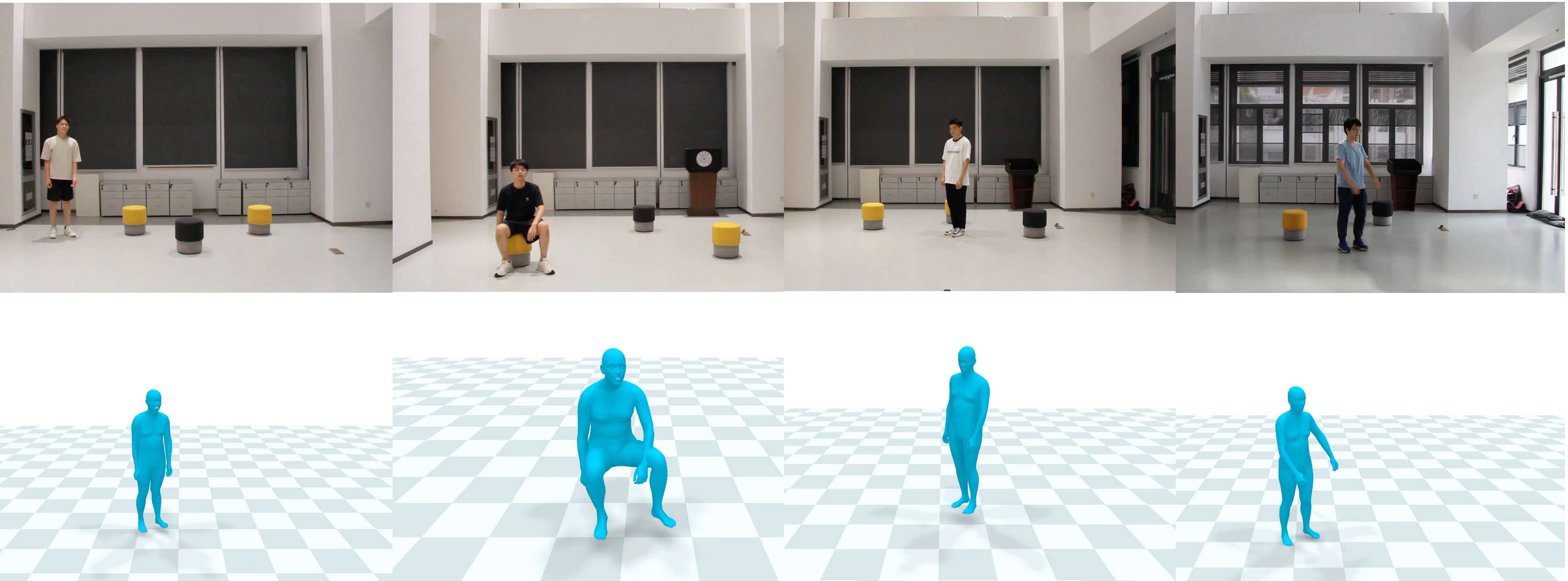
IDEA400 Motion Dataset

- For each motion, the actor performs 3 standing, 3 walking, 4 sitting lower-body actions (10 times in total).



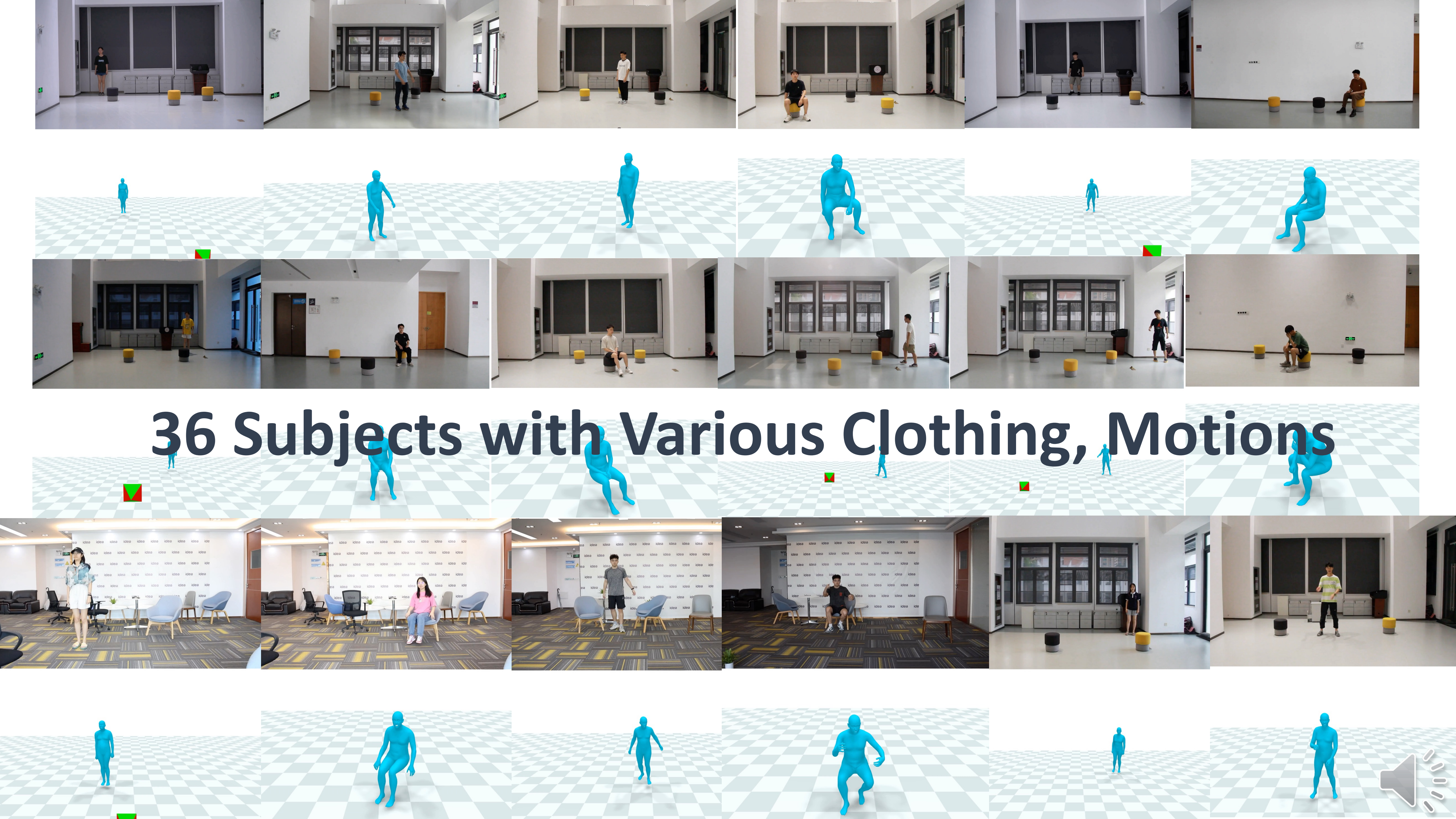
IDEA400 Motion Dataset

- Currently, we provide **12K** motions in total (400*10 times*3 rounds).



36 Subjects with Various Clothing, Motions

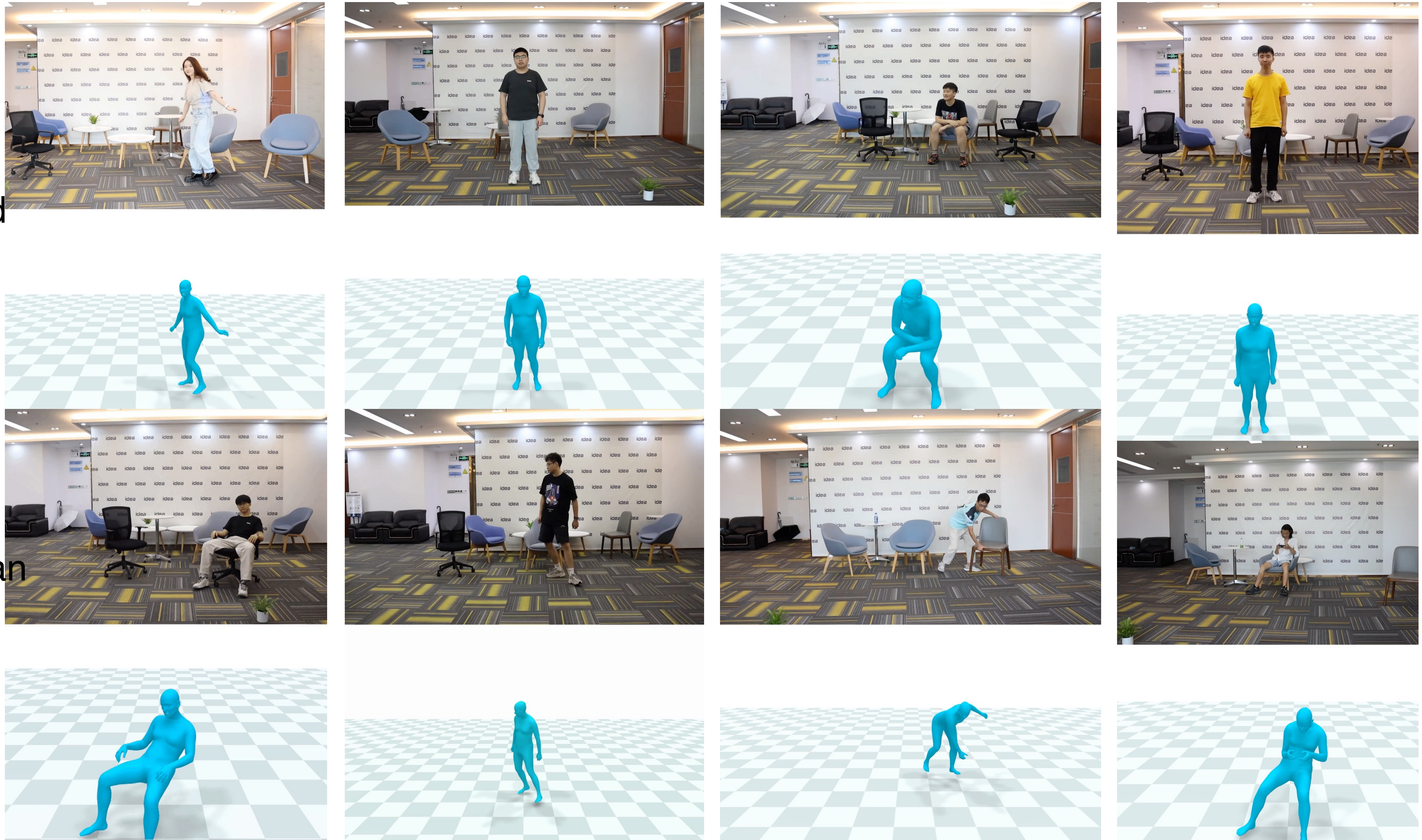
36 Subjects with Various Clothing, Motions

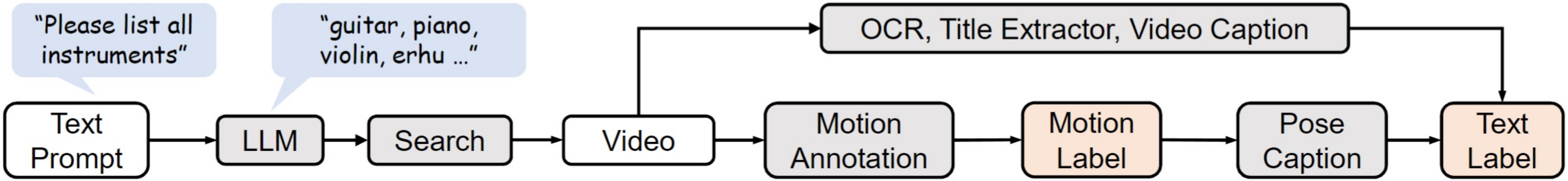


Some Interesting Scenes…

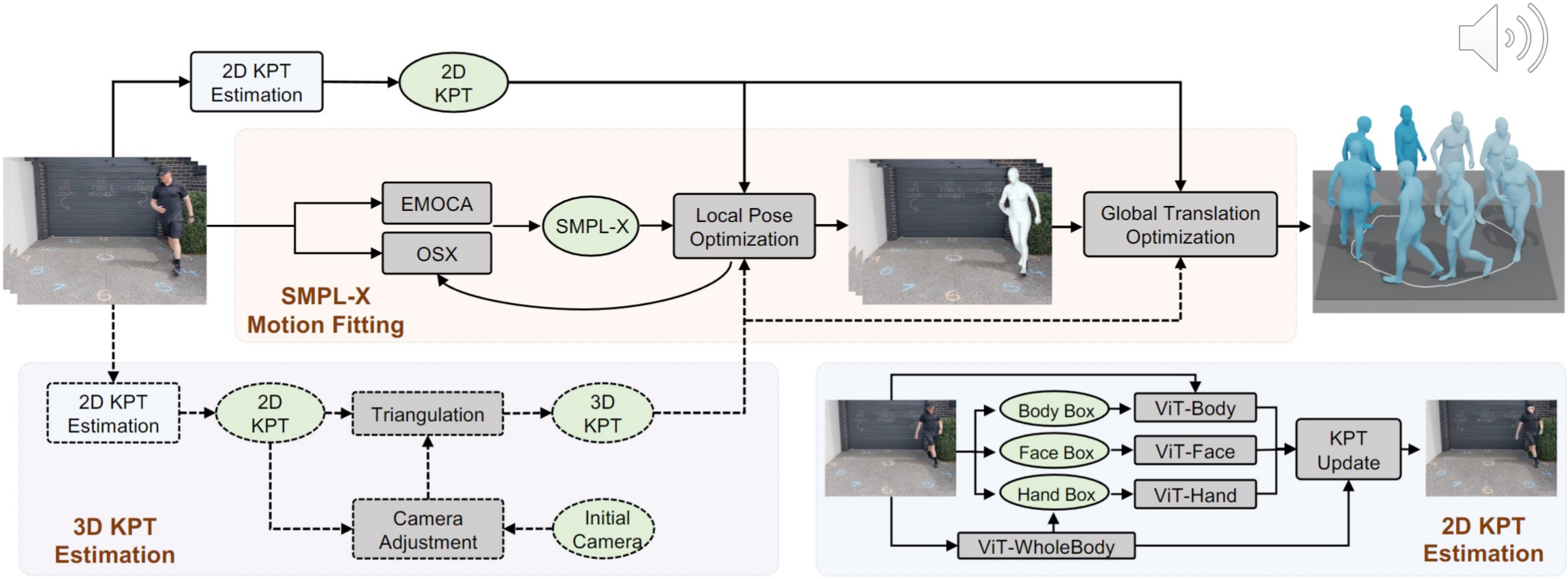
1. Rich facial expressions and hand gestures

2. Diverse human self-contact and human-object contact

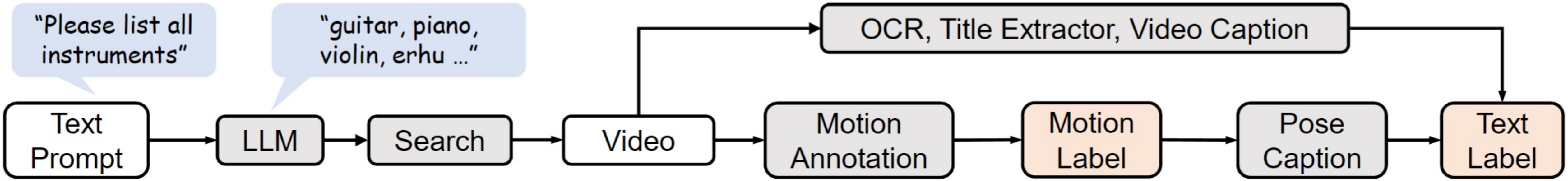




"Please list all instruments"
"guitar, piano, violin, erhu …"
Text Prompt
LLM
Search
Video
OCR, Title Extractor, Video Caption
Motion Annotation
Motion Label
Pose Caption
Text Label



2D KPT Estimation
2D KPT
EMOCA
OSX
SMPL-X
Local Pose Optimization
Global Translation Optimization
SMPL-X Motion Fitting
2D KPT Estimation
2D KPT
Triangulation
3D KPT
Camera Adjustment
Initial Camera
3D KPT Estimation
Body Box
Face Box
Hand Box
ViT-Body
ViT-Face
ViT-Hand
ViT-WholeBody
KPT Update
2D KPT Estimation

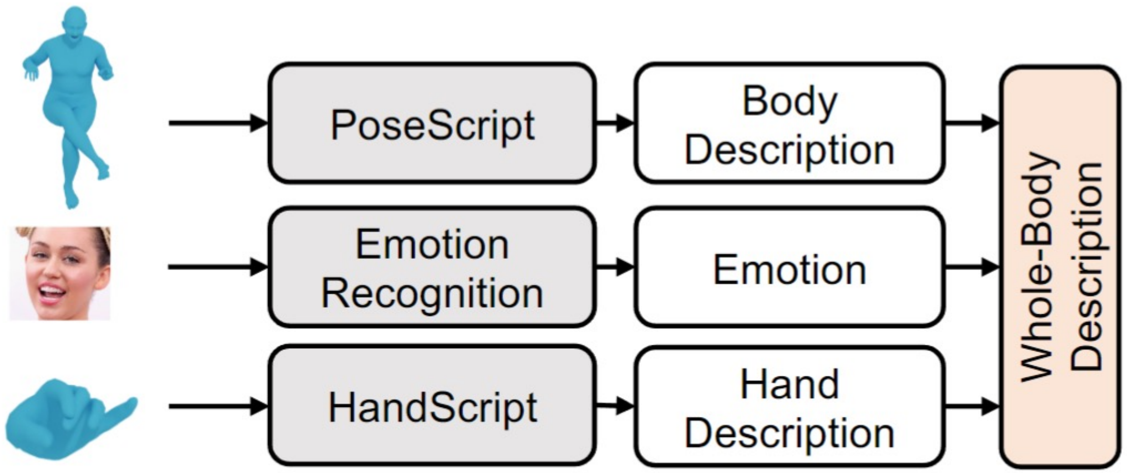


"Please list all instruments"

"guitar, piano, violin, erhu …"

Text Prompt → LLM → Search → Video → Motion Annotation → Motion Label → Pose Caption → Text Label

OCR, Title Extractor, Video Caption



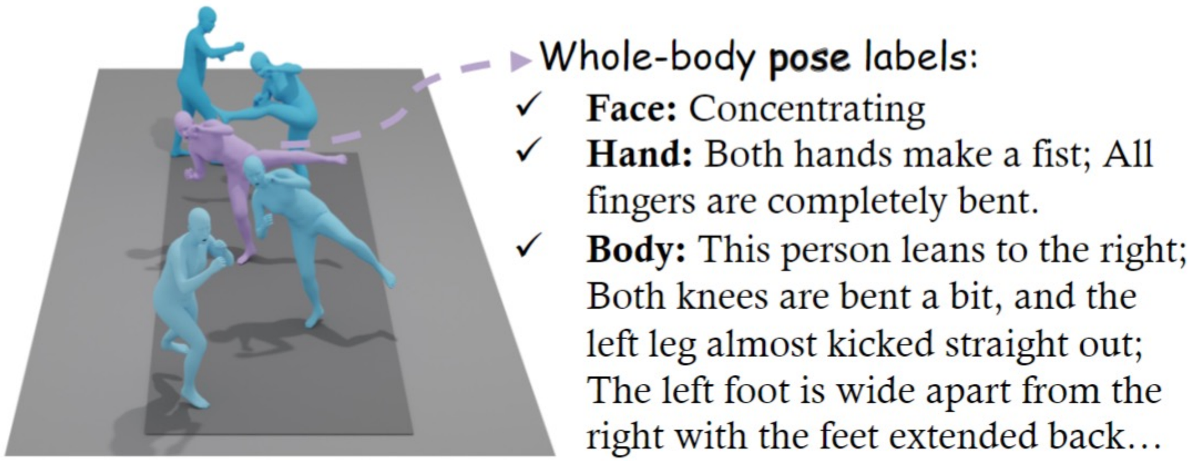
PoseScript → Body Description

Emotion Recognition → Emotion

HandScript → Hand Description

→ Whole-Body Description

(a) Whole-body Pose Caption



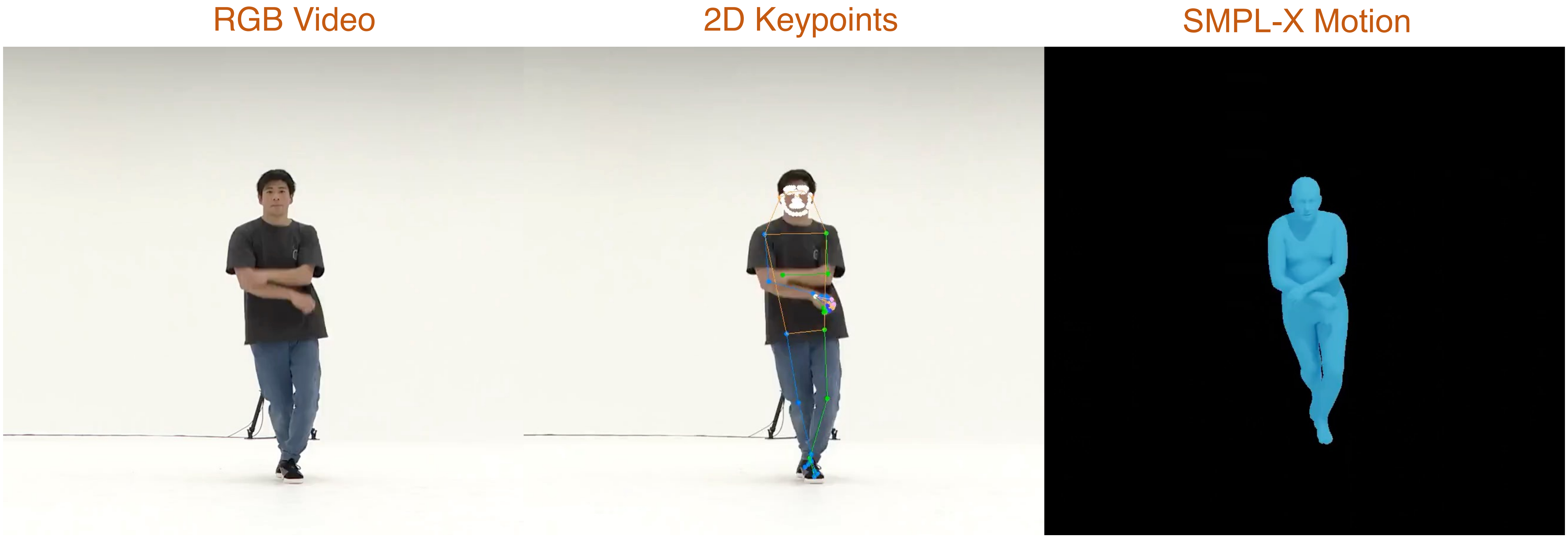
Whole-body **pose** labels:

- ✓ **Face:** Concentrating
- ✓ **Hand:** Both hands make a fist; All fingers are completely bent.
- ✓ **Body:** This person leans to the right; Both knees are bent a bit, and the left leg almost kicked straight out; The left foot is wide apart from the right with the feet extended back…

Sequence label: Shaolin Kung Fu Wushu Tsunami Kick

(b) Example of the Annotation Result

Motion-Text Examples


RGB Video

2D Keypoints

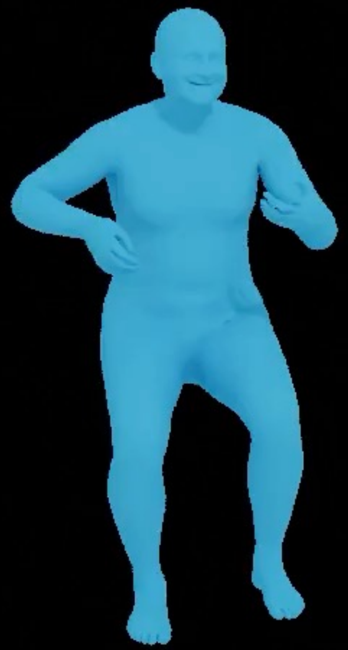
SMPL-X Motion

A man is doing break advanced dance.

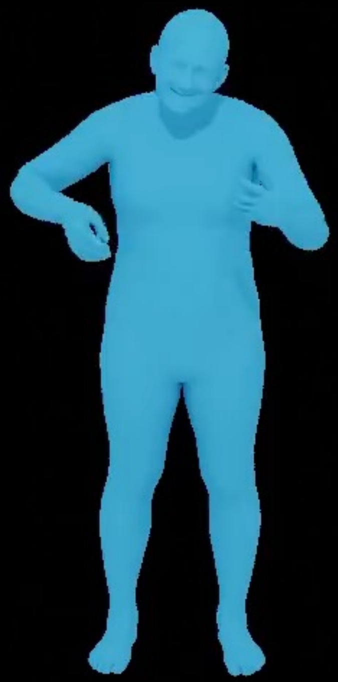
Rich diverse text-to-motion data

A person is performing ballet.

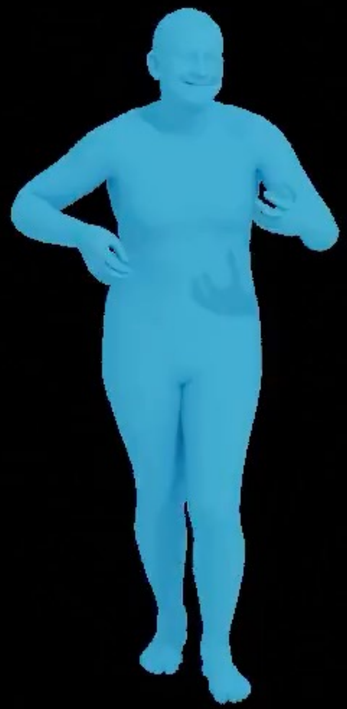
Motion Augmentation



Sit and play guitar

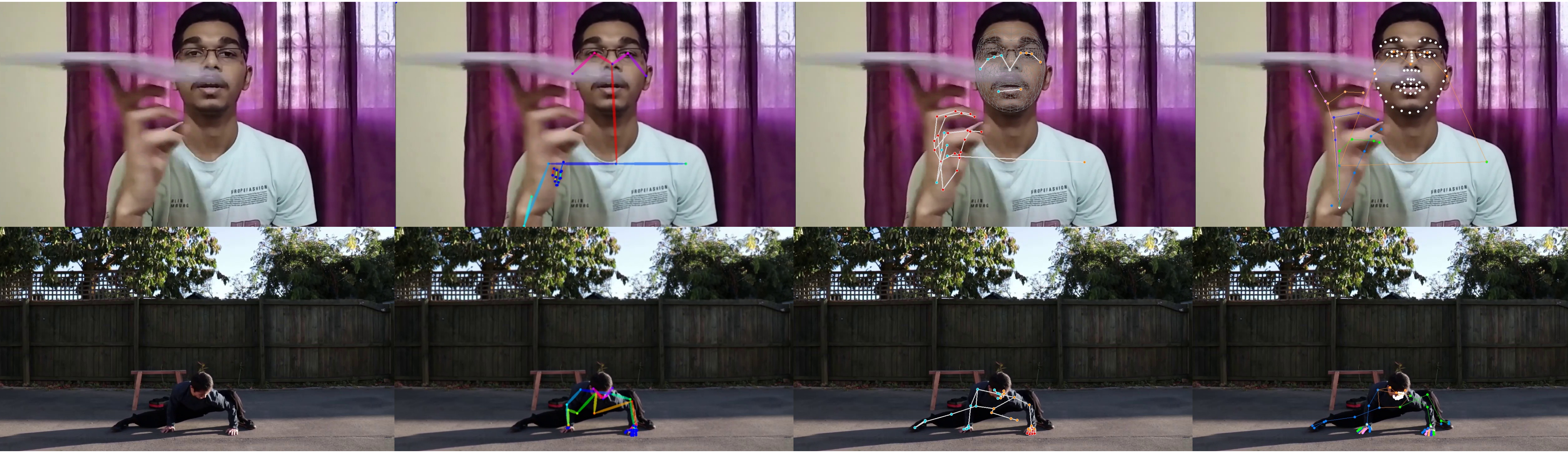


Stand and play guitar



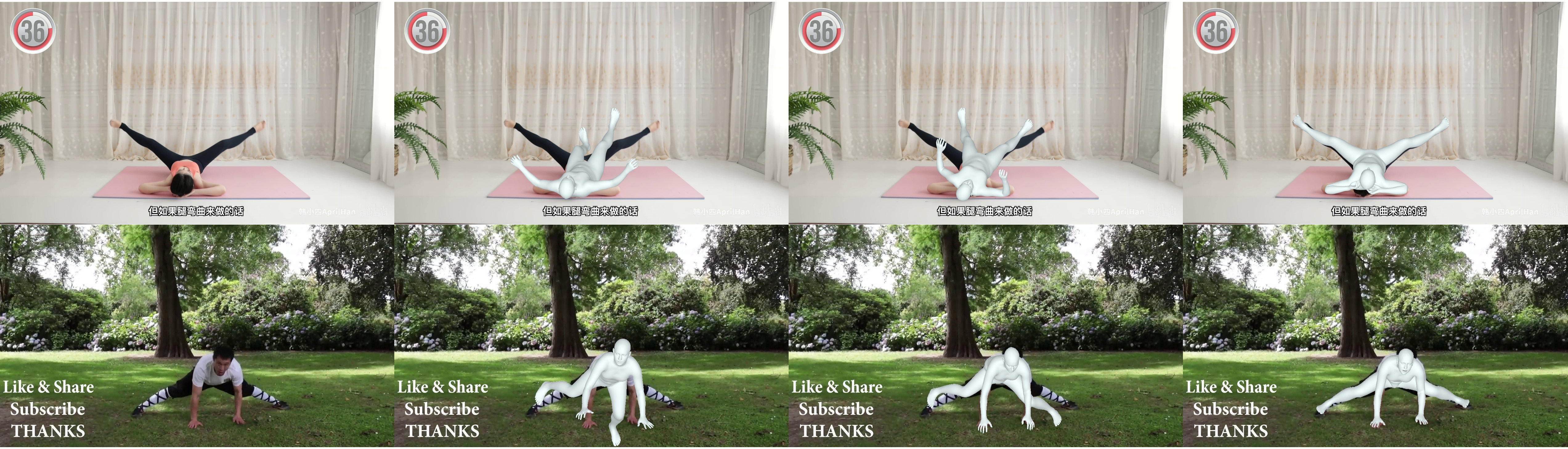
Walk and play guitar

2D Whole-body Keypoints Annotation



(a) Input Image (b) Openpose (c) MediaPipe (d) Ours

3D Whole-body SMPL-X Annotation



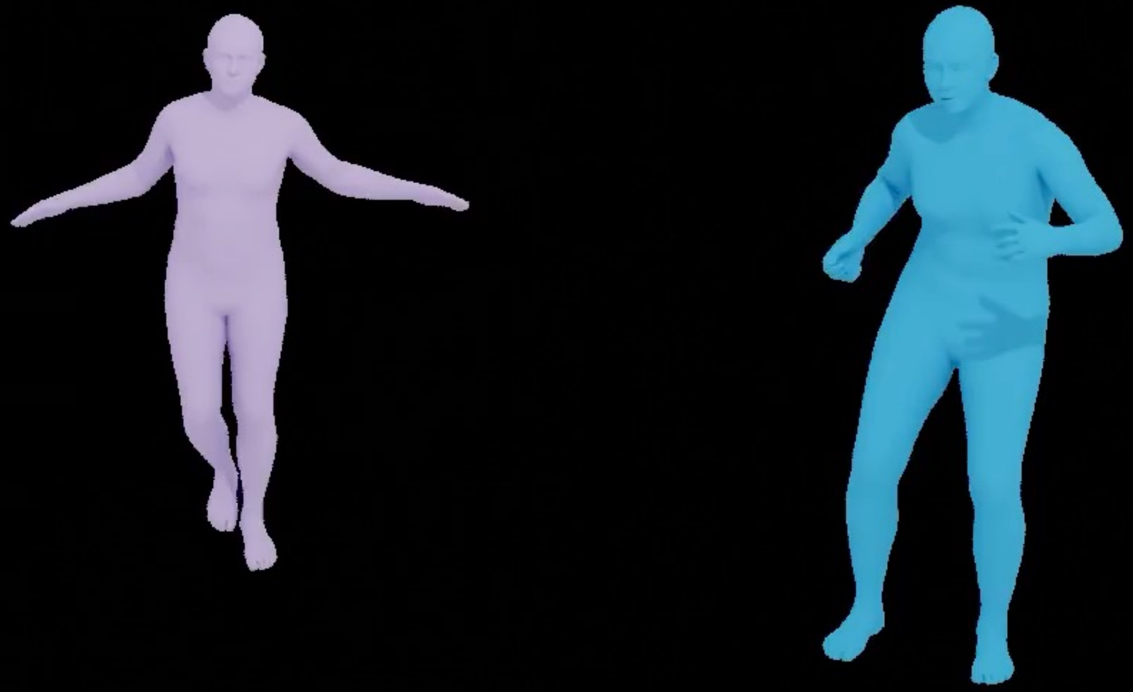
(a) Input Image

(b) Hand4Whole

(c) OSX

(d) Ours

Text-driven Motion Generation

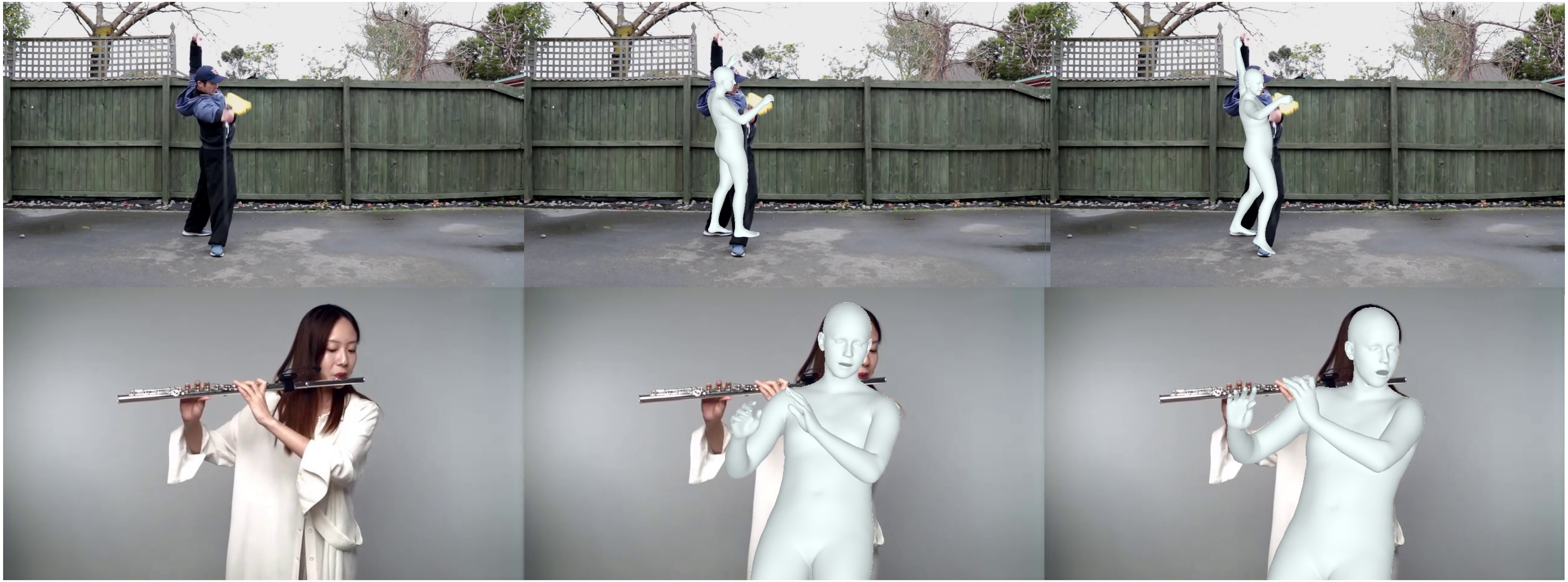


(a) w/o Motion-X

(b) w/ Motion-X

A man is playing erhu.

3D Whole-body Human Mesh Recovery

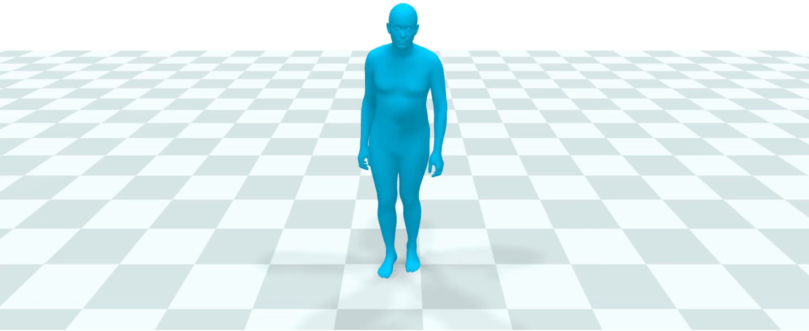


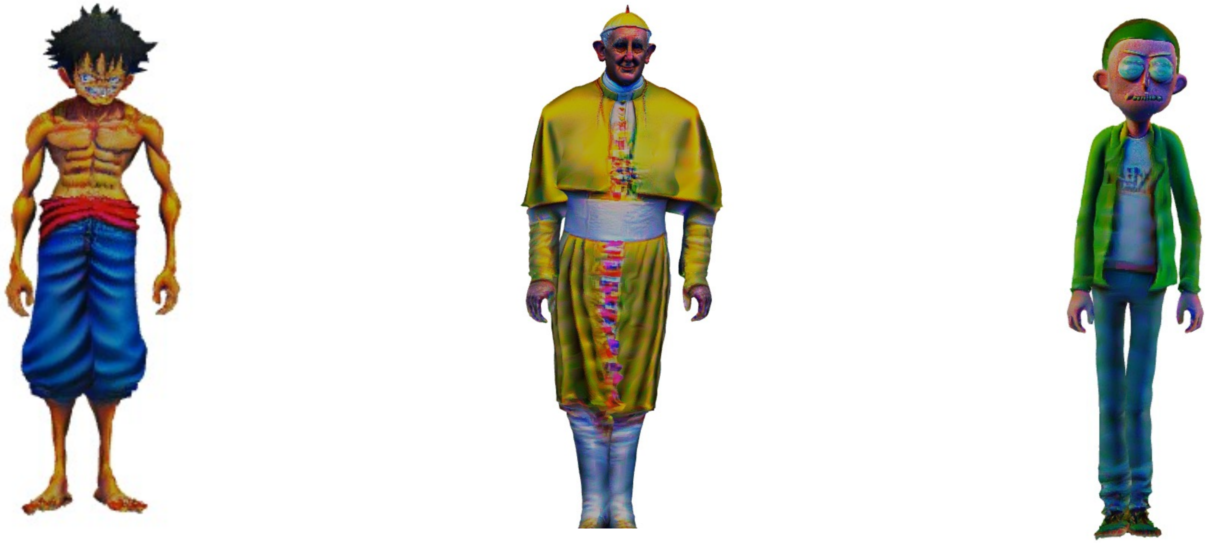
(a) Input Image

(b) w/o Motion-X

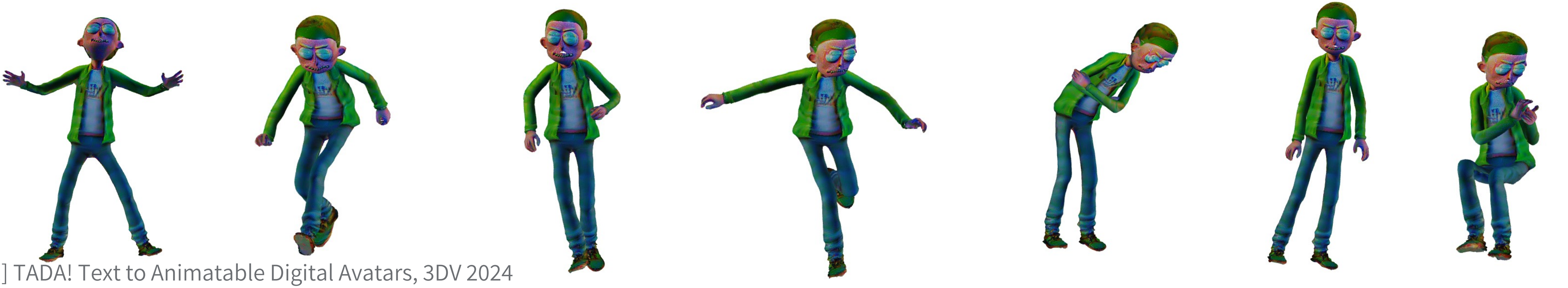
(c) w/ Motion-X

👬 Motion-X with TADA![1]





Given SMPL-X sequences in Motion-X,
we can animate various characters from TADA!



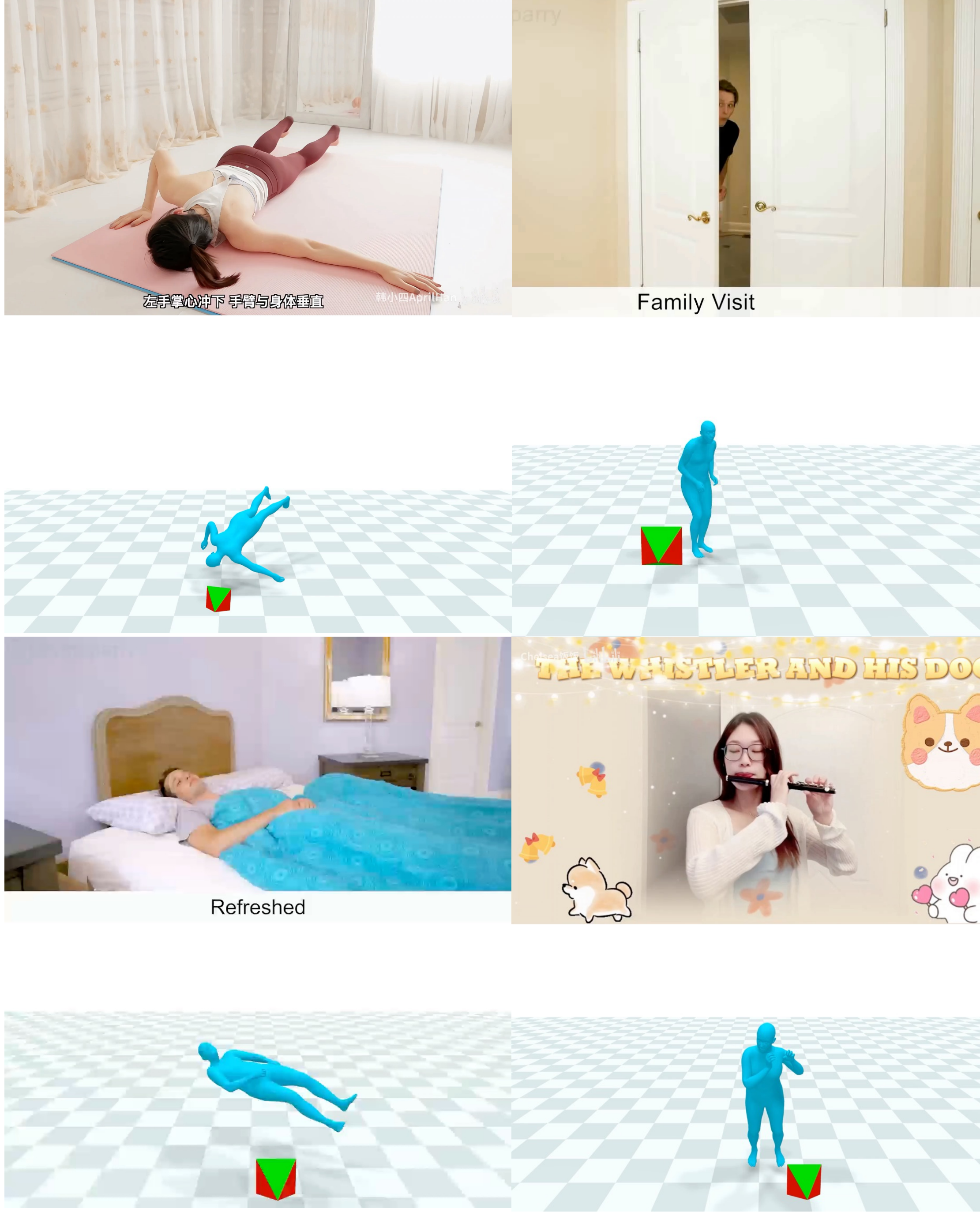
[1] TADA! Text to Animatable Digital Avatars, 3DV 2024

Limitation

- The quality of annotated videos are key.
- Heavy truncation, occlusion scenes make the invisible parts hard to annotate.
- Multi-person interaction are still challenging.

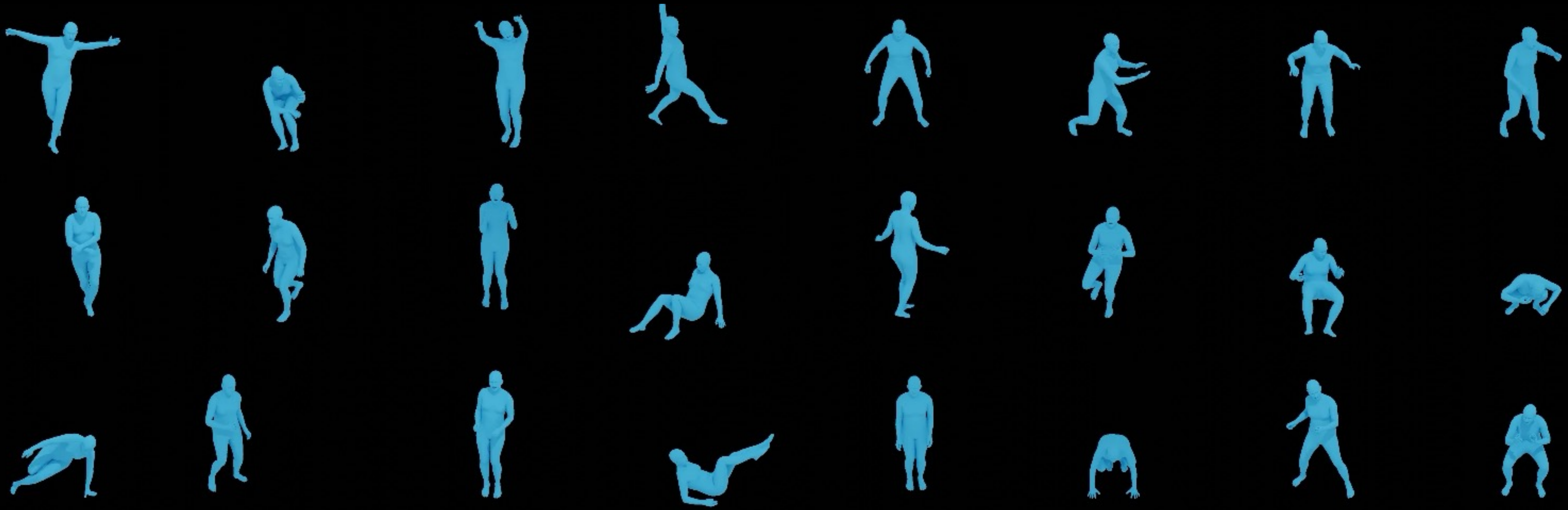
Future Work

- Since noisy labels are inevitable, learning from noisy labels for generation and understanding tasks would be quite important.
- We will continue to improve the motion and text labels' quality.



Summary

- A multi-modality, large-scale whole-body human motion dataset
- A novel, automatic whole-body motion and text annotation pipeline
- Effective in motion generation and human mesh recovery tasks



💖 Acknowledgement

- Thanks to all **video owners** for providing excellent videos.
- Thanks to all **IDEAers** and **THUers** who participated in the IDEA400 performance!
- Thanks to **Tingting Liao, Yuliang Xiu,** and **Tianze Zheng** for character animation by TADA!
- **And thanks for watching!**