NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

The Path from Marionette to Autonomous 3D Characters

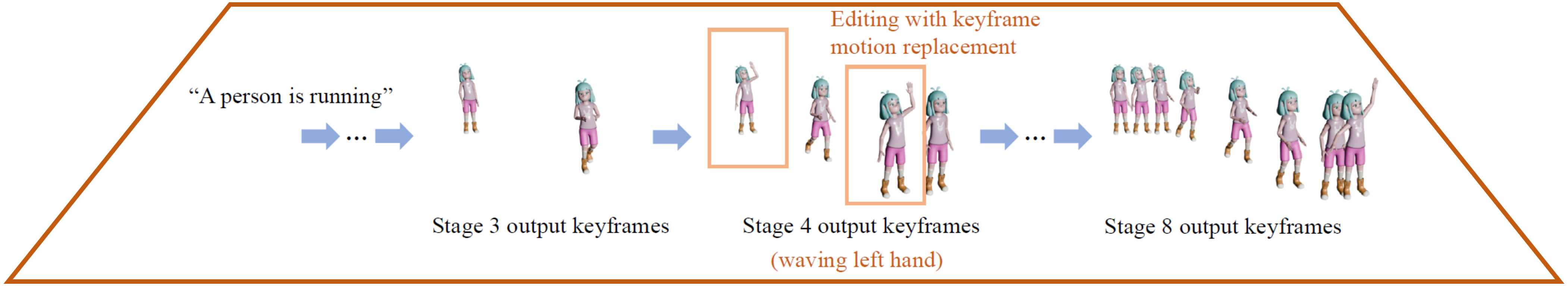
Ziwei Liu 刘子纬

Nanyang Technological University

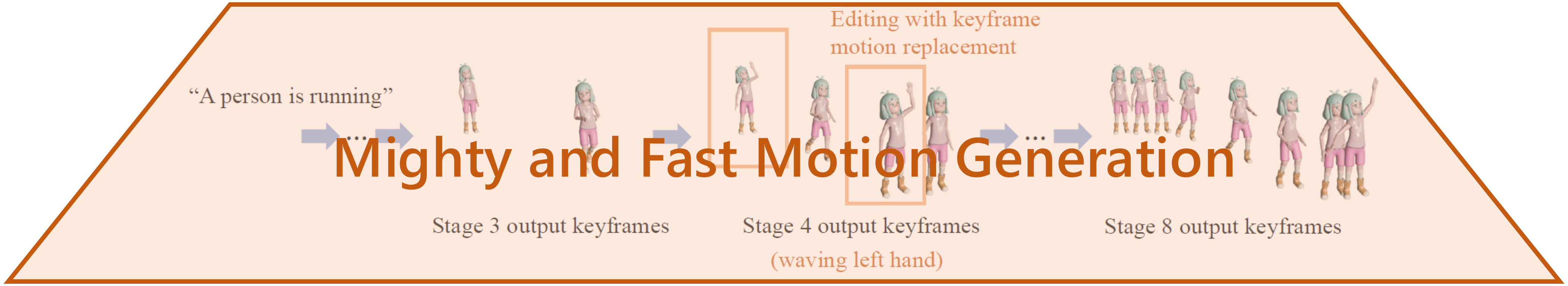
https://liuziwei7.github.io

Autonomous 3D Characters

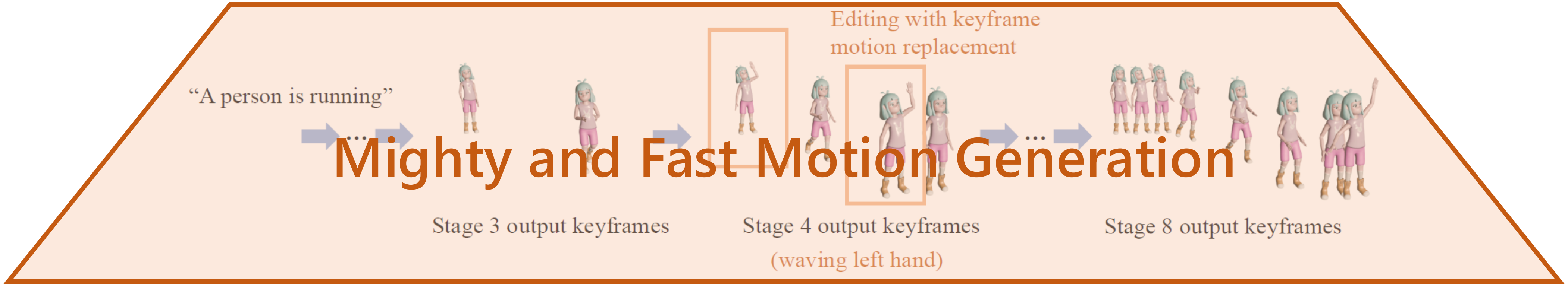
Source: The movie <Free Guy> https://www.youtube.com/watch?v=X2m-08cOAbc


Editing with keyframe motion replacement
"A person is running"
Stage 3 output keyframes
Stage 4 output keyframes
(waving left hand)
Stage 8 output keyframes

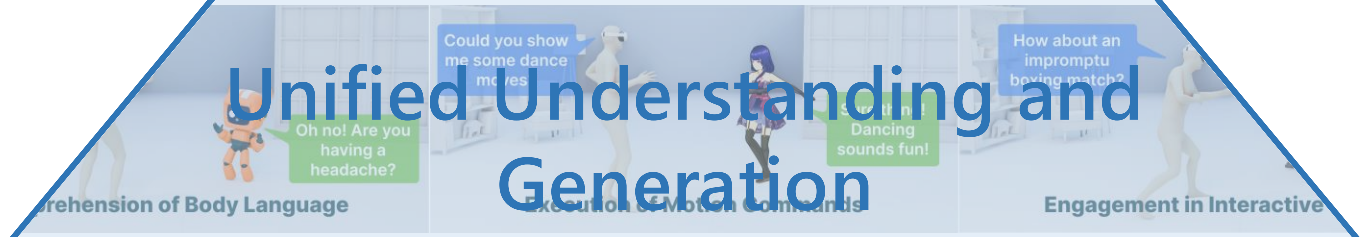
Mighty and Fast Motion Generation



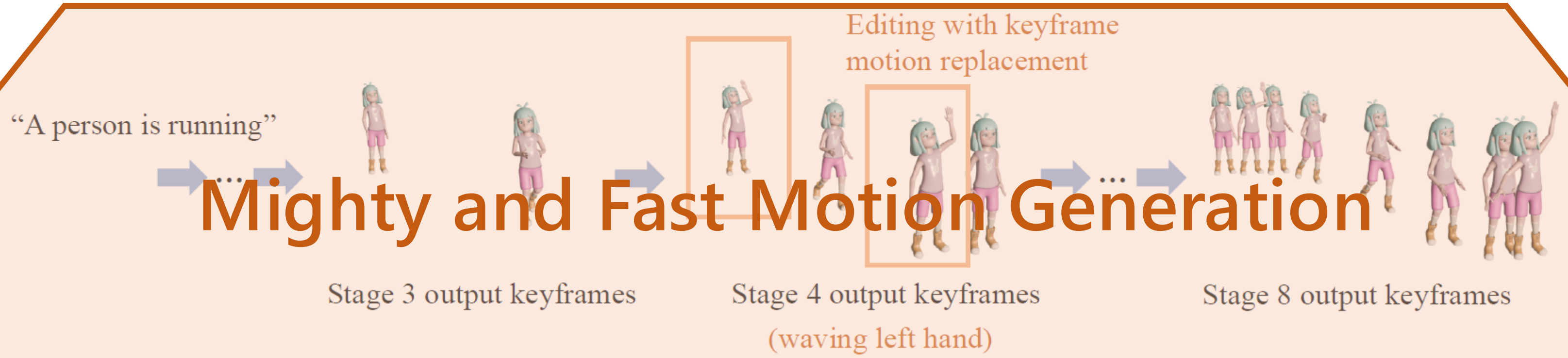
Mighty and Fast Motion Generation

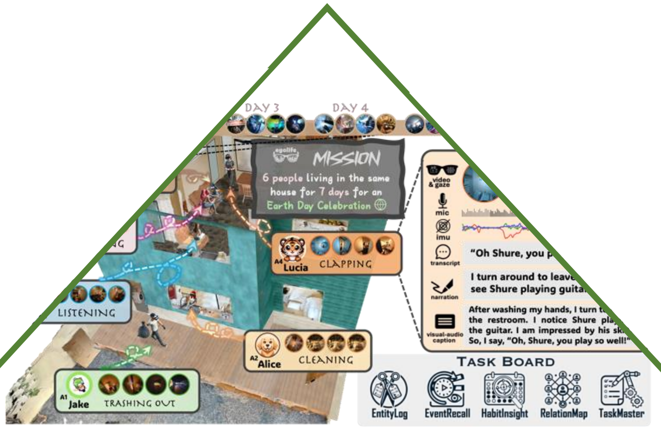


Unified Understanding and Generation

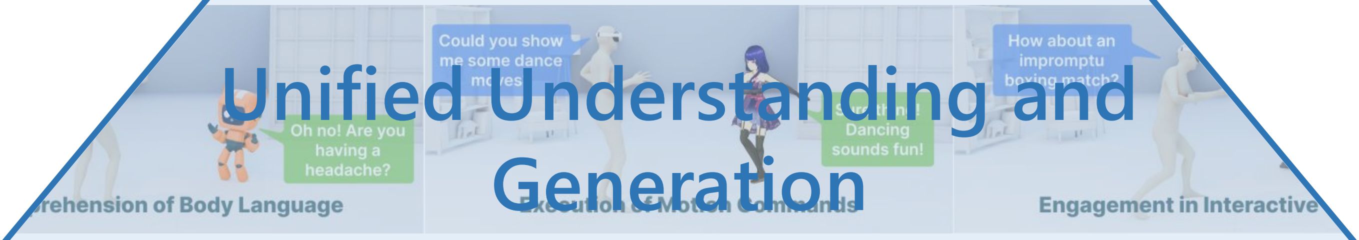


Mighty and Fast Motion Generation

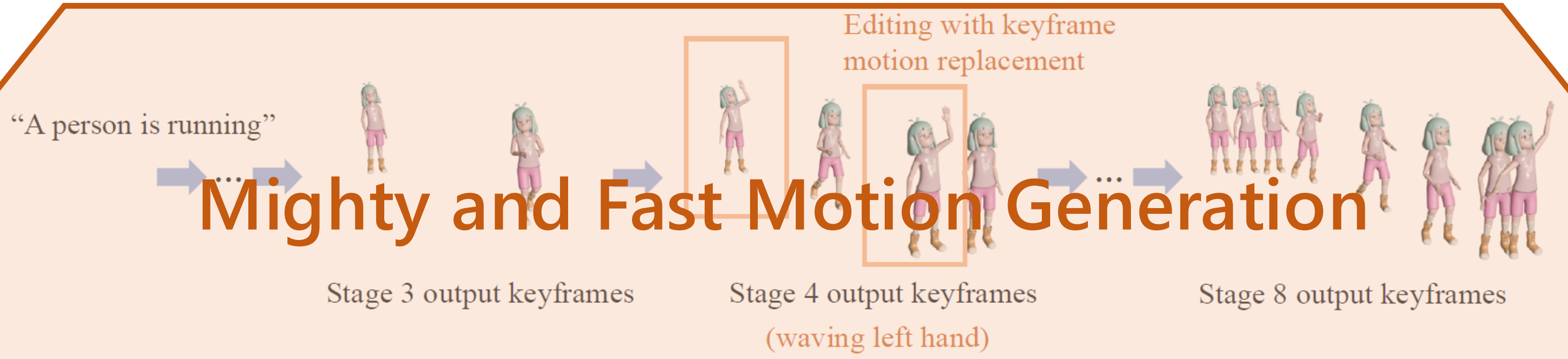




Unified Understanding and Generation



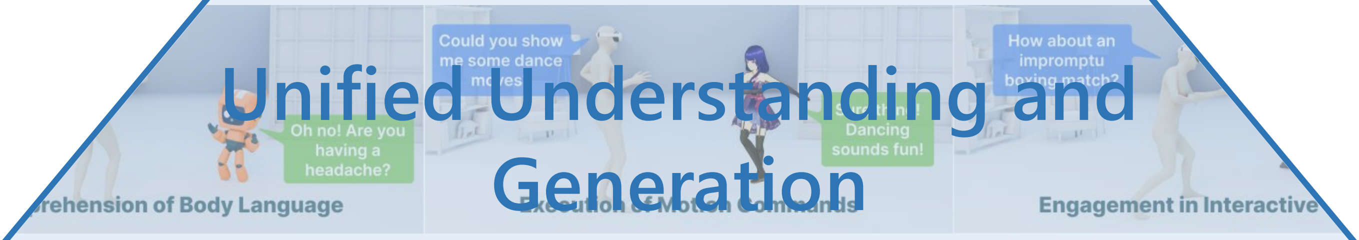
Mighty and Fast Motion Generation



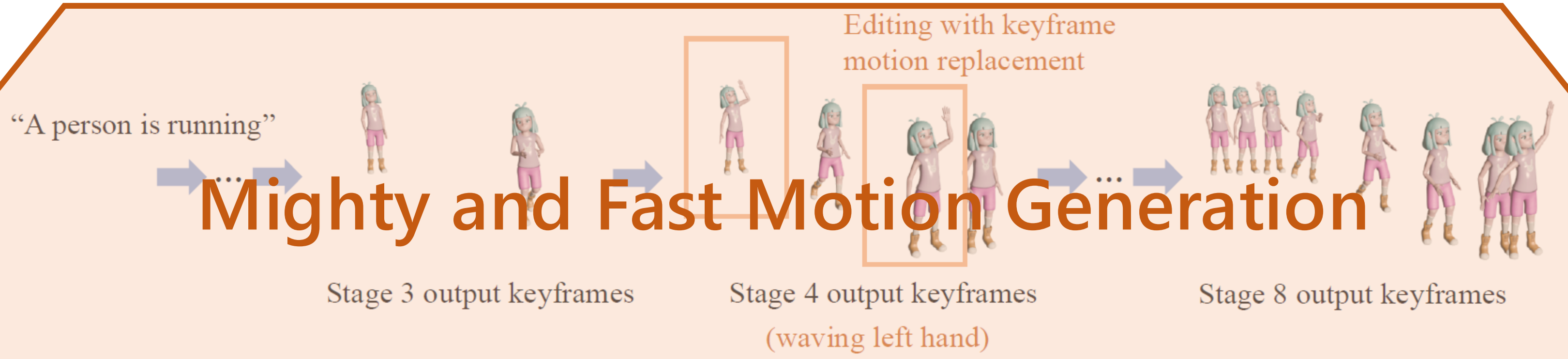
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

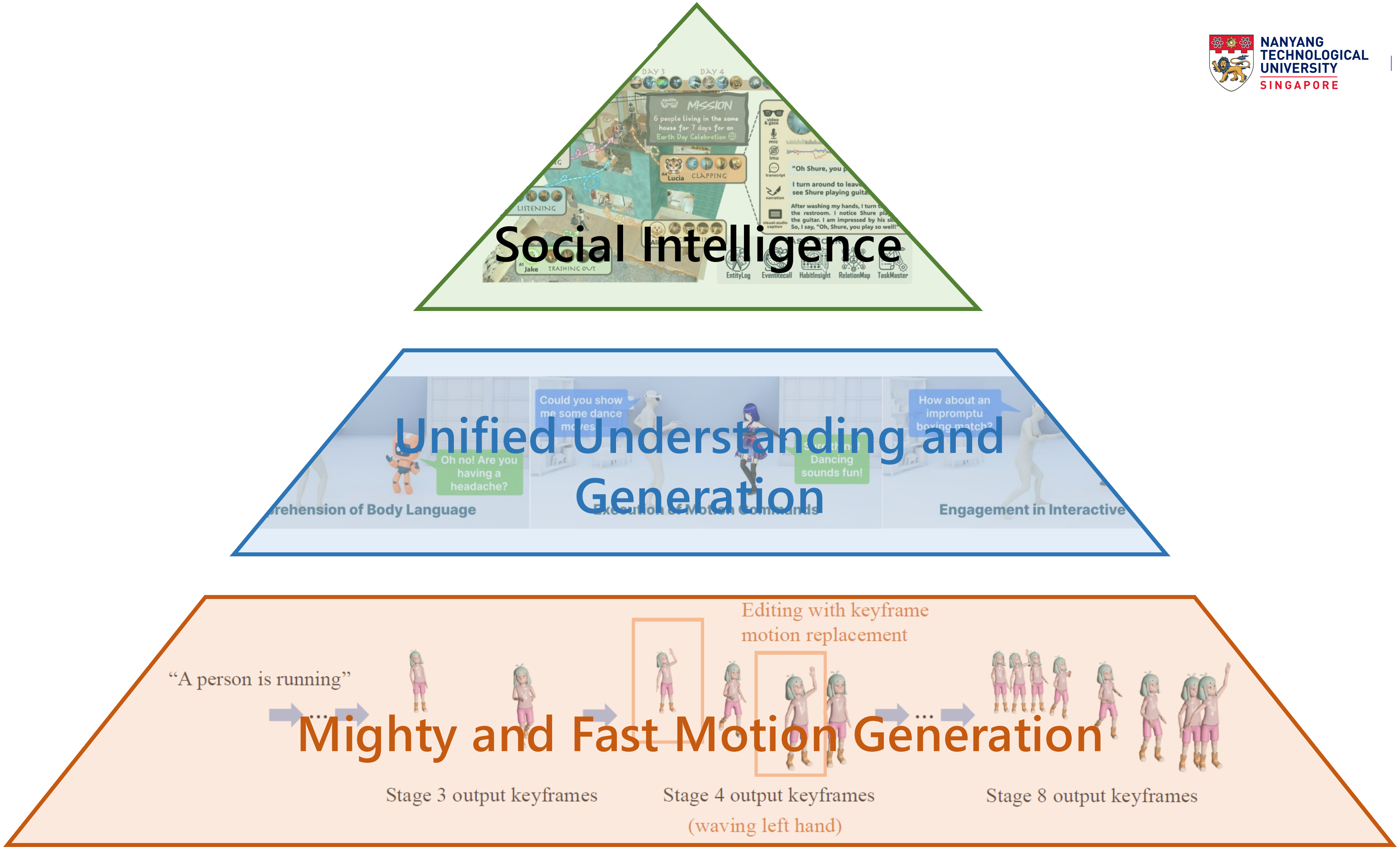
Social Intelligence

Unified Understanding and Generation



Mighty and Fast Motion Generation



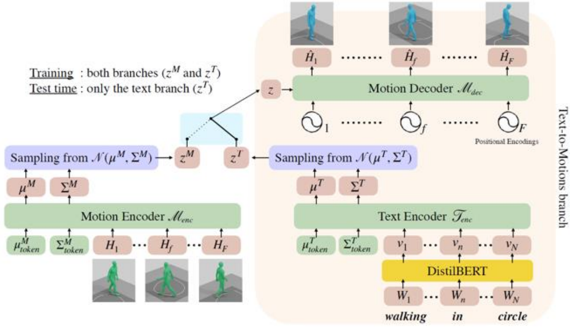

Social Intelligence
DAY 3
DAY 4
MISSION
6 people living in the same house for 7 days for an Earth Day Celebration
LISTENING
Lucia CLAPPING
Jake TRASHING OUT
"Oh Shure, you p
I turn around to leave see Shure playing guita
EntityLog
EventRecall
HabitInsight
RelationMap
TaskMaster
Unified Understanding and Generation
Could you show me some dance moves
Oh no! Are you having a headache?
Dancing sounds fun!
How about an impromptu boxing match?
Engagement in Interactive
Mighty and Fast Motion Generation
"A person is running"
Editing with keyframe motion replacement
Stage 3 output keyframes
Stage 4 output keyframes
(waving left hand)
Stage 8 output keyframes

Mighty and Fast Motion Generation - FracMoGen

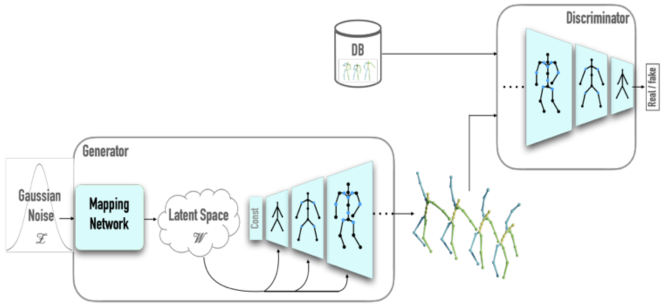
Fractal Human Motion Generative Model

Mingyuan Zhang, Chenyang Gu, Haozhe xie, Zhongang Cai, Ziwei Liu

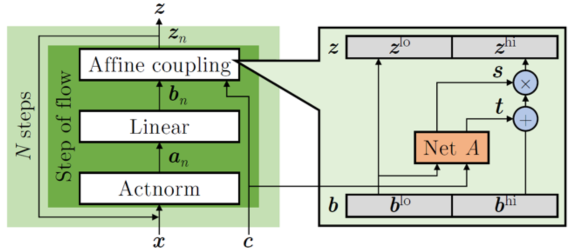
Existing Motion Generative Model



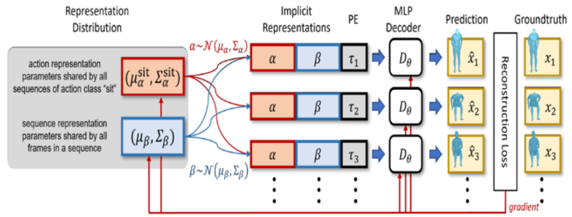
TEMOS[1]: VAE



MoDi[2]: GAN

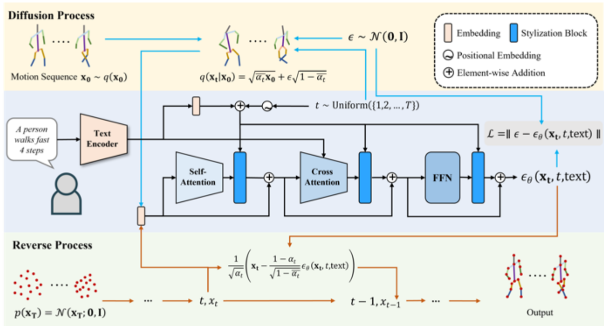


MoGlow[3]: Normalization Flow

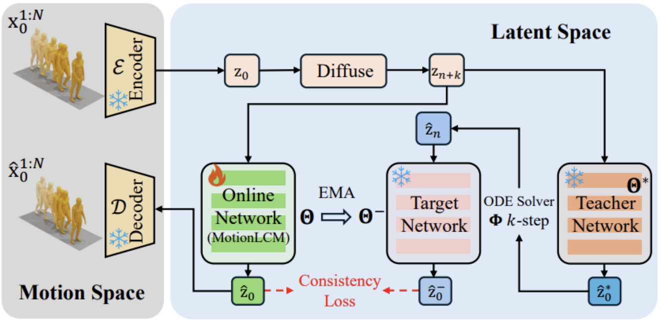


INR[4]: Implicit Function

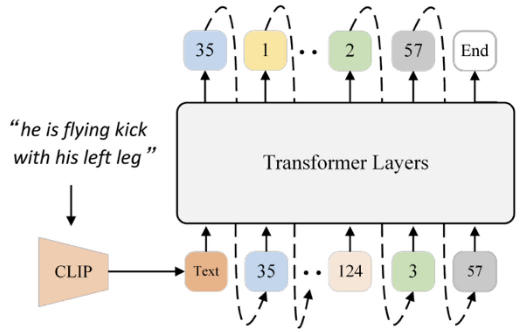
2023



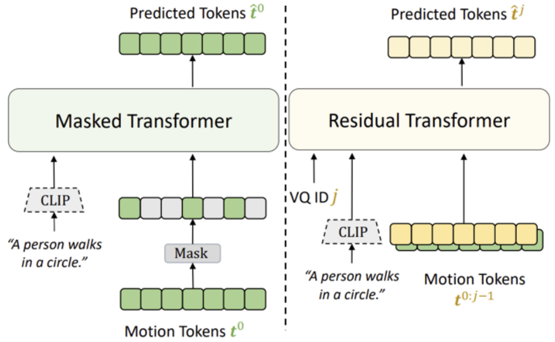
MotionDiffuse[5]: Diffusion Model



MotionLCM[6]: Latent Diffusion Model



T2M-GPT[7]: Auto-Regressive Model



MoMask[8]: Masked Decoder

[1] Petrovich M, *et al.* Temos: Generating diverse human motions from textual descriptions. ECCV 2022
[2] Sigal R, *et al.* MoDi: Unconditional Motion Synthesis from Diverse Data. CVPR 2023
[3] Henter GE, *et al.* Moglow: Probabilistic and controllable motion synthesis using normalising flows. TOG 2020
[4] Cervantes P, *et al.* Implicit neural representations for variable length human motion generation. ECCV 2022
[5] Zhang M, *et al.* MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. TPAMI 2024
[6] Dai W, *et al.* MotionLCM: Real-Time Controllable Motion Generation via Latent Consistency Model. ECCV 2024
[7] Zhang J, *et al.* T2M-GPT: Generating Human Motion from Textual Description with Discrete Representation CVPR 2023
[8] Guo C, *et al.* MoMask: Generative Masked Modeling of 3D Human Motions. CVPR 2024

Fractal Human Motion Generative Model

Raw Continuous Space

- Motion editing
- Long motion generation
- Body-parts modeling
- Low-level control

- Inference efficiency

Latent Continuous Space

- Motion editing
- Low-level control

- Long motion generation
- Body-parts modeling

Discrete Token

- Motion editing
- Long motion generation
- Inference efficiency

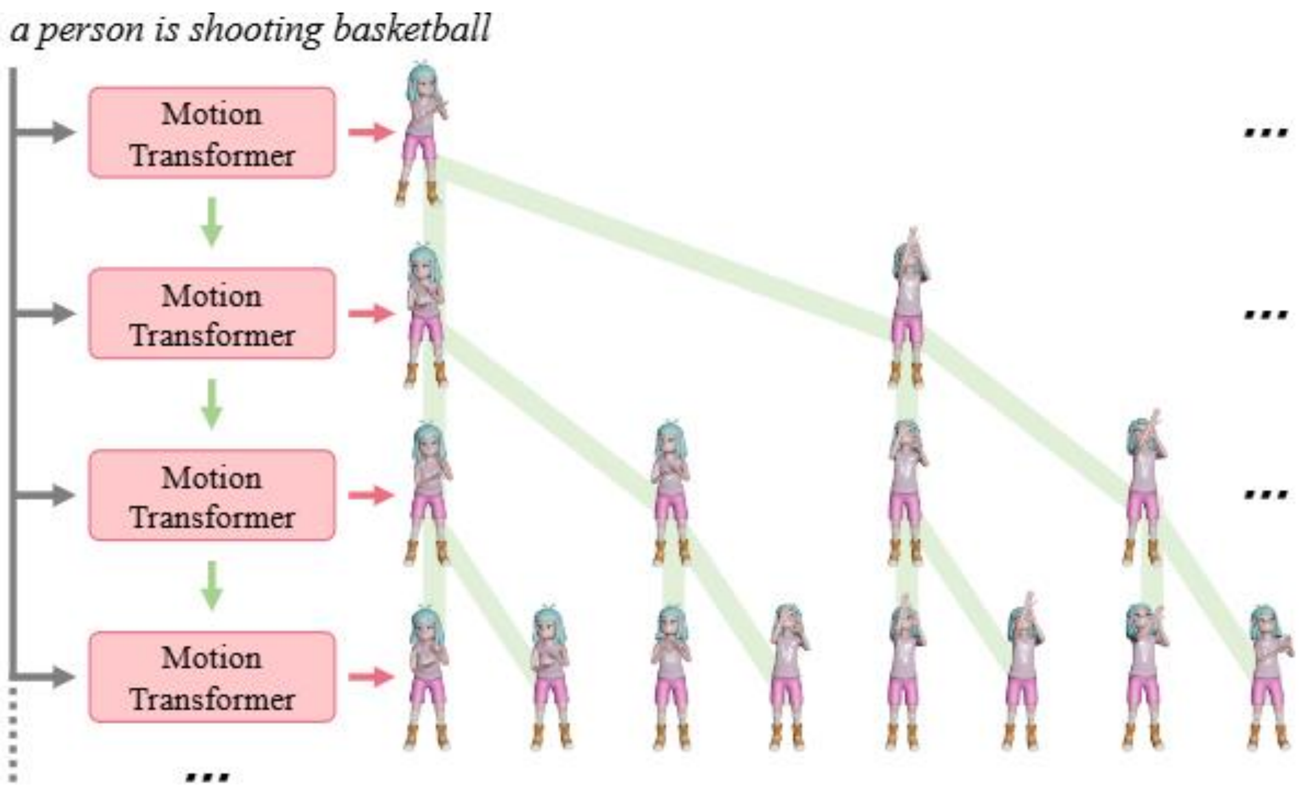
- Body-parts modeling
- Low-level control

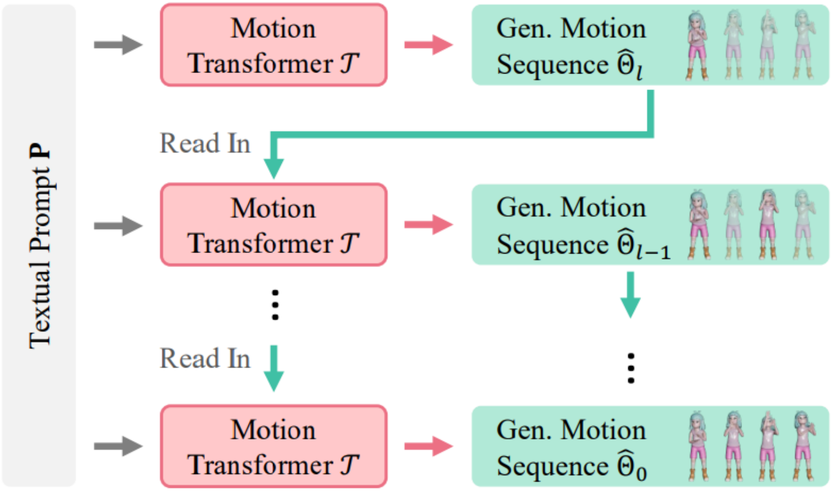
Model	Generation Process	#Stages	Continuity	Compressed	Controllability
Diffusion	$q(\mathbf{x}_{t-1}\|\mathbf{x}_t), t \in [1, T]$	1000/50	Continuous	No	High
Latent Diffusion	$q(\widehat{\mathbf{x}}_{t-1}\|\widehat{\mathbf{x}}_t), t \in [1, T]$	1000/50/4	Continuous	Yes	Medium
Auto-Regressive	$q(\widehat{\theta}_i\|\widehat{\theta}_{i-1}, \ldots, \widehat{\theta}_1), i \in [1, F/r]$	F/r	Discrete	Yes	Low
Masked Decoders	$q(\mathbf{x}_T\|\mathbf{x}_S), S \subset T, T \in \{0,1\}^{F/r}$	$[1, F/r]$	Discrete	Yes	Low
FracMoGen	$q(\widehat{\Theta}_j\|\widehat{\Theta}_i), i \geq j, i \in [0, \lceil \log_2 F \rceil]$	$[1, \infty]$	Continuous	No	High

An efficient motion generative model on raw continuous space!

Build a Generative Model from Scratch

Step 1: Establish your generation process (Fractal Modeling)



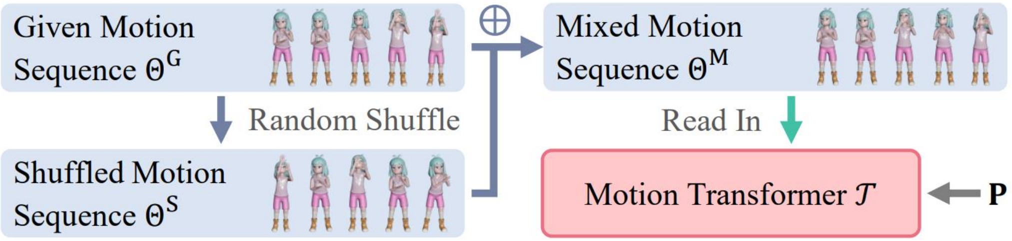


- $q(\widehat{\Theta}_j|\widehat{\Theta}_i), i \geq j, i \in [0, \lceil \log_2 F \rceil]$
- Raw continuous space (highest controllability)
- Flexible inference strategy (via different chains of $\widehat{\Theta}_i$)

Build a Generative Model from Scratch

Step 2: Introduce noise into your training (Intra-group Input Mixing)

Q: Why we need noise during training?
A: Bridge gap between training and inference.



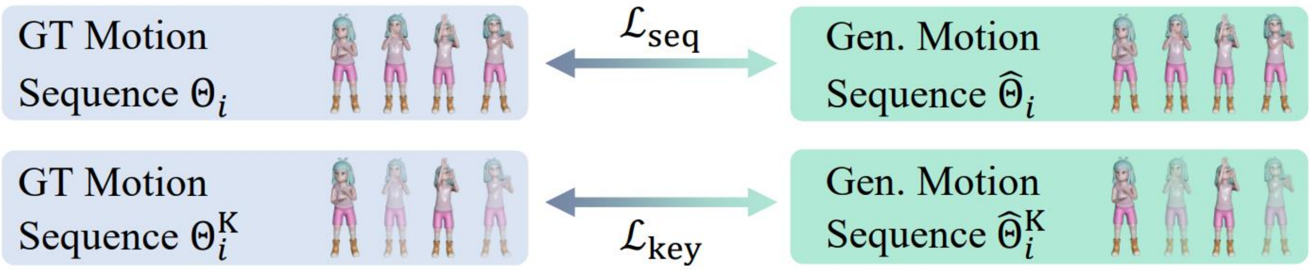
- Data distribution of raw motion representation is far from Gaussian distribution *(why not diffusion noise)*.
- Introduce Noise via frame mixing can better capture the data distribution *(why input mixing)*.

Table 5. **Comparison of different noise schedule on the KIT-ML test set.** The term #Levels refers to the number of different noise levels. In the Diffusion Model, a common setting is 1000 levels. To ensure a fair comparison, we also evaluate the two methods under the condition of the same noise levels.

Noise Type	#Levels	Top 1↑	FID↓
None	-	0.239	2.592
Diffusion	10	0.210	4.015
Diffusion	100	0.375	0.409
Diffusion	1000	0.362	0.451
Ours	10	0.451	0.181
Ours	100	0.435	0.217

Build a Generative Model from Scratch

Step 3: Set your training objectives



$$\mathcal{L} = \lambda \cdot \mathcal{L}_{key} + (1-\lambda)\mathcal{L}_{seq}$$

$$\mathcal{L}_{key} = \left\| \widehat{\Theta}_j^{\mathrm{K}} - \Theta_j^{\mathrm{K}} \right\|$$ Focus on local modeling

$$\mathcal{L}_{seq} = \left\| \widehat{\Theta}_j - \Theta \right\|$$ Focus on global modeling

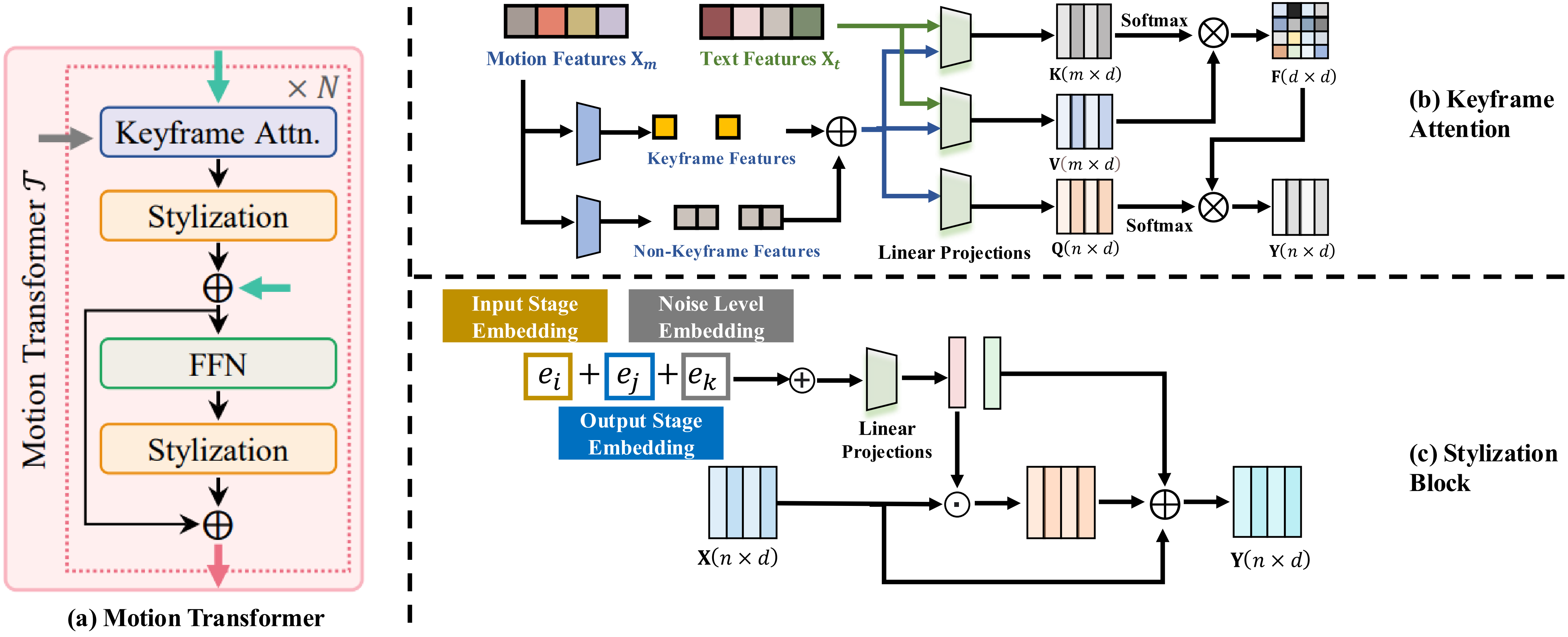
Early stages should focus more on **global** modeling, while **later** stages should focus more on **local** modeling.

Table 6. **Comparison of different configuration of balanced target loss on the KIT-ML test set.** There are two types of experiments here. One uses a constant, meaning the same λ coefficient is applied to all stages. The other uses a linearly increasing λ value. For example, in the case of $0.2 \rightarrow 0.8$, the λ coefficient for stage 6 is 0.2, for stage 5 it is 0.3, and so on, with the λ coefficient for stage 0 being 0.8.

λ	Top 1↑	FID↓
0	0.421	0.246
0.2	0.431	0.217
0.4	0.442	0.195
0.6	0.419	0.250
0.8	0.384	0.319
1.0	0.326	0.584
$0 \rightarrow 0.6$	0.428	0.230
$0.2 \rightarrow 0.8$	0.451	0.181
$0.4 \rightarrow 1.0$	0.405	0.317

Build a Generative Model from Scratch

Step 4: Design your backbone

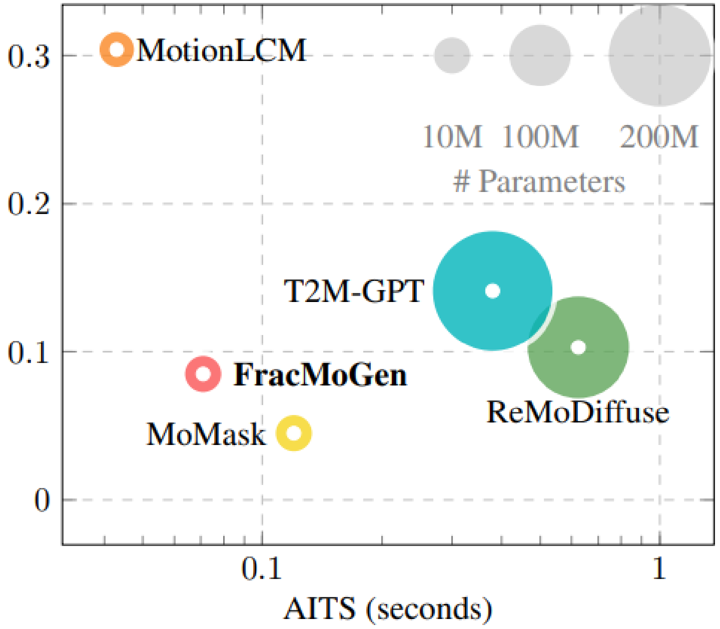


Build a Generative Model from Scratch

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

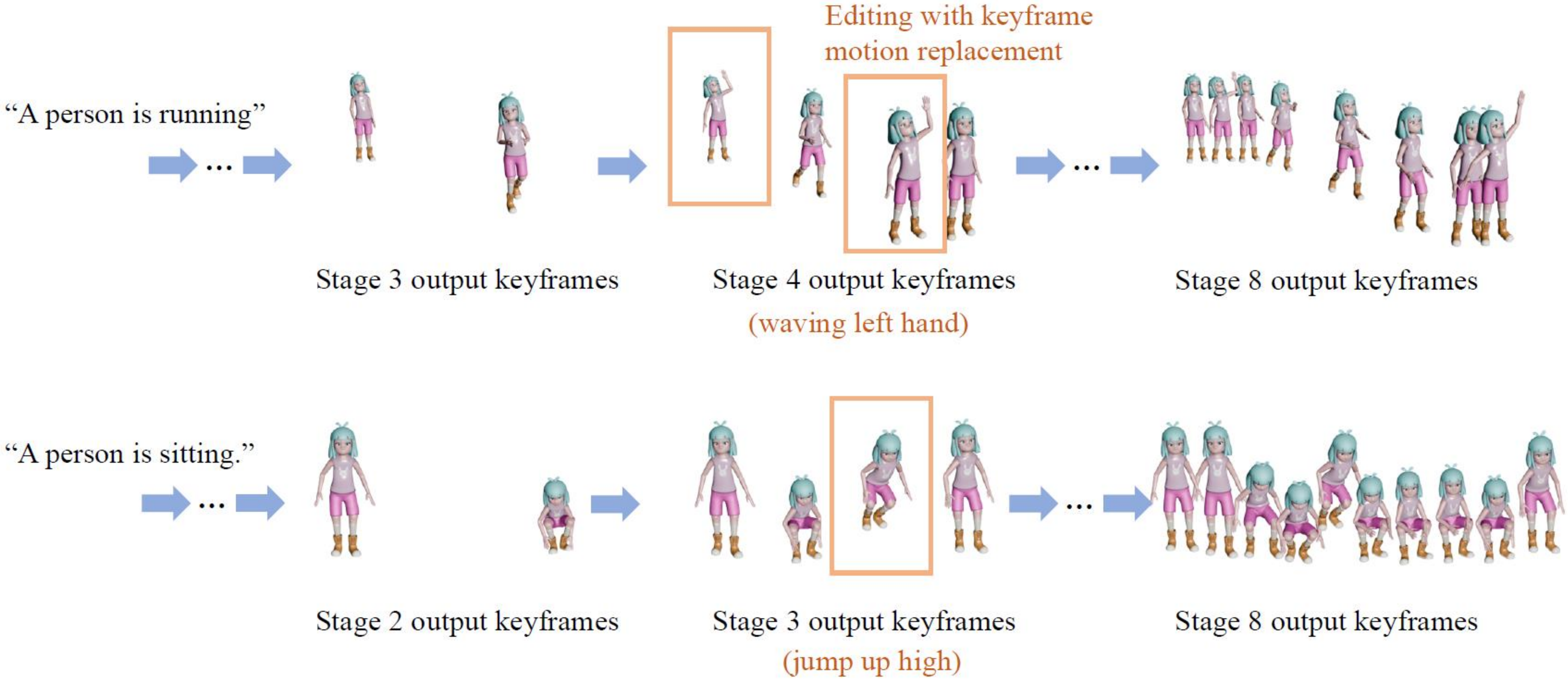
Step 5: Enjoy your results

	Method	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity →	MModality ↑
		Top-1 ↑	Top-2 ↑	Top-3 ↑				
HumanML3D	Real Motion	0.511$^{\pm.003}$	0.703$^{\pm.003}$	0.797$^{\pm.002}$	0.002$^{\pm.000}$	2.974$^{\pm.008}$	9.503$^{\pm.065}$	-
	MotionDiffuse [36]	0.491$^{\pm.001}$	0.681$^{\pm.001}$	0.782$^{\pm.001}$	0.630$^{\pm.001}$	3.113$^{\pm.001}$	**9.410**$^{\pm.049}$	1.553$^{\pm.042}$
	ReMoDiffuse [34]	0.510$^{\pm.005}$	0.698$^{\pm.006}$	0.795$^{\pm.004}$	0.103$^{\pm.004}$	2.974$^{\pm.016}$	9.018$^{\pm.075}$	1.795$^{\pm.043}$
	T2M-GPT [33]	0.492$^{\pm.003}$	0.679$^{\pm.002}$	0.775$^{\pm.002}$	0.141$^{\pm.005}$	3.121$^{\pm.009}$	9.722$^{\pm.082}$	**1.831**$^{\pm.048}$
	MoMask [11]	**0.521**$^{\pm.002}$	**0.713**$^{\pm.002}$	**0.807**$^{\pm.002}$	**0.045**$^{\pm.002}$	**2.958**$^{\pm.008}$	-	1.241$^{\pm.040}$
	MotionLCM [6]	0.502$^{\pm.003}$	0.698$^{\pm.002}$	0.798$^{\pm.002}$	0.304$^{\pm.012}$	3.012$^{\pm.007}$	**9.607**$^{\pm.066}$	**2.259**$^{\pm.092}$
	FracMoGen (Ours)	**0.515**$^{\pm.003}$	**0.703**$^{\pm.005}$	**0.802**$^{\pm.003}$	**0.085**$^{\pm.008}$	**2.946**$^{\pm.013}$	9.632$^{\pm.031}$	1.417$^{\pm.056}$
KIT-ML	Real Motion	0.424$^{\pm.005}$	0.649$^{\pm.006}$	0.779$^{\pm.006}$	0.031$^{\pm.004}$	2.788$^{\pm.012}$	11.08$^{\pm.097}$	-
	MotionDiffuse [36]	0.417$^{\pm.004}$	0.621$^{\pm.004}$	0.739$^{\pm.004}$	1.954$^{\pm.062}$	2.958$^{\pm.005}$	**11.10**$^{\pm.143}$	0.730$^{\pm.013}$
	ReMoDiffuse [34]	0.427$^{\pm.014}$	0.641$^{\pm.004}$	0.765$^{\pm.055}$	**0.155**$^{\pm.006}$	2.814$^{\pm.012}$	10.80$^{\pm.105}$	**1.239**$^{\pm.028}$
	T2M-GPT [33]	0.416$^{\pm.006}$	0.627$^{\pm.006}$	0.745$^{\pm.006}$	0.514$^{\pm.029}$	3.007$^{\pm.023}$	10.92$^{\pm.108}$	**1.570**$^{\pm.039}$
	MoMask [11]	**0.433**$^{\pm.007}$	**0.656**$^{\pm.005}$	**0.781**$^{\pm.005}$	0.204$^{\pm.011}$	**2.779**$^{\pm.022}$	-	1.131$^{\pm.043}$
	FracMoGen (Ours)	**0.451**$^{\pm.009}$	**0.688**$^{\pm.008}$	**0.810**$^{\pm.009}$	**0.181**$^{\pm.010}$	**2.668**$^{\pm.019}$	**11.01**$^{\pm.115}$	1.047$^{\pm.046}$



Build a Generative Model from Scratch

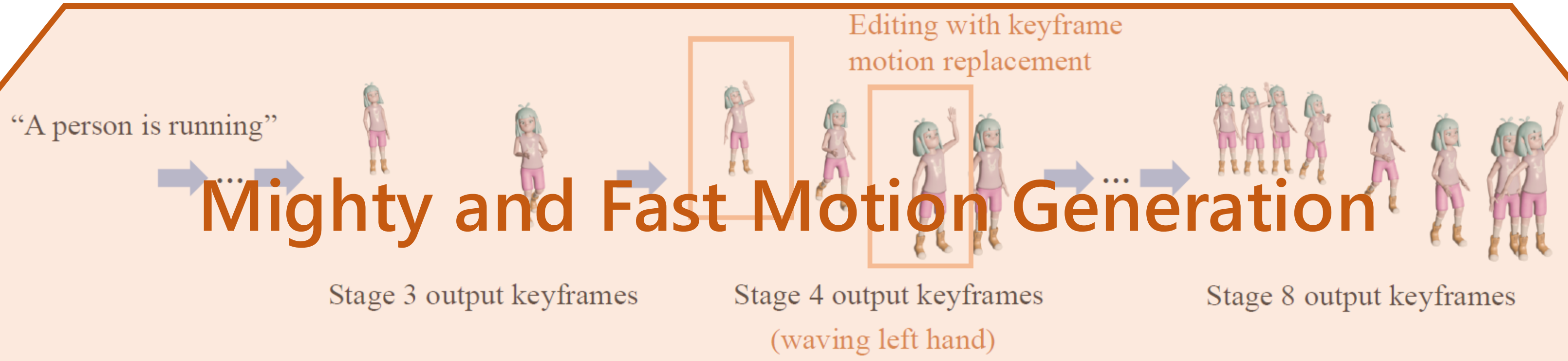
Step 5: Enjoy your results



Social Intelligence

Unified Understanding and Generation

Mighty and Fast Motion Generation



Unified Understanding and Generation - SOLAMI

SOLAMI: Social Vision-Language-Action Modeling for Immersive Interaction with 3D Autonomous Characters

Jianping Jiang, Weiye Xiao, Zhengyu Lin, Huaizhong Zhang, Tianxiang Ren, Yang Gao, Zhiqian Lin, Zhongang Cai, Lei Yang, Ziwei Liu

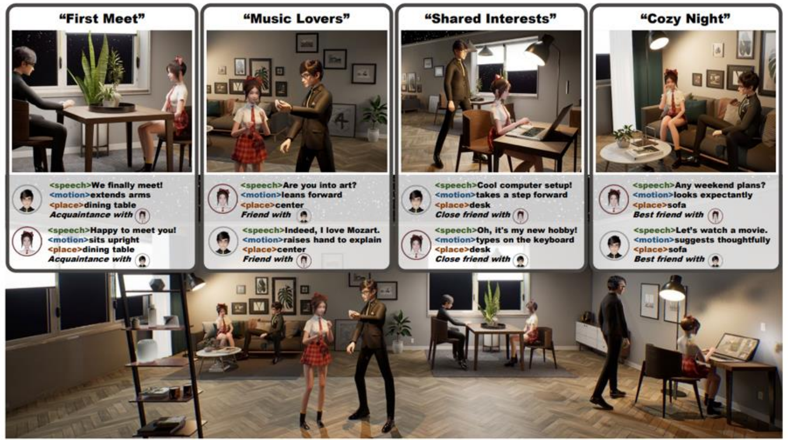
CVPR 2025

3D Characters with Social Intelligence

- Modeling with LLM-Agent Framework

Generative Agents [1]

Digital Life Project [2]

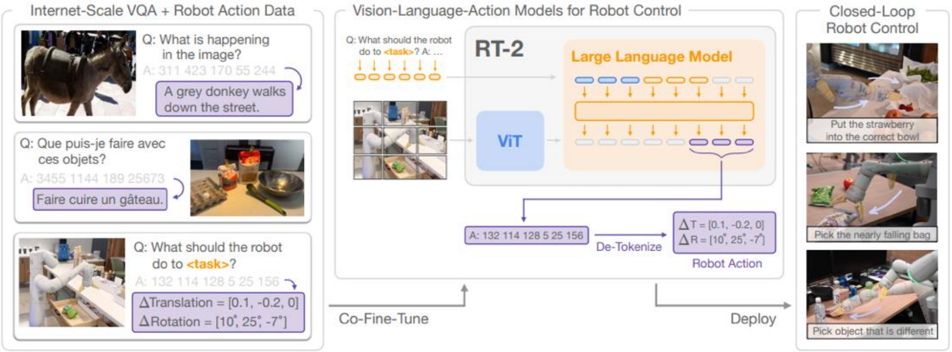


- Limitations
  - Scalable Formulation
  - Multimodal Coherence
  - Latency

[1] Generative Agents: Interactive Simulacra of Human Behavior. UIST 2023.
[2] Digital Life Project: Autonomous 3D Characters with Social Intelligence. CVPR 2024.

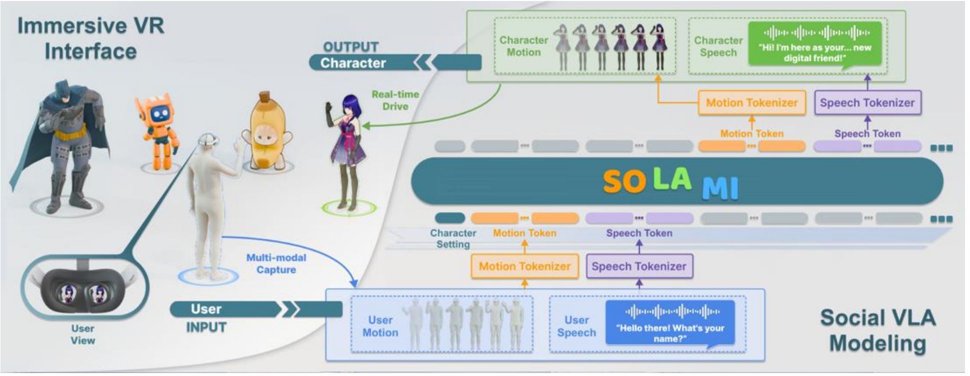
Motivation: Avatar as Virtual Robot

Robot
3D Agent with Real Embodiment
(Real-world Task & Interaction)



RT-2 [1]:Vision-Language-Action Models

3D Avatar
3D Agent with Virtual Embodiment
(Natural Appearance & Behavior)

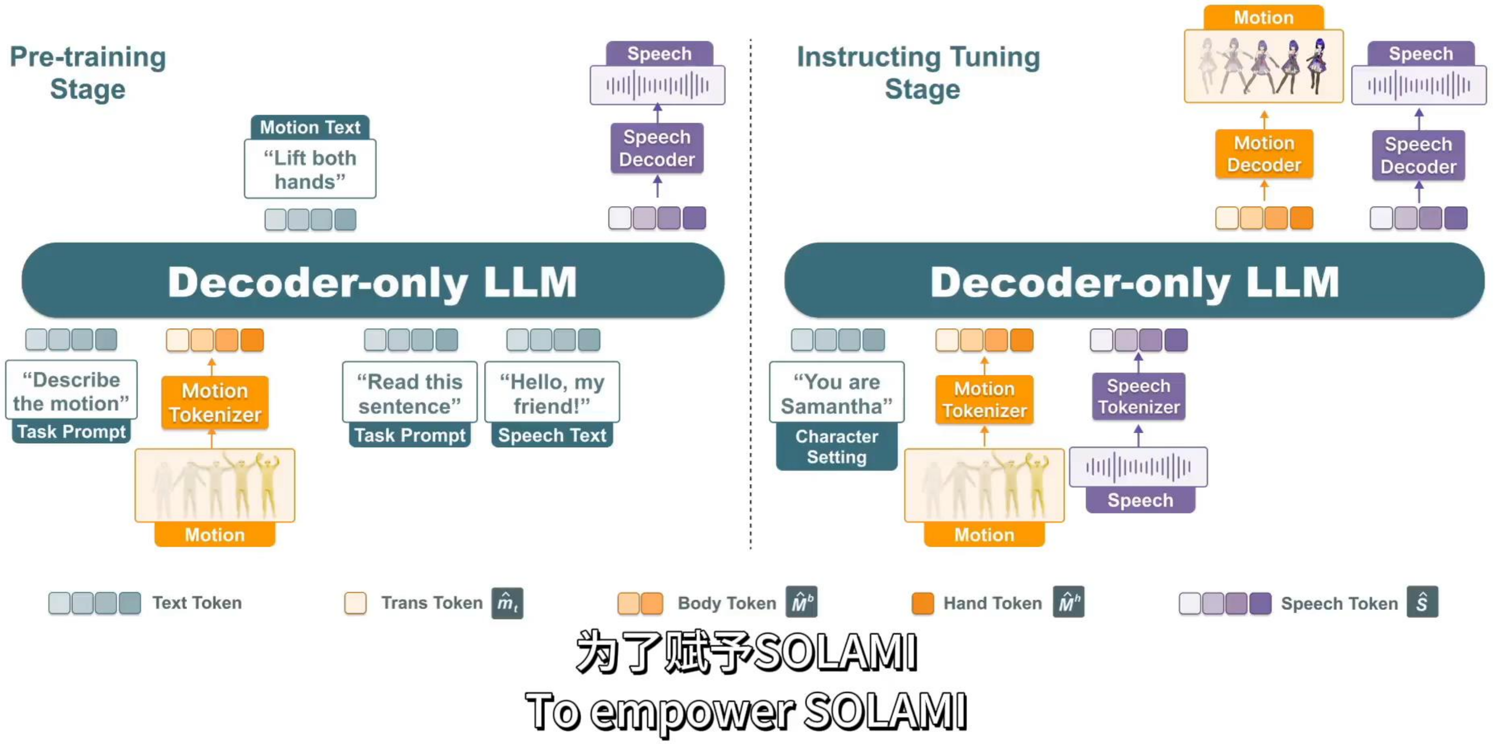


Social VLA for Immersive Interaction with 3D Characters

[1] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. CoRL 2023.

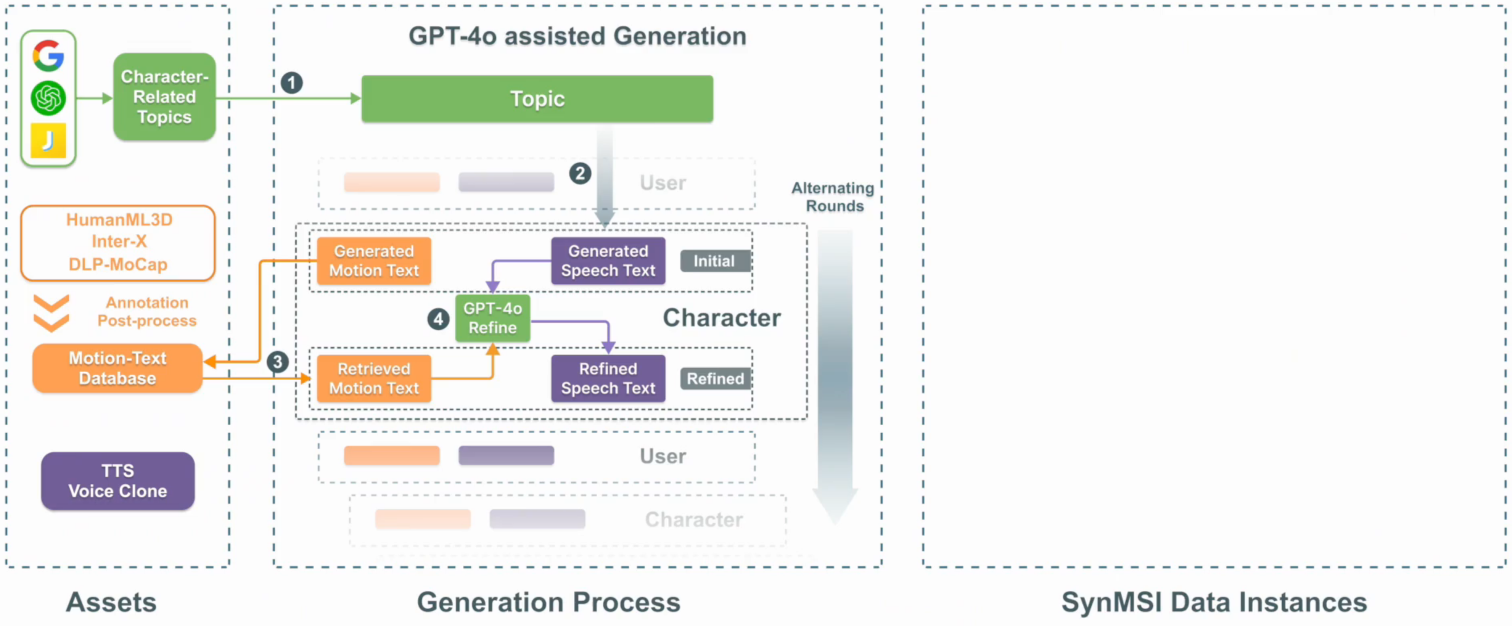
Training Recipe

- Training Stages
  - Stage1: Motion & Speech Tokenizer Training
  - Stage2: Motion-Text-Speech Alignment with Multi-Task Pretraining
  - Stage3: Instruction Tuning for Multimodal Chat



Data Generation

- Multimodal Chat Data Synthesize
  - LLM-Generated Scripts
    - Diverse Topics
    - Refined Process
  - Motion-Text Dataset
    - Large-Scale



Evaluation: Quantitative & Qualitative

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

- Compared to Speech-Only Method
  - Better User Experience
- Compared to LLM-Agent Framework
  - Low Latency & Multimodal Coherence
  - Alignment Tax on Text

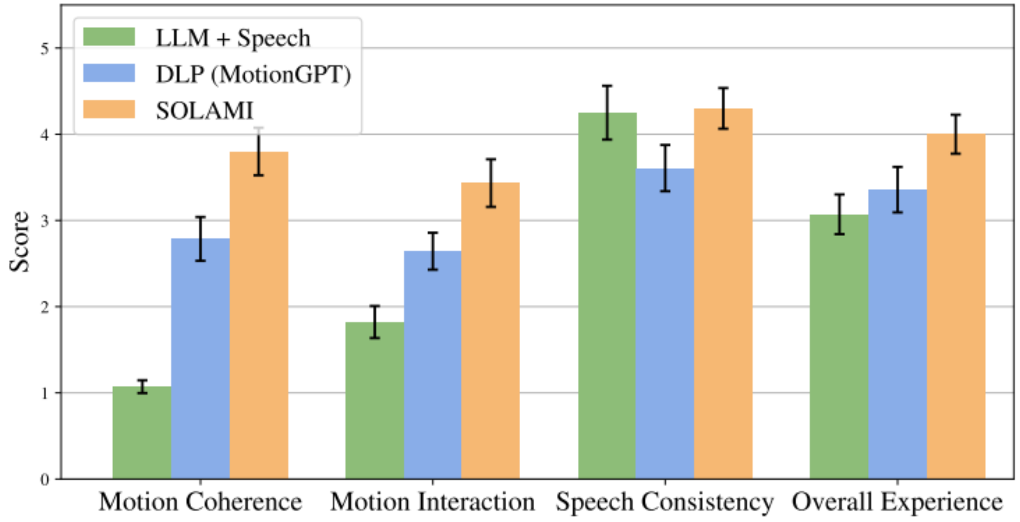
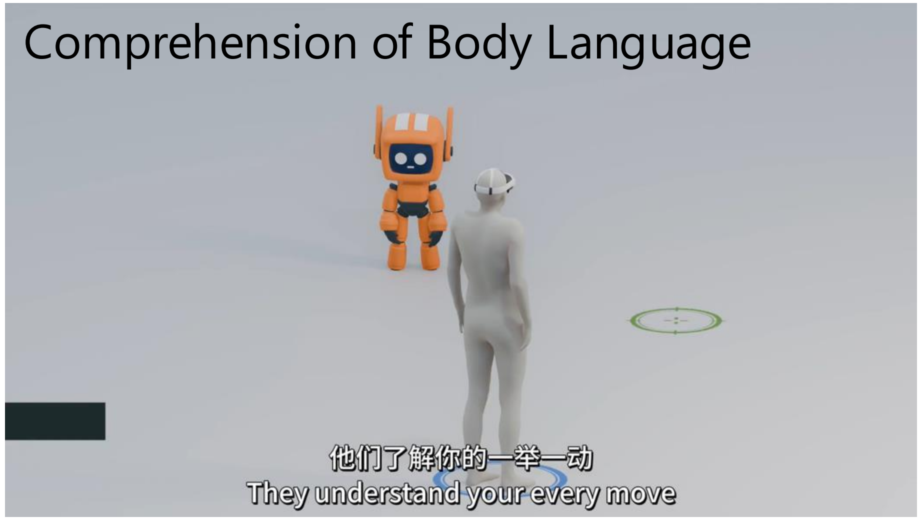
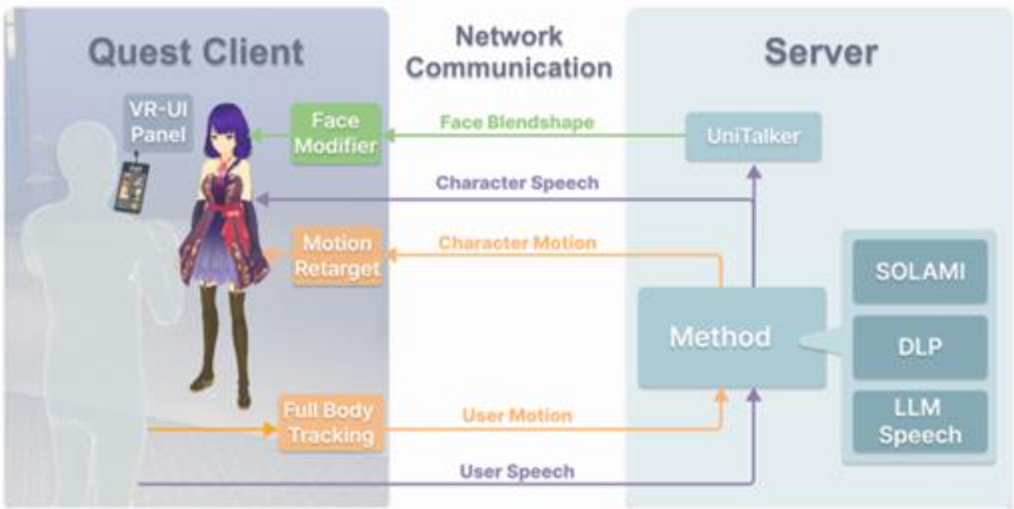


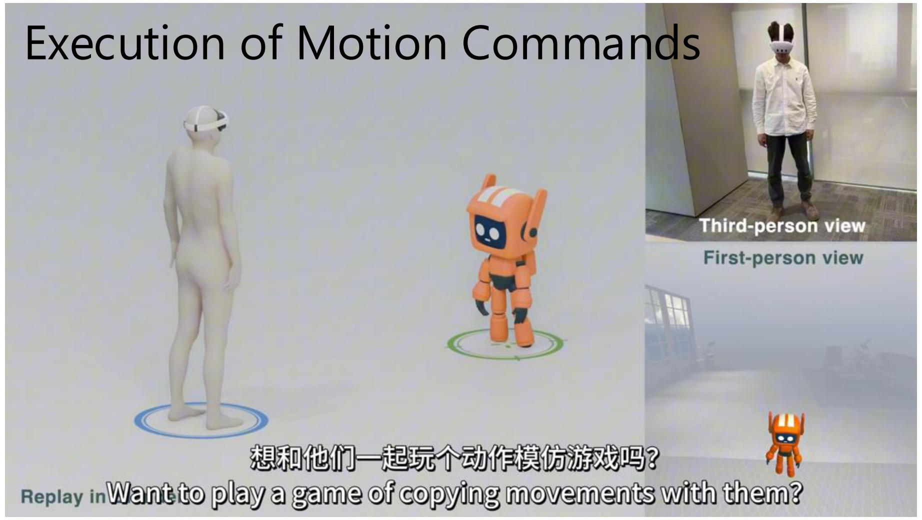
Table 1. **Quantitative results of baselines and SOLAMI**. '↑'('↓') indicates that the values are better if the metrics are larger (smaller). We run all the evaluations 5 times and report the average metric. The best results are in bold and the second best results are underlined.

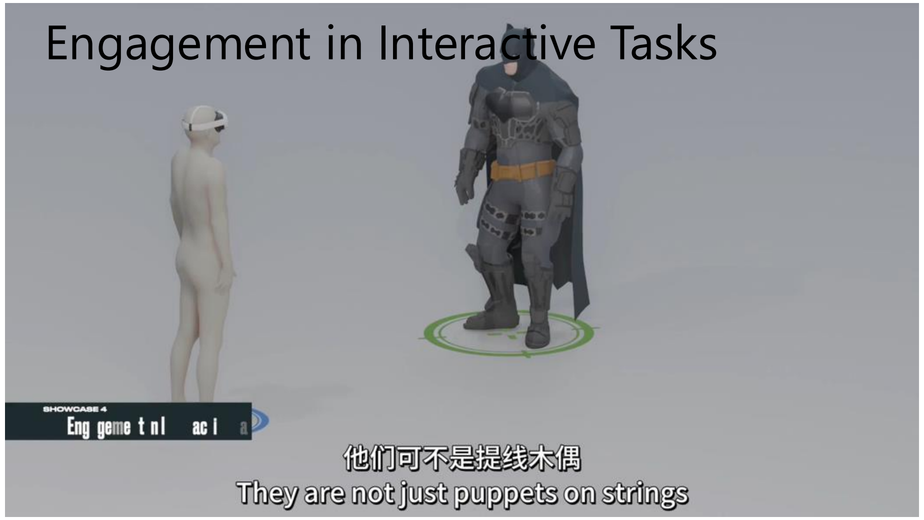
Methods	Motion Metrics				Speech Metrics			Inference Latency ↓
	FID↓	Diversity↑	PA-MPJPE↓	Angle Error↓	VC Similarity↑	Context Relevance↑	Character Consistency↑	
SynMSI Dataset	-	9.136	-	-	-	4.888	4.893	-
LLM+Speech (Llama2) [69]	-	-	-	-	0.818	3.527	**3.859**	3.157
AnyGPT (fine-tune) [81]	-	-	-	-	0.819	3.502	3.803	**2.588**
DLP (MotionGPT) [17]	4.254	8.259	165.053	0.495	0.812	3.577	3.785	5.518
SOLAMI (w/o pretrain)	5.052	8.558	159.709	0.387	0.820	3.541	3.461	2.657
SOLAMI (LoRA)	15.729	8.145	167.149	0.400	0.770	3.251	3.423	2.710
SOLAMI (full params)	**3.443**	**8.853**	**151.500**	**0.360**	**0.824**	**3.634**	3.824	2.639

Demo: VR Interface









Unified Understanding and Generation - EgoLM

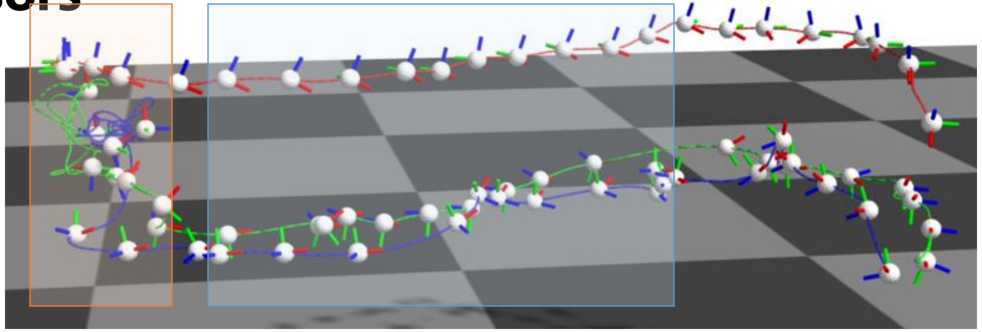
EgoLM: Multi-Modal Language Model of Egocentric Motions

Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, Lingni Ma

CVPR 2025, Oral Presentation

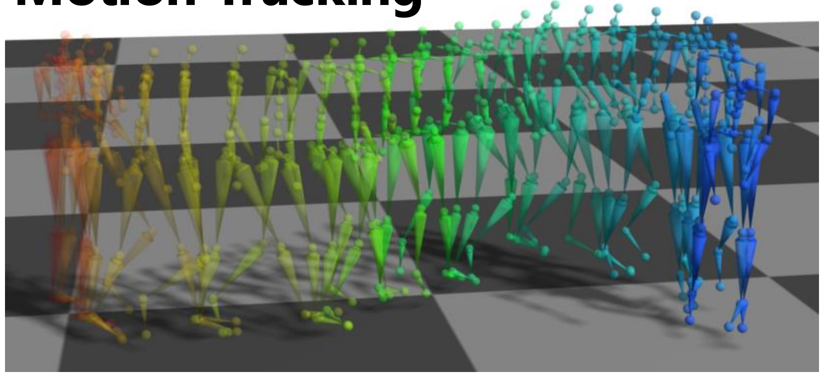
Egocentric Motion Tracking and Understanding

Sparse Motion Sensors



Egocentric Videos

Motion Tracking

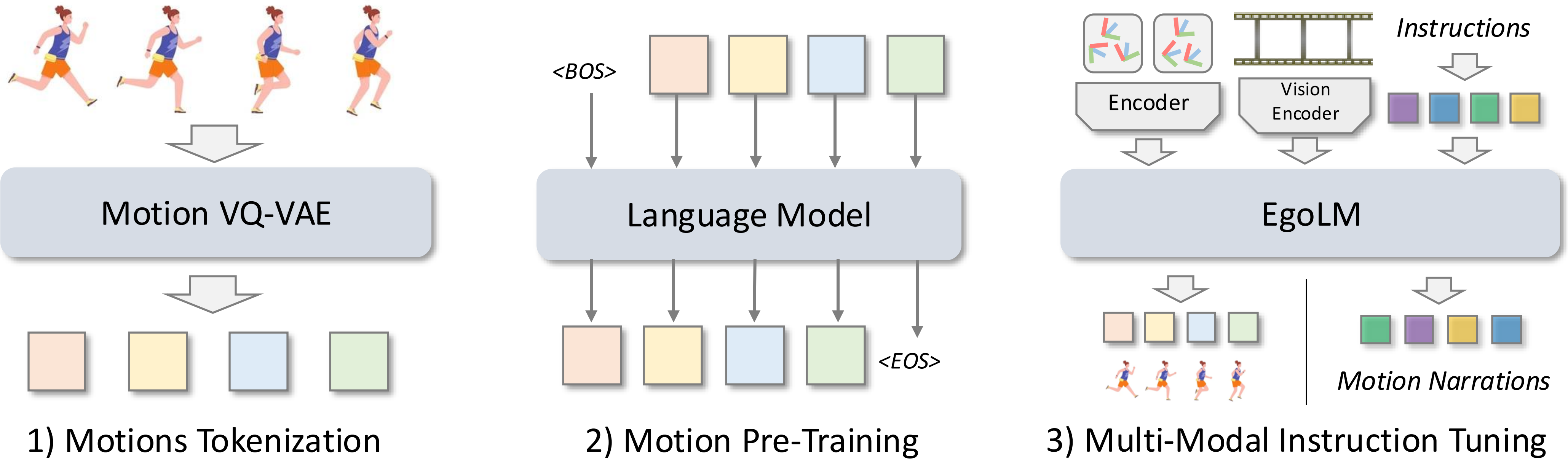


Motion Understanding

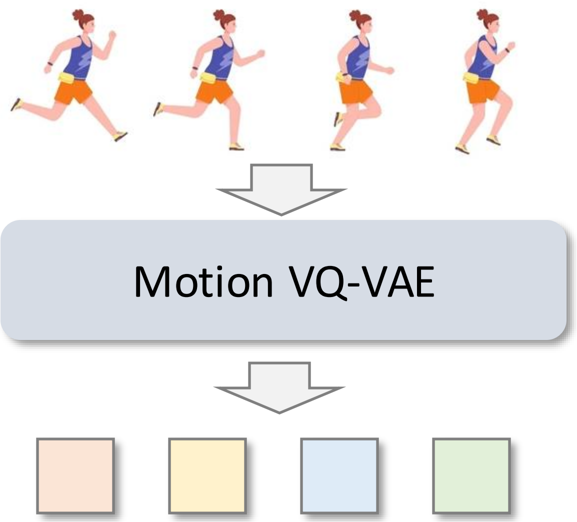
"The person is standing straight as she puts the piece of clothing on the hanger."

"The person turns around then walks out of the bedroom."

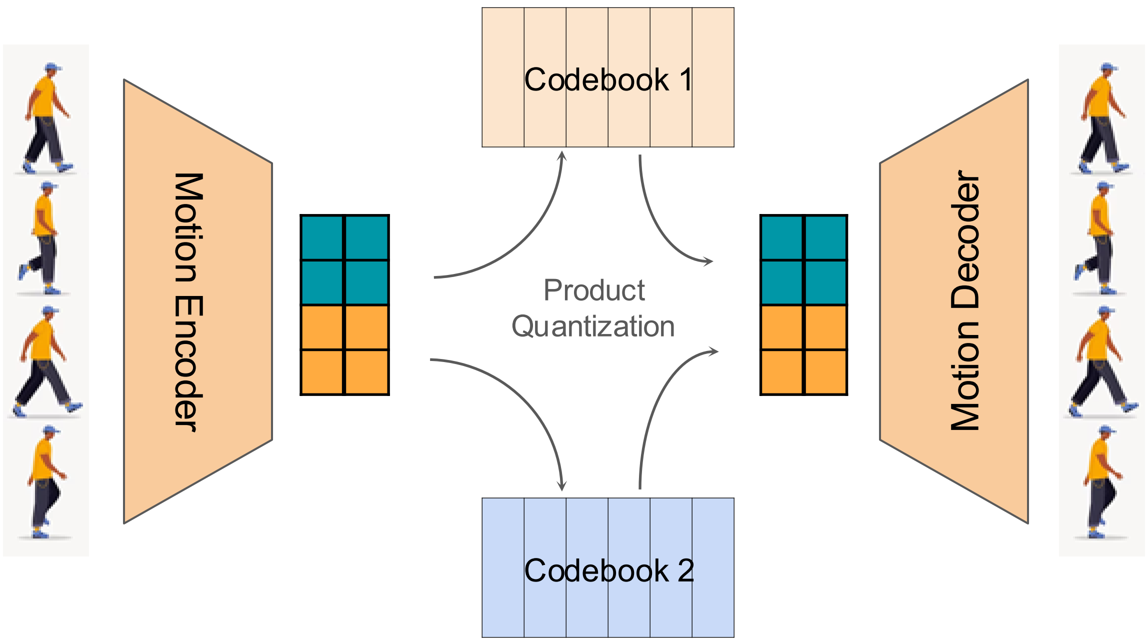
Multi-Modal Multi-Tasking LM for Ego Motion



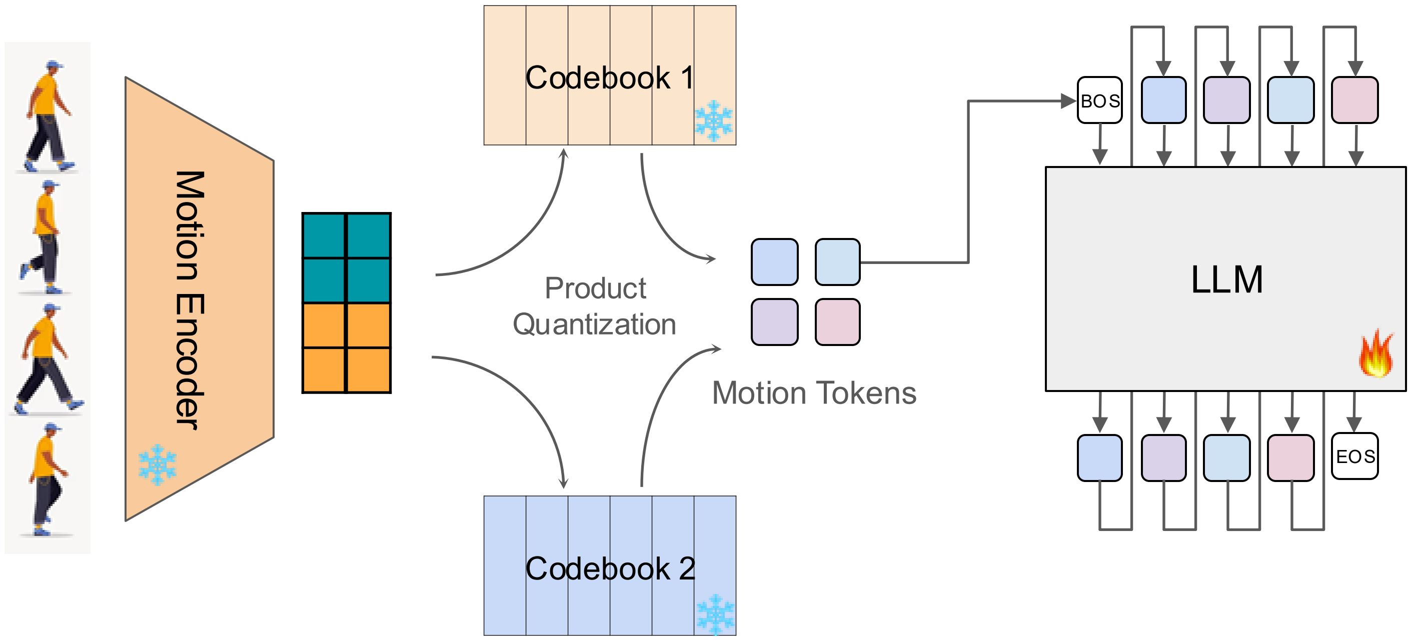
Step 1: Motion VQ-VAE



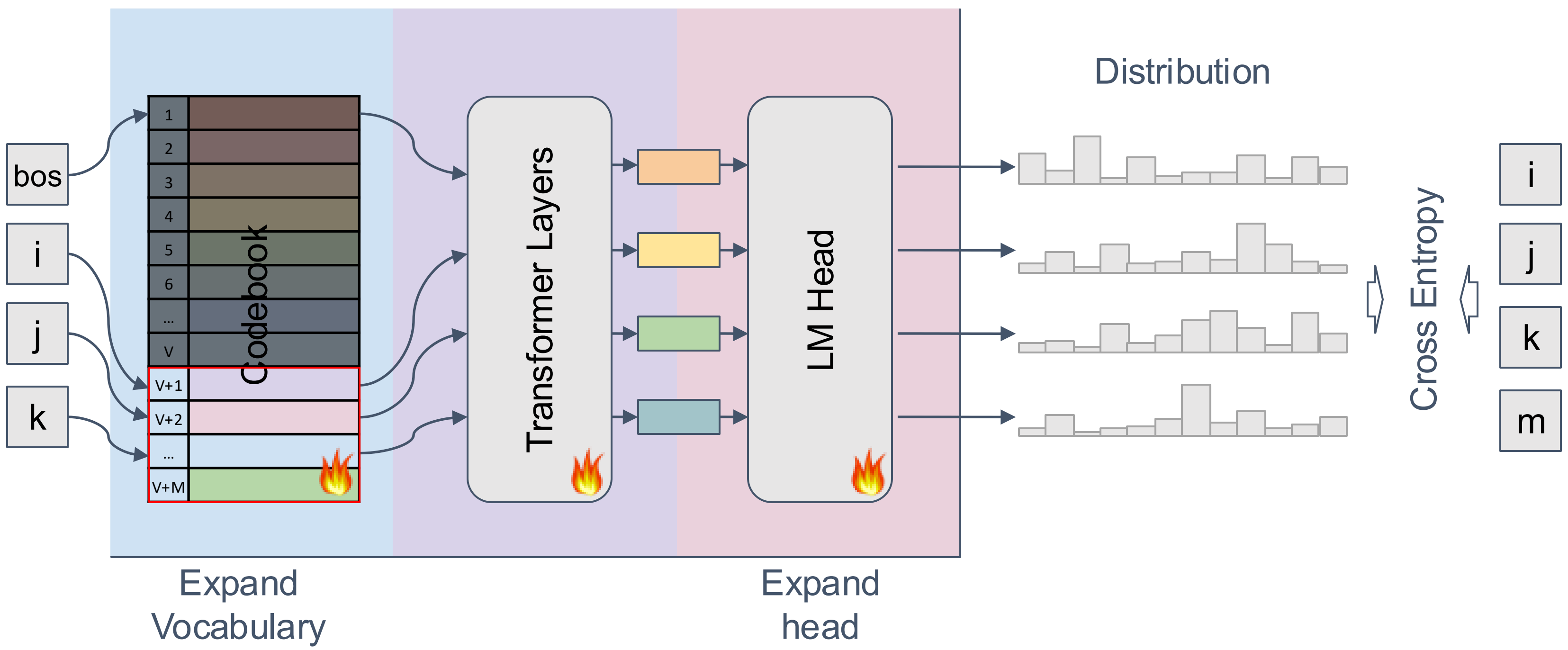
1) Motions Tokenization



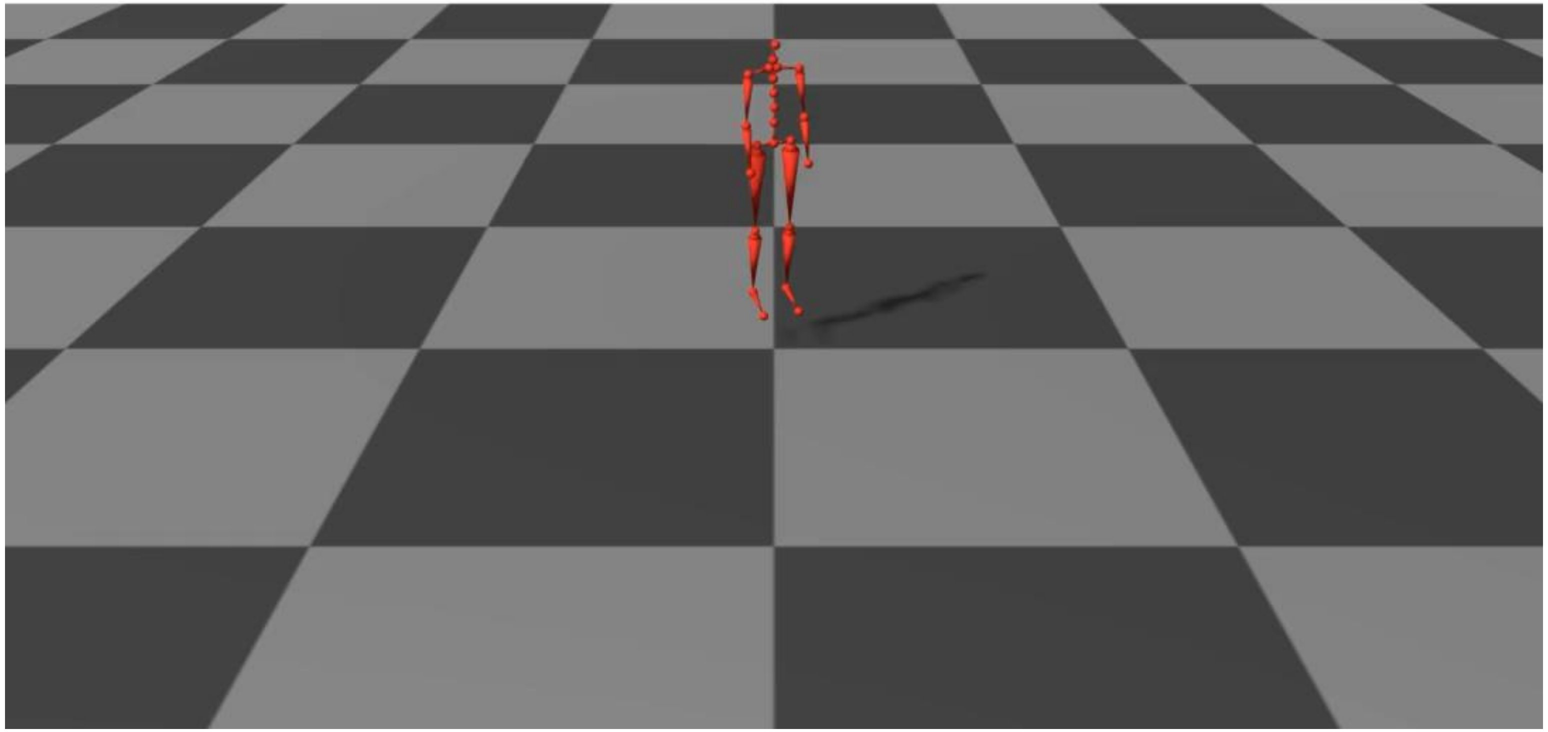
Step 2: Motion Pre-Training



Step 2: Motion Pre-Training

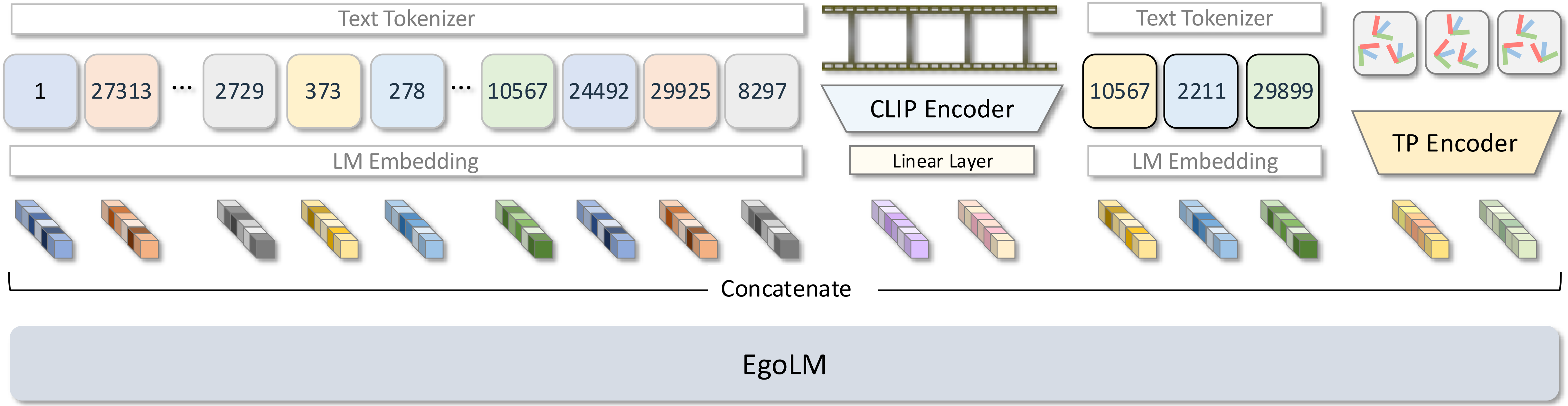


By-Product: Unconditional Motion Generator



Step 3: Instruction Tuning

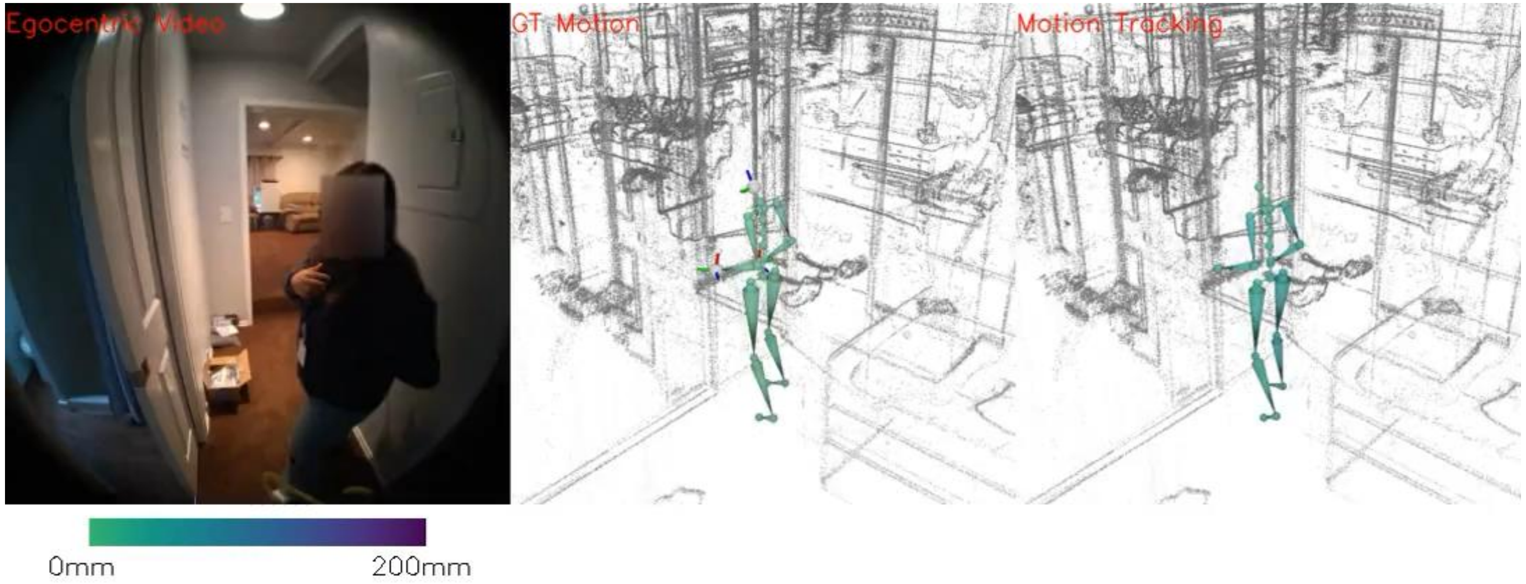
“<s> Perform ... based on the given ... Input CLIP embeddings: <CLIP_Placeholder>*. Input three-points:* <TP_Placeholder>*”*



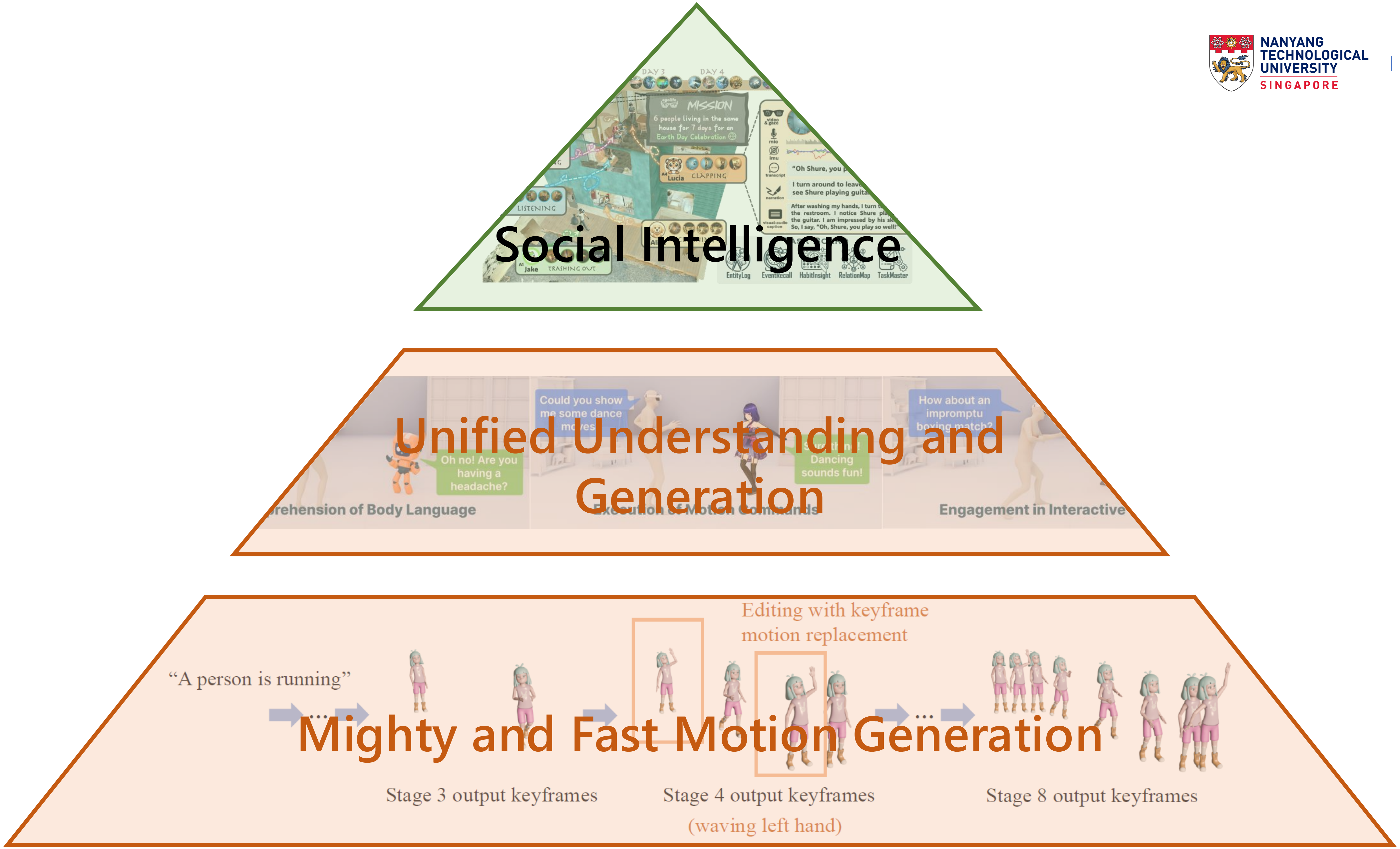
Experiments

Task: *Motion Understanding*
Instruction: *Describe the human motion based on the given three-points and CLIP embeddings.*
Input: *Input CLIP embeddings:* `<CLIP_Placeholder>`. *Input three-points feature:* `<TP_Placeholder>`
Output: `<Narration_Placeholder>`

Task: *Motion Tracking*
Instruction: *Perform motion tracking based on the given three-points and CLIP embeddings.*
Input: *Input CLIP embeddings:* `<CLIP_Placeholder>`. *Input three-points feature:* `<TP_Placeholder>`
Output: `<Motion_Placeholder>`



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE



Social Intelligence - CrowdMoGen

CrowdMoGen: Event-Driven Collective Human Motion Generation

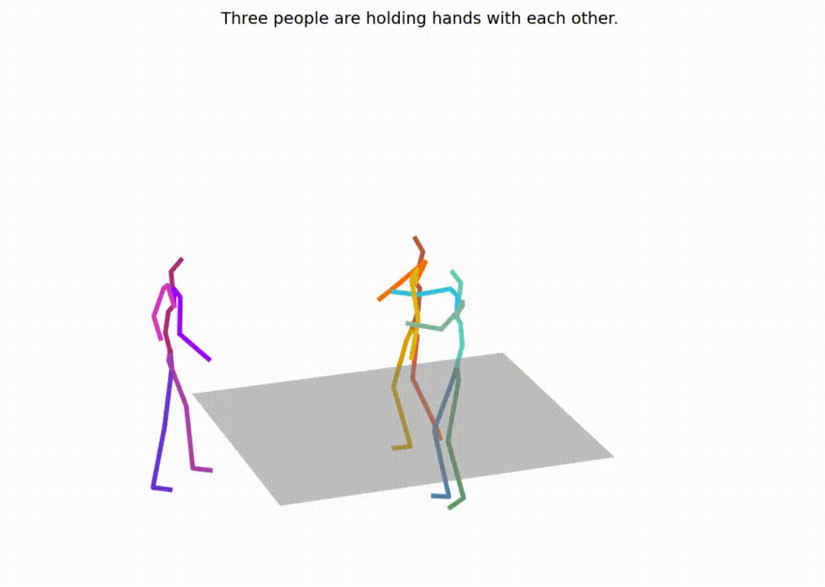
Yukang Cao, Xinying Guo, Mingyuan Zhang, Haozhe Xie, Chenyang Gu, Ziwei Liu

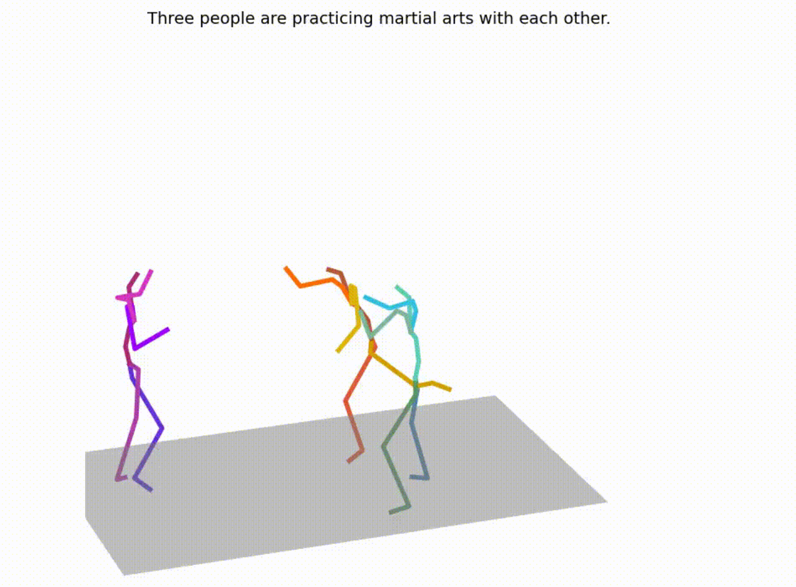
Challenges

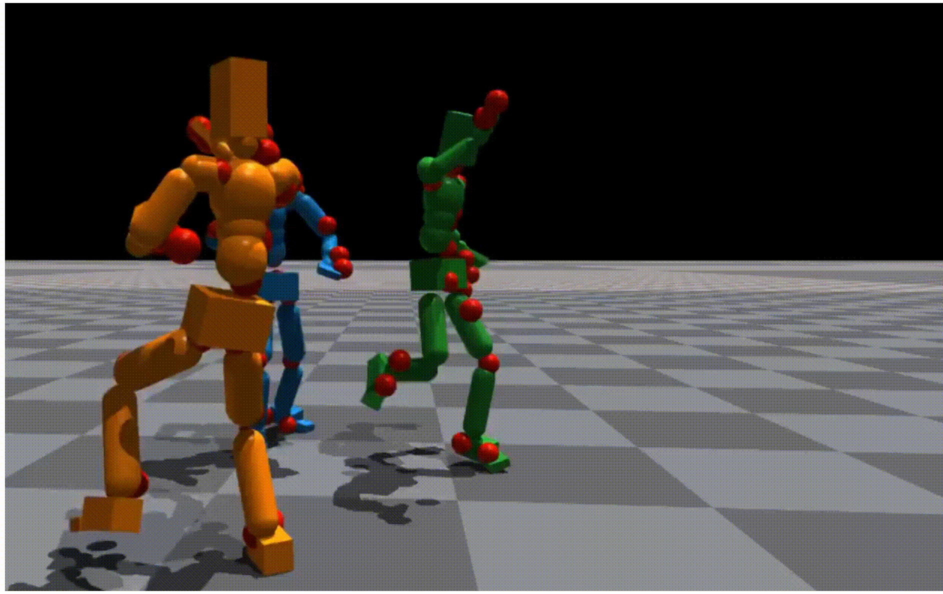
Three people are holding hands together.

Two people are fighting with another person, leading to a 2v1 fighting game.

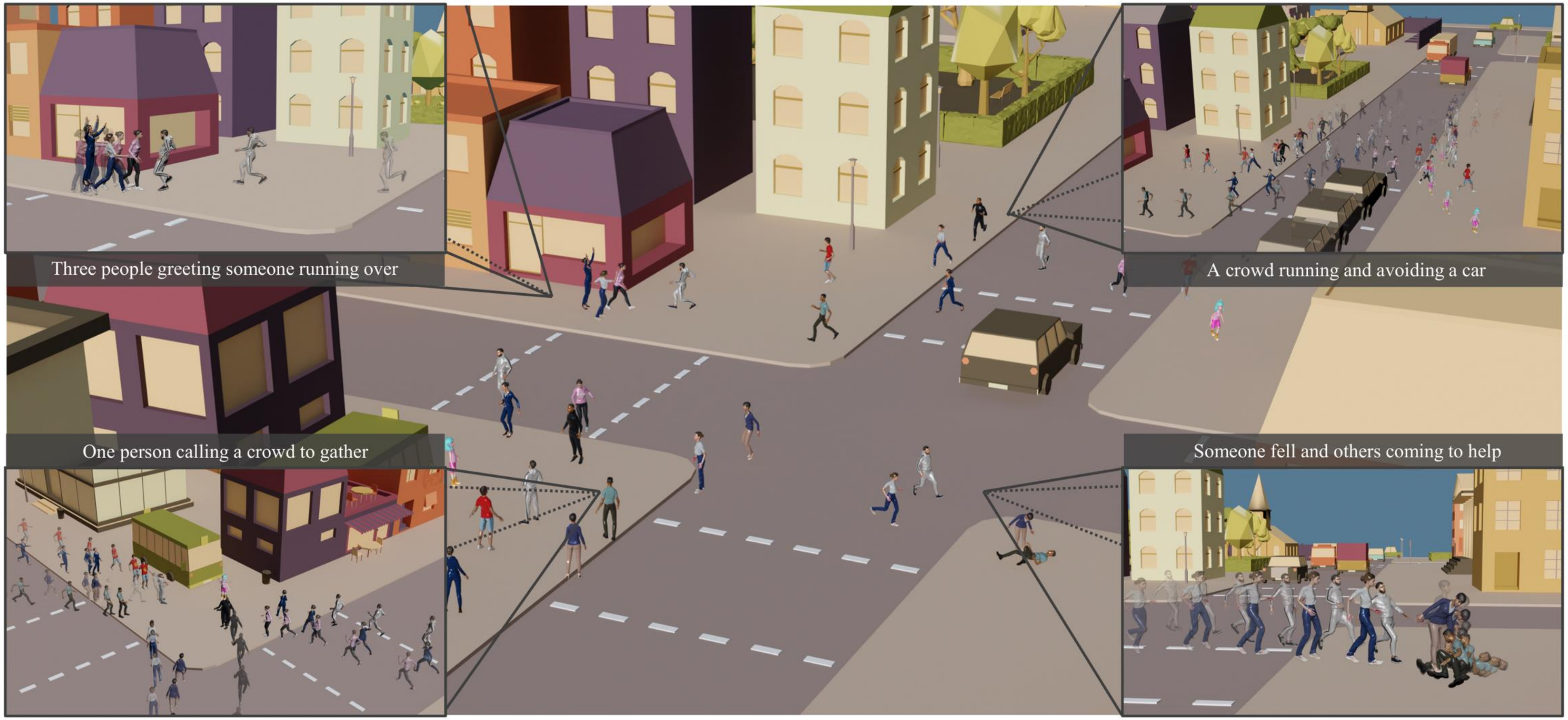
Character animation version of 2v1 fighting in a physics simulator.







CrowdMoGen target

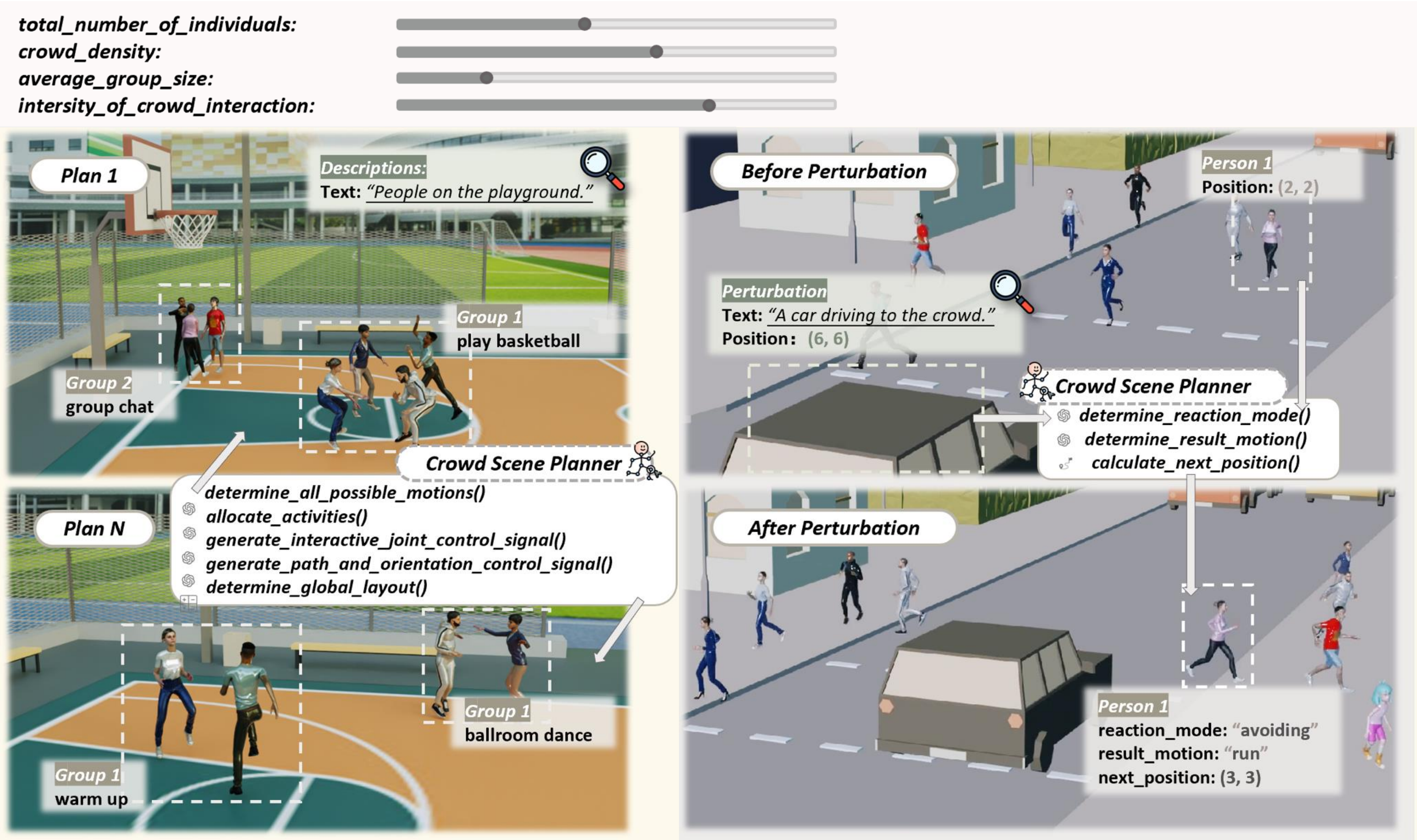


CrowdMoGen pipeline

CrowdMoGen pipeline

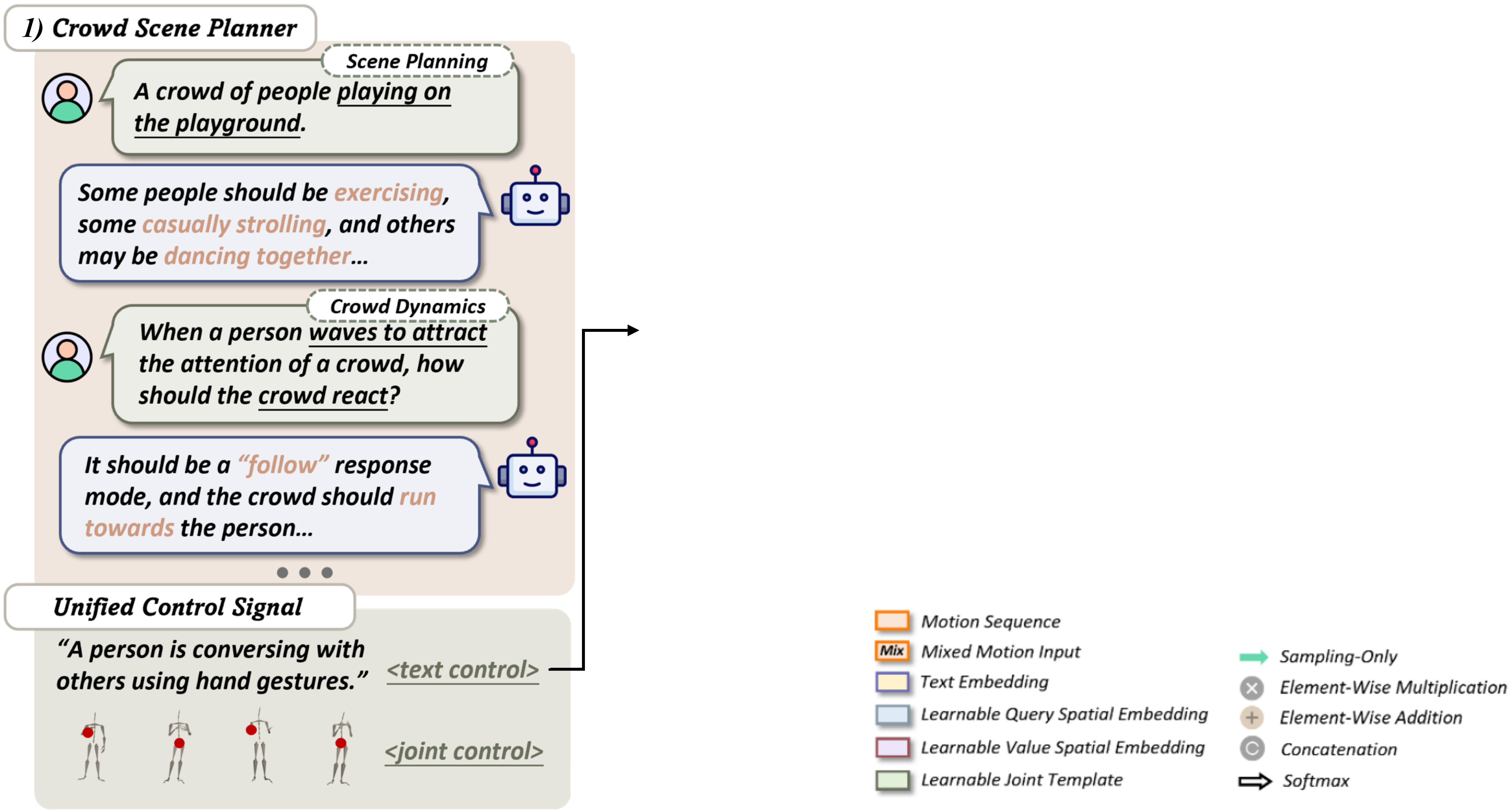
a) Scene-guided activities

CrowdMoGen pipeline

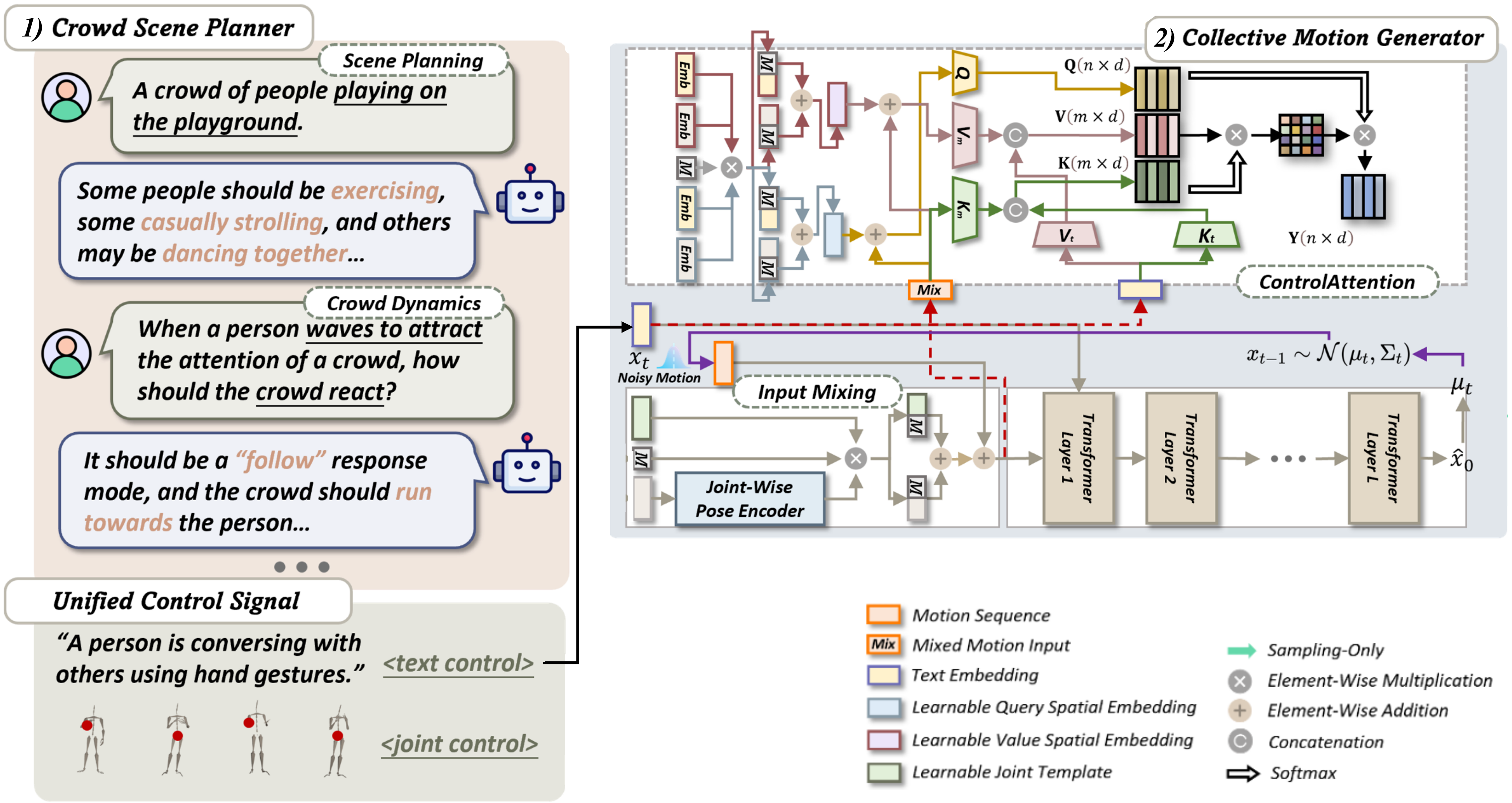


a) Scene-guided activities *b) Event-driven activities*

CrowdMoGen pipeline

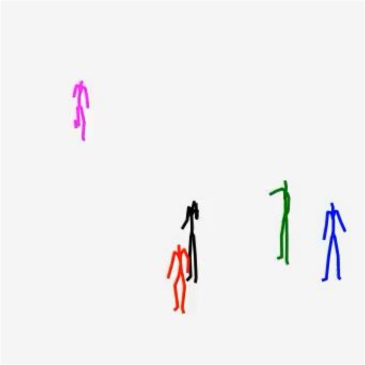


CrowdMoGen pipeline

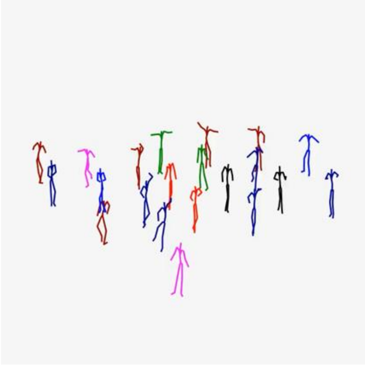


Experiment results

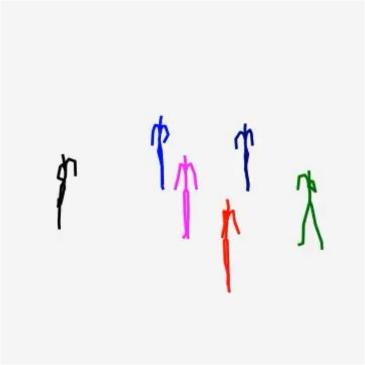
Experiment results



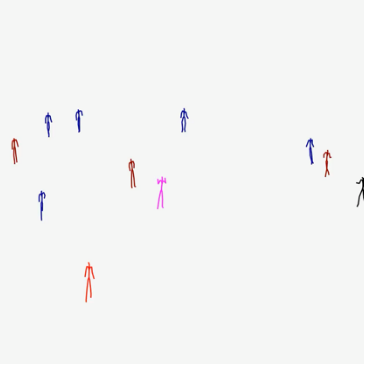
A person run to others and say hello to each other



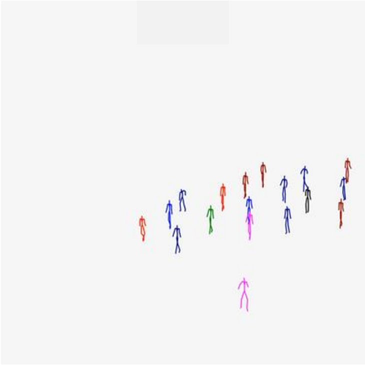
A group of people dancing together



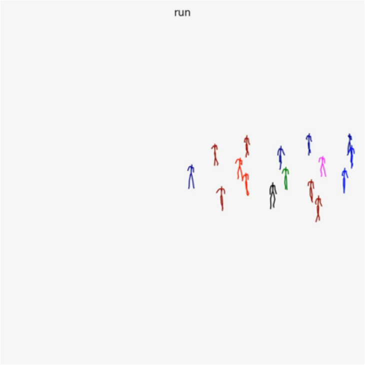
A group of people join to dance together



A person waves the hands to make others gathering around him



People is walking while a high-speed person is running towards the crowd



People is running while a high-speed person is running towards them

Experiment results

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

egolife

Social Intelligence: EgoLife

EvolvingLMMs-Lab/EgoLife

EgoLife: Towards Egocentric Life Assistant

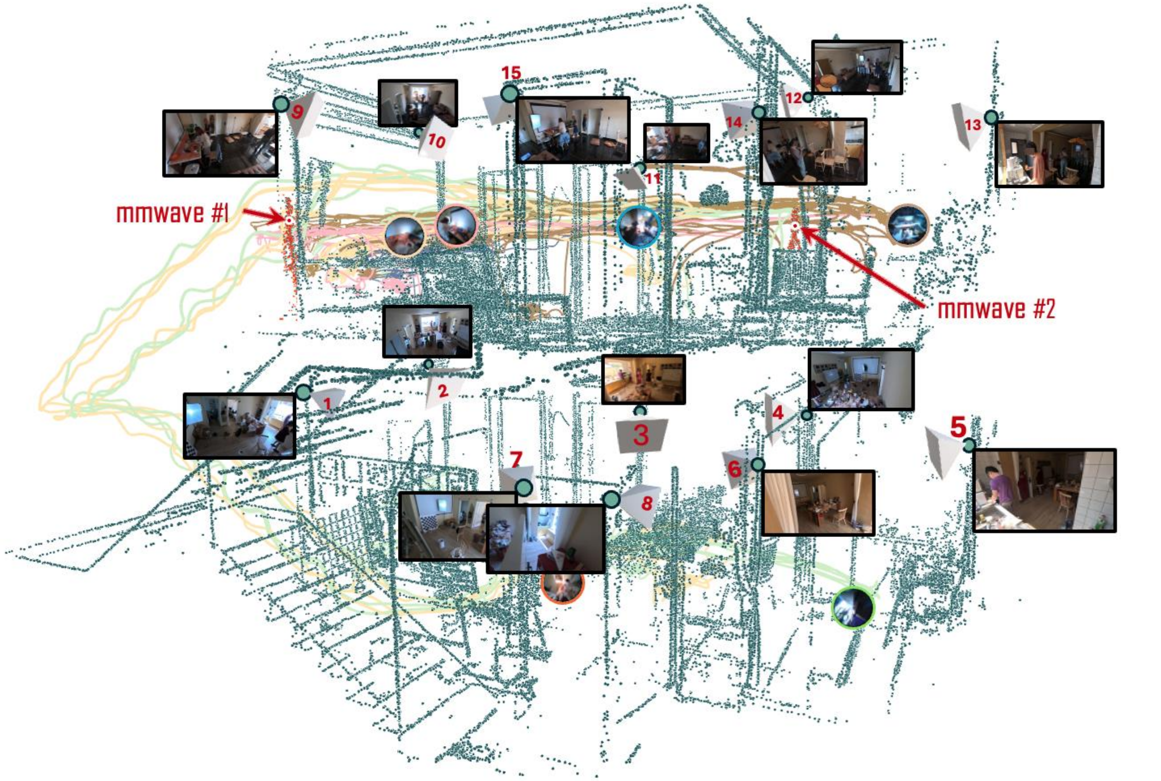
Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, Ziwei Liu

CVPR 2025

We invited 6 people living together for 7 days in egolife

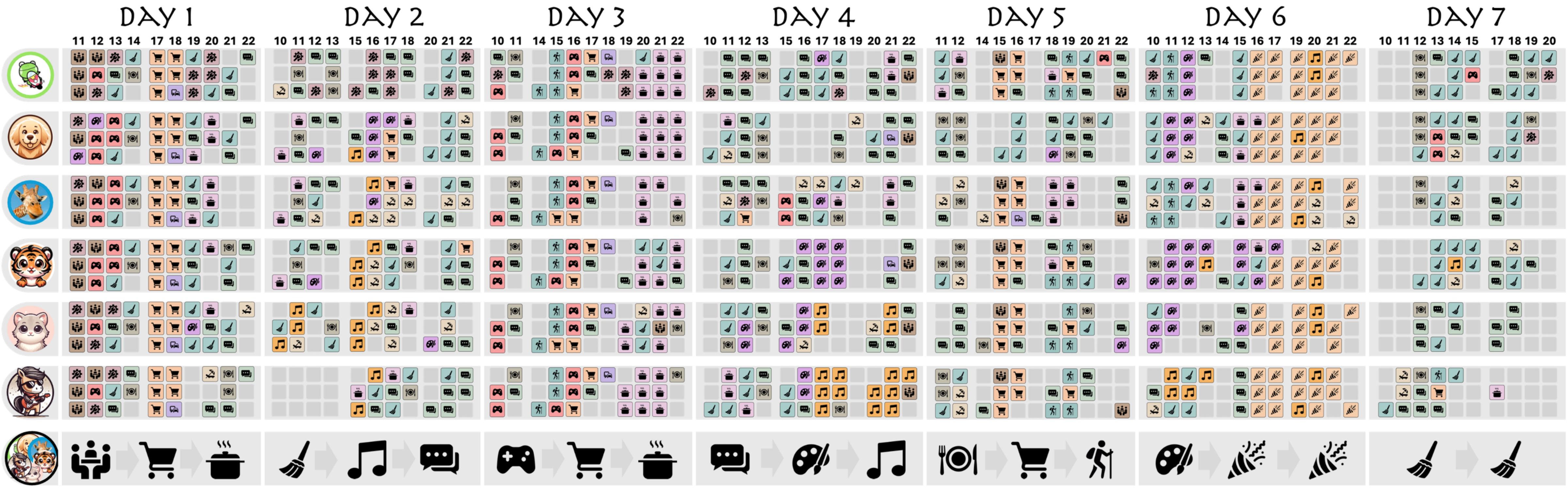
Each one wearing Meta Aria glasses (almost) all day long.

The EgoLife Collected Data



Ego video, audio, mmwave, wifi, Ego/Exo signals synchronization.

The EgoLife Timeline



Social ×183 | Housekeeping ×145 | Cooking ×86 | Shopping ×74 | Dining ×67 | Party ×64 | Arts & Craftwork ×57

Leisure ×49 | Games ×46 | Music & Dance ×45 | Outing ×40 | Setup ×35 | Meeting ×31 | Commuting ×15

The EgoLifeQA Benchmark

6 x 500 = 3000 QAs

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

EventRecall — Past Events of Interest

Day 1: 21:48:21.200

What was the first song mentioned after planning to dance?

A. Why Not Dance B. Mushroom
C. I Wanna Dance with Somebody
D. Never Gonna Give You Up

Answer: A. Evidence: @ Day 1 11:46:59.050
Shure sang after Jake asked us to dance.

EntityLog — Past Objects of Interest

Day 4: 11:34:05.400

Which price is closest to what we paid for one yogurt?

A. RMB 2 B. RMB 3
C. RMB 4 D. RMB 5

Answer: B. Evidence:
The yogurt is on sale, RMB19.9 for 6 cups
@ Day 3: 17:00:04.450

TaskMaster — Tasks Assignment and Review

Many things are in my cart already. What items that we previously discussed have I not bought yet?

A. Milk
B. Chicken wings
C. Strawberries
D. Bananas

Day 5: 16:20:46.350

Answer: A. Evidence:
I made a shopping list, and already got fruit, etc., but ...
D5-15:10 15:57 ... 16:14

Day 4: 12:08:50.600

What activity do I usually do while drinking coffee?

A. Scrolling through TikTok
B. Texting on the phone
C. Tidying up the room
D. Doing Craftwork

Answer: D. Evidence:
D1-16:14 D2-10:40 D2-10:52 ... D4-11:39
I had coffee a total of five times, three of which were while doing crafts...

HabitInsight — Personal Habit Patterns

Day 6: 19:50:19.750

Shure is playing the guitar now. Who else usually joins us playing guitar together?

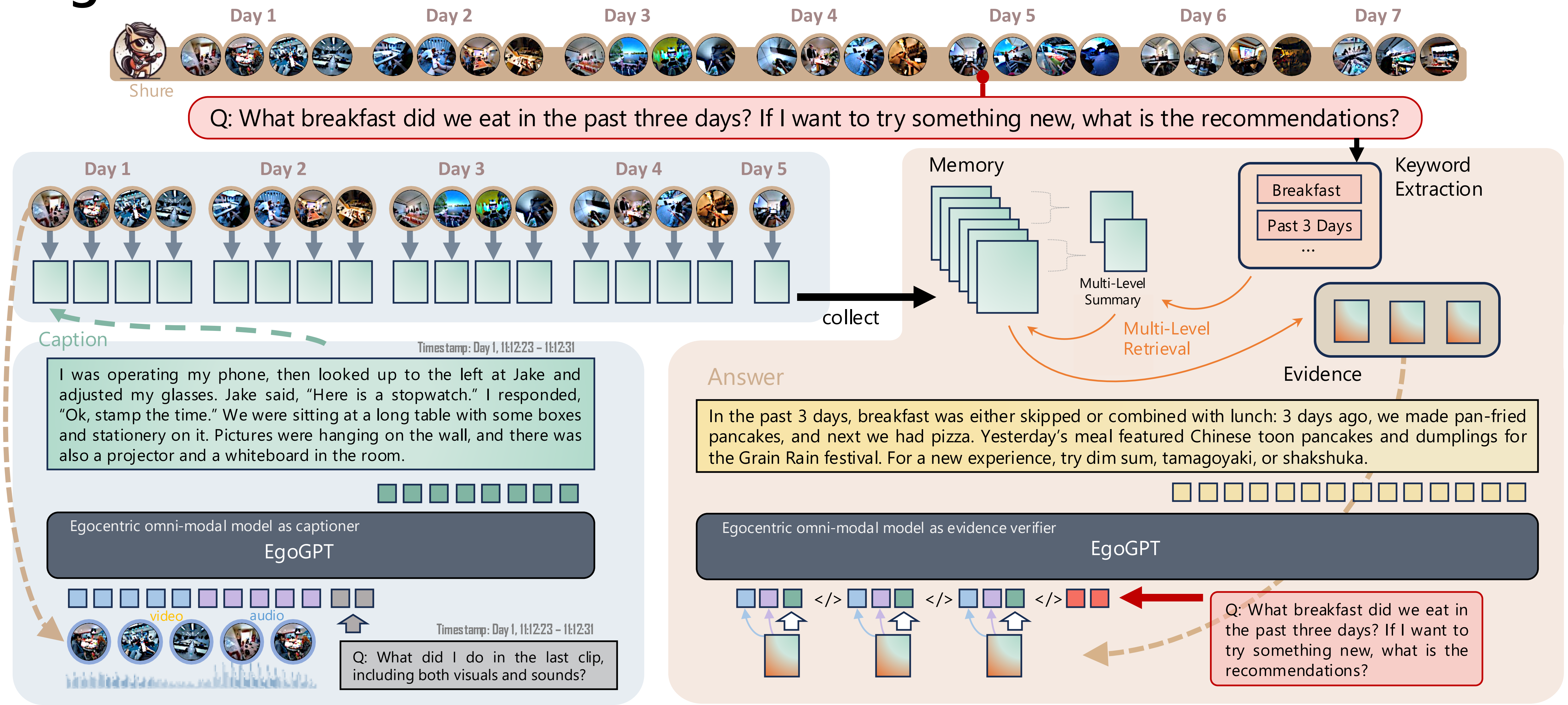
A. Choizst
B. Jake
C. Nicous
D. Lucia

Answer: C. Evidence:
D4-17:19 D4-17:22 D4-22:00 D5-22:52
Nicous played the guitar with Shure and me twice, more frequently than anyone else.

RelationMap — Interpersonal Interaction Patterns

The EgoLifeQA Benchmark

EgoButler



(a) Captioning Stage with EgoGPT

(b) Question Answering Stage with EgoRAG

EgoButler – The EgoGPT Component

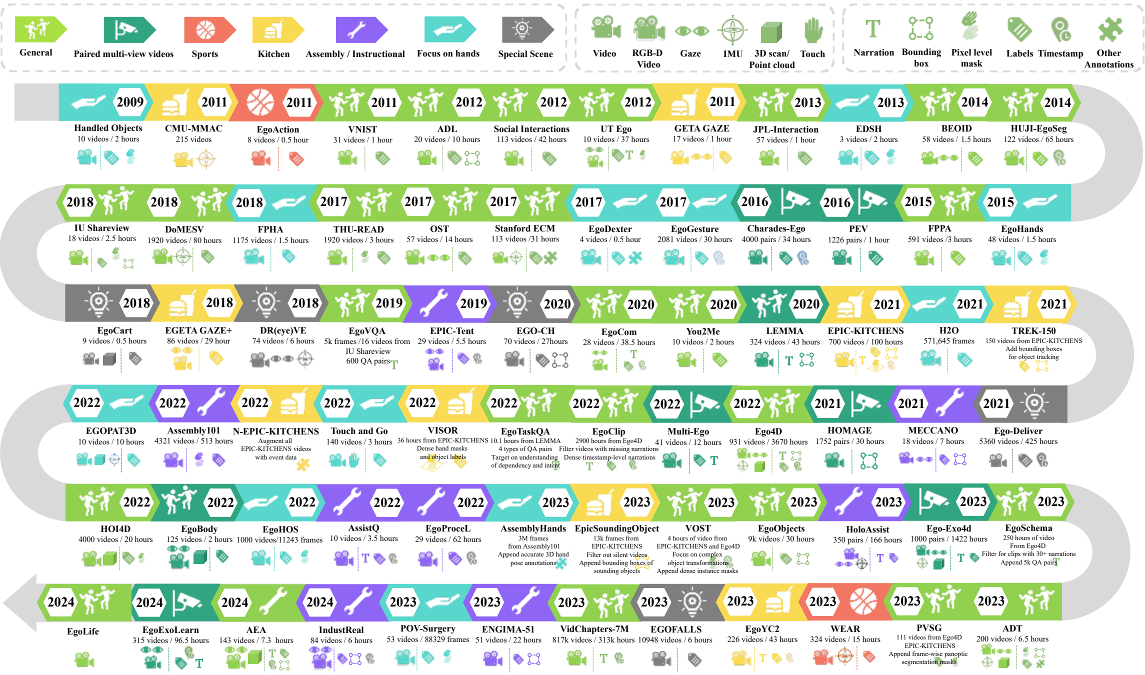
LLaVA-OneVision (Qwen2 as LLM)

Whisper as audio encoder, SFT an audio projector on Qwen2 with ASR datasets

LLaVA-OneVision that supports audio

SFT on EgoIT and EgoLife

EgoGPT



Overview of Classic Egocentric Dataset

Performance of EgoGPT-7B. The table presents a comprehensive comparison of EgoGPT against state-of-the-art commercial and open-source models on existing egocentric benchmarks. With EgoIT and EgoLife Day 1 data, EgoGPT achieve impressive performance on ego setting.

Model	#Param	#Frames	EgoSchema	EgoPlan	EgoThink
GPT-4v [95]	-	32	56.6	38.0	65.5
Gemini-1.5-Pro [96]	-	32	72.2	31.3	62.4
GPT-4o [97]	-	32	72.2	32.8	65.5
LLaVA-Next-Video [98]	7B	32	49.7	29.0	40.6
LongVA [99]	7B	32	44.1	29.9	48.3
IXC-2.5 [100]	7B	32	54.6	29.4	56.0
InternVideo2 [101]	8B	32	55.2	27.5	43.9
Qwen2-VL [94]	7B	32	66.7	34.3	59.3
Oryx [57]	7B	32	56.0	33.2	53.1
LLaVA-OV [55]	7B	32	60.1	30.7	54.2
LLaVA-Videos [102]	7B	32	57.3	33.6	56.4
EgoGPT (EgoIT)	7B	32	73.2	32.4	61.7
EgoGPT (EgoIT+EgoLifeD1)	7B	32	75.4	33.4	61.4

EgoButler – The EgoGPT Component

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

LLaVA-OneVision (Qwen2 as LLM) → Whisper as audio encoder, SFT an audio projector on Qwen2 with ASR datasets → LLaVA-OneVision that supports audio → SFT on EgoIT and EgoLife → EgoGPT

Dataset Composition of EgoIT-99K. We curated 9 classic egocentric video datasets and utilized their annotations to generate captioning and QA instruction-tuning data for fine-tuning EgoGPT, #AV indicates the number of videos with audio used for training.

Dataset	Duration	#Videos (#AV)	#QA
Ego4D [5]	3.34h	523 (458)	1.41K
Charades-Ego [25]	5.04h	591 (228)	18.46K
HoloAssist [29]	9.17h	121	33.96K
EGTEA Gaze+ [26]	3.01h	16	11.20K
IndustReal [28]	2.96h	44	11.58K
EgoTaskQA [93]	8.72h	172	3.59K
EgoProceL [27]	3.11h	18	5.90K
Epic-Kitchens [4]	4.15h	36	10.15K
ADL [24]	3.66h	8	3.23K
Total	**43.16h**	**1529 (686)**	**99.48K**

Performance of EgoGPT-7B. The table presents a comprehensive comparison of EgoGPT against state-of-the-art commercial and open-source models on existing egocentric benchmarks. With EgoIT and EgoLife Day 1 data, EgoGPT achieve impressive performance on ego setting.

Model	#Param	#Frames	EgoSchema	EgoPlan	EgoThink
GPT-4v [95]	-	32	56.6	38.0	65.5
Gemini-1.5-Pro [96]	-	32	72.2	31.3	62.4
GPT-4o [97]	-	32	72.2	32.8	65.5
LLaVA-Next-Video [98]	7B	32	49.7	29.0	40.6
LongVA [99]	7B	32	44.1	29.9	48.3
IXC-2.5 [100]	7B	32	54.6	29.4	56.0
InternVideo2 [101]	8B	32	55.2	27.5	43.9
Qwen2-VL [94]	7B	32	66.7	34.3	59.3
Oryx [57]	7B	32	56.0	33.2	53.1
LLaVA-OV [55]	7B	32	60.1	30.7	54.2
LLaVA-Videos [102]	7B	32	57.3	33.6	56.4
EgoGPT (EgoIT)	7B	32	73.2	32.4	61.7
EgoGPT (EgoIT+EgoLifeD1)	7B	32	75.4	33.4	61.4

EgoButler – The EgoRAG Component

Boosted by EgoGPT, EgoButler achieves SOTA:

- In-depth egocentric video familiarity
- Omni-modal comprehension — effectively integrating both visual and audio signals

Powered by EgoRAG, EgoGPT enables:

- Week-long memory retrieval, answering complex, long-horizon questions
- Robust grounding and context-aware reasoning, where others often fail

Limitations

- ❗ One-Time Retrieval → Agentic Search
- 🧠 Better Person Identification Modeling
- 🔁 Pattern Tracker: Building a habit and behavior pattern engine for continuous insight generation

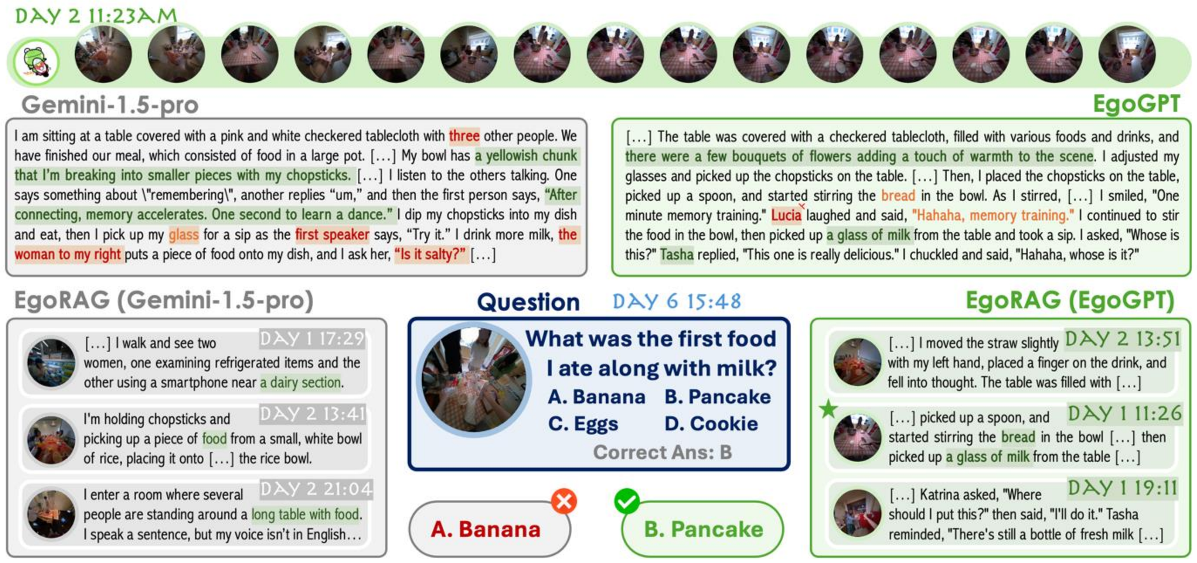


Table 5. **Performance comparison of EgoGPT with state-of-the-art models on EgoLifeQA benchmarks.** For a fair comparison on EgoLifeQA, EgoGPT was replaced with the corresponding models in the EgoButler pipeline to evaluate their performance under the same conditions. Models that provide captions for EgoLifeQA use 1 FPS for video sampling.

Model	#Frames	Audio	Identity	EgoLifeQA EntityLog	EventRecall	HabitInsight	RelationMap	TaskMaster	Average
Gemini-1.5-Pro [95]	-	✓	✗	36.0	37.3	45.9	30.4	34.9	36.9
GPT-4o [96]	1 FPS	✗	✗	34.4	42.1	29.5	30.4	44.4	36.2
LLaVA-OV [55]	1 FPS	✗	✗	36.8	34.9	31.1	22.4	28.6	30.8
EgoGPT (EgoIT-99K)	1 FPS	✓	✗	35.2	36.5	27.9	29.6	36.5	33.1
EgoGPT (EgoIT-99K+D1)	1 FPS	✓	✓	39.2	36.5	31.1	33.6	39.7	36.0

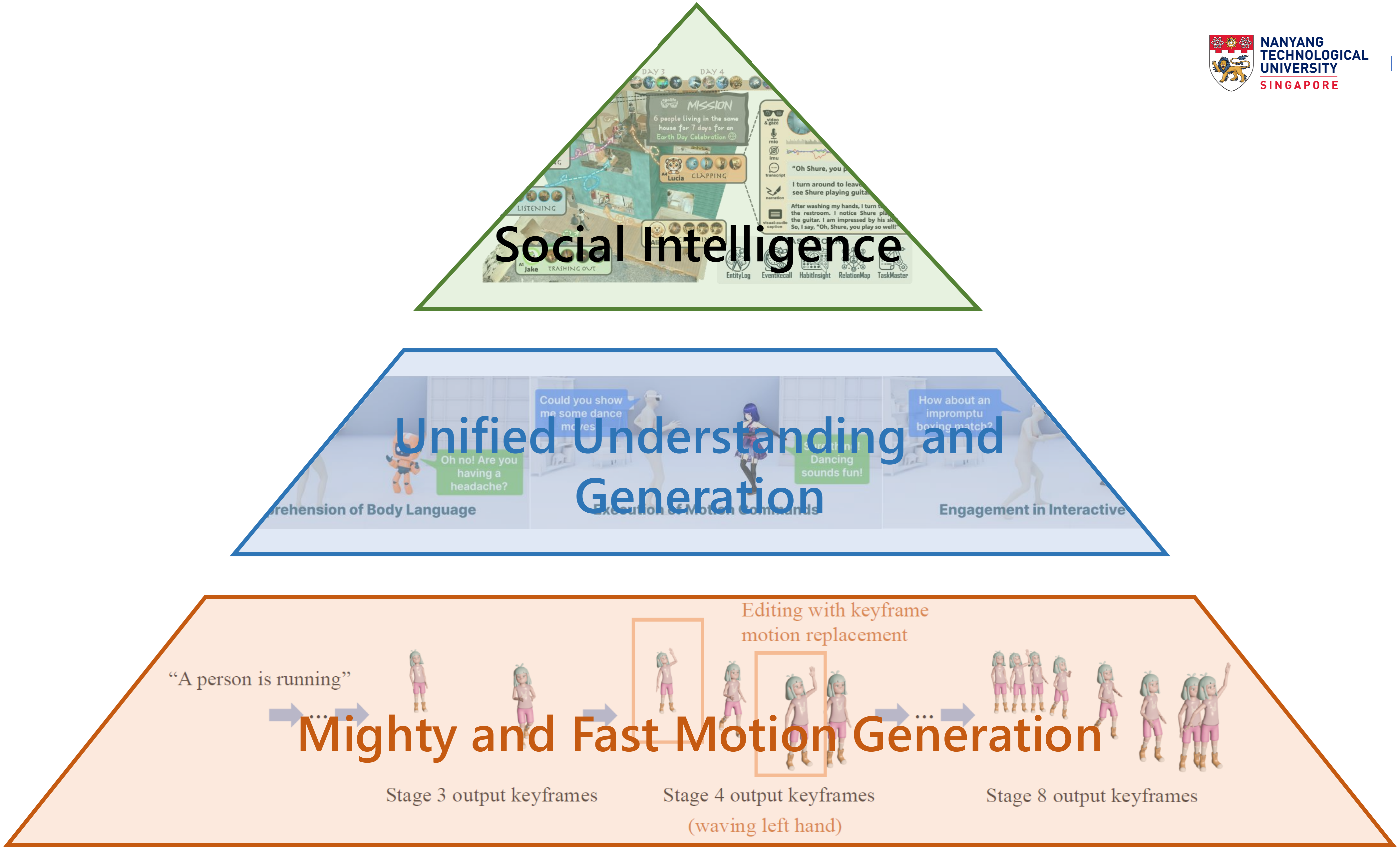
Towards

Extremely Long, Egocentric, Interpersonal, Multi-view, Multi-modal, Daily Life Video Understanding

More to explore:

Dense Caption, Transcript, Gaze, Multiple Third-Person View, SLAM

egolife-ai.github.io


Social Intelligence
MISSION
6 people living in the same house for 7 days for an Earth Day Celebration
LISTENING
Lucia
CLAPPING
Jake
TRASHING OUT
"Oh Shure, you
EntityLog
EventRecall
HabitInsight
RelationMap
TaskMaster
Unified Understanding and Generation
Could you show me some dance moves
Oh no! Are you having a headache?
Dancing sounds fun!
How about an impromptu boxing match?
Engagement in Interactive
Mighty and Fast Motion Generation
"A person is running"
Editing with keyframe motion replacement
Stage 3 output keyframes
Stage 4 output keyframes
(waving left hand)
Stage 8 output keyframes

Thank You

Ziwei Liu 刘子纬

Nanyang Technological University