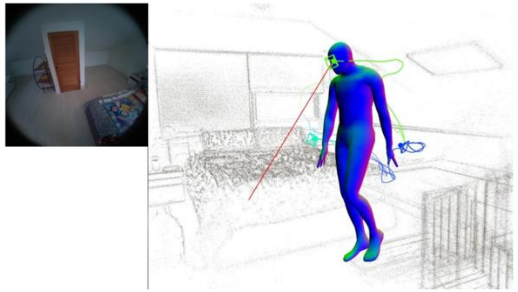
From Egocentric Perception to Embodied Intelligence: Building the World in First Person

Ziwei Liu 刘子纬
Nanyang Technological University

Why Egocentric Perception?

Egocentric Perspective Provides:

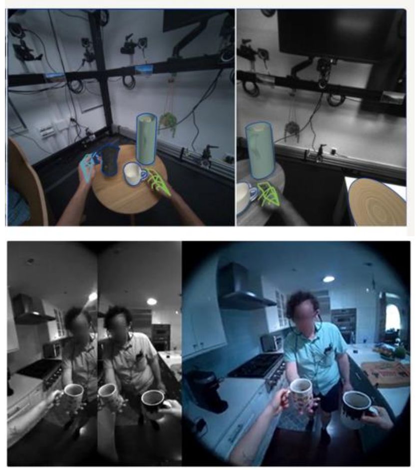
1. Natural Human Experience and Cognition



Gaze, Attention

Navigation, Spatial Awareness

2. Better Context for Interaction

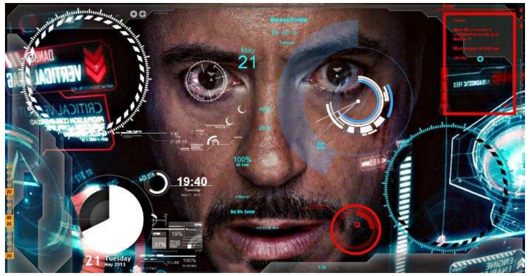


Object Manipulation
Hand-Eye Coordination

Social Interaction

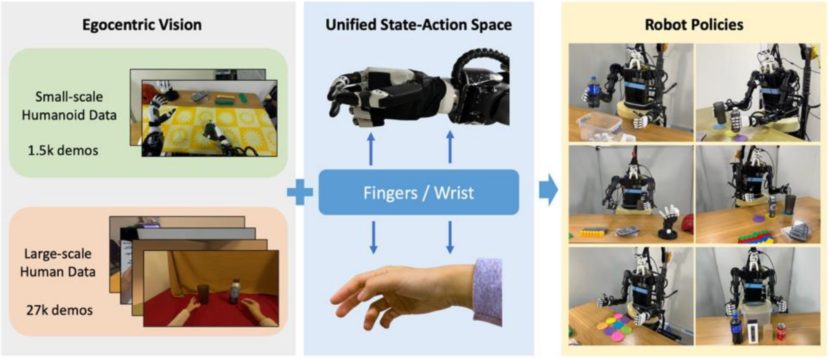
Egocentric Perception Enables:

1. Personal AI Assistant



EgoLife, CVPR2025

2. Embodied AI Learning



DexCap, EgoMimic, PH^2D

Perception & Understanding

Learn to see and reason from the first person

cognitive foundation of
egocentric understanding

egolife

Egocentric
Life Assistant AI

Perception & Understanding
Life Assistant AI

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

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

egolife

Towards Egocentric Life Assistant

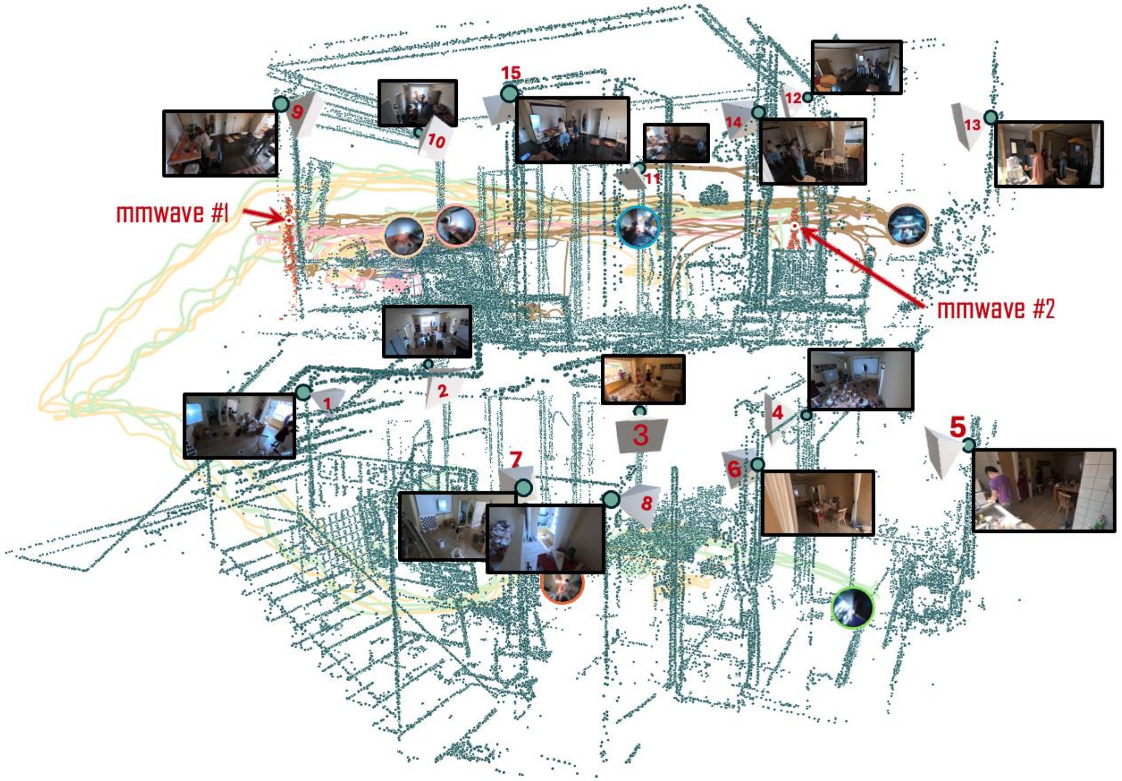
(Inspired by some reality show)

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

We invited 6 (strangers) people living together for 7 days in egolife

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

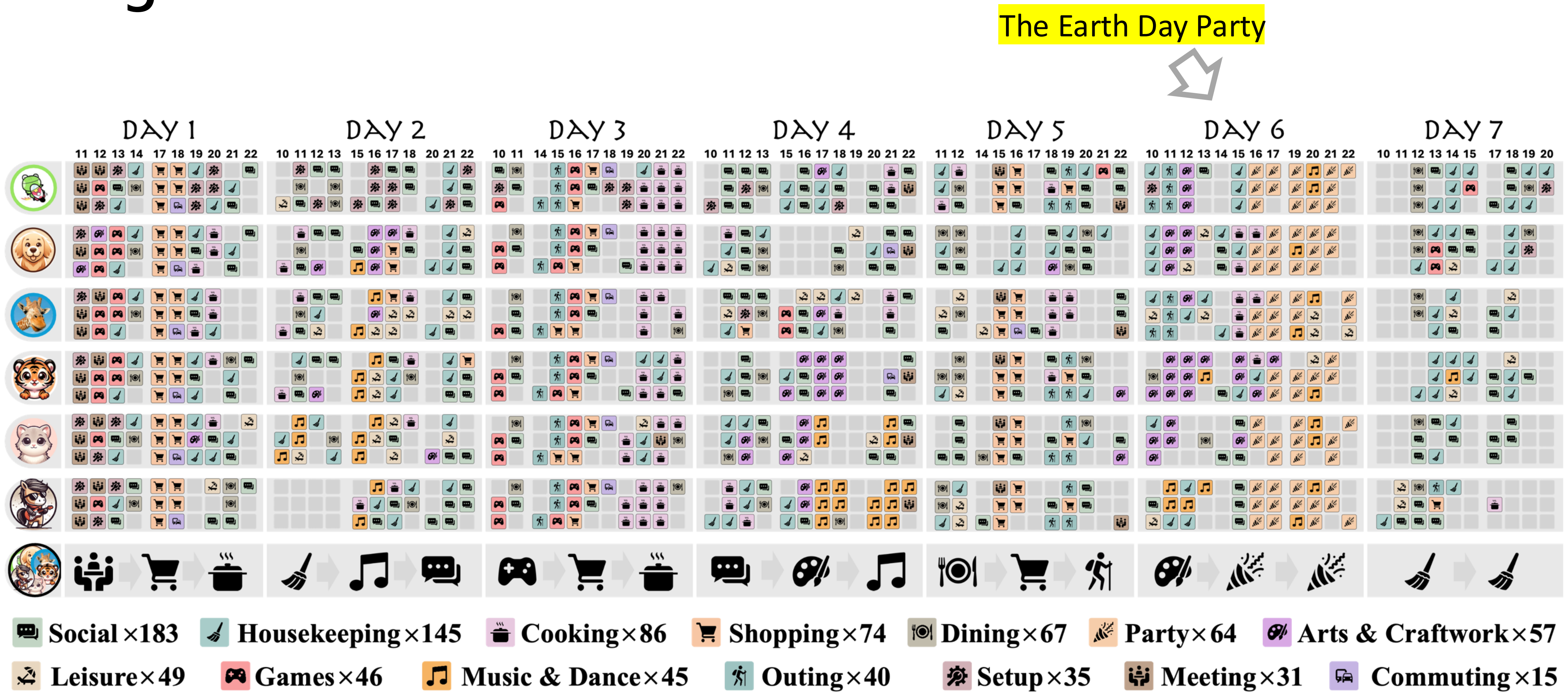
The EgoLife Collected Data



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

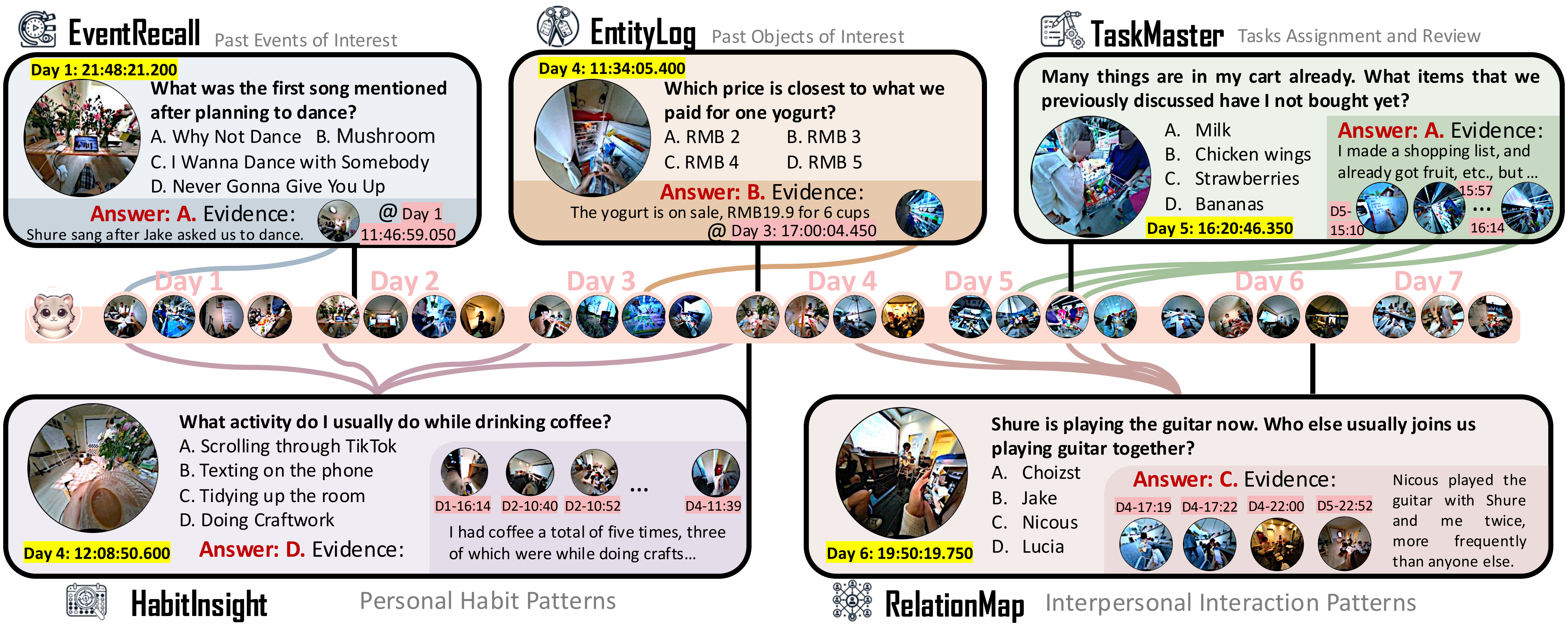
The EgoLife Timeline

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE



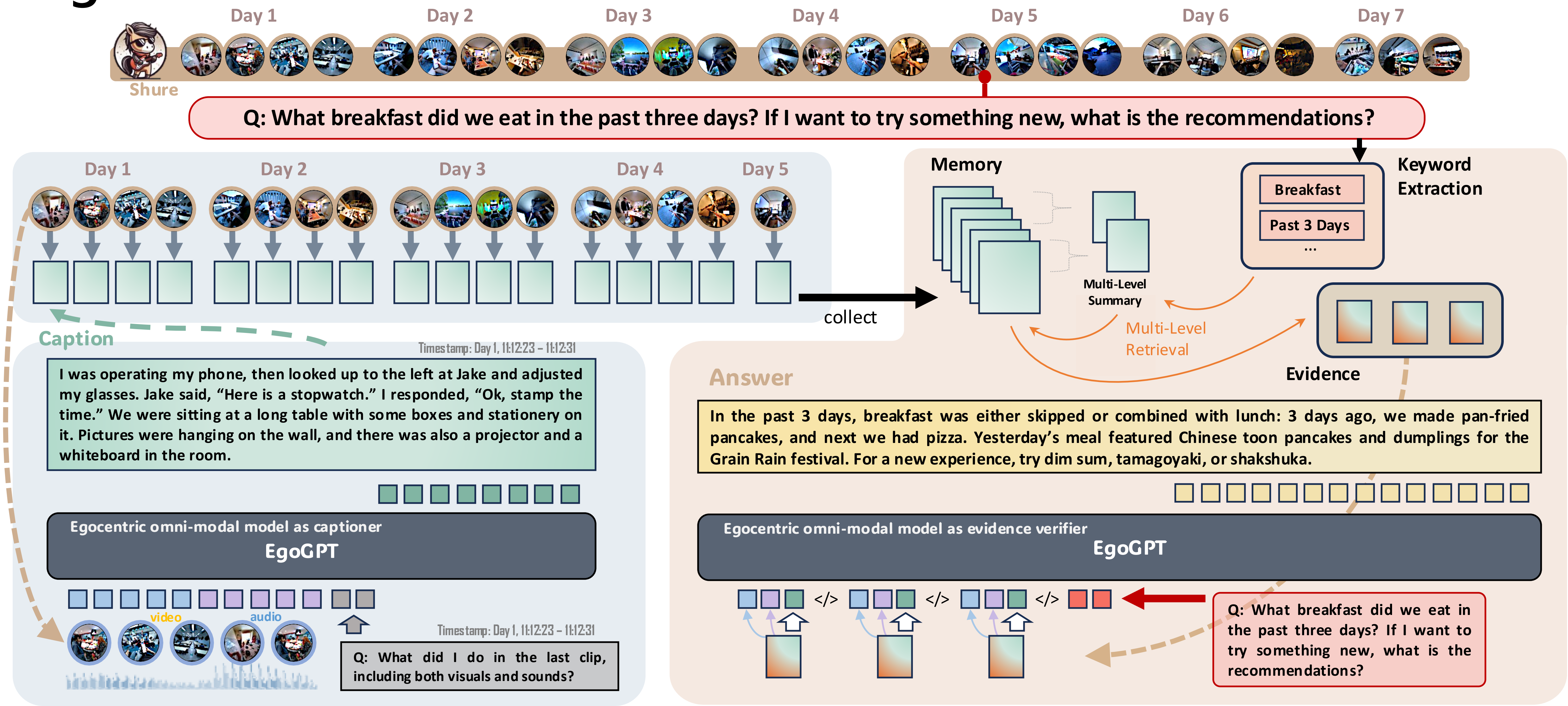
The EgoLifeQA Benchmark

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE



The EgoLifeQA Benchmark

EgoButler

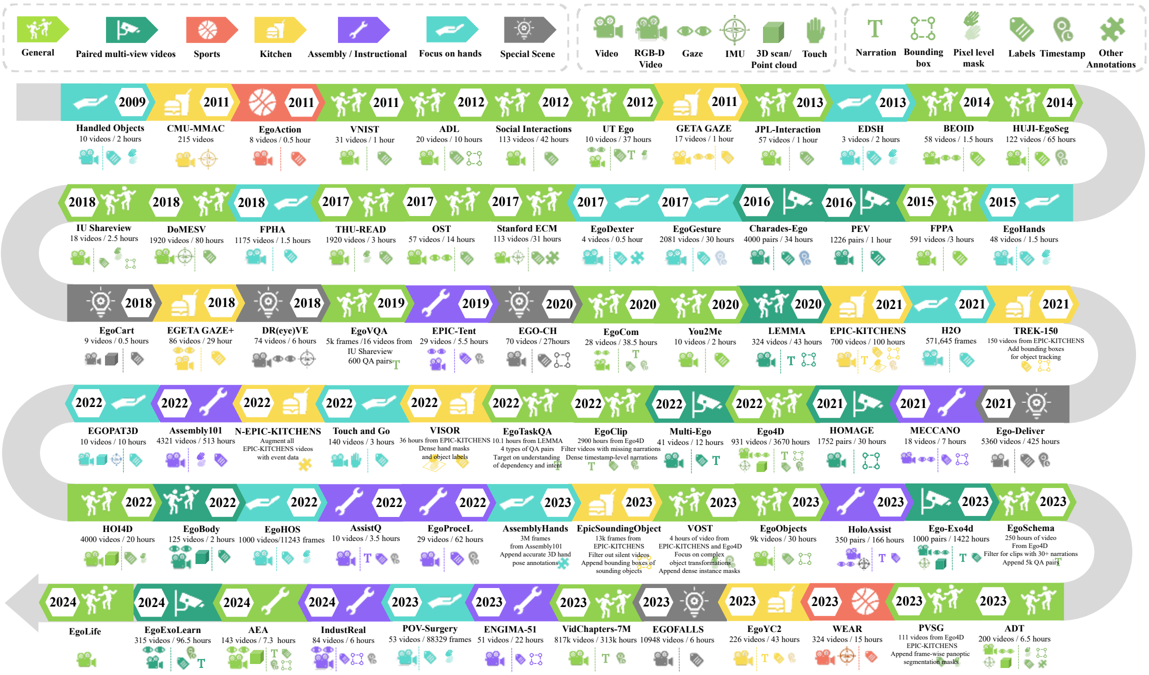


(a) Captioning Stage with EgoGPT

(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

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 through:

- 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

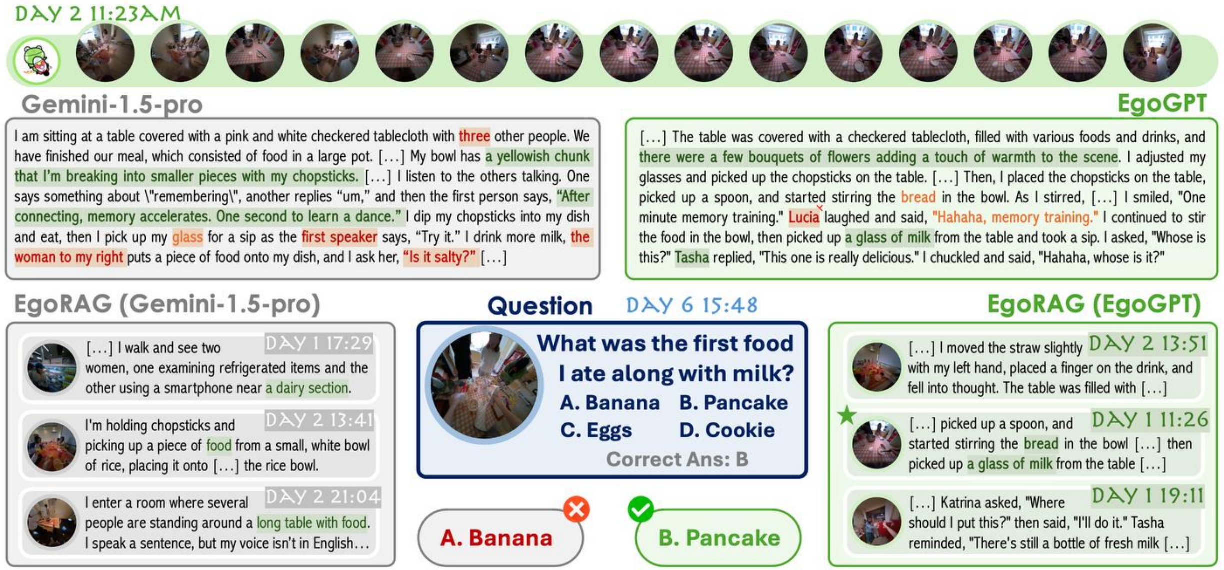


Table 5. **Performance comparison of EgoGPT with state-of-the-art models on EgoLifeQA benchmarks.** For a fair comparison on EgoLifeQA, EgoGPT was replaced with the corresponding models in the EgoButler pipeline to evaluate their performance under the same conditions. Models that provide captions for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

Limitation

❗ One-Time Retrieval → Agentic Search

🧠 Better Person Identification Modeling

🔁 Pattern Tracker: Building a habit and behavior pattern engine for continuous insight generation

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

Perception & Understanding

Learn to see and reason from the first person

cognitive foundation of
egocentric understanding

egolife

Egocentric
Life Assistant AI

Perception & Understanding

Learn to see and reason from the first person

cognitive foundation of
egocentric understanding

better long-term reasoning!

Egocentric
Life Assistant AI

Perception & Understanding
Life Assistant AI - long-term reasoning

Ego-R1: Chain-of-Tool-Thought for Ultra-Long Egocentric Video Reasoning

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, Ziwei Liu

Challenges

- Ultra-Long Event-Based Egocentric Videos
  - **Long-horizon**
    - Spanning from days to weeks
  - **Multimodal Complexity**
    - Cross-modality linkage
  - **Sparse event cues**
    - Randomly distributed evidence
- Modelling of Long Video
  - **Rigid Pipeline**
    - End-to-end
    - Sampling

Why it's hard

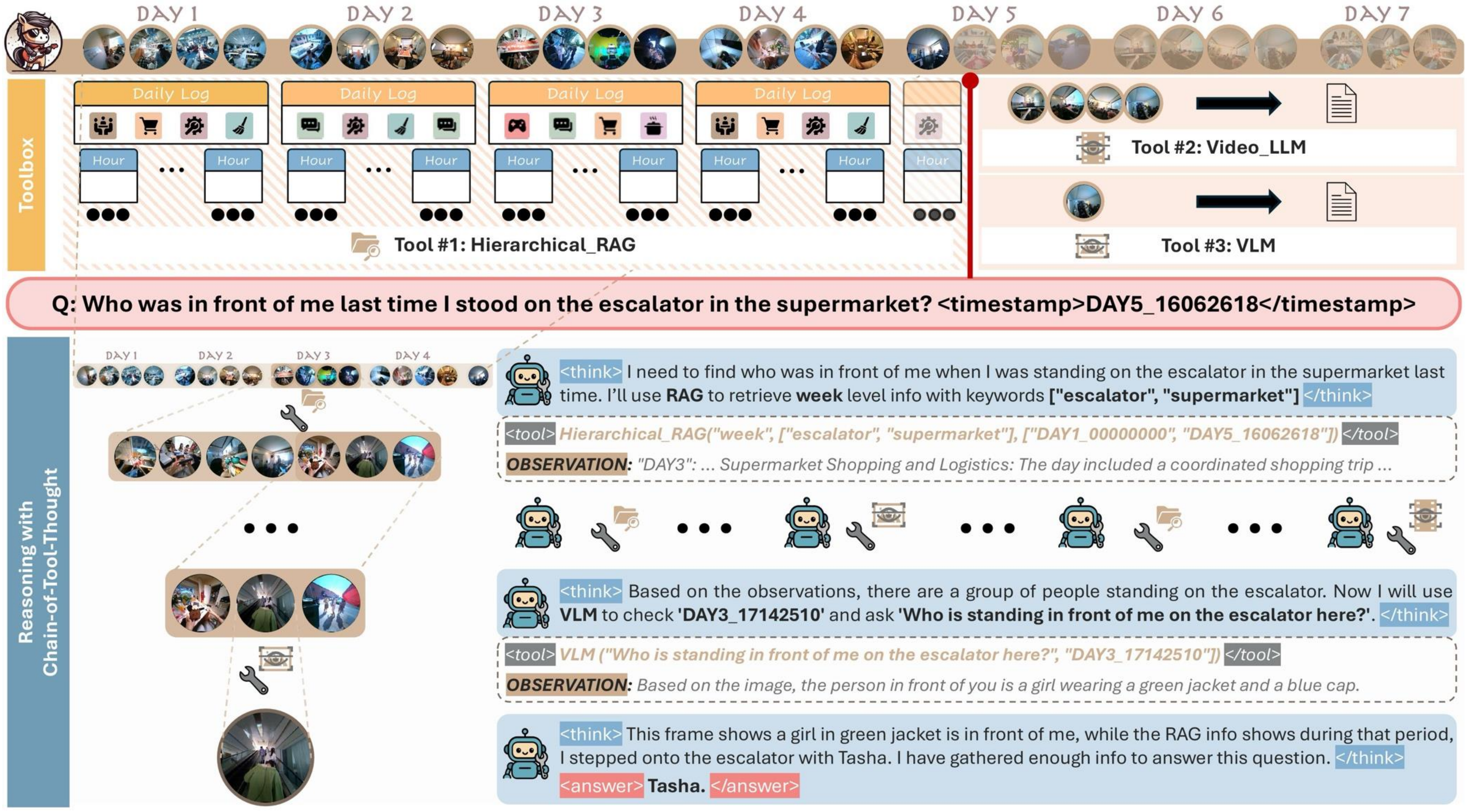
Long horizon
Events span multiple days

Multi-modal
Vision, audio & text

Sparse cues
Key events are rare

Rigid pipelines
Lack adaptability

Ego-R1: Tool-Use Agent for Ultra-Long Egocentric Videos



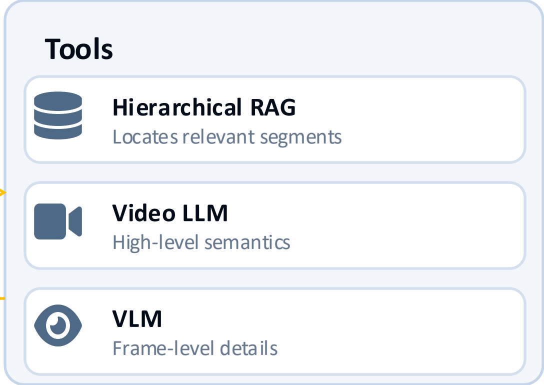
A sample workflow of Ego-R1.

Key Idea: Chain-of-Tool-Thought (CoTT)

$$C = (S^0, S^1, \ldots, S^n), \quad S^i = \left(T_i^{\mathrm{th}},\, T_i^{\mathrm{to}},\, o_i\right)$$

- C is a sequence of n reasoning steps.
- T_i^{th} : thought
- T_i^{to} : tool call

- Action space: $\mathcal{A} = F_j$
- Observation space: $\left(o_i^{\mathrm{rag}},\, o_i^{\mathrm{vid}},\, o_i^{\mathrm{vlm}}\right) \in \mathcal{O}$



Stage I: Data Generation – Raw QA

❶ Raw QA Data (for RL)

500 h Ego-Vid

↓

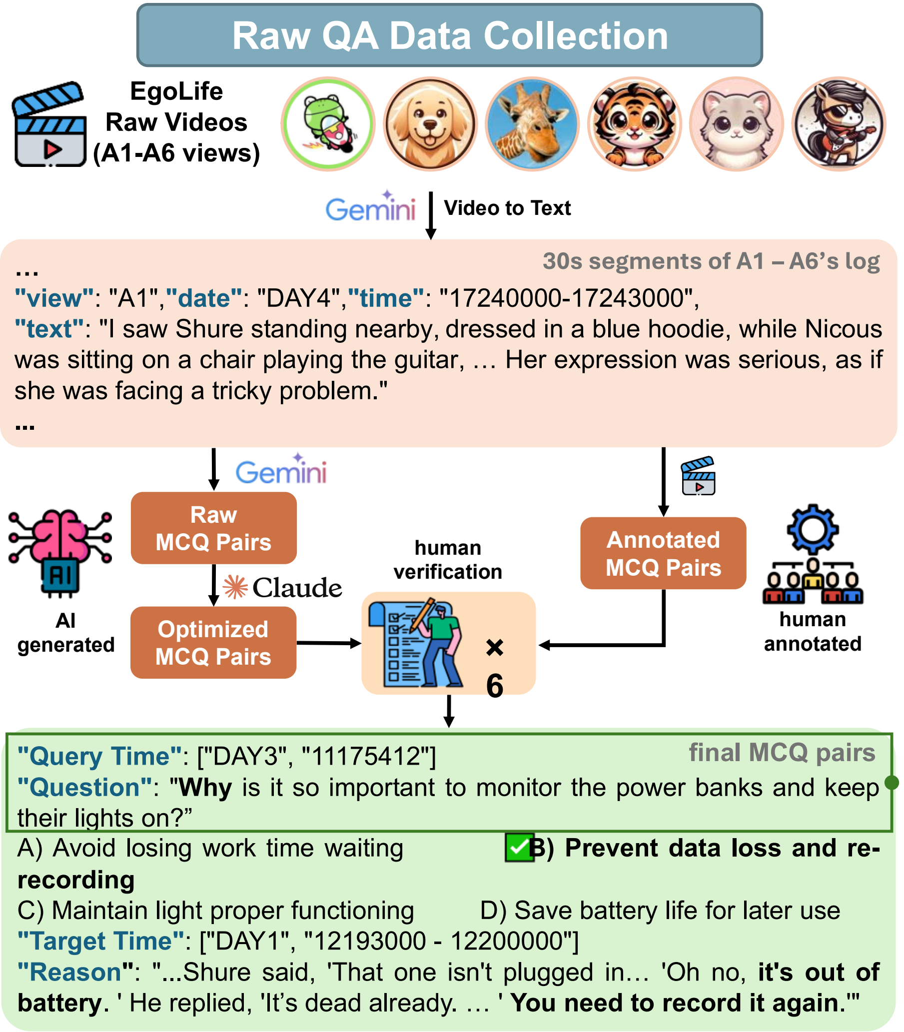
ASR + 30 s captions

↓

Candidate QA

↓

Ego-QA-4.4K



Stage I: Data Generation – CoTT

❷ CoTT Data (for SFT)

2.9 K high-quality raw QA

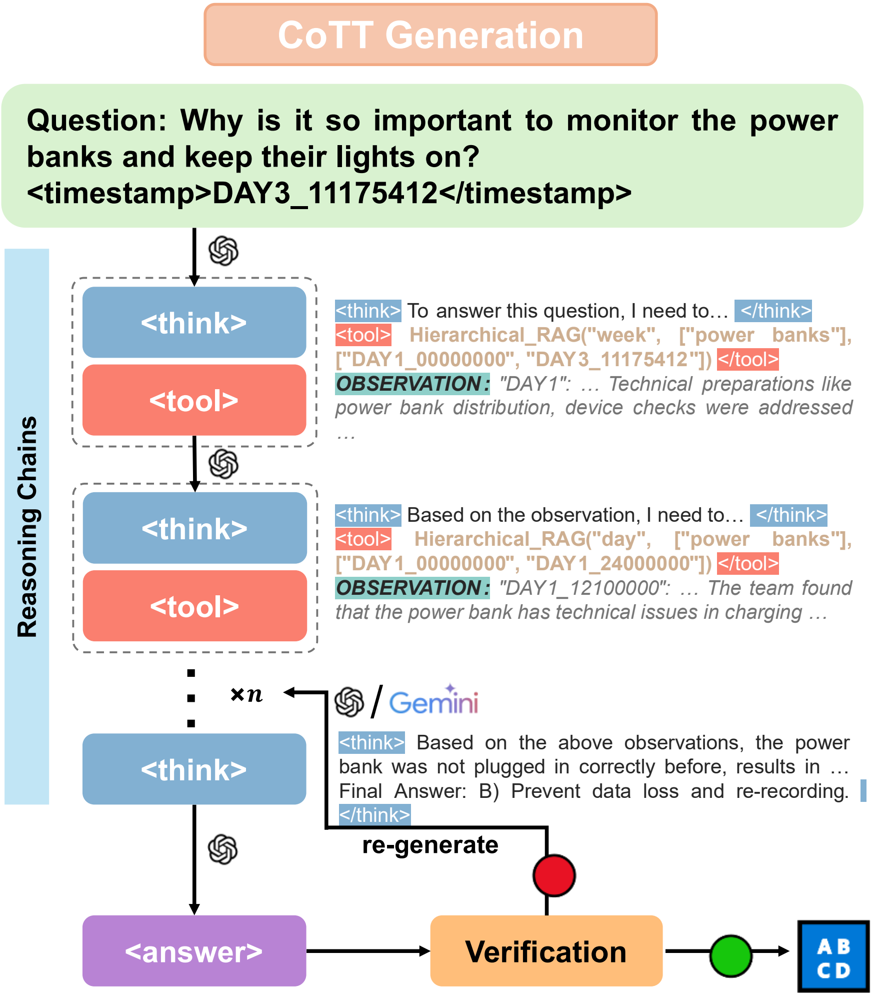
↓

LLM-driven CoTT engine

↓

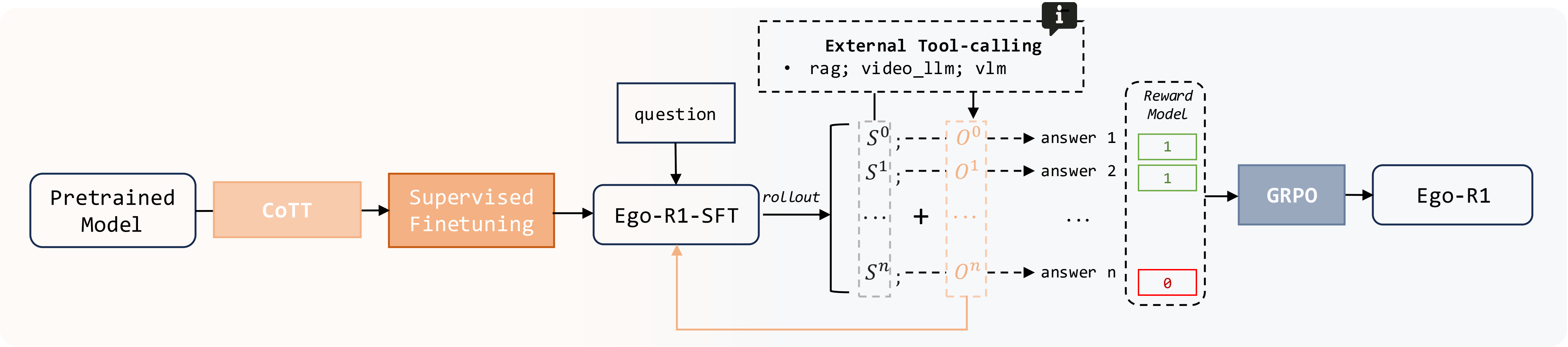
Ego-CoTT-25K

- *avg 7.42 tool calls / task*
- *observation loop* ✱



Stage II: Training (SFT+RL)

Ego-CoTT-25K **Ego-QA-4.4K**



(1) Stage 1: SFT with CoTT (2) Stage 2: GRPO for Ego-R1

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{[q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)]}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\sum_{y=1}^{T}\frac{1}{|S_i^y|}\sum_{t=1}^{|S_i^y|}\Bigg\{\min\Bigg[\frac{\pi_\theta(S_{i,t}|q,I_y,S_{i,<t})}{\pi_{\theta_{\text{old}}}(S_{i,t}|q,I_y,S_{i,<t})}\hat{A}_{i,t}^{y},$$
$$\text{clip}\Big(\frac{\pi_\theta(S_{i,t}|q,I_y,S_{i,<t})}{\pi_{\theta_{\text{old}}}(S_{i,t}|q,I_y,S_{i,<t})},1-\varepsilon,1+\varepsilon\Big)\hat{A}_{i,t}^{y} - \beta\mathbb{D}_{\text{KL}}[\pi_\theta\|\pi_0]\Bigg]\Bigg\}\Bigg]$$

Results: Compact Model, Competitive Results

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Table 2: **Quantitative results on video question-answering benchmarks.** The proposed Ego-R1 model demonstrates superior performance across multiple metrics. Bold indicates best performance, underscored values show second best. The results from the 72B version of the model or using less frames are marked in gray. As some of the QA pairs in EgoLifeQA were used for CoTT generation and training, we excluded these from evaluation and retained only a clean subset for fair testing.

Method	Size	Frames	Exocentric	Egocentric		
			VideoMME (long)	EgoSchema	EgoLifeQA	Ego-R1 Bench
Average durations			41 min	3 min	44.3 h	44.3 h
MLLMs						
LongVA [81]	7B	64	45.0	44.1	33.0	23.0
LLaVA-Video [82]	7B	64	61.5	57.3	<u>36.4</u>	29.0
LLaVA-OneVision [28]	7B	1 FPS	60.0	60.1	30.8	31.6
InternVideo2.5 [64]	8B	512	53.4	63.9	33.0	34.0
Gemini-1.5-Pro [58]	-	-	**67.4**	**72.2**	**36.9**	38.3
RAG Methods						
LLaVA-Video + Video-RAG [37]	7B	64	46.0	66.7	30.0	29.3
LongVA + Video-RAG [37]	7B	64	55.7	41.0	26.0	31.0
Reasoning Models						
Video-R1 [16]	7B	64	50.8	-	34.0	20.0
Video Agents						
VideoAgent [63]	-	8	50.8	54.1	29.2	32.6
LLaVA-OneVision + T^* [79]	7B	8	46.3	66.6	35.4	35.6
Ours						
Ego-R1	**3B**	-	<u>64.9</u>	<u>68.2</u>	36.0*	**46.0**

Perception & Understanding

Learn to see and reason from the first person

cognitive foundation of
egocentric understanding

egolife

ego-r1 Ego-R1

From Seeing to Acting

Perception & Understanding

Learn to see and reason from the first person

→

Action & Embodiment

Learn to move and act as you see

cognitive foundation of
egocentric understanding

egolife

ego-r1 Ego-R1

embodied simulation

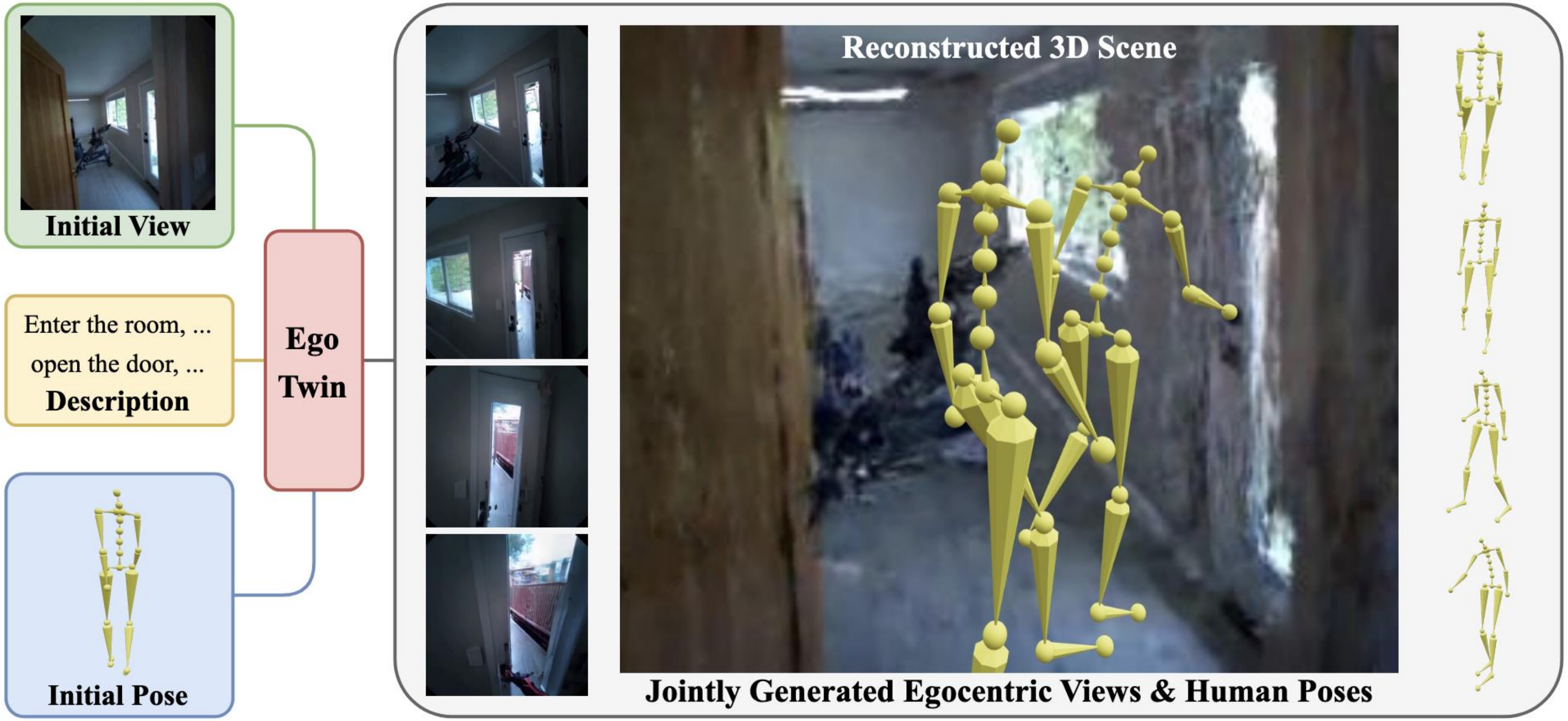
Action & Embodiment
Learn to move and act as you see

EgoTwin: Dreaming Body and View in First Person

Jingqiao Xiu, Fangzhou Hong, Yicong Li, Mengze Li, Wentao Wang, Sirui Han, Liang Pan, Ziwei Liu

EgoTwin: Dreaming Body and View in First Person

Jingqiao Xiu, Fangzhou Hong, Yicong Li, Mengze Li, Wentao Wang, Sirui Han, Liang Pan, Ziwei Liu



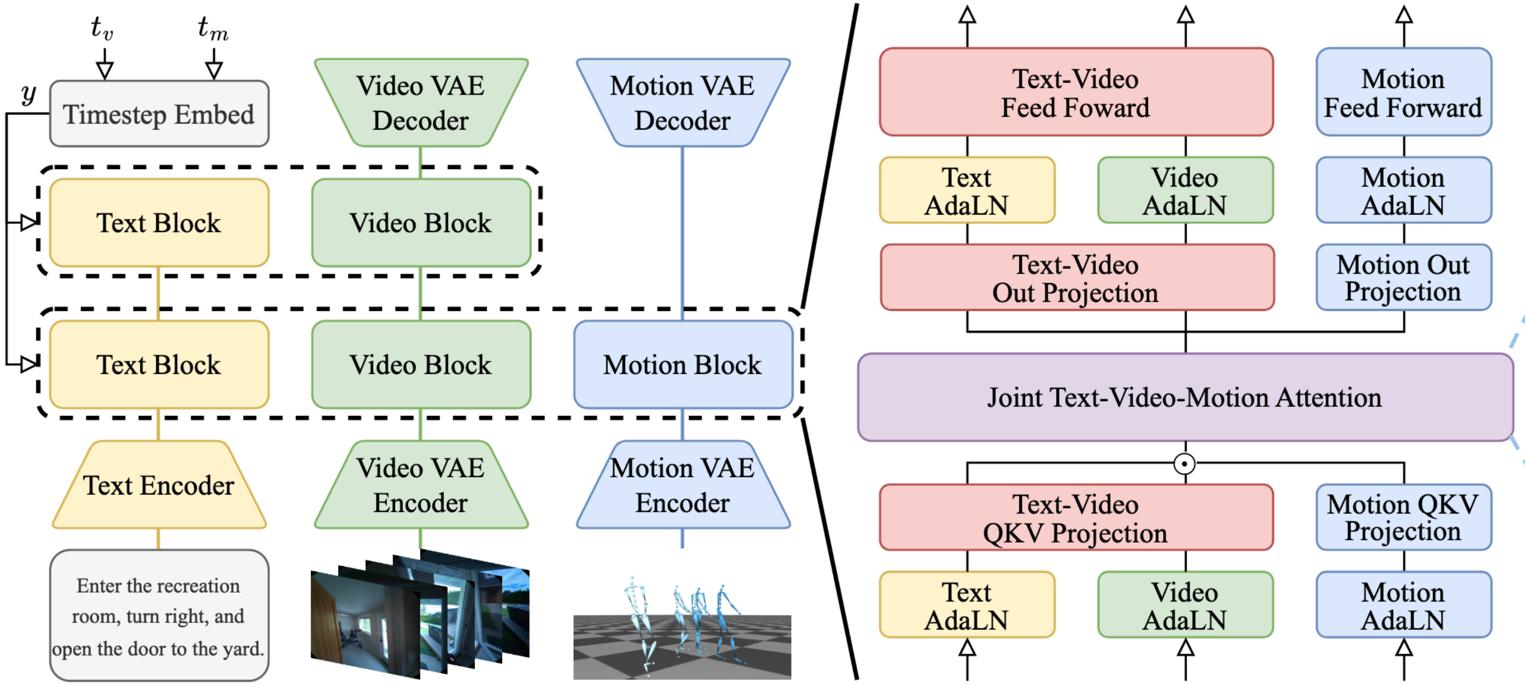
Challenges

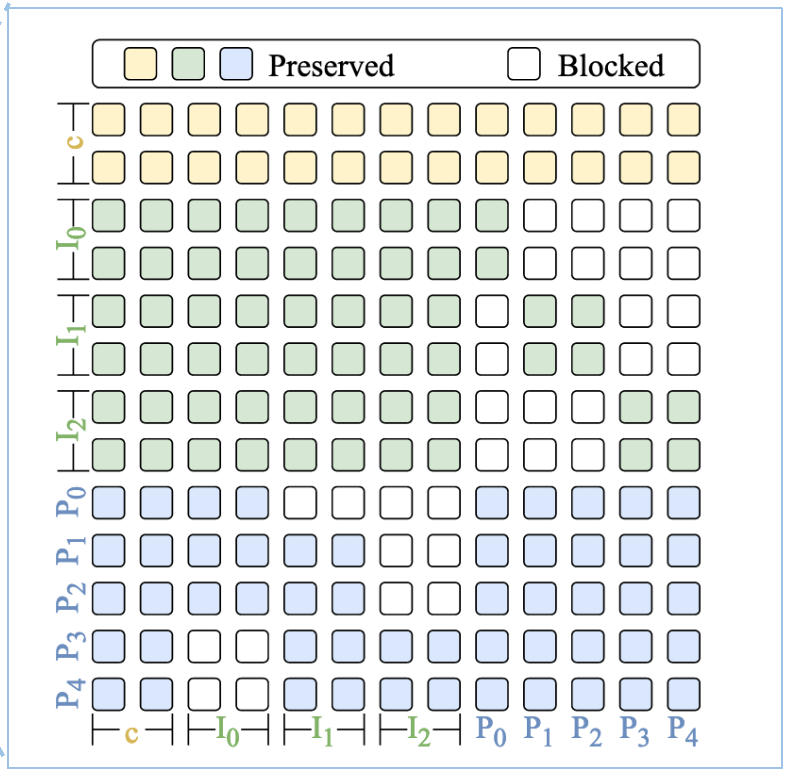
- **Viewpoint Alignment**
  - Existing video generators rely on preset camera parameters, unsuitable for egocentric views.
  - Traditional Motion representations centered on the pelvis cannot accurately align with egocentric viewpoints.

- **Causal Interaction**
  - Each visual frame provides **spatial context** guiding human actions *(e.g., seeing a door handle → reaching out)*.
  - Newly generated actions, in turn, **alter subsequent visual observations** *(e.g., opening the door changes scene layout and camera view)*.
  - Modeling this **observation–action causal loop** is essential for temporal coherence and realism.

EgoTwin Framework

Text-Video-Motion Joint Training





$$\mathcal{L}_{\text{DiT}} = \mathbb{E}_{\epsilon_v,\epsilon_m,c,t_v,t_m}\left[\left\|\epsilon_v - \epsilon_\theta^v(z_v^{t_v}, z_m^{t_m}, c, t_v, t_m)\right\|_2^2 + \left\|\epsilon_m - \epsilon_\theta^m(z_m^{t_m}, z_v^{t_v}, c, t_m, t_v)\right\|_2^2\right]$$

Bidirectional Causal Attention Mechanism

Quantitative Results

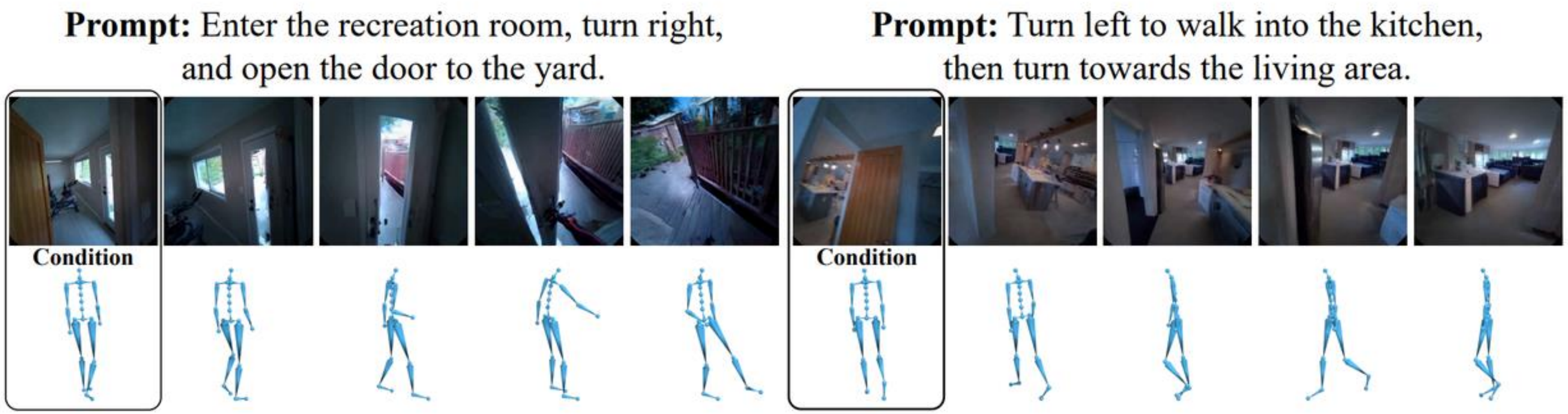
Method	Video Quality			Motion Quality			Video-Motion Consistency		
	I-FID ↓	FVD ↓	CLIP-SIM ↑	M-FID ↓	R-Prec ↑	MM-Dist ↓	TransErr ↓	RotErr ↓	HandScore ↑
VidMLD	157.86	1547.28	25.58	45.09	0.47	19.12	1.28	1.53	0.36
EgoTwin	98.17	1033.52	27.34	41.80	0.62	15.05	0.67	0.46	0.81

Main Results

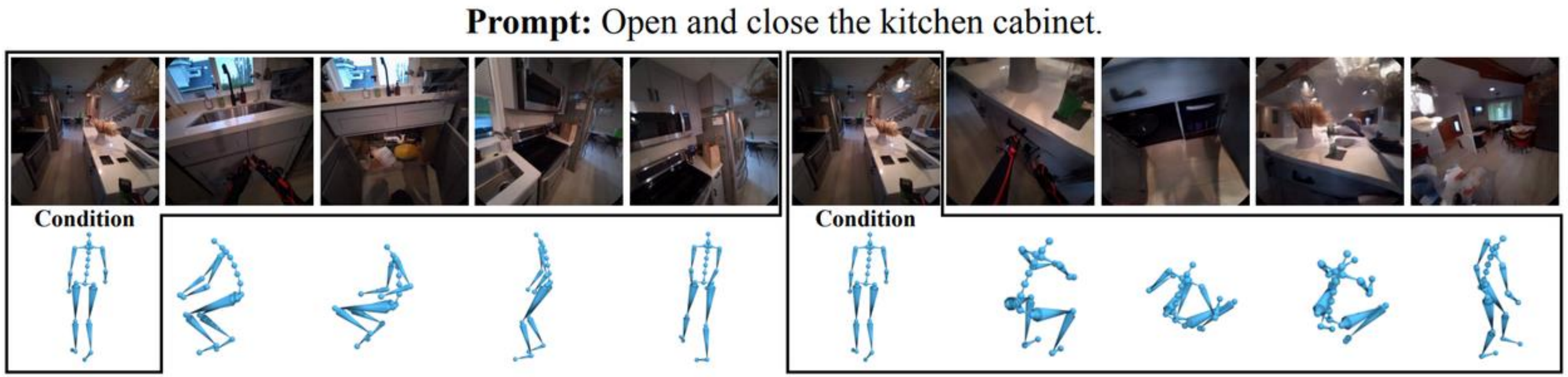
Variant	Video Quality			Motion Quality			Video-Motion Consistency		
	I-FID ↓	FVD ↓	CLIP-SIM ↑	M-FID ↓	R-Prec ↑	MM-Dist ↓	TransErr ↓	RotErr ↓	HandScore ↑
w/o MR	134.27	1356.81	26.36	43.65	0.56	17.31	0.96	1.22	0.44
w/o IM	117.54	1237.58	27.10	44.01	0.59	15.87	0.85	0.89	0.57
w/o AD	109.73	1124.19	26.91	42.58	0.53	16.48	0.74	0.62	0.73
EgoTwin	98.17	1033.52	27.34	41.80	0.62	15.05	0.67	0.46	0.81

Ablation Study

Qualitative Results



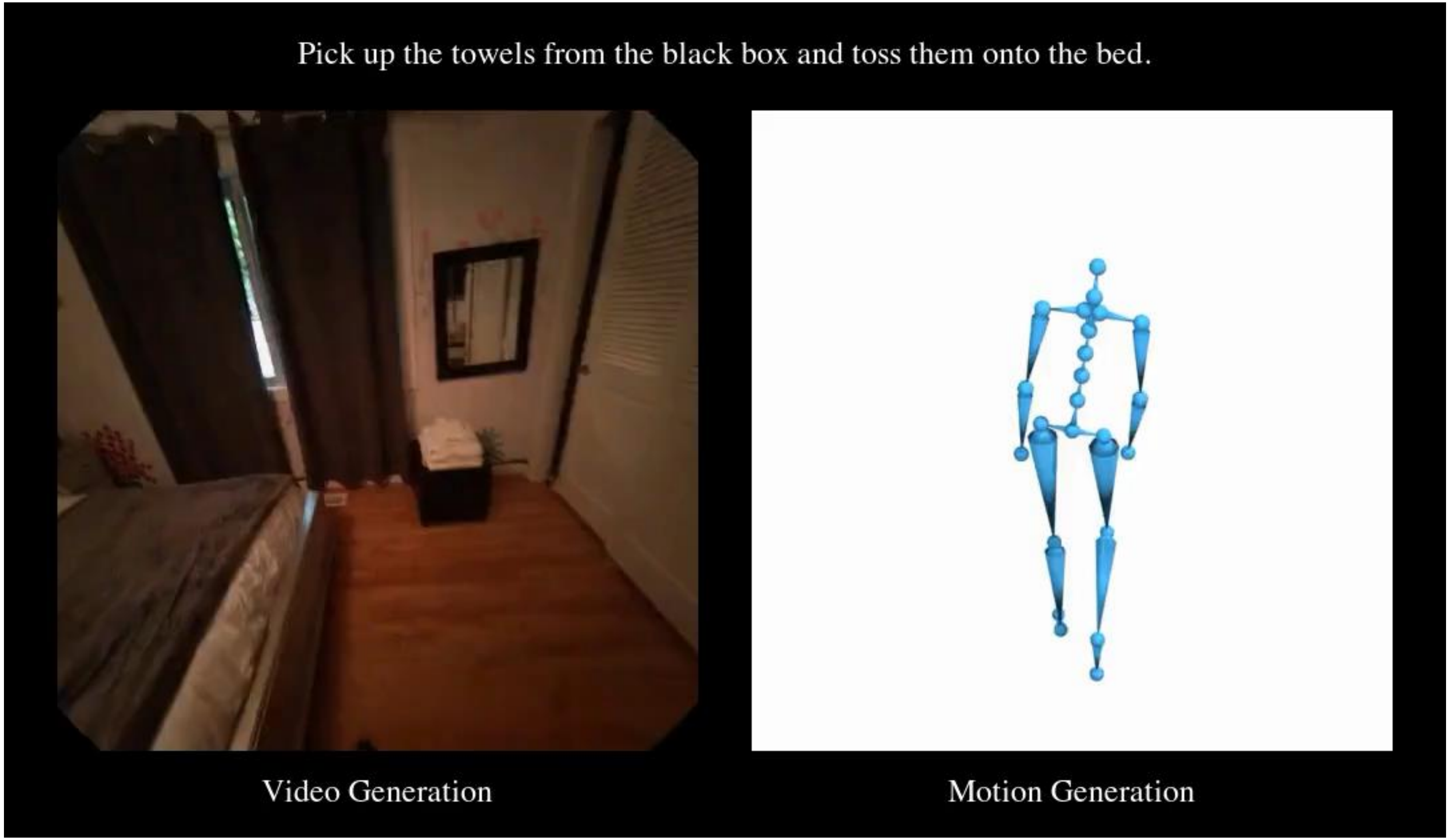
Text-to-Motion & Video



Text & Motion-to-Video

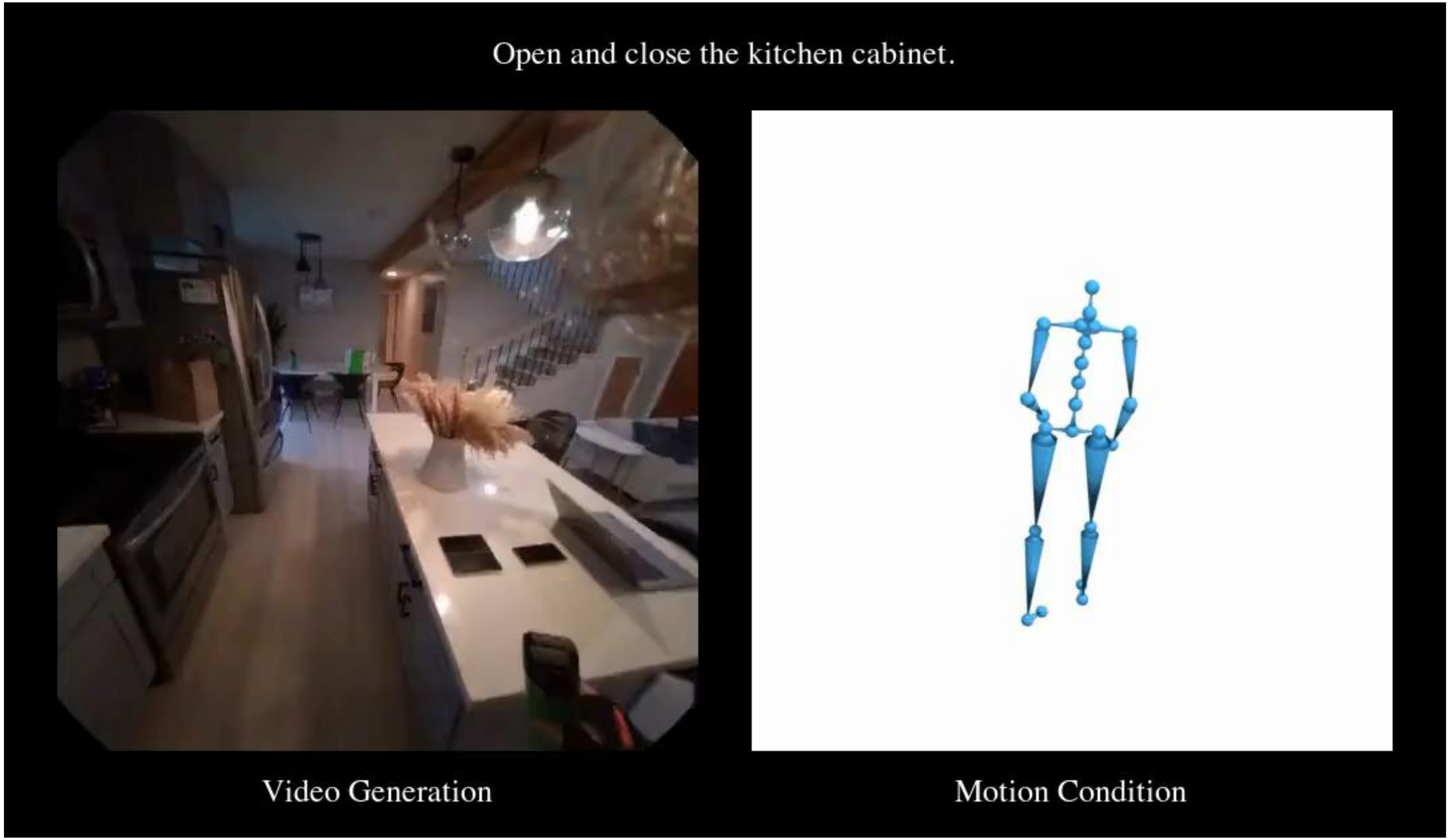
Text & Video-to-Motion

Demo: Text-to-Motion&Video

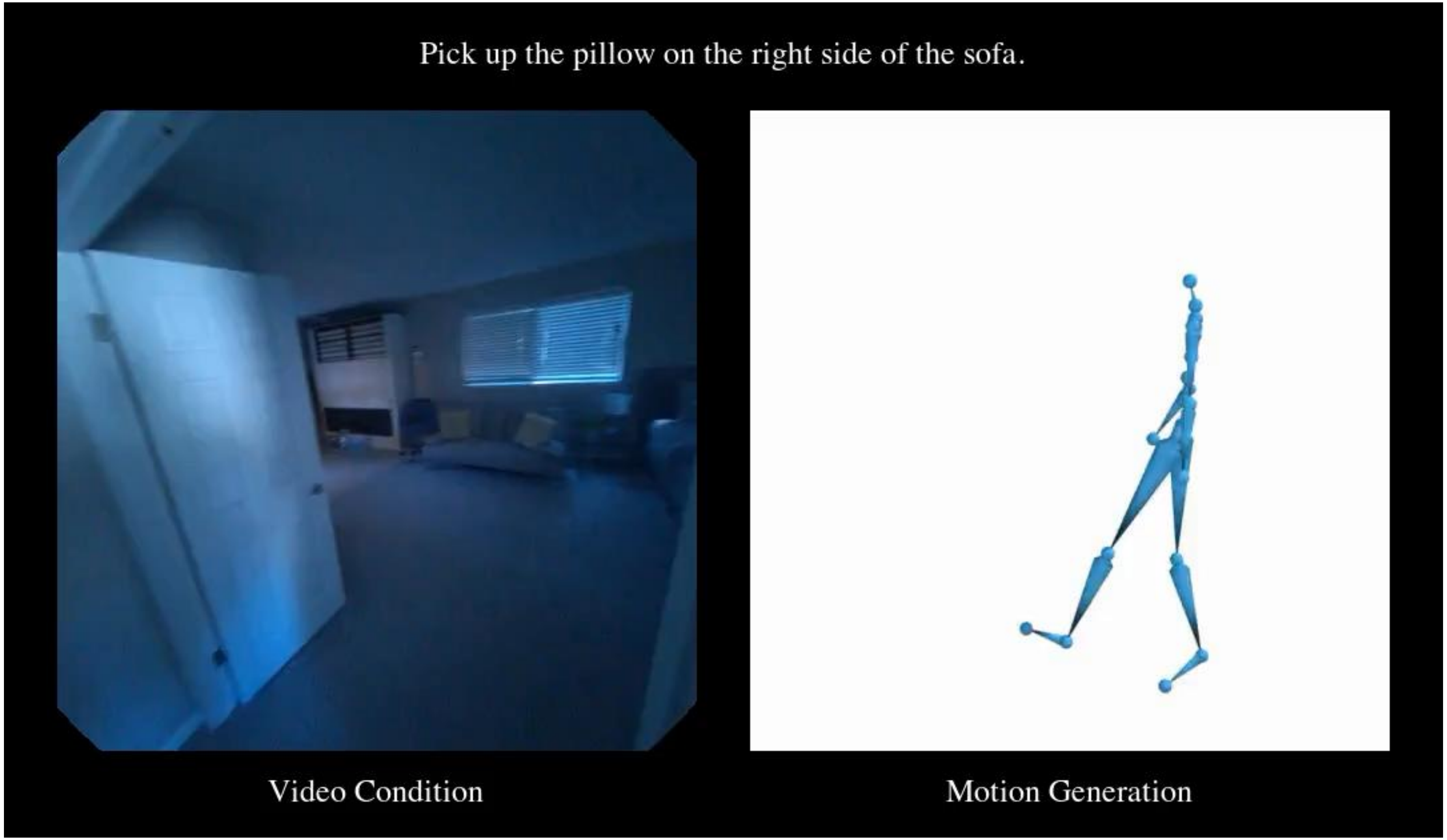


Demo: Text-to-Motion&Video

Demo: Text&Motion-to-Video



Demo: Text&Video-to-Motion



From Seeing to Acting

Perception & Understanding

Learn to see and reason from the first person

Action & Embodiment

Learn to move and act as you see

cognitive foundation of egocentric understanding

egolife

ego-r1

Ego-R1

embodied simulation

From Seeing to Acting

Perception & Understanding

Learn to see and reason from the first person

Action & Embodiment

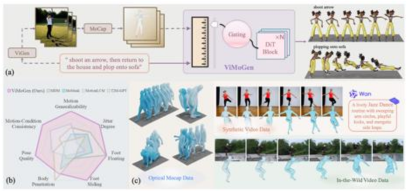
Learn to move and act as you see

cognitive foundation of egocentric understanding

egolife

ego-r1 Ego-R1

embodied simulation and generalizable motion intelligence



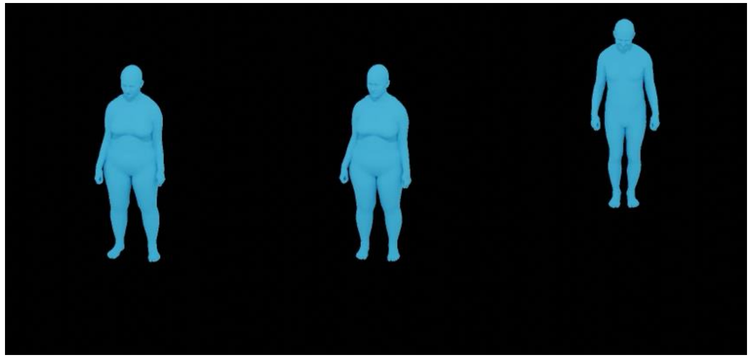
Action & Embodiment
Learn to move and act as you see - Generalization

ViMoGen: The Quest for Generalizable Human Motion Generation

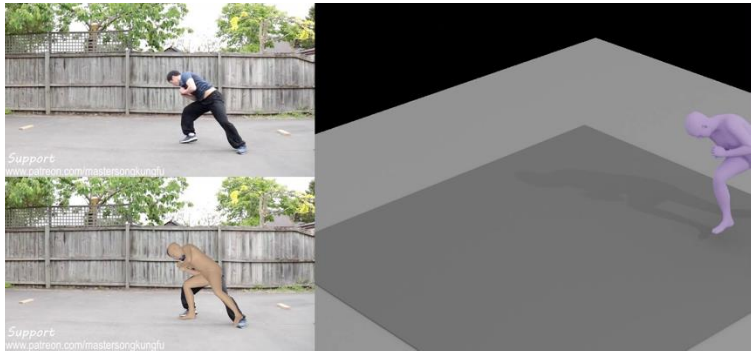
Data, Model, and Evaluation

Challenges Towards Generalizable Motion Generation

- Data Scarcity
  - **Optical MoCap Dataset**
    - Limited in scale and semantic diversity
  - **Web Video-based Datasets**
    - Compromise motion quality
    - Exhibit semantic biases
- Transfer Knowledge from Other Modalities
  - Video generation models have rich world knowledge
  - Have limited motion quality and robustness.
- Evaluation Benchmark
  - Lack of a benchmark for comprehensive evaluation of motion generation algorithms, with a particular emphasis on **generalization** capability



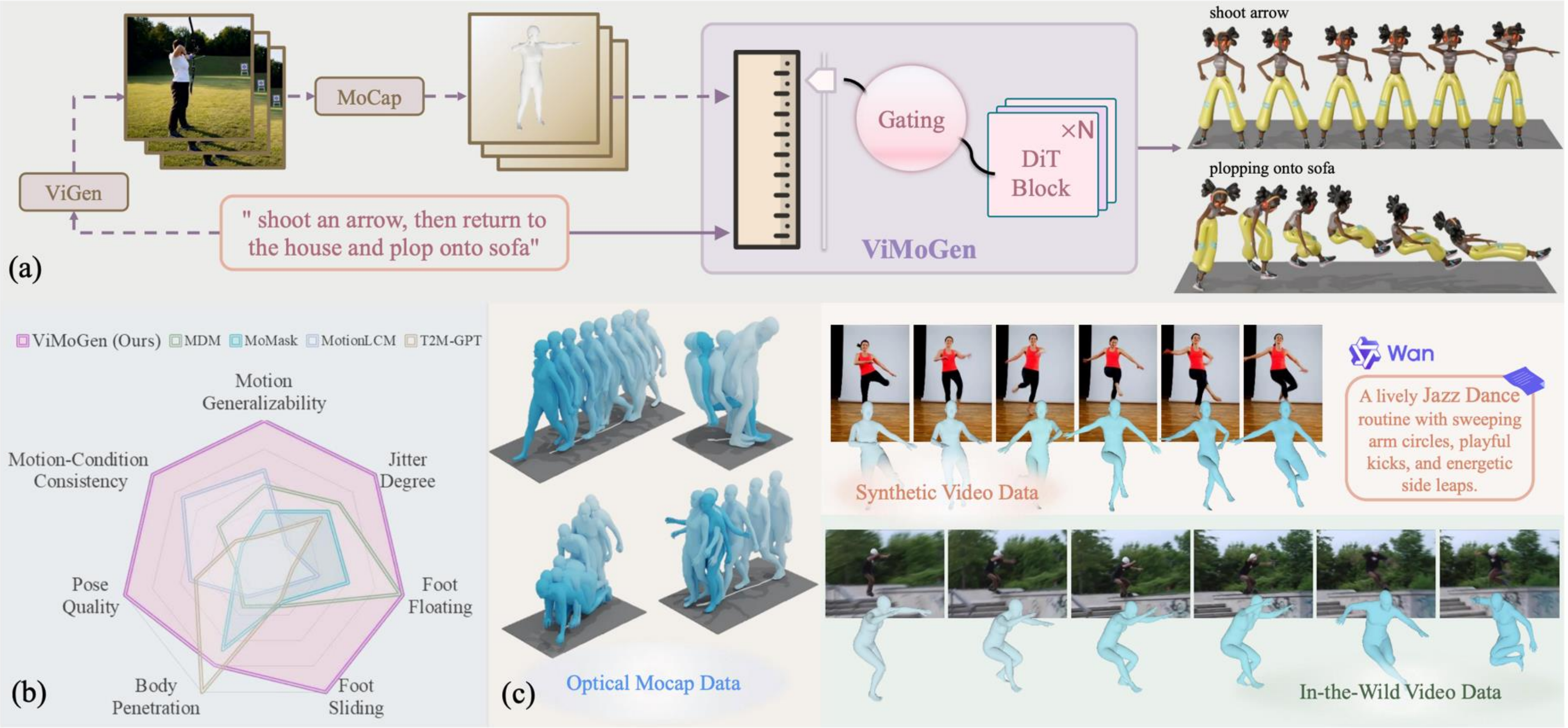
MoCap Dataset



Web Video-based Dataset

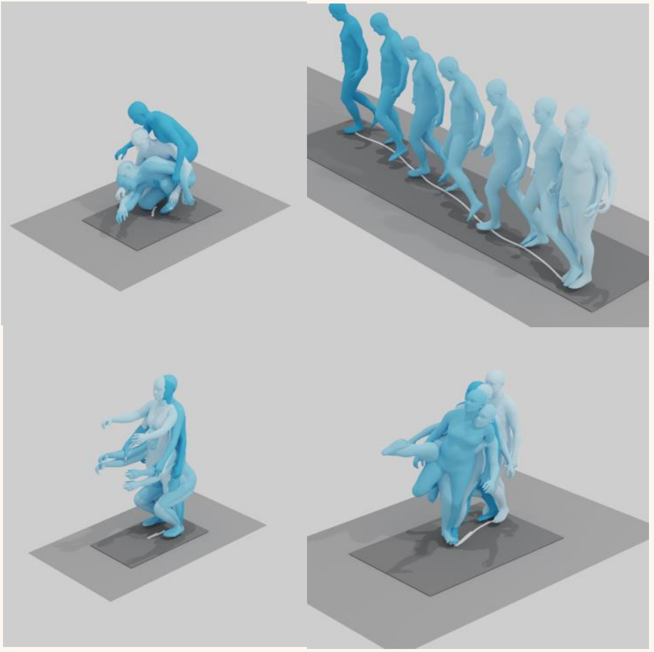
The Quest for Generalizable Motion Generation

We address generalization by focusing on three fundamental components: Data, Model, and Evaluation

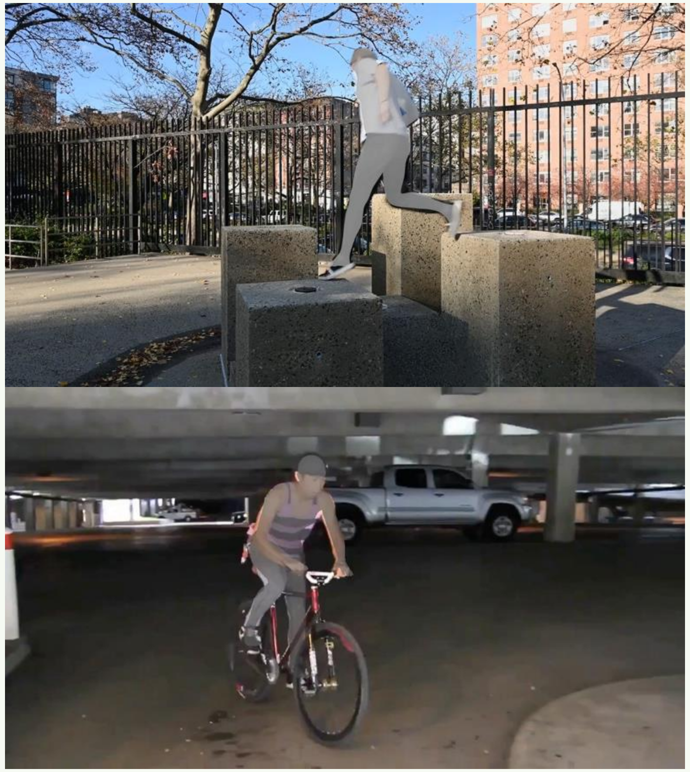


ViMoGen-228K: a Large-scale, Diverse dataset

(a) Unify **30 Mocap dataset** and augment them with **text** captions.



(b) Collect **millions** of **web-videos**, annotate, and select 1% high-quality **motions**.



(c) Construct semantically rich action prompts, **generate** videos, and annotate **motion** labels.

Vocabulary Construction, Prompts Expansion

↓

SOTA Video Generation Model

↓

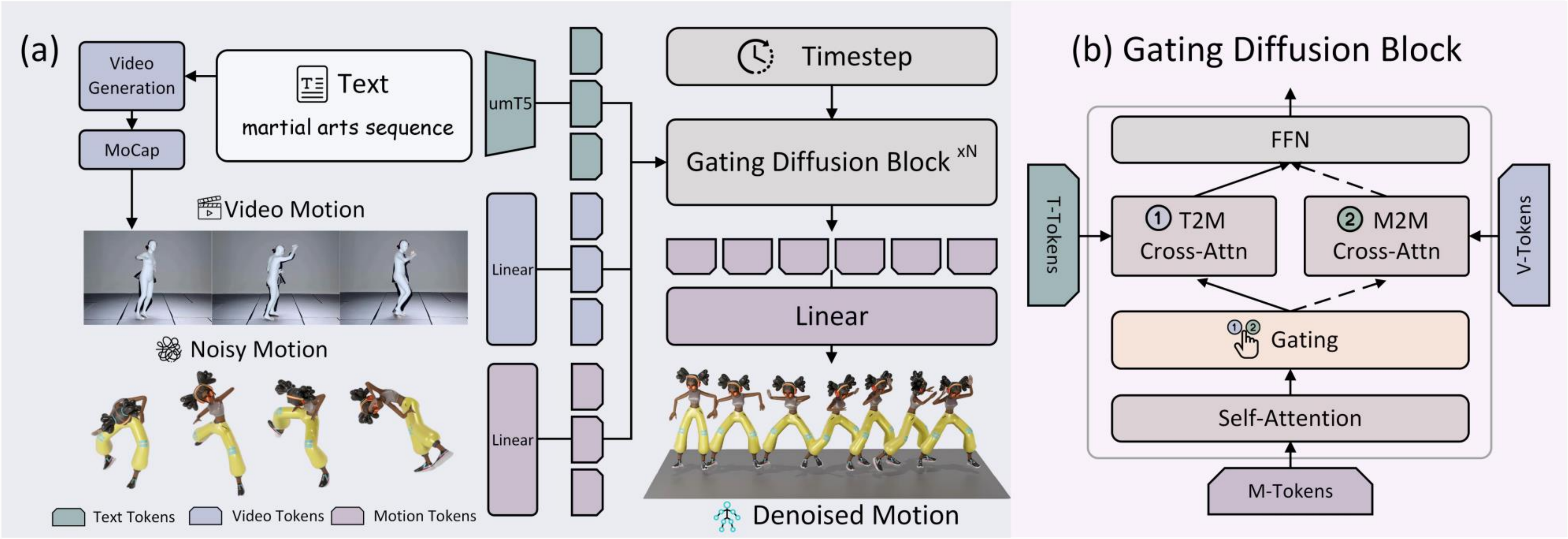
↓

Motion Annotation, Filtering, and Refinement

↓

ViMoGen: Unifying Video and Motion Generation Model Prior

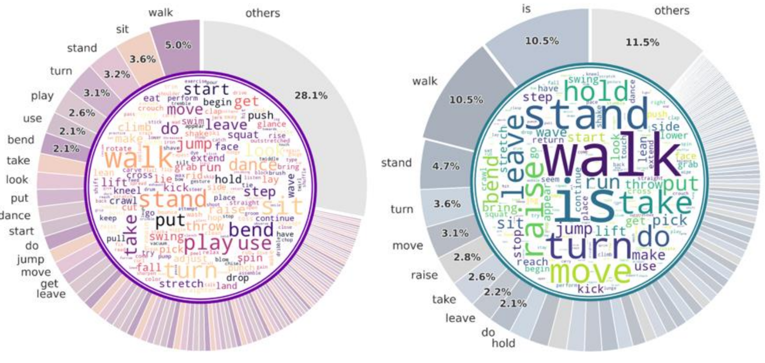
- MoGen Model: High-quality motion but poor generalization ability.
- ViGen Model: Good generalization but unsatisfied motion quality
- ViMoGen: Combine the semantic richness of video models with the high fidelity of motion-specific synthesis



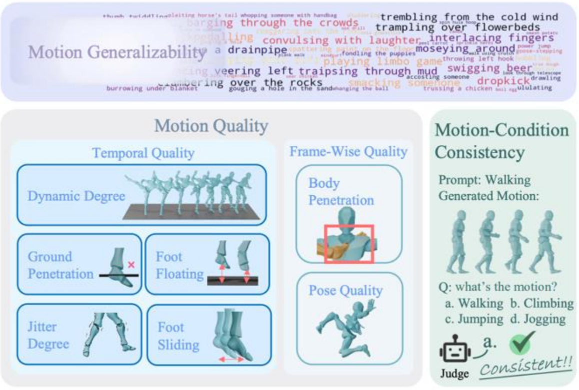
MBench: Hierarchical Benchmark with Multifaceted Assessment

Compared to existing benchmark, MBench featured with:

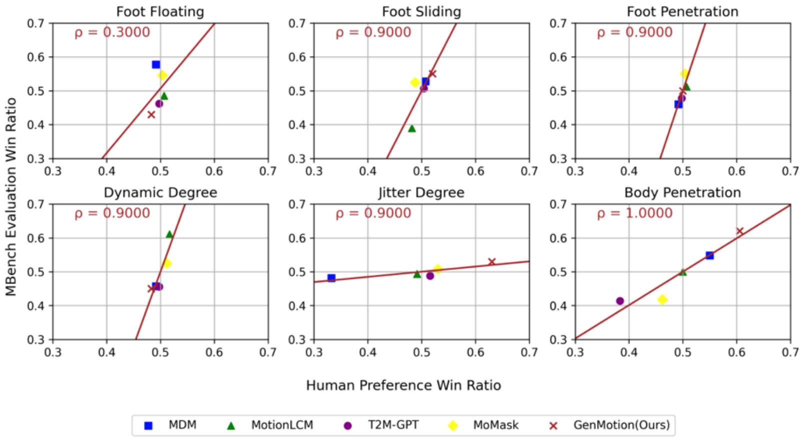
- More balanced distribution, semantically diverse prompts
- Granular and multifaceted assessment with nine dimensions
- Highly aligned with human perception



(a) Top 100 Verbs (Left: MBench. Right: HumanML3D)

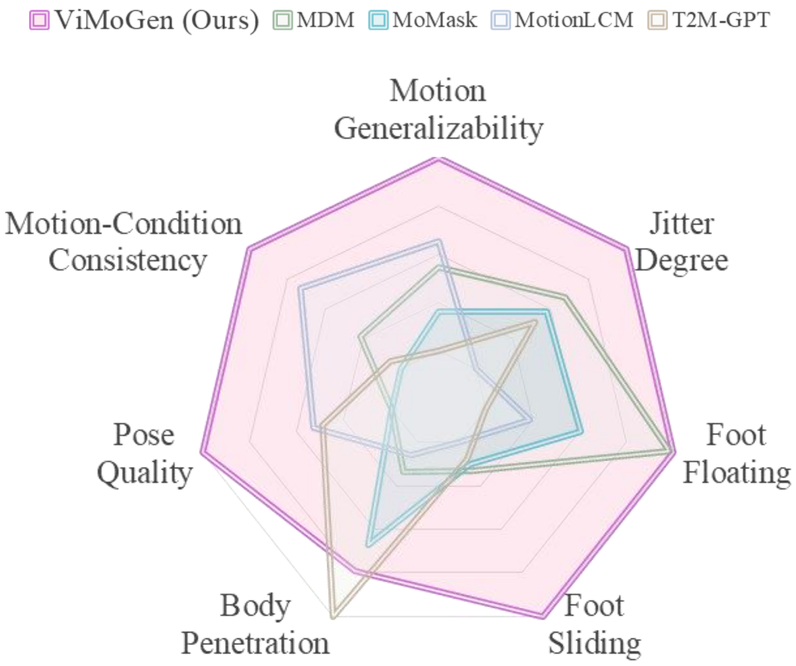


(b) Evaluation Dimensions of MBench



Performance of ViMoGen

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

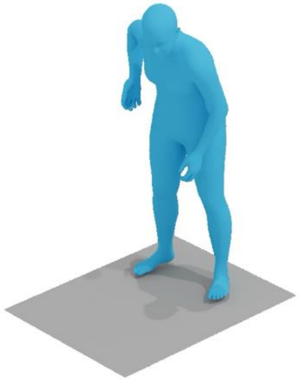


Methods	R Precision↑ Top 1	R Precision↑ Top 2	R Precision↑ Top 3	FID↓	MultiModal Dist↓	MultiModality↑
TM2T (Guo et al., 2022c)	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.002}$	$1.501^{\pm.017}$	$3.467^{\pm.011}$	$2.424^{\pm.093}$
T2M (Guo et al., 2022b)	$0.455^{\pm.003}$	$0.636^{\pm.003}$	$0.736^{\pm.002}$	$1.087^{\pm.021}$	$3.347^{\pm.008}$	$2.219^{\pm.074}$
MDM (Tevet et al., 2023)	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$\mathbf{2.799}^{\pm.072}$
MotionDiffuse (Zhang et al., 2024b)	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$1.553^{\pm.042}$
T2M-GPT (Zhang et al., 2023a)	$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$	$0.141^{\pm.005}$	$3.121^{\pm.009}$	$1.831^{\pm.048}$
MoMask (Guo et al., 2024)	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$\mathbf{0.045}^{\pm.002}$	$2.958^{\pm.008}$	$1.241^{\pm.040}$
Motion-LCM (Dai et al., 2024)	$0.502^{\pm.003}$	$0.698^{\pm.002}$	$0.798^{\pm.002}$	$0.304^{\pm.012}$	$3.012^{\pm.007}$	$2.259^{\pm.092}$
MLD (Chen et al., 2023)	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$2.413^{\pm.079}$
MLD + ViMoGen-light (Ours)	$\mathbf{0.542}^{\pm.003}$	$\mathbf{0.733}^{\pm.002}$	$\mathbf{0.825}^{\pm.002}$	$0.114^{\pm.005}$	$\mathbf{2.826}^{\pm.007}$	$1.973^{\pm.074}$

ViMoGen exhibits excellent generalization capability while remaining comparable motion quality on MBench.

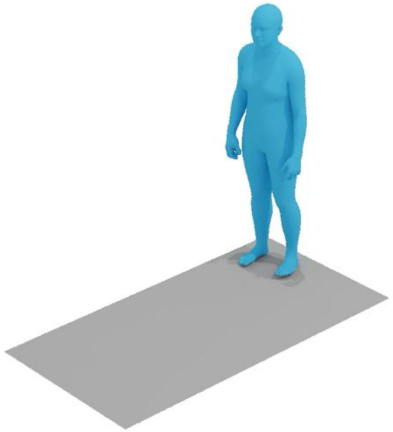
Our distillated version ViMoGen-light also exhibits completive performance on traditional benchmark.

Performance of ViMoGen

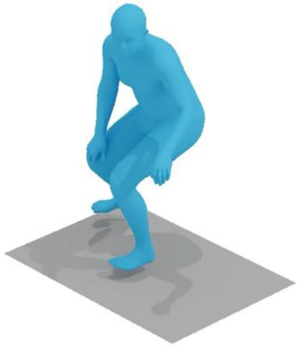


dig out dirt

wear headphones



climbing ladder



surfing

What are Crucial for Generalization Ability?

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

- Knowledge from ViGen model improves generalizability.

Branch Selection	Motion Condition Consistency	Motion Generalizability	Jitter Degree	Foot Sliding
Video Generation Baseline	0.51	0.58	0.0193	0.0161
T2M Only	0.46	0.54	0.0111	**0.0039**
M2M Only	0.51	0.59	0.0145	0.0113
Adaptive Gating	**0.53**	**0.68**	**0.0108**	0.0064

- Compact but semantically rich synthetic data is critical.

Training Datasets	Motion Clip Number	Motion Condition Consistency	Motion Generalizability	Foot Sliding
HumanML3D	89k	0.41	0.44	**0.0032**
+ Other Optical Mocap Data	83k	0.44	0.48	0.0033
+ Visual Mocap Data	42k	0.43	0.50	0.0042
+ Synthetic Video Data	14k	**0.47**	**0.55**	0.0051

- A powerful text encoder is needed.

Text Encoder	Motion Condition Consistency	Motion Generalizability	Foot Sliding	Body Penetration
CLIP	0.32	0.35	**0.0023**	1.39
T5-XXL	**0.41**	0.44	0.0032	**1.05**
MLLM	0.38	**0.46**	0.0032	1.51

- Text augmentation during training helps.

Training Text Style	Testing Text Style	Motion Condition Consistency	Motion Generalizability	Foot Sliding
Motion	Motion	0.36	0.40	0.0032
Motion	Video	0.32	0.39	**0.0031**
Video	Motion	**0.43**	**0.48**	0.0033
Video	Video	0.41	0.44	0.0032

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

From Seeing to Acting

Perception & Understanding → **Action & Embodiment**

Learn to see and reason from the first person

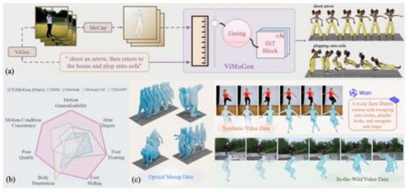
Learn to move and act as you see

cognitive foundation of egocentric understanding

egolife

ego-r1 Ego-R1

embodied simulation and generalizable motion intelligence



→ see, act, and learn from the first person, just like us

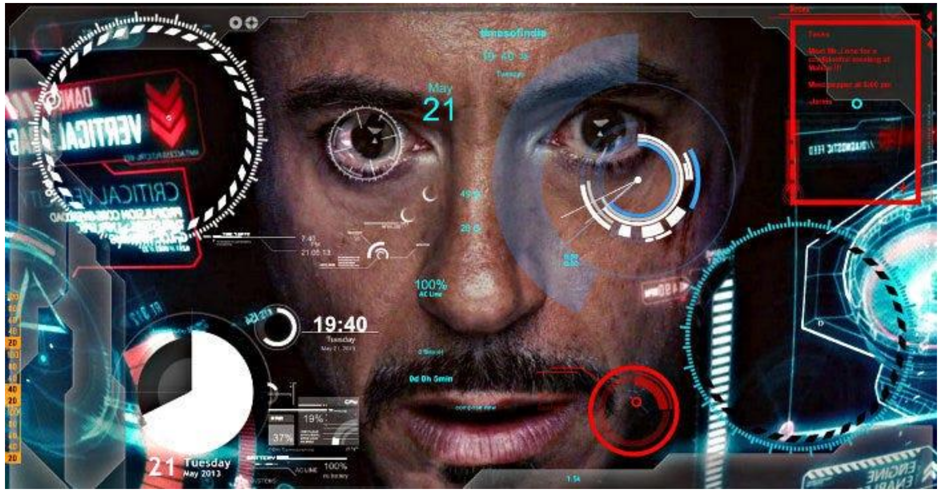
Dragon Ball (1984)

The Terminator (1984)

Mission Impossible 2 (2000)

Detective Conan (1994)

Spy Kids (2001)



Iron Man (2008)

Spider-Man (2019)

Thank You

Ziwei Liu 刘子纬
Nanyang Technological University