动作捕捉与动作生成的相遇还有多远?

Towards the Union of Motion Capture and Motion Generation

Zhongang Cai 蔡中昂

Ph.D. Student
S-Lab, Nanyang Technological University

19 Apr 2024

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

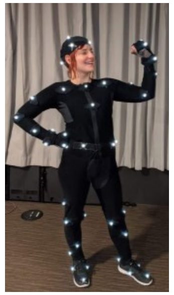
Background

Movies

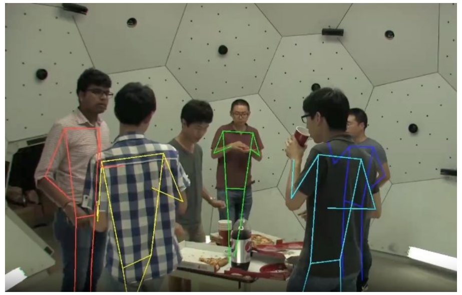
Games

3D Cartoon / Anime

VTubers


Optical (Marker-based) MoCap

IMU-based MoCap


Vision-based (Markerless) MoCap

Overview

* Equal Contributions, † Corresponding Author

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

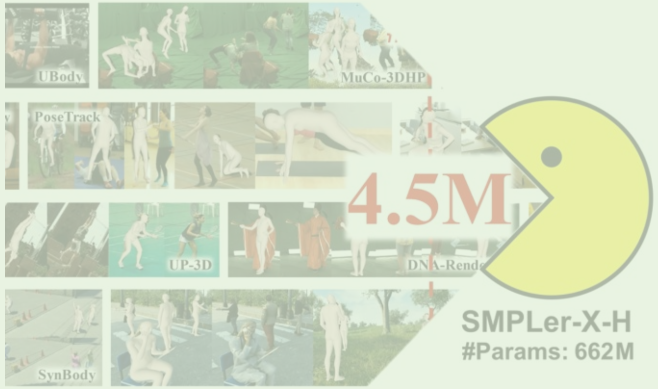
Data

Cai* et al., *HuMMan*, 2022

Cai* et al., *GTA-Human*, 2021

Yang*, **Cai*** et al., *SynBody*, 2023

Algorithms



Cai* et al., *SMPLer-X*, 2023

Sun*, Wang*, **Cai**† et al., *AiOS*, 2023

Yin*, **Cai*** et al., *WHAC*, 2024

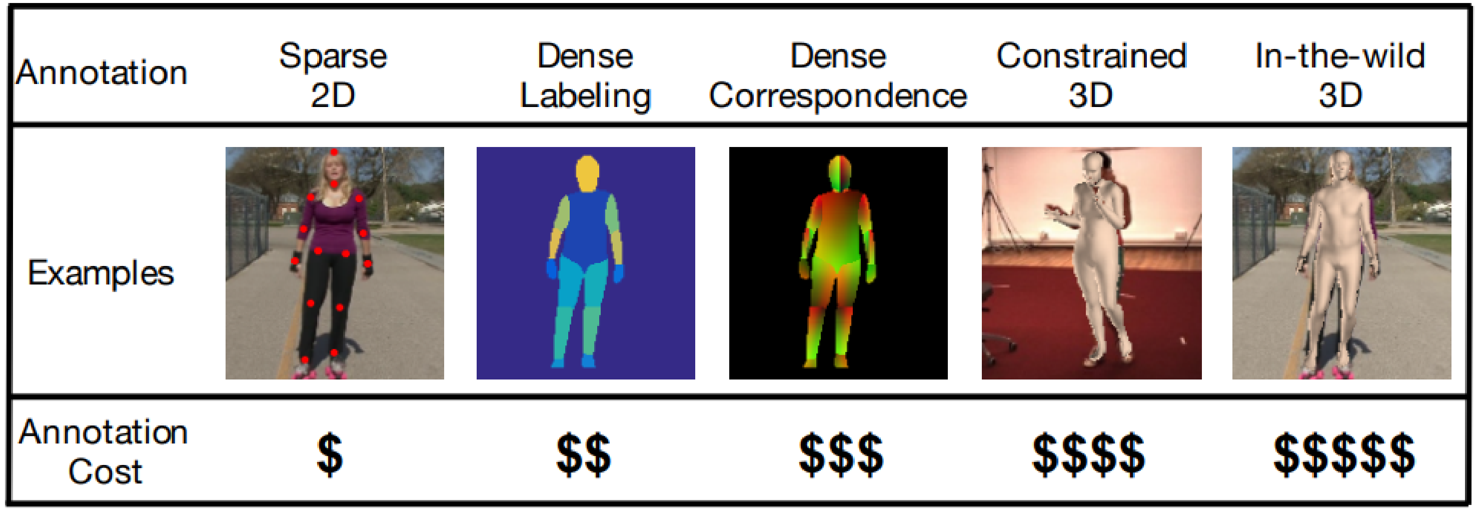
Applications

Zhang*, **Cai*** et al., *MotionDiffuse*, 2022

Qing, **Cai** et al., *Story-to-Motion*, 2023

Cai* et al., Digital Life Project, 2023

Data | 3D Human Data is Expensive



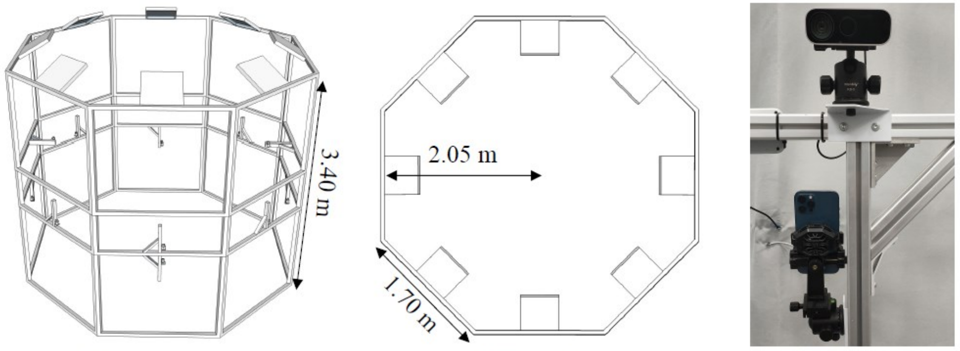
Annotation	Sparse 2D	Dense Labeling	Dense Correspondence	Constrained 3D	In-the-wild 3D
Examples					
Annotation Cost	$	$$	$$$	$$$$	$$$$$

In-the-Wild 3D Human Data is Expensive [1]

[1] Y. Rong et al., Delving deep into hybrid annotations for 3d human recovery in the wild, ICCV 2019

Data | Reduce Setup Cost!

HuMMan v1.0: Recon/MoGen/Point Subsets



a) Perspective view

b) Top view

c) Sensors

Microsoft
Azure Kinect

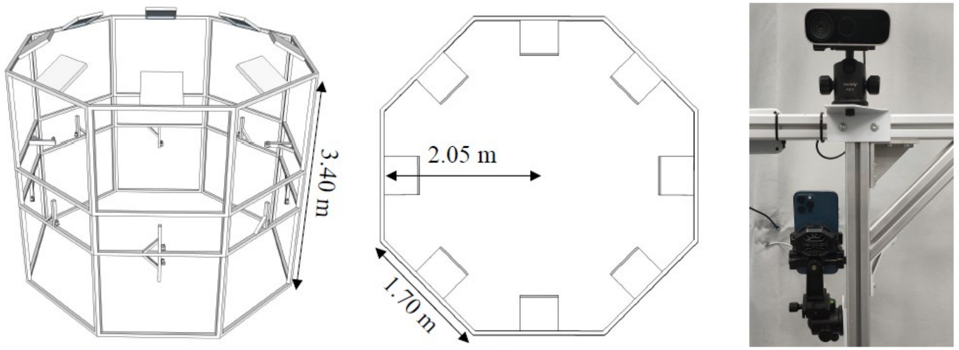
iPhone 12 Pro Max
(with LiDAR)

High-resolution
Scanner

Low-cost Setup with Commercial Sensors

Z Cai*, D Ren*, A Zeng*, Z Lin*, T Yu*, W Wang*, X Fan, Y Gao, Y Yu, L Pan, F Hong, M Zhang, CC Loy, L Yang, Z Liu. *HuMMan: Multi-Modal 4D Human Dataset for Versatile Sensing and Modeling*. European Conference on Computer Vision (ECCV) 2022, *Oral* Presentation.

Data | Reduce Setup Cost!



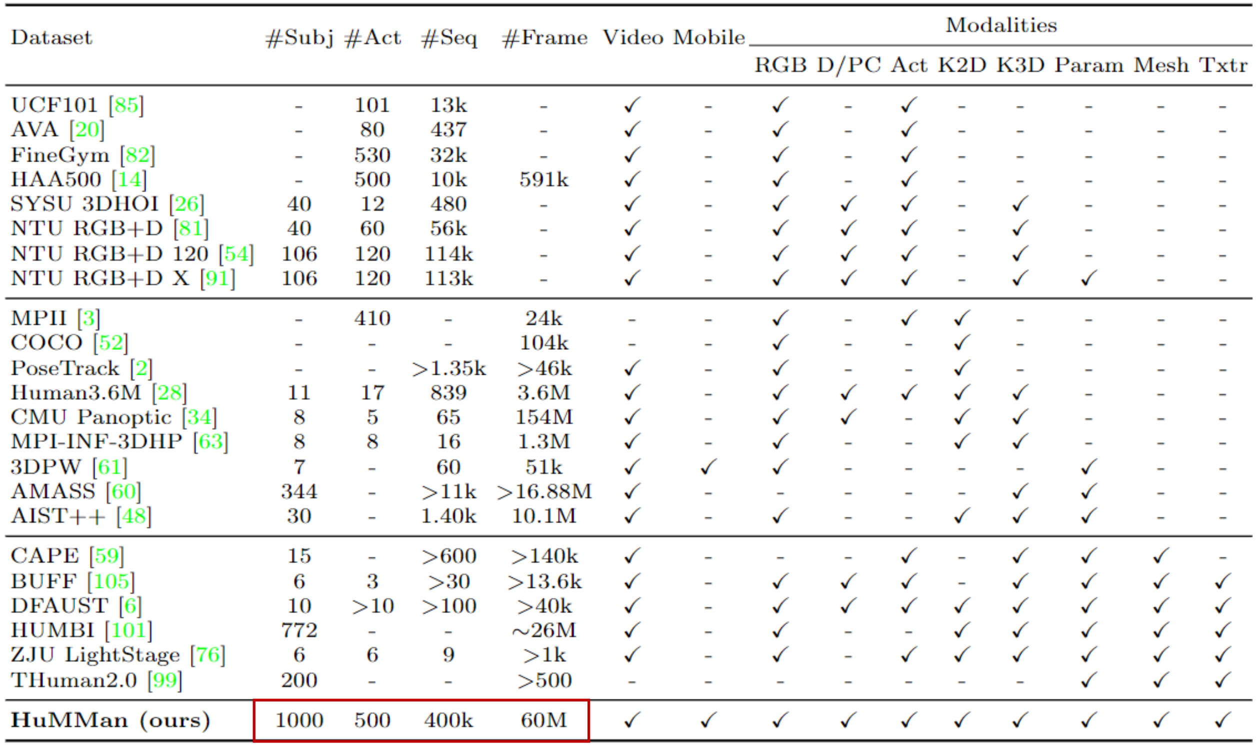
a) Perspective view b) Top view c) Sensors

Microsoft Azure Kinect

iPhone 12 Pro Max (with LiDAR)

High-resolution Scanner

Low-cost Setup with Commercial Sensors



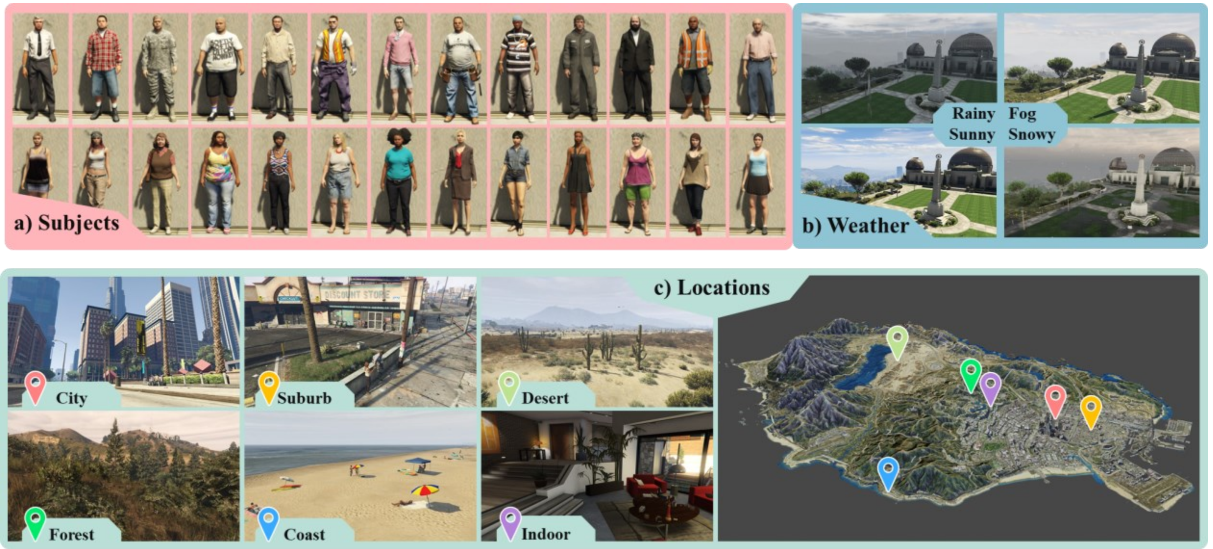
Dataset	#Subj	#Act	#Seq	#Frame	Video	Mobile	Modalities							
							RGB	D/PC	Act	K2D	K3D	Param	Mesh	Txtr
UCF101 [85]	-	101	13k	-	✓	-	✓	-	✓	-	-	-	-	-
AVA [20]	-	80	437	-	✓	-	✓	-	✓	-	-	-	-	-
FineGym [82]	-	530	32k	-	✓	-	✓	-	✓	-	-	-	-	-
HAA500 [14]	-	500	10k	591k	✓	-	✓	-	✓	-	-	-	-	-
SYSU 3DHOI [26]	40	12	480	-	✓	-	✓	✓	✓	-	✓	-	-	-
NTU RGB+D [81]	40	60	56k	-	✓	-	✓	✓	✓	-	✓	-	-	-
NTU RGB+D 120 [54]	106	120	114k	-	✓	-	✓	✓	✓	-	✓	-	-	-
NTU RGB+D X [91]	106	120	113k	-	✓	-	✓	✓	✓	-	✓	✓	-	-
MPII [3]	-	410	-	24k	-	-	✓	-	✓	✓	-	-	-	-
COCO [52]	-	-	-	104k	-	-	✓	-	-	✓	-	-	-	-
PoseTrack [2]	-	-	>1.35k	>46k	✓	-	✓	-	-	✓	-	-	-	-
Human3.6M [28]	11	17	839	3.6M	✓	-	✓	✓	✓	✓	✓	-	-	-
CMU Panoptic [34]	8	5	65	154M	✓	-	✓	✓	-	✓	✓	-	-	-
MPI-INF-3DHP [63]	8	8	16	1.3M	✓	-	✓	-	-	✓	✓	-	-	-
3DPW [61]	7	-	60	51k	✓	✓	✓	-	-	-	-	✓	-	-
AMASS [60]	344	-	>11k	>16.88M	✓	-	-	-	-	-	✓	✓	-	-
AIST++ [48]	30	-	1.40k	10.1M	✓	-	✓	-	-	✓	✓	✓	-	-
CAPE [59]	15	-	>600	>140k	✓	-	-	-	✓	-	✓	✓	✓	-
BUFF [105]	6	3	>30	>13.6k	✓	-	✓	✓	✓	-	✓	✓	✓	✓
DFAUST [6]	10	>10	>100	>40k	✓	-	✓	✓	✓	✓	✓	✓	✓	✓
HUMBI [101]	772	-	-	~26M	✓	-	✓	-	-	✓	✓	✓	✓	✓
ZJU LightStage [76]	6	6	9	>1k	✓	-	✓	-	✓	✓	✓	✓	✓	✓
THuman2.0 [99]	200	-	-	>500	-	-	-	-	-	-	-	✓	✓	✓
HuMMan (ours)	1000	500	400k	60M	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Mega-scale Multimodal 4D Human Dataset

Z Cai*, D Ren*, A Zeng*, Z Lin*, T Yu*, W Wang*, X Fan, Y Gao, Y Yu, L Pan, F Hong, M Zhang, CC Loy, L Yang, Z Liu. *HuMMan: Multi-Modal 4D Human Dataset for Versatile Sensing and Modeling*. European Conference on Computer Vision (ECCV) 2022, *Oral* Presentation.

Data | Synthetic Data is Nearly Free!

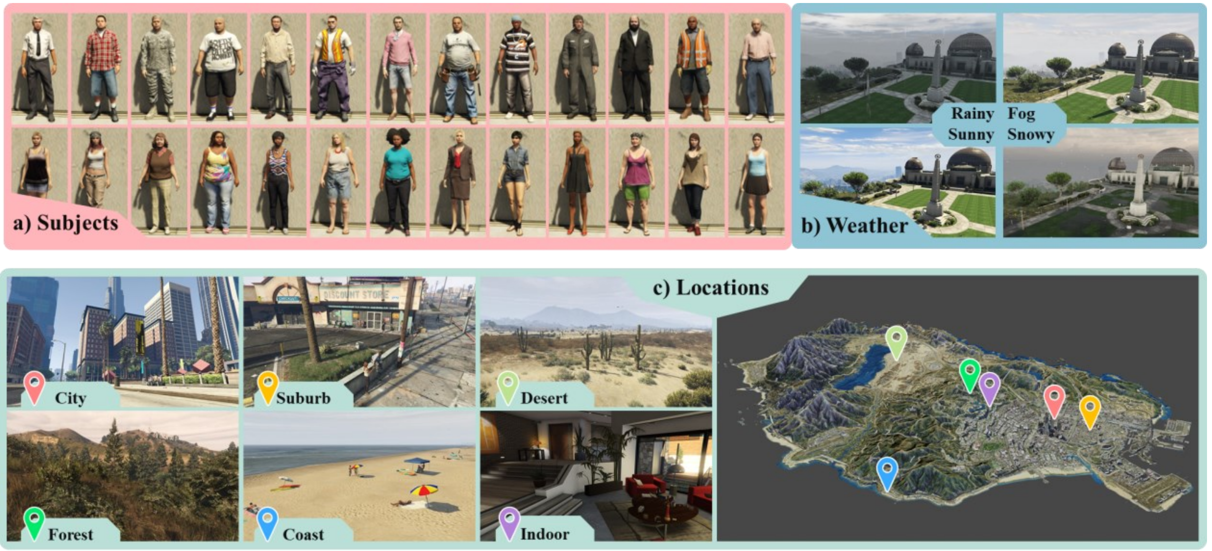
GTA-Human



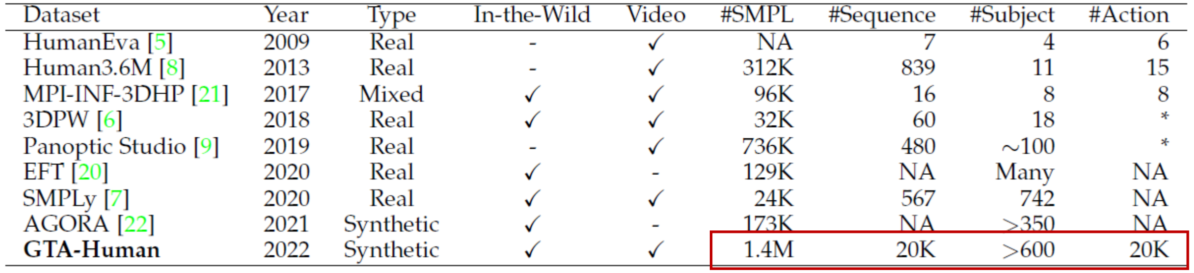
Diverse Data (Subjects, Locations, Weathers, and Light Conditions)

Z Cai*, M Zhang*, J Ren*, C Wei, D Ren, J Li, Z Lin, H Zhao, S Yi, L Yang, CC Loy, Z Liu. *Playing for 3D Human Recovery*. 2022.

Data | Synthetic Data is Nearly Free!



Diverse Data (Subjects, Locations, Weathers, and Light Conditions)



Dataset	Year	Type	In-the-Wild	Video	#SMPL	#Sequence	#Subject	#Action
HumanEva [5]	2009	Real	-	✓	NA	7	4	6
Human3.6M [8]	2013	Real	-	✓	312K	839	11	15
MPI-INF-3DHP [21]	2017	Mixed	✓	✓	96K	16	8	8
3DPW [6]	2018	Real	✓	✓	32K	60	18	*
Panoptic Studio [9]	2019	Real	-	✓	736K	480	~100	*
EFT [20]	2020	Real	✓	-	129K	NA	Many	NA
SMPLy [7]	2020	Real	✓	✓	24K	567	742	NA
AGORA [22]	2021	Synthetic	✓	-	173K	NA	>350	NA
GTA-Human	2022	Synthetic	✓	✓	1.4M	20K	>600	20K

Large-scale Game-playing Synthetic Data

Z Cai*, M Zhang*, J Ren*, C Wei, D Ren, J Li, Z Lin, H Zhao, S Yi, L Yang, CC Loy, Z Liu. *Playing for 3D Human Recovery*. In submission to IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

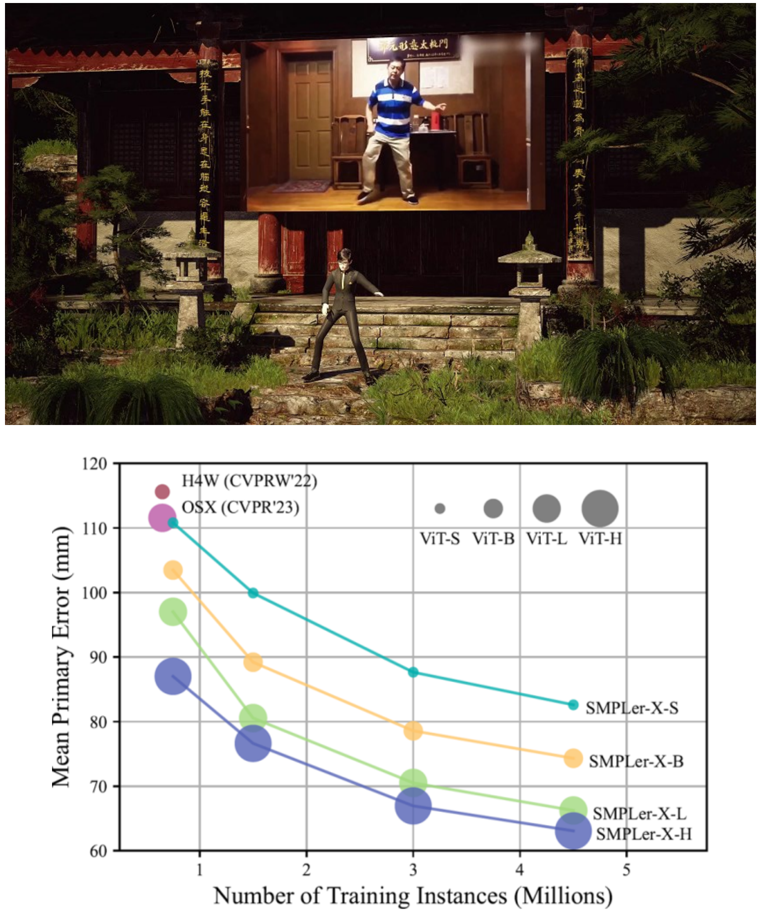
Data | Fully Controllable Synthesis

Unreal Engine-Empowered Synthetic Data Synthesis

Layered Human Model with Procedural Clothing / Accessories

Z Yang*, **Z Cai***, H Mei*, S Liu*, Z Chen*, W Xiao, Y Wei, Z Qing, C Wei, B Dai, W Wu, C Qian, D Lin, Z Liu, L Yang. *SynBody: Synthetic Dataset with Layered Human Models for 3D Human Perception and Modeling.* International Conference on Computer Vision (ICCV) 2023.

Algorithm | Faster – Higher – Stronger

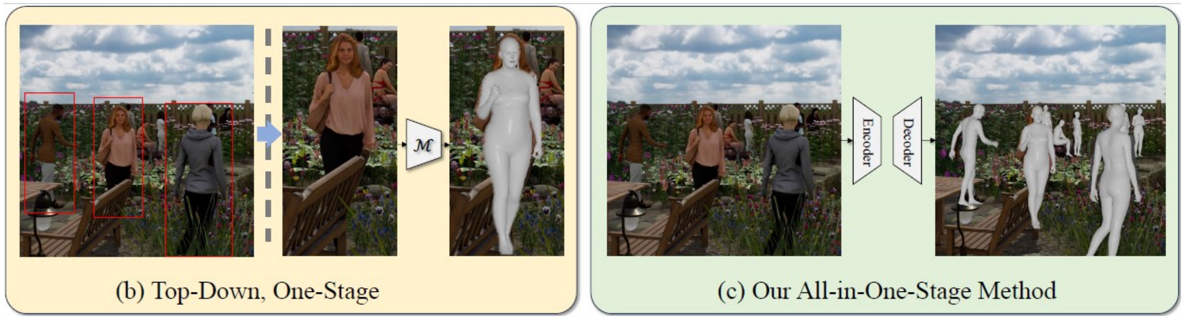


Model & Data Scaling

Animation & Film Making

Z Cai*, W Yin*, A Zeng, C Wei, Q Sun, Y Wang, HE Pang, H Mei, M Zhang, L Zhang, CC Loy, L Yang, Z Liu. *SMPLer-X: Scaling Up Expressive Human Pose and Shape Estimation.* Conference on Neural Information Processing Systems (NeurIPS) 2023.

Algorithm | Faster – Higher – Stronger



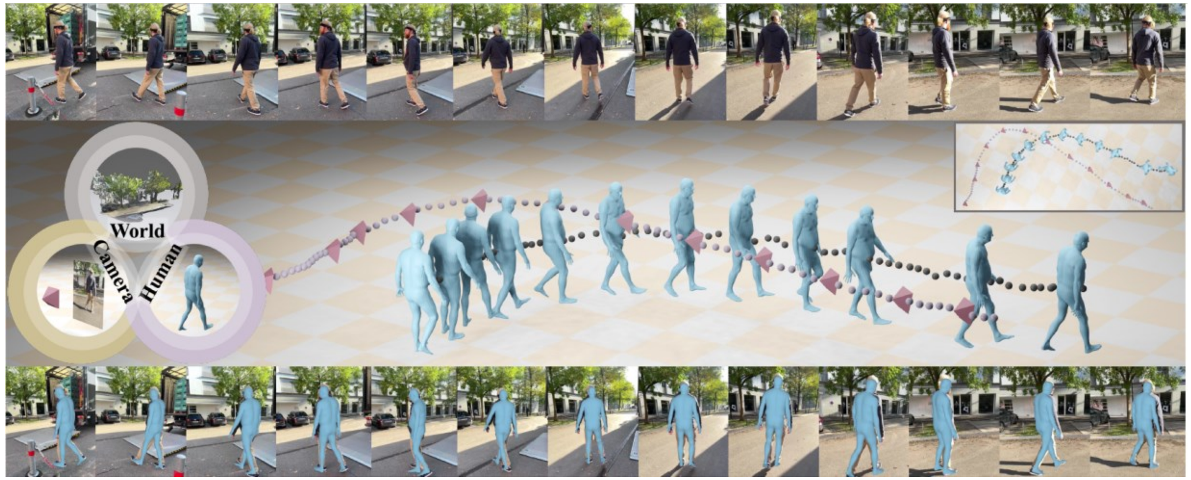
(b) Top-Down, One-Stage

(c) Our All-in-One-Stage Method

猛男舞团IconX bilibili

All-in-One-Stage: Detection + Motion Capture

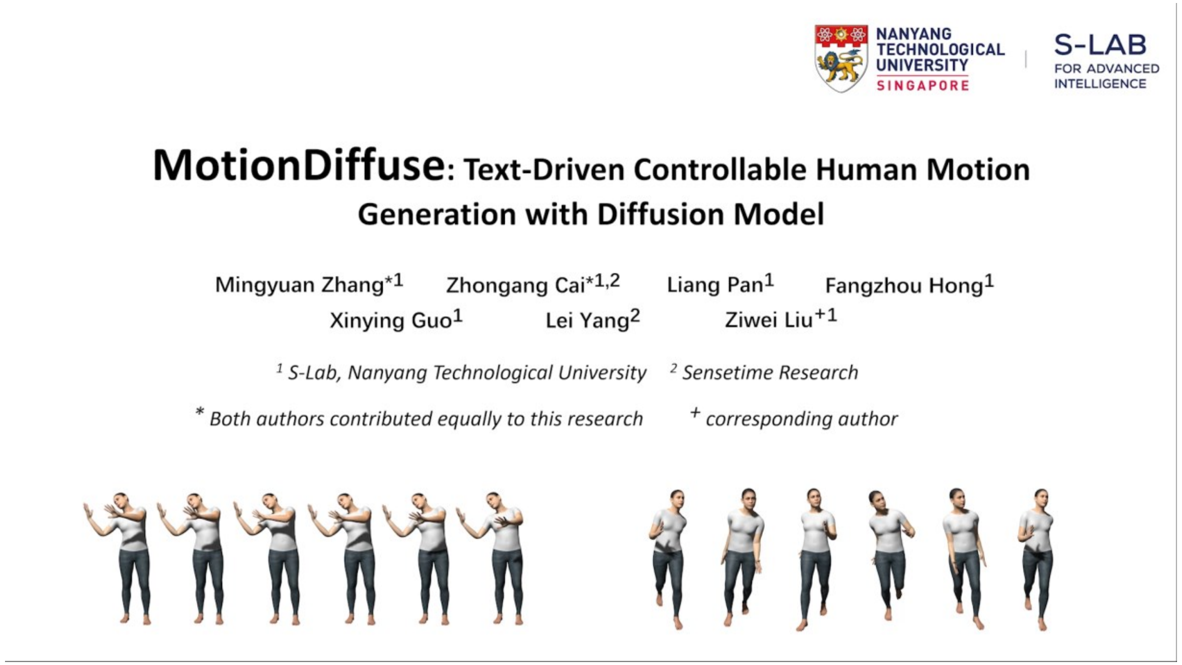
Q Sun*, Y Wang*, A Zeng, W Yin, C Wei, W Wang, H Mei, CS Leung, Z Liu, L Yang, **Z Cai**†. *AiOS: All-in-One-Stage Expressive Human Pose and Shape Estimation*. Conference on Computer Vision and Pattern Recognition (CVPR) 2024.



World-space Motion Capture

W Yin*, **Z Cai***, R Wang, F Wang, C Wei, H Mei, W Xiao, Z Yang, Q Sun, A Yamashita, Z Liu, L Yang. *WHAC: World-grounded Humans and Cameras*. ArXiv, 2024.

Applications | Autonomous Characters



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

MotionDiffuse: Text-Driven Controllable Human Motion Generation with Diffusion Model

Mingyuan Zhang*[1] Zhongang Cai*[1,2] Liang Pan[1] Fangzhou Hong[1]
Xinying Guo[1] Lei Yang[2] Ziwei Liu+[1]

[1] *S-Lab, Nanyang Technological University* [2] *Sensetime Research*

* *Both authors contributed equally to this research* + *corresponding author*

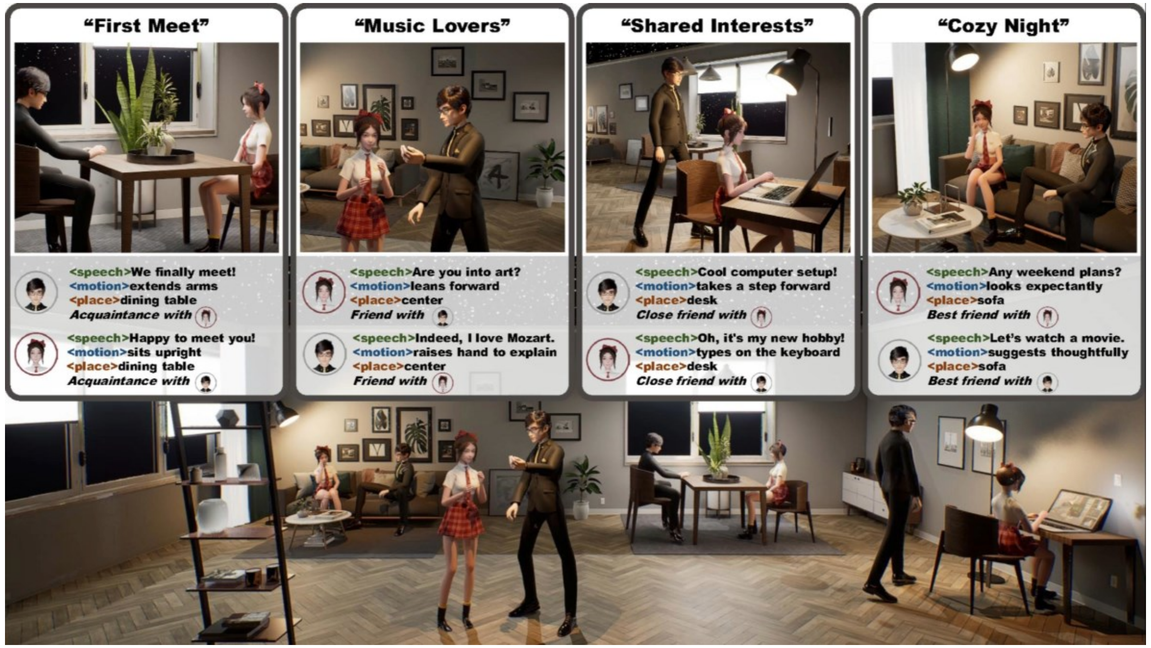
Text-to-Motion

M Zhang*, **Z Cai***, L Pan, F Hong, X Guo, L Yang, Z Liu. *MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model.* IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

Story-to-Motion

Z Qing, **Z Cai**, Z Yang, L Yang. *Story-to-Motion: Human Motion Synthesis using Trajectories and Semantic Descriptions*. SIGGRAPH Asia (Technical Communications) 2023

Applications | Autonomous Characters



<speech>We finally meet!
<motion>extends arms
<place>dining table
Acquaintance with

<speech>Happy to meet you!
<motion>sits upright
<place>dining table
Acquaintance with

<speech>Are you into art?
<motion>leans forward
<place>center
Friend with

<speech>Indeed, I love Mozart.
<motion>raises hand to explain
<place>center
Friend with

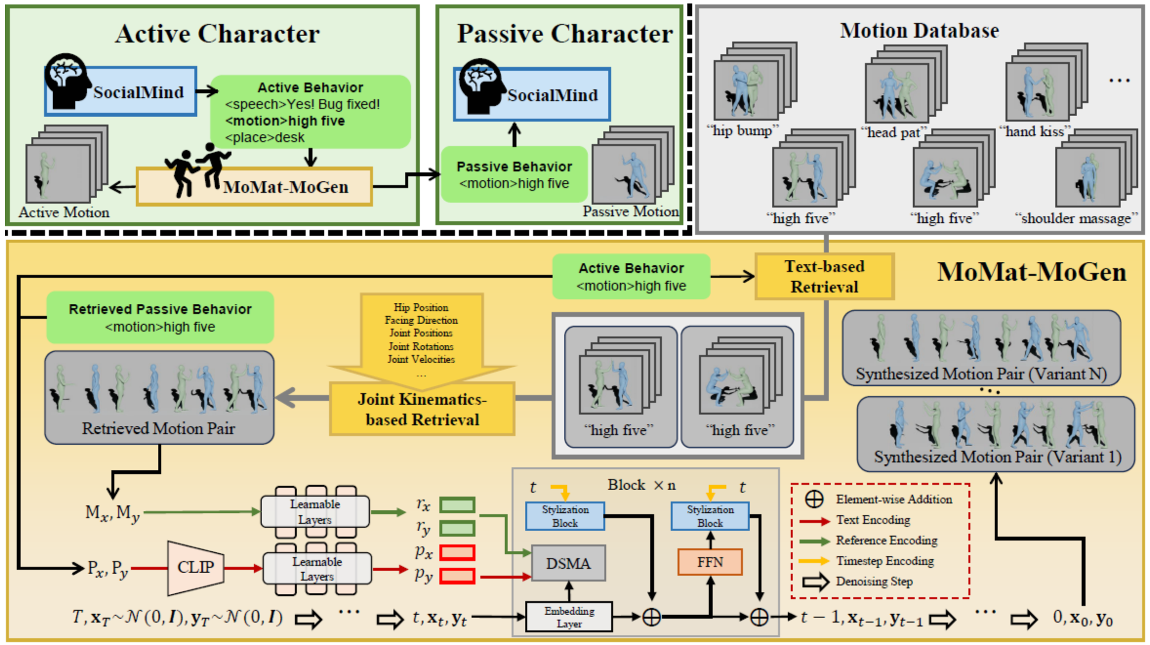
<speech>Cool computer setup!
<motion>takes a step forward
<place>desk
Close friend with

<speech>Oh, it's my new hobby!
<motion>types on the keyboard
<place>desk
Close friend with

<speech>Any weekend plans?
<motion>looks expectantly
<place>sofa
Best friend with

<speech>Let's watch a movie.
<motion>suggests thoughtfully
<place>sofa
Best friend with

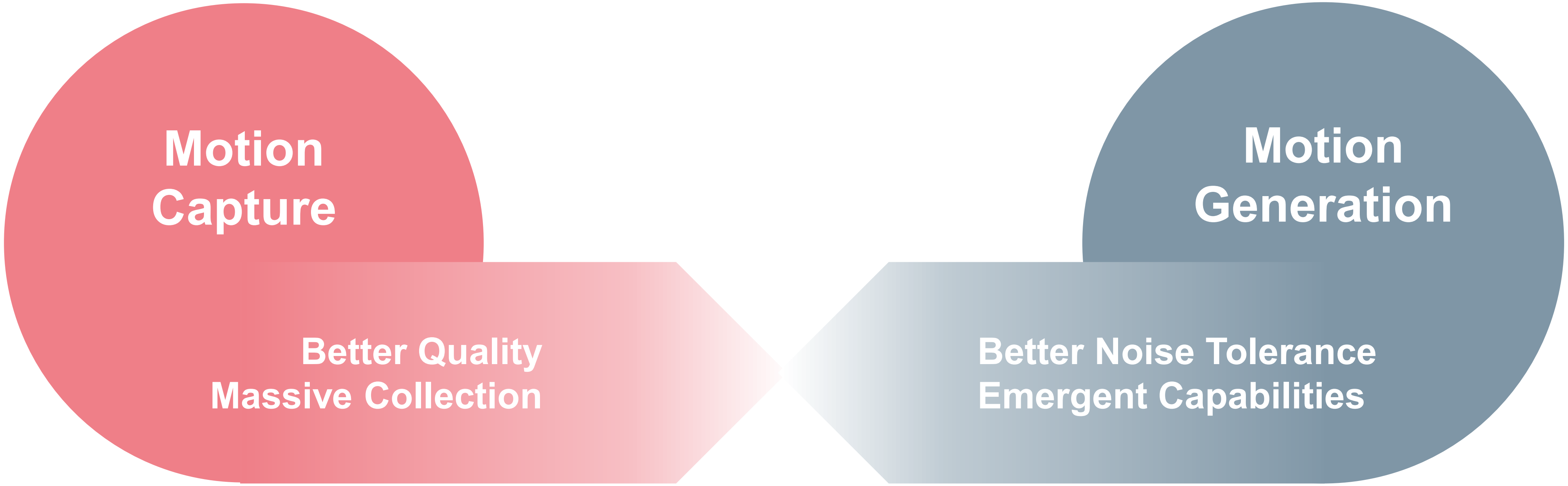
Socially Intelligent 3D Characters



MoMat-MoGen: Motion Matching + Motion Generation

Z Cai*, J Jiang*, Z Qing*, X Guo*, M Zhang*, Z Lin, H Mei, C Wei, R Wang, W Yin, L Pan, X Fan, H Du, P Gao, Z Yang, Y Gao, J Li, T Ren, Y Wei, X Wang, CC Loy, L Yang, Z Liu. *Digital Life Project: Autonomous 3D Characters with Social Intelligence.* Conference on Computer Vision and Pattern Recognition (CVPR) 2024.

What's Next?



Motion Capture

Better Quality
Massive Collection

Motion Generation

Better Noise Tolerance
Emergent Capabilities

Thank you!