Multi-Modal Generative AI with Foundation Models

Ziwei Liu 刘子纬

Nanyang Technological University

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

2024

By ~~2027~~, creators won't have to be technical, just creative, thanks to automation tools.

AI-Generated Content

Movie

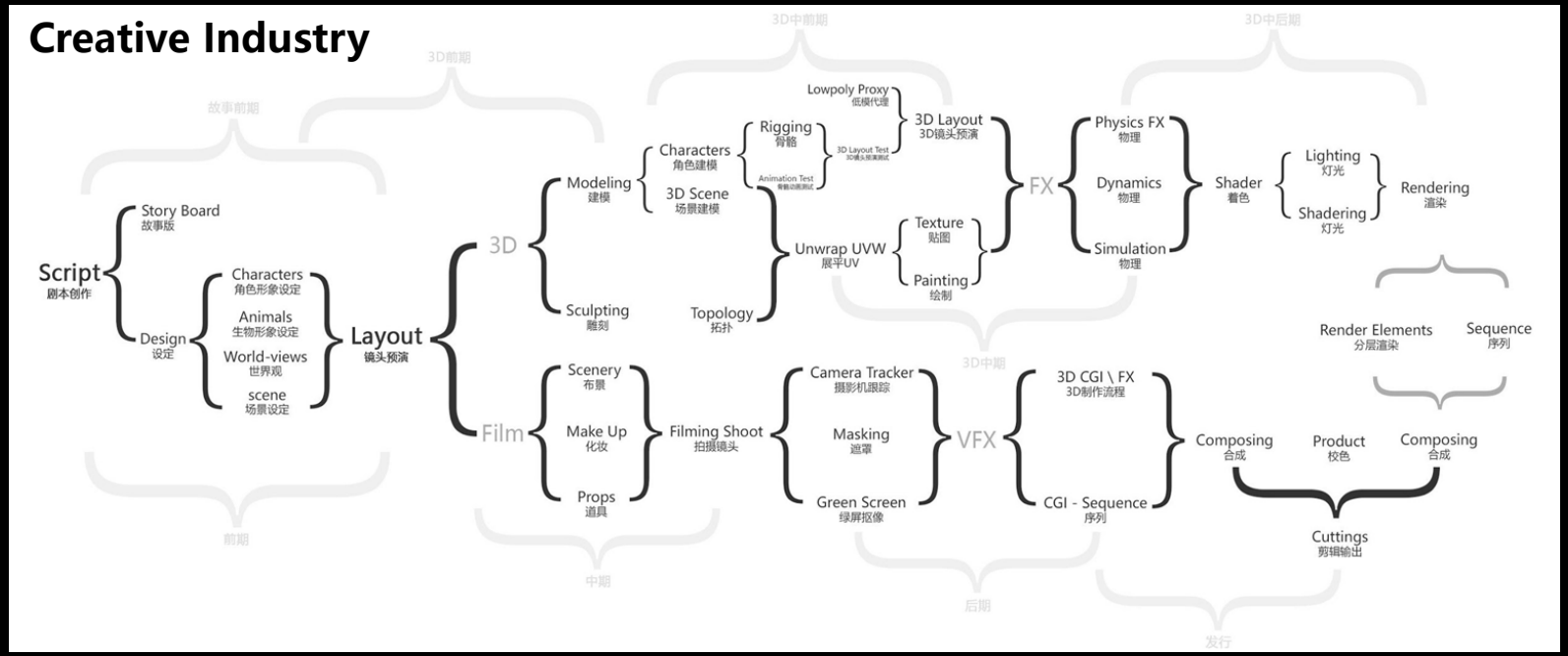
Game

Anime

VTuber

Virtual Beings

Creative Industry



Visual AIGC

Avatar

Foundation Model

Object

Scene

Visual AIGC

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

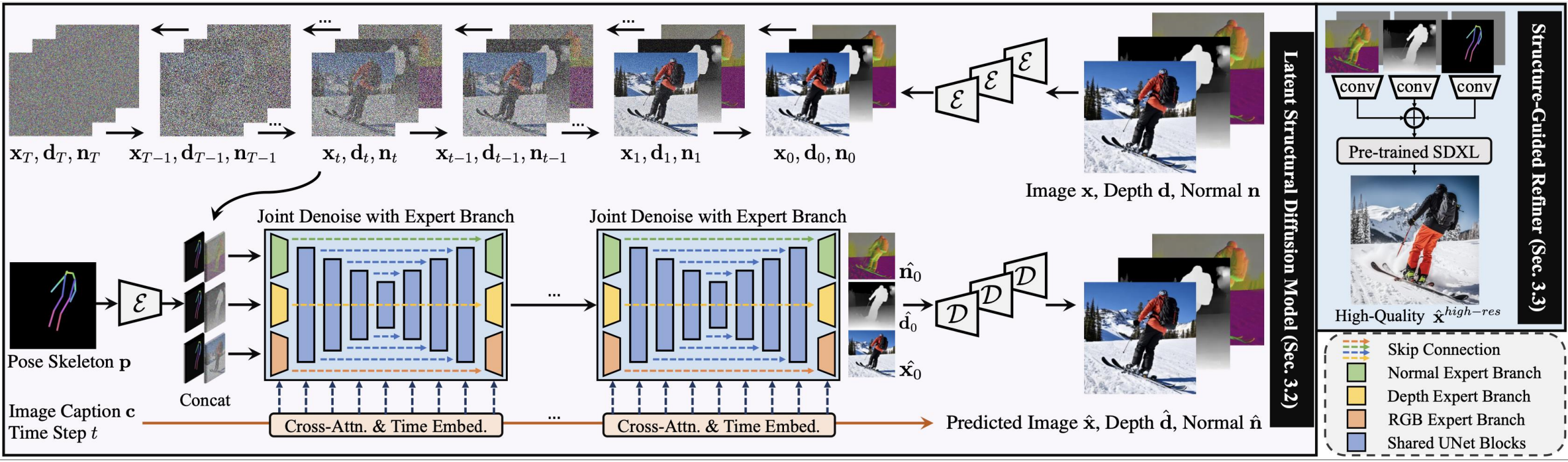
Avatar

Foundation Model

Object

Scene

HyperHuman | High-Quality 2D Text-to-Human



Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, Sergey Tulyakov.
HyperHuman: Hyper-Realistic Human Generation with Latent Structural Diffusion.
International Conference on Learning Representations (ICLR) 2024.

HyperHuman | High-Quality 2D Text-to-Human

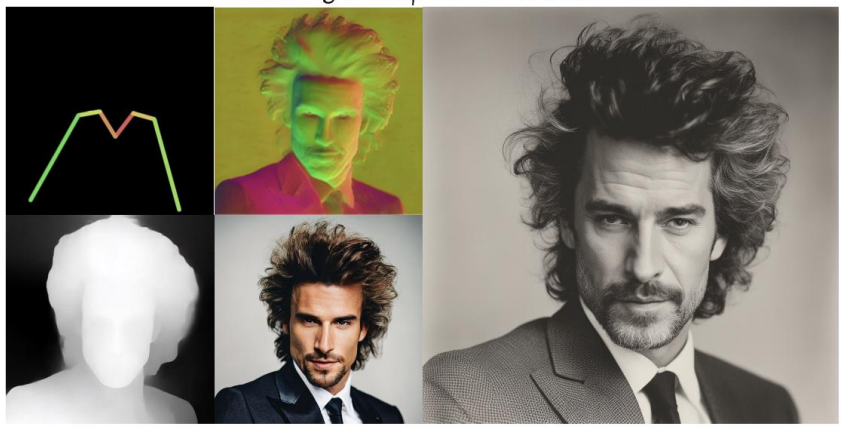
A man sitting down with a brown teddy bear on his shoulders.

Two women holding surfboards while smiling at the camera.

A guy in a brown jacket standing near a sign holding a cellphone to his ear.

A woman poses with avocado sandwich lunch at an outdoor restaurant.

An elderly woman looks to the side as she sits in front of a cheese pizza in a restaurant.



A picture of a man with suit, tie and wild hair.

Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, Sergey Tulyakov.
HyperHuman: Hyper-Realistic Human Generation with Latent Structural Diffusion.
International Conference on Learning Representations (ICLR) 2024.

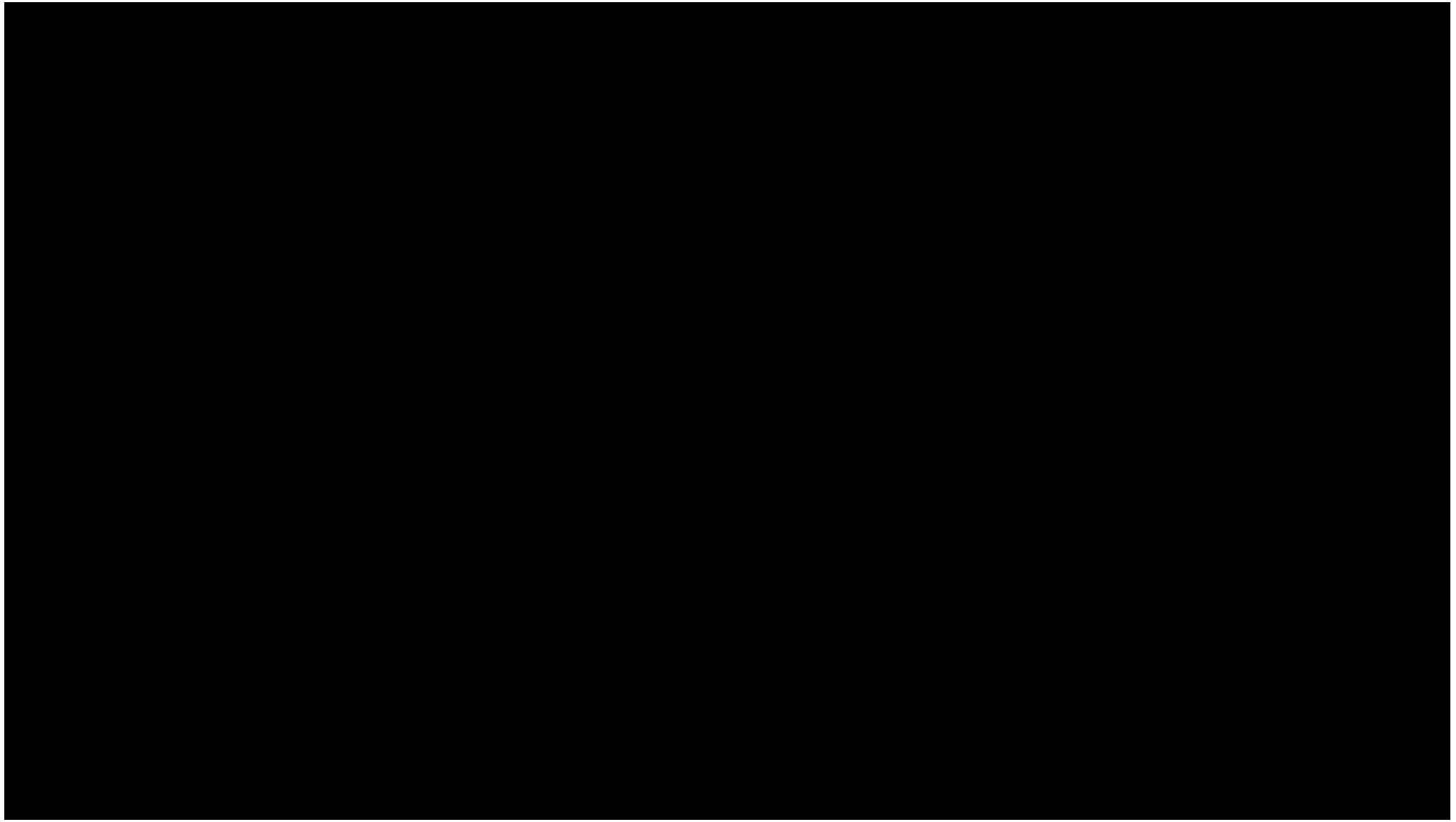
HyperHuman | High-Quality 2D Text-to-Human

HyperHuman | High-Quality 2D Text-to-Human

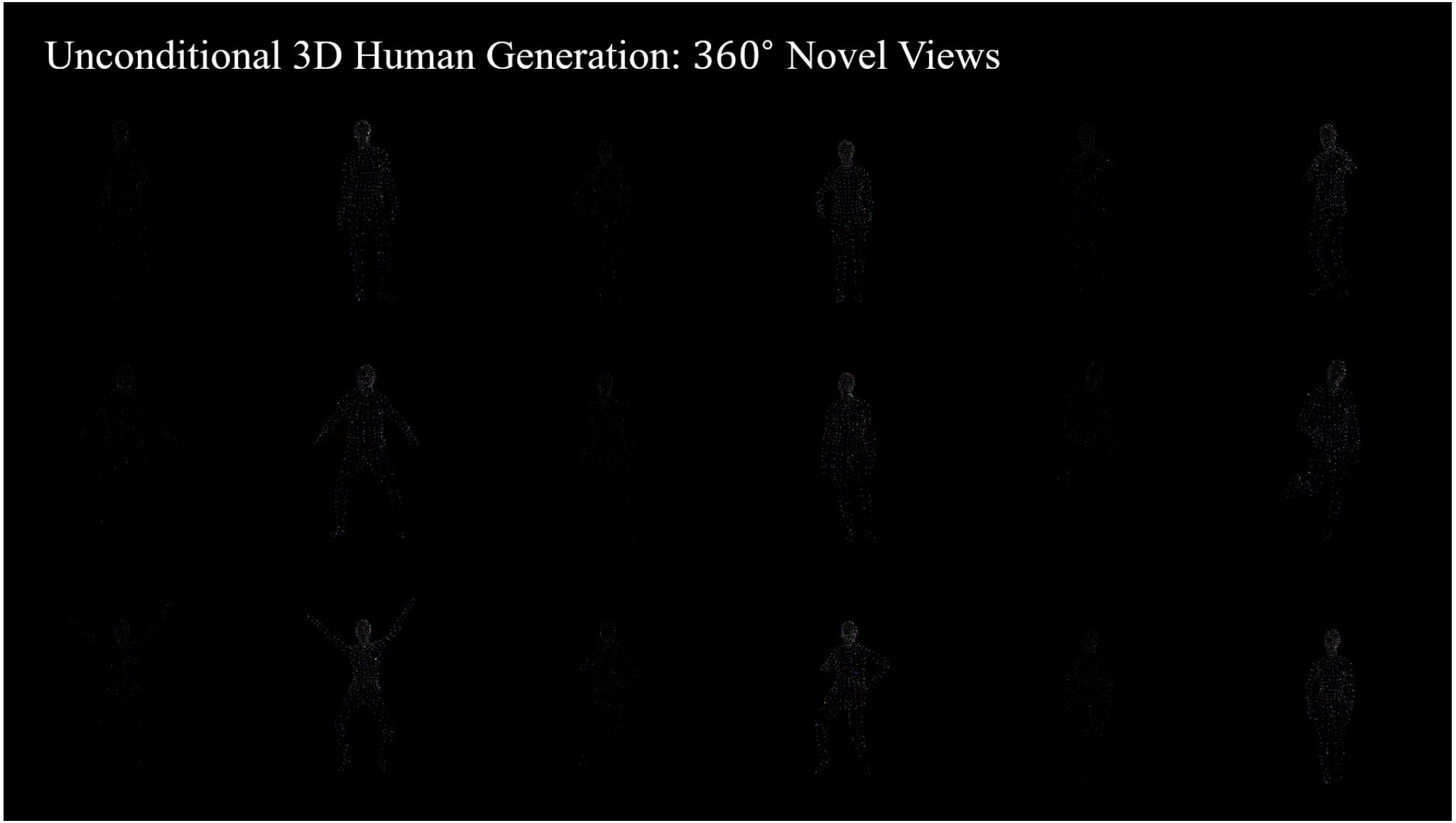
A baby girl with beautiful blue eyes standing next to a brown teddy bear.

A little girl with wavy hair and smile holding a teddy bear.

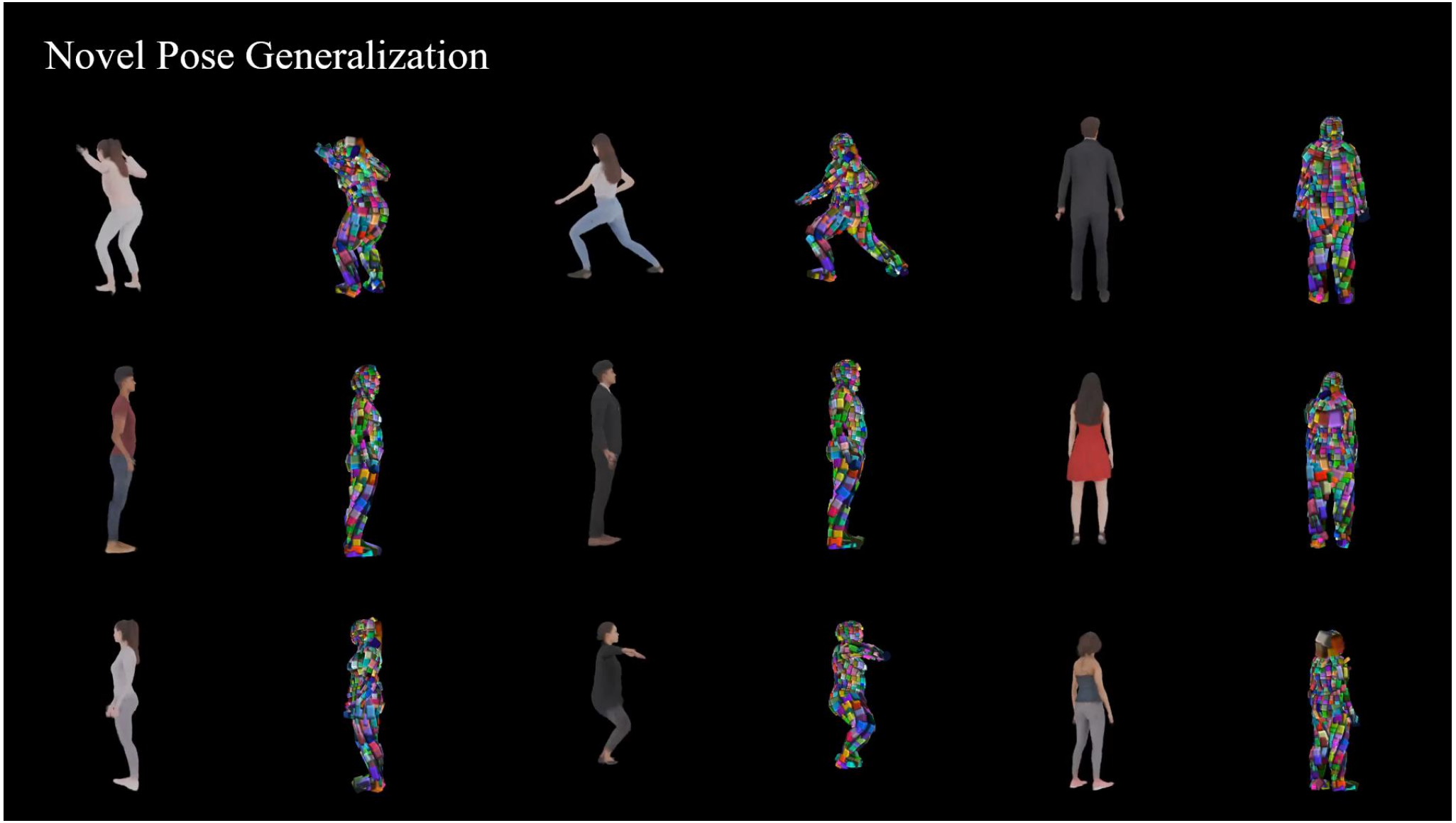
PrimDiffusion: Feedforward 3D Human Diffusion



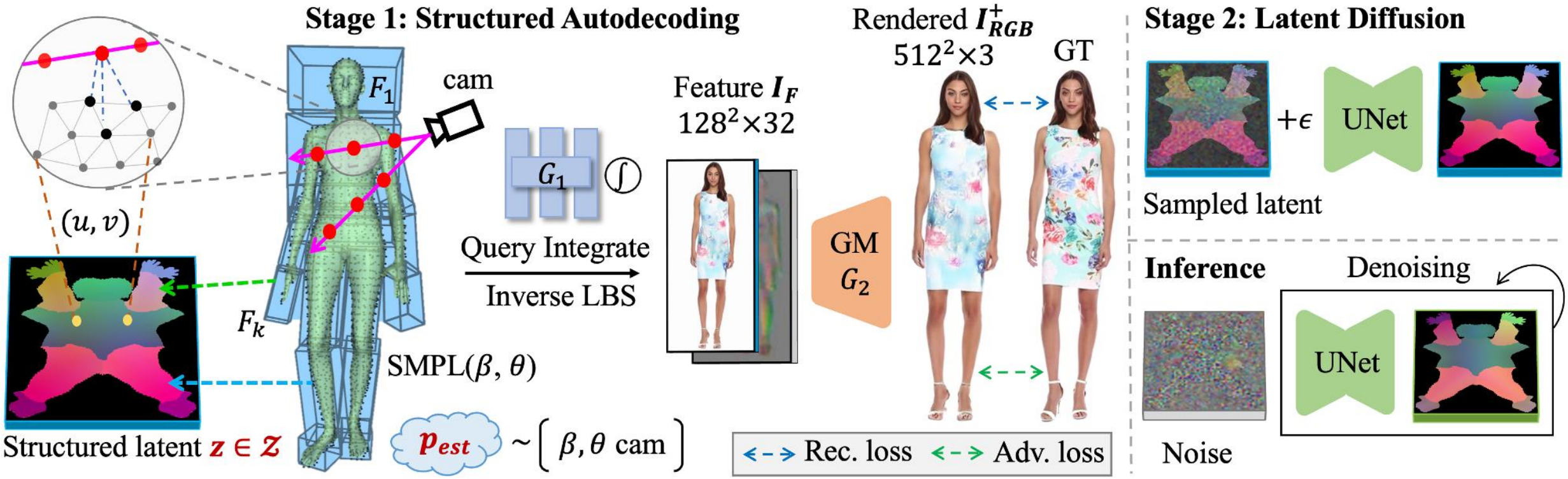
PrimDiffusion: Feedforward 3D Human Diffusion



PrimDiffusion: Feedforward 3D Human Diffusion



StructLDM: Structured Latent Diffusion



Tao Hu, Fangzhou Hong, Ziwei Liu. StructLDM: Structured Latent Diffusion for 3D Human Generation. arXiv 2024.

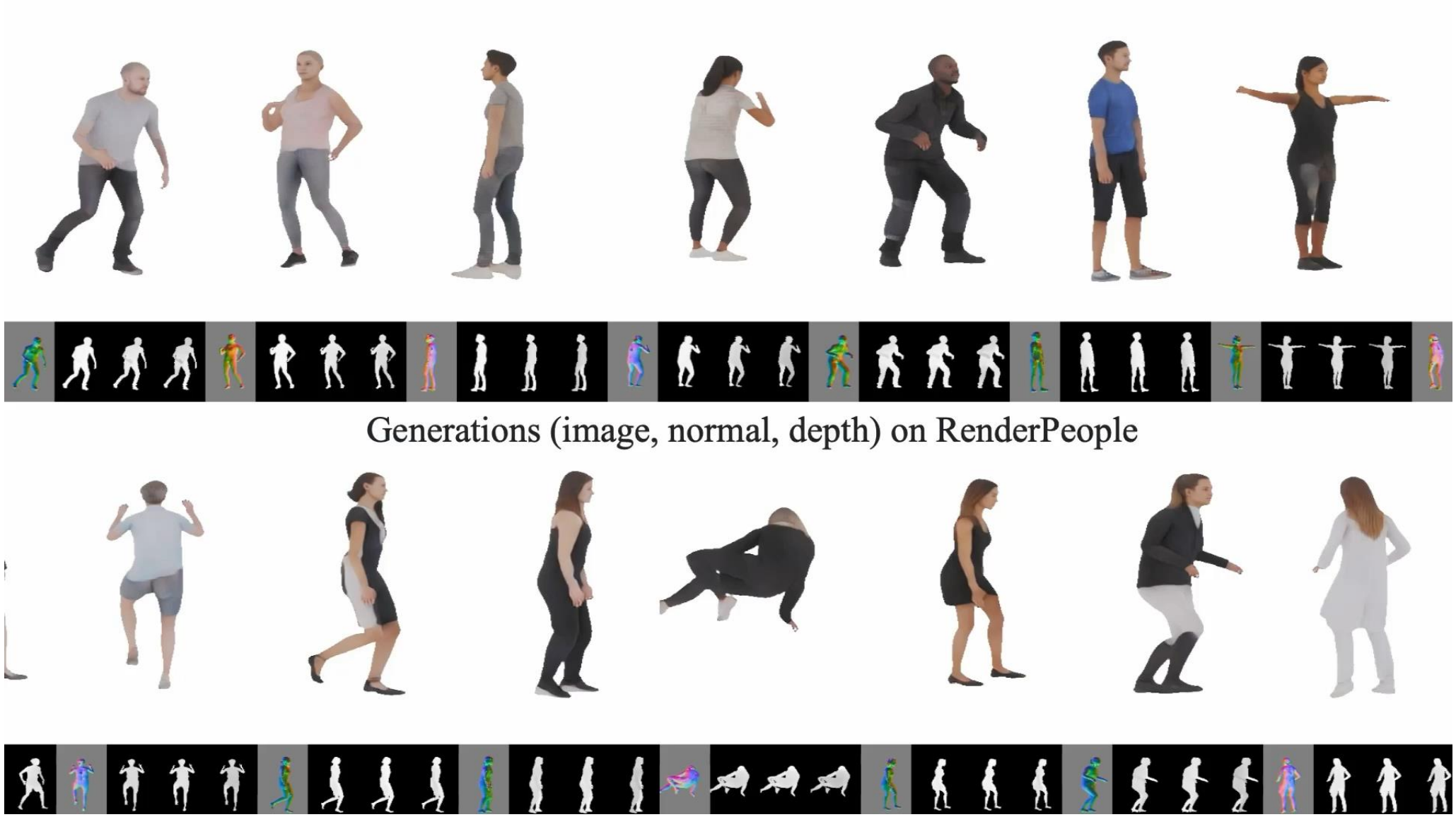
StructLDM: Structured Latent Diffusion

Tao Hu, Fangzhou Hong, Ziwei Liu. StructLDM: Structured Latent Diffusion for 3D Human Generation. arXiv 2024.

StructLDM: 3D Human Editing

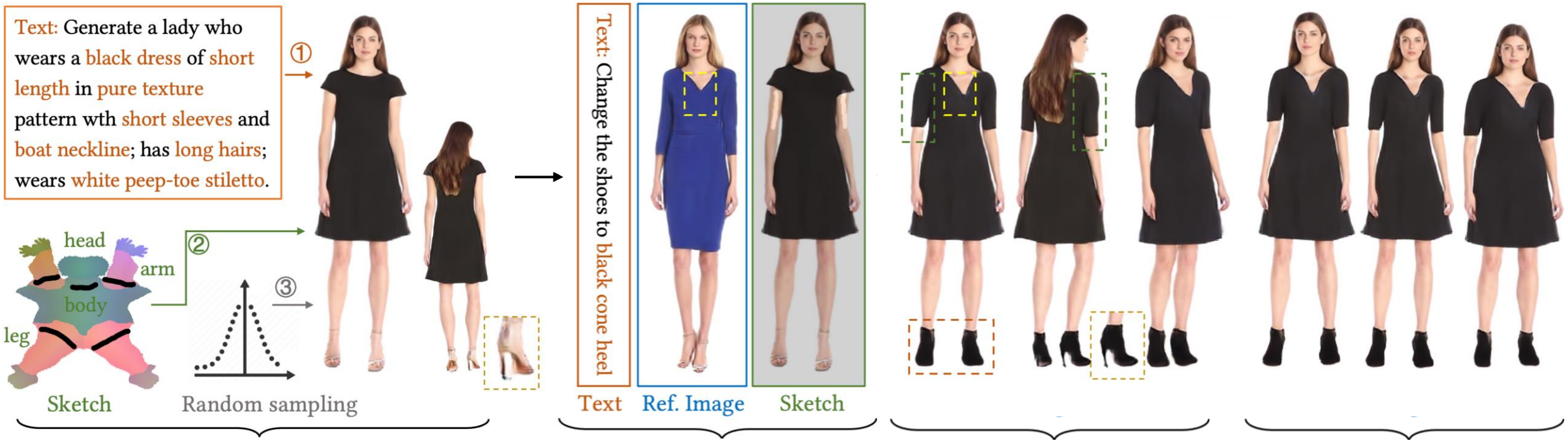
Tao Hu, Fangzhou Hong, Ziwei Liu. StructLDM: Structured Latent Diffusion for 3D Human Generation. arXiv 2024.

StructLDM: Structured Latent Diffusion



Tao Hu, Fangzhou Hong, Ziwei Liu. StructLDM: Structured Latent Diffusion for 3D Human Generation. arXiv 2024.

FashionEngine: Interactive & Multimodal Editing



Text-, Sketch- based generation

Multimodal controls

Consistent editing

Pose/shape/view ctrl.

Tao Hu, Fangzhou Hong, Ziwei Liu. FashionEngine: Interactive 3D Human Generation and Editing via Multimodal Controls. arXiv 2024.

FashionEngine: Interactive Generation and Editing of 3D Clothed Humans

Sketch-based Editing

Asset

UBCFashion

Latent

Seed 482315781

Random Style

Sketch

Apply

Reset

Sketch to Human

Lock

Diffusion

Steps 100

η 0.5

Local Diffusion

Image

Style Transfer

Text

Appearance Description

<clothing color>, <length of lower clothing>, <sleeve length>, <clothing pattern>, <hairstyle>, <neckline>, <shoe>.

Examples

A lady wears a black dress of short length, with short sleeve. The texture of it is solid color.

Appearar

Style

Text to Human

Generation

Editing

Matching Style

Pose Corpus

View 0

Shape 0

Animation

Undo

Reset

FashionEngine: Interactive Generation and Editing of 3D Clothed Humans

Asset

UBCFashion

Latent

Seed 482315781

Random Style

Sketch

Apply

Reset

Sketch to Human

Lock

Diffusion

Steps 100

η 0.5

Local Diffusion

Image

Style Transfer

Text

Appearance Description

<clothing color>, <length of lower clothing>, <sleeve length>, <clothing pattern>, <hairstyle>, <neckline>, <shoe>.

Examples

A lady wears a black dress of short length, with short sleeve. The texture of it is solid color.

Appearar

Style

Text to Human

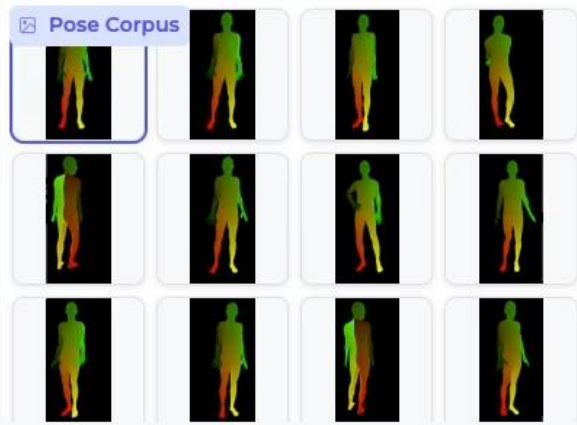
Generation

Start drawing

Editing

Matching Style

Pose Corpus



View 0

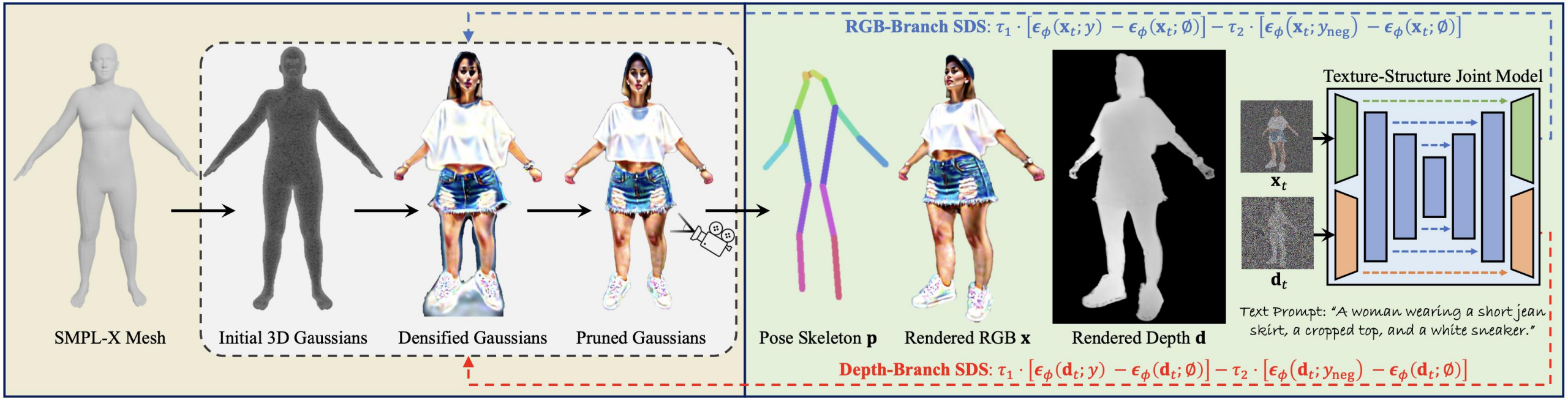
Shape 0

Animation

Undo

Reset

HumanGaussian | Animatable 3D Text-to-Human



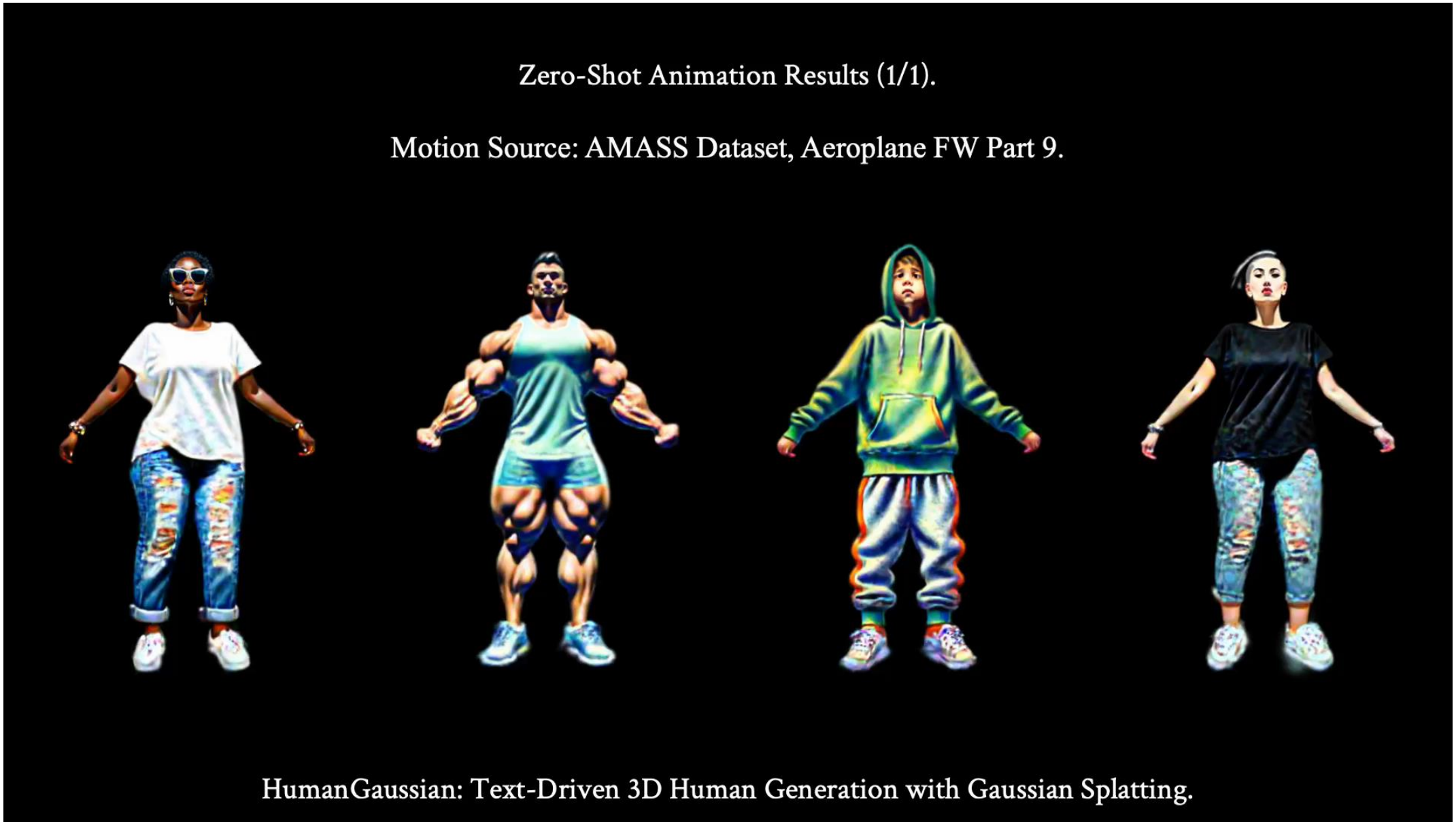
Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, Ziwei Liu.
HumanGaussian: Text-Driven 3D Human Generation with Gaussian Splatting.
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024. (Highlight, Top 2.8%)

HumanGaussian | Animatable 3D Text-to-Human

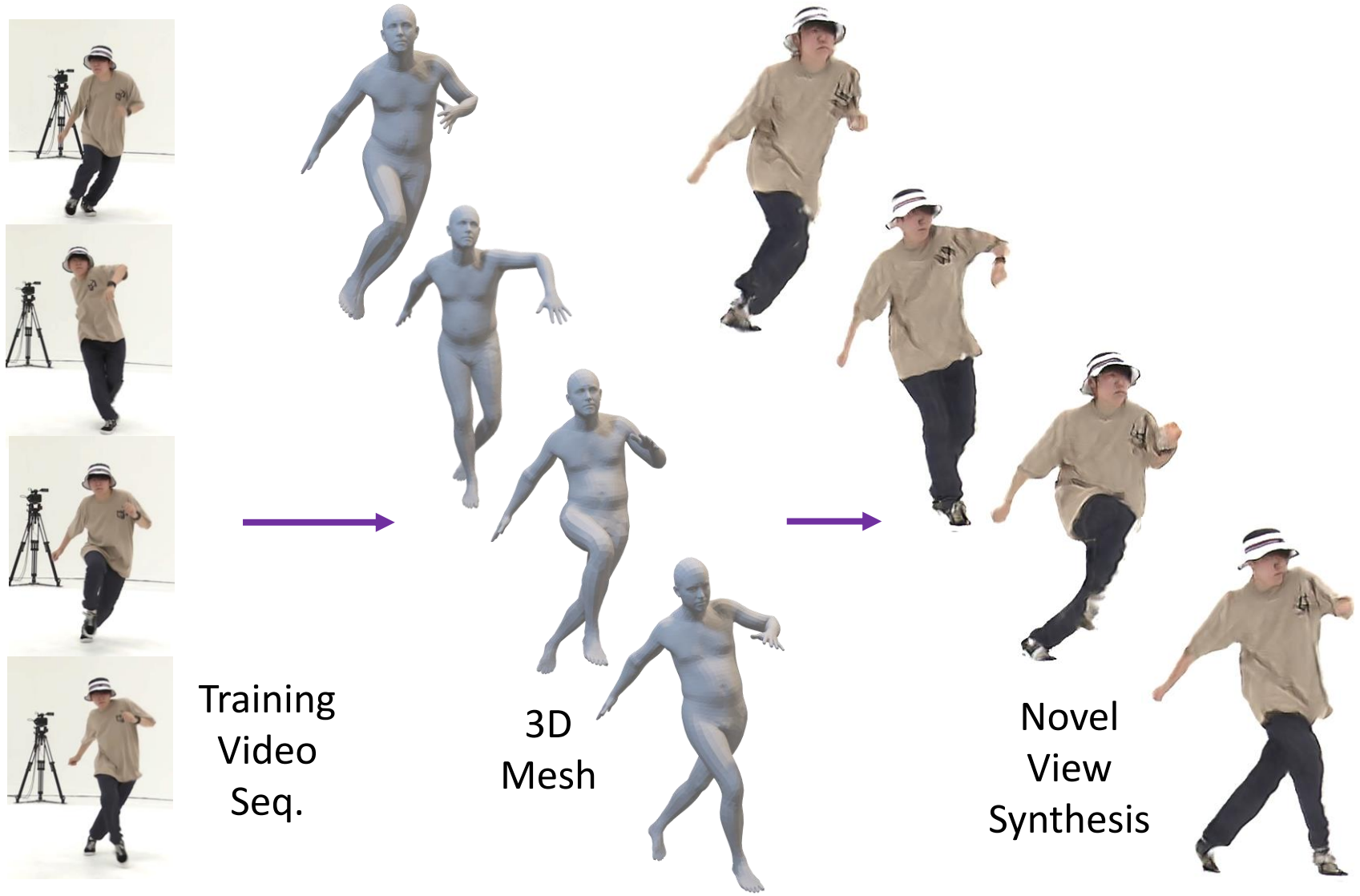
Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, Ziwei Liu.
HumanGaussian: Text-Driven 3D Human Generation with Gaussian Splatting.
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024. (Highlight, Top 2.8%)

HumanGaussian | Animatable 3D Text-to-Human

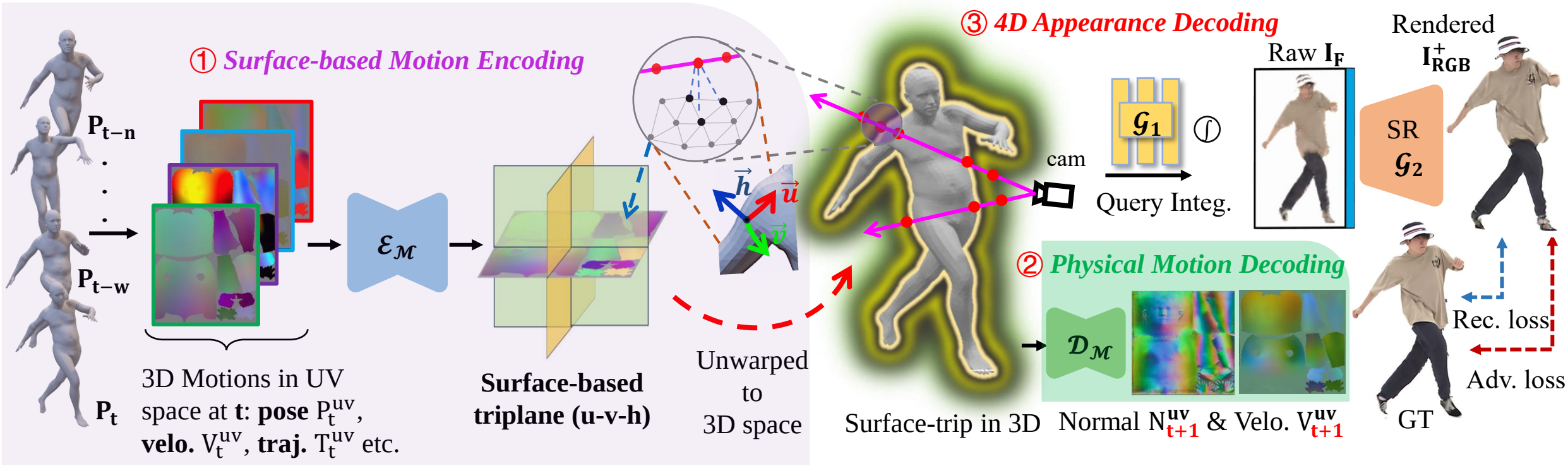
HumanGaussian | Animatable 3D Text-to-Human



SurMo: Surface-based 4D Motion Modeling



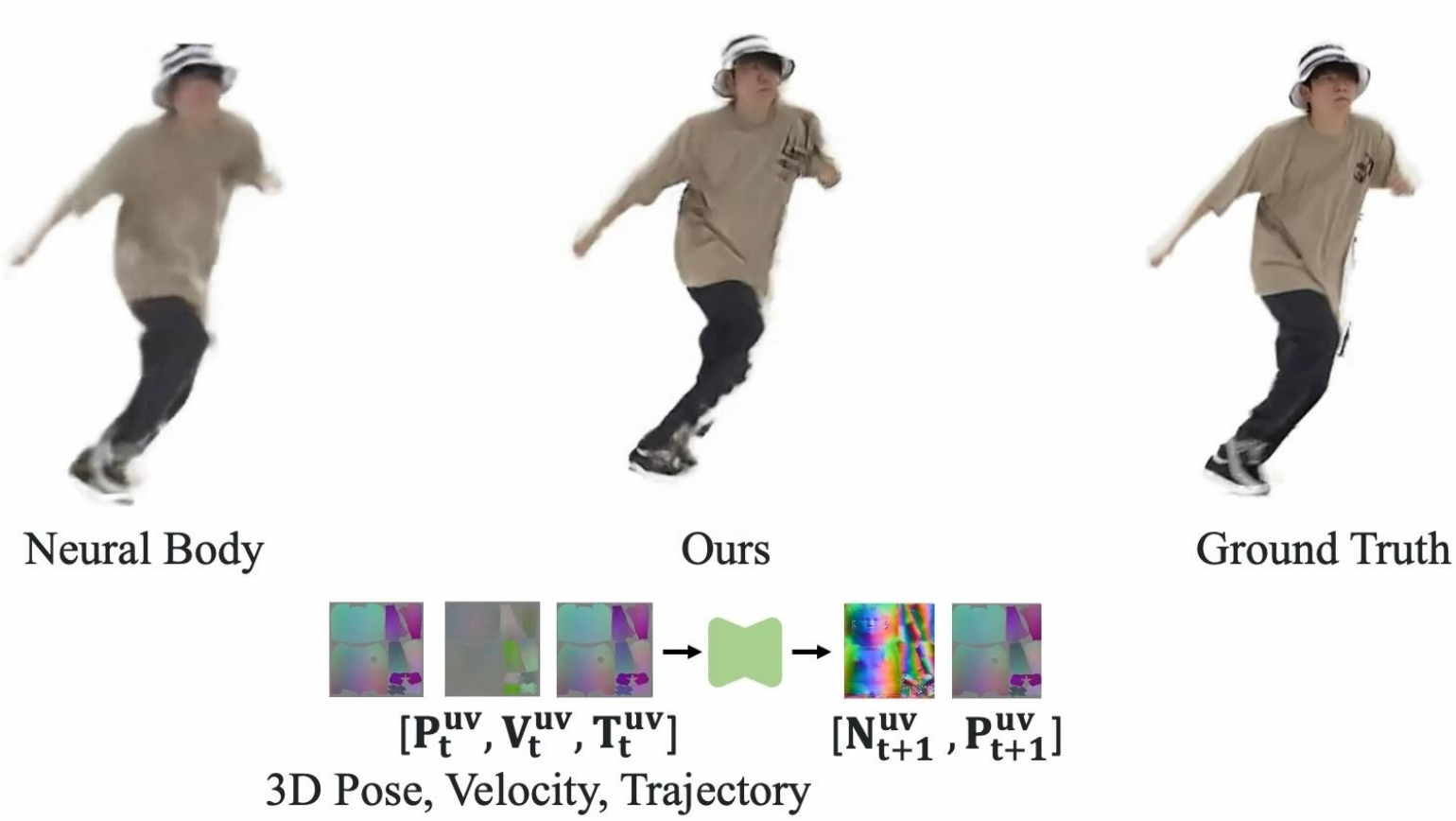
SurMo: Surface-based 4D Motion Modeling



New Paradigm: Motion Encoding → Physical Motion Decoding, Appearance Decoding

Tao Hu, Fangzhou Hong, Ziwei Liu. SurMo: Surface-based 4D Motion Modeling for Dynamic Human Rendering. CVPR 2024.

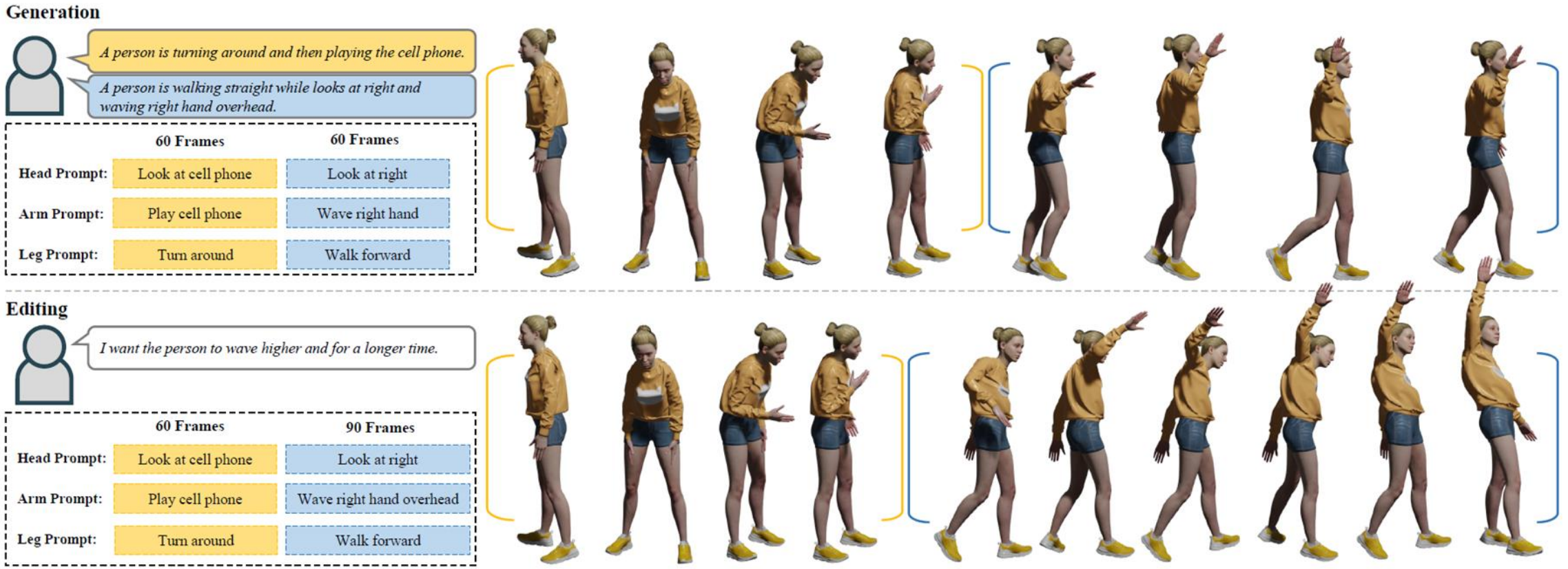
SurMo: Surface-based 4D Motion Modeling



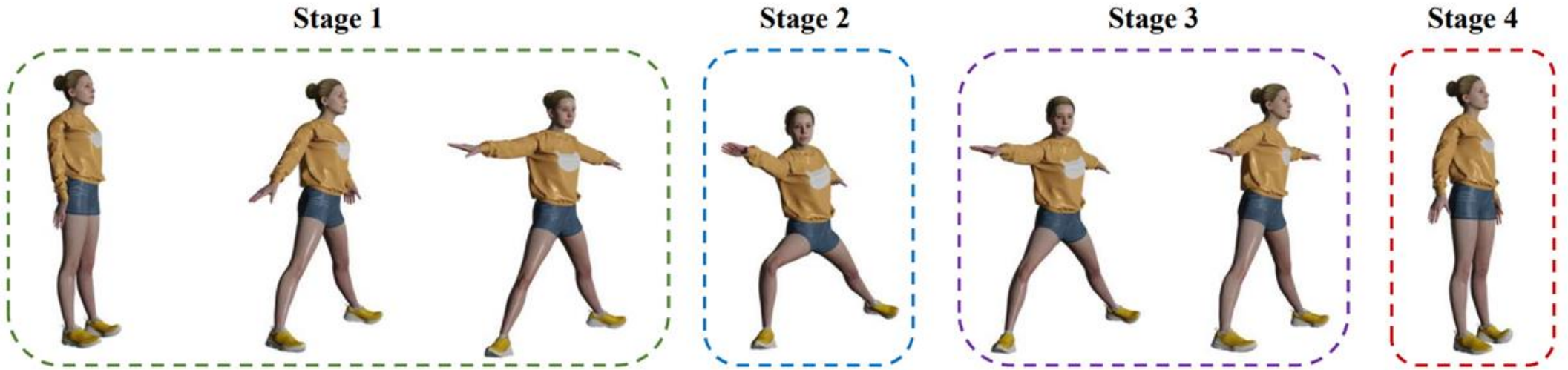
Tao Hu, Fangzhou Hong, Ziwei Liu. SurMo: Surface-based 4D Motion Modeling for Dynamic Human Rendering. CVPR 2024.

FineMoGen | Fine-Grained Motion Generation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

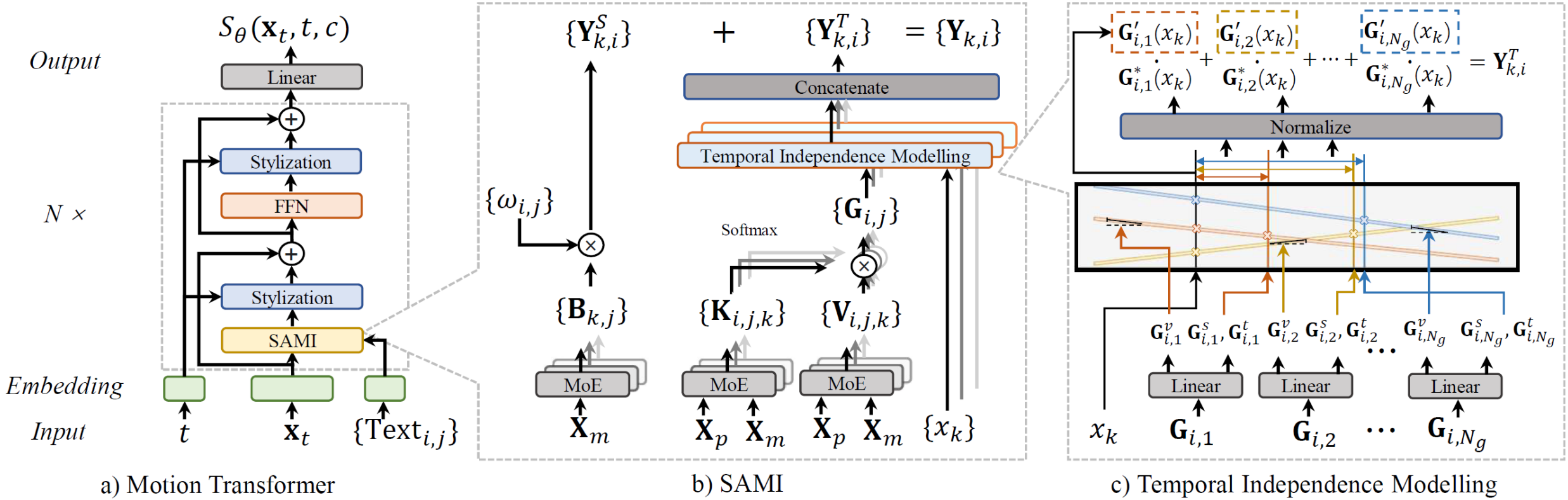


FineMoGen | Fine-Grained Description



Stage 1	Stage 2	Stage 3	Stage 4
Left & Right upper limb: Raise the arms to body sides until parallel to the ground. **Left & Right lower limb:** Shift the feet alternately to open the legs until the feet are 3-4 feet apart. **Pelvis:** Shifted downward.	**Head:** Turn to the right. **Right lower limb:** Point the right foot to the right. Bend the right knee to the right until the thigh is parallel to the ground. **Pelvis:** Shifted downward and right.	**Head:** Turn to the front. **Right lower limb:** Point the right foot to the front. Unbend the right knee and straighten the right leg. **Pelvis:** Shifted upward and left.	**Left upper limb:** Put down the arms to the sides of the body. **Left & Right lower limb:** Shift the feet alternately to move the legs inwards until they touch each other. **Pelvis:** Shifted upward.

FineMoGen | Spatio-Temporal Modeling



a) Motion Transformer

b) SAMI

c) Temporal Independence Modelling

FineMoGen: Fine-Grained Spatio-Temporal Motion Generation and Editing

-- Demo Video

Visual AIGC

Avatar

Foundation Model

Object

Scene

OmniObject3D: 3D Object Dataset

OmniObject3D is a **large-vocabulary** 3D dataset for **real-world scanned objects**.

- ✓ **6k** high-quality 3D models
- ✓ **190** categories
- ✓ **4** modalities: textured mesh, point cloud, real-captured video, synthetic multi-view images.
- ✓ Many down-stream tasks

	Dataset	Year	Real	Full 3D	Video	Num Objs	Num Cats
	ShapeNet	2015		√		51k	55
	ModelNet	2014		√		12k	40
	3D-Future	2020		√		16k	34
	ABO	2021		√		8k	63
	Toys4K	2021		√		4k	105
	CO3D	2021	√		√	19k	50
Real-world 3D scans	DTU	2014	√	√		124	NA
	GSO	2021	√	√		1k	17
	AKB-48	2022	√	√		2k	48
	Ours	2022	√	√	√	**6k**	**190**

OmniObject3D: 3D Object Dataset

Summary

It's a teacup.

Appearance

This is a relatively small teacup with a brownish-red exterior and white interior, featuring a blue line pattern at the top and a rounded white bump on the bottom, structured in an overall axisymmetric manner.

Material

Ceramic, hard, reflective, smooth surface.

Style

Simplicity.

Function

Water storage.

Summary

It's a teapot.

Appearance

This teapot is white with a gray handle positioned perpendicular to the spout, and a small round gray handle at the top of the lid; the body of the teapot is adorned with a pattern of pink lotuses, gray lotus leaves, and red buds, all structured in an asymmetric manner.

Material

Ceramic, rough surface, hard, slightly reflective.

Style

Simplicity.

Function

Tea making, water storage.

Summary

It's a glasses case.

Appearance

Overall purple, the box features a pink LinaBell on the surface wearing a dark purple flower and blue eyes, complemented by a row of purple and pink letters underneath, all structured in an axisymmetric manner.

Material

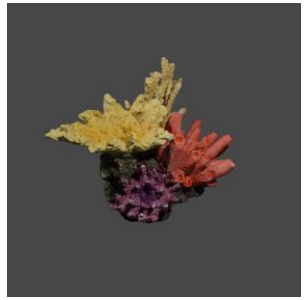
Leather, rubber, metal, smooth surface, hard, slightly reflective, metallic.

Style

Cartoon.

Function

Store glasses, decoration.



Summary

It's a coral simulation model.

Appearance

The upper part of this coral simulation model is yellow, below the yellow section, there are pink and purple corals, the purple corals have white attachments on their surfaces, several colors of corals are on a brown reef, and the entire model is asymmetrical.

Material

Plastic, rough surface, hard, slightly reflective.

Style

Reality.

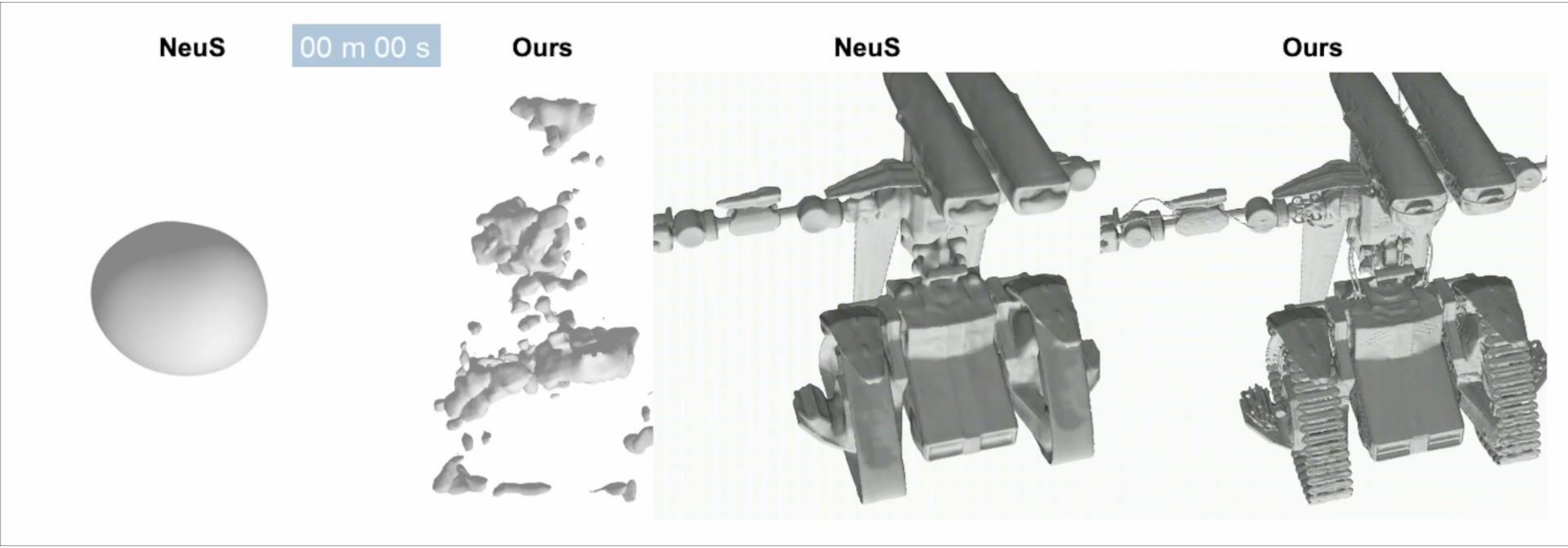
Function

Entertainment, decoration.

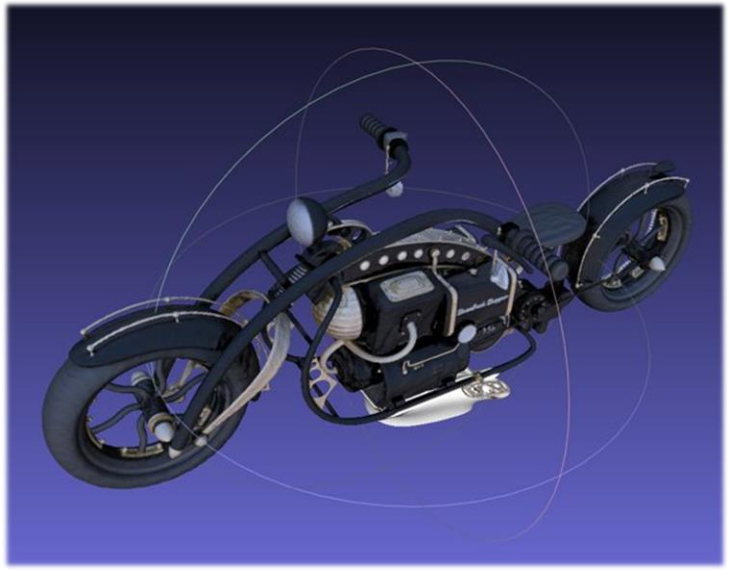
DiffTF: 3D Diffusion Transformer

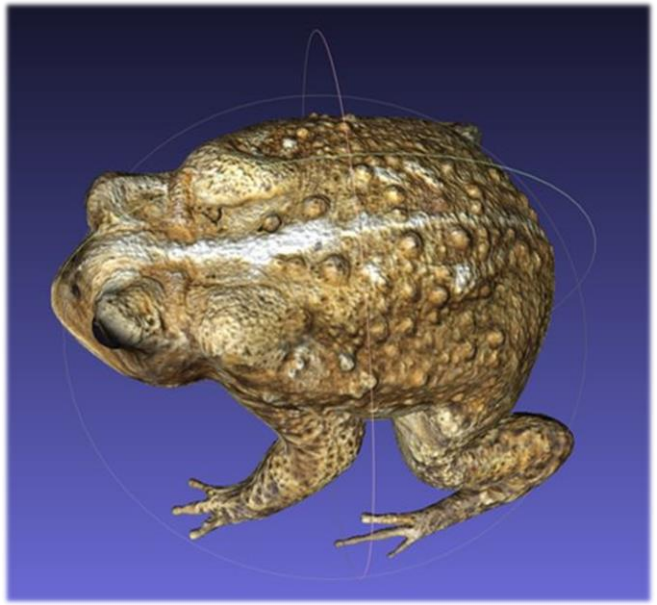
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Voxurf: Fast 3D Object Reconstruction



Voxurf: Fast 3D Object Reconstruction




Visual AIGC

Avatar

Foundation Model

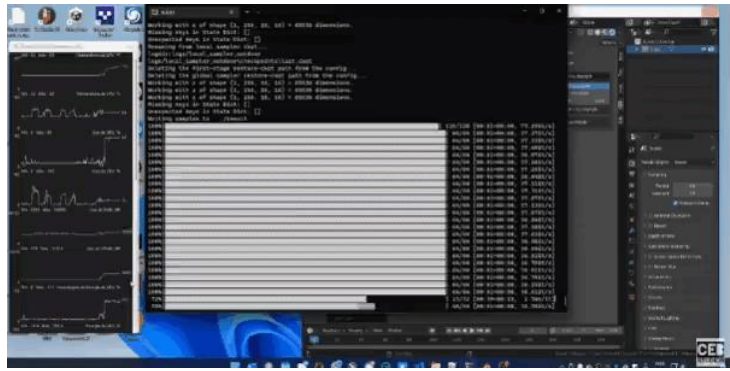
Object

Scene

Text2Light: Text-to-3D Environment

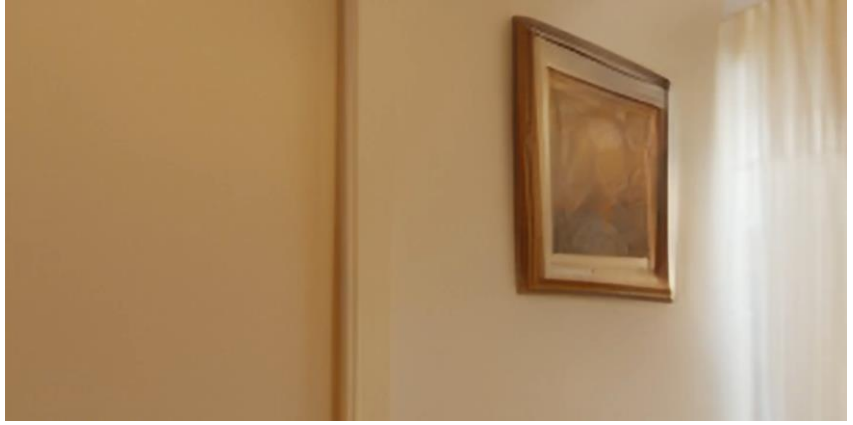
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

"brown wooden dock on lake surrounded by green trees during daytime"

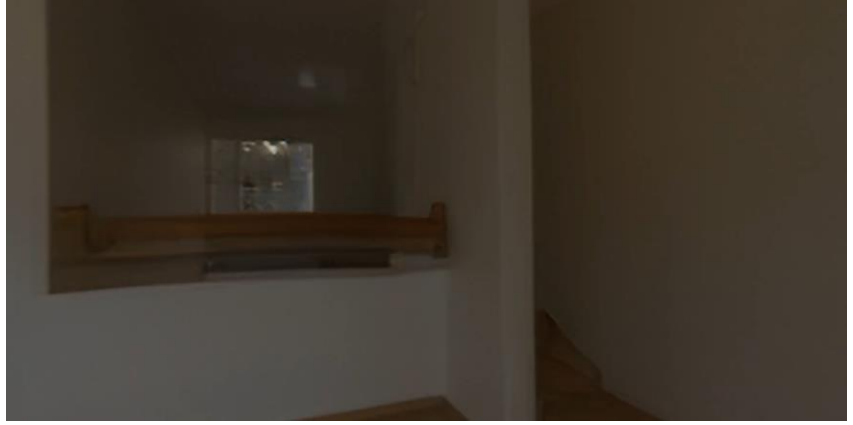


4K+ Resolution with High Dynamic Range

“white bed linen with white pillow”

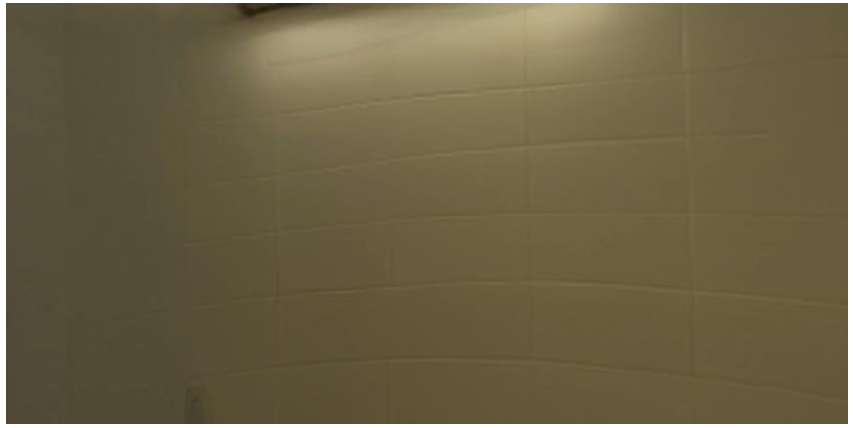


“brown wooden floor with white wall”

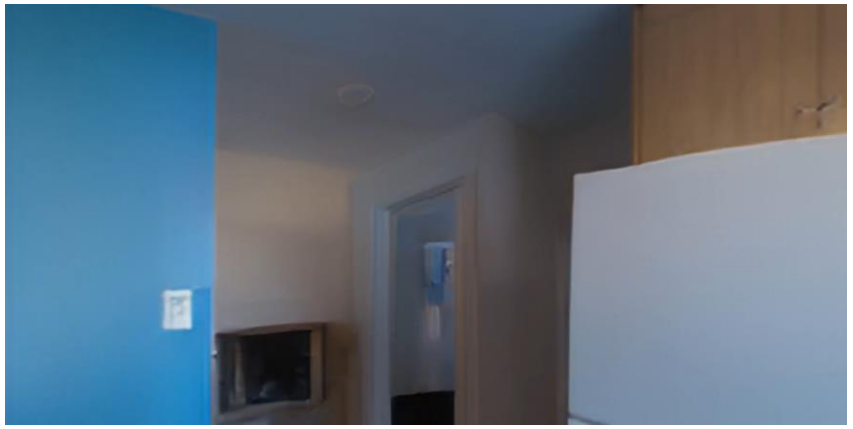


“gray concrete pathway with wall signages”

“closeup photo of concrete stair surrounded by white painted wall”



“blue and brown wooden counter”



“empty parking lot during daytime”

Suzanne Monkey: glossy Shader balls: glass, diffuse, glossy, mixture of diffuse and glossy

Text2Light: Text-to-3D Environment

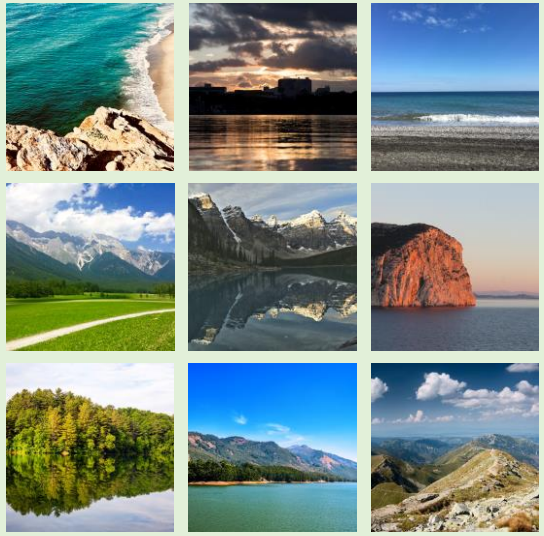
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Describe Your Scene

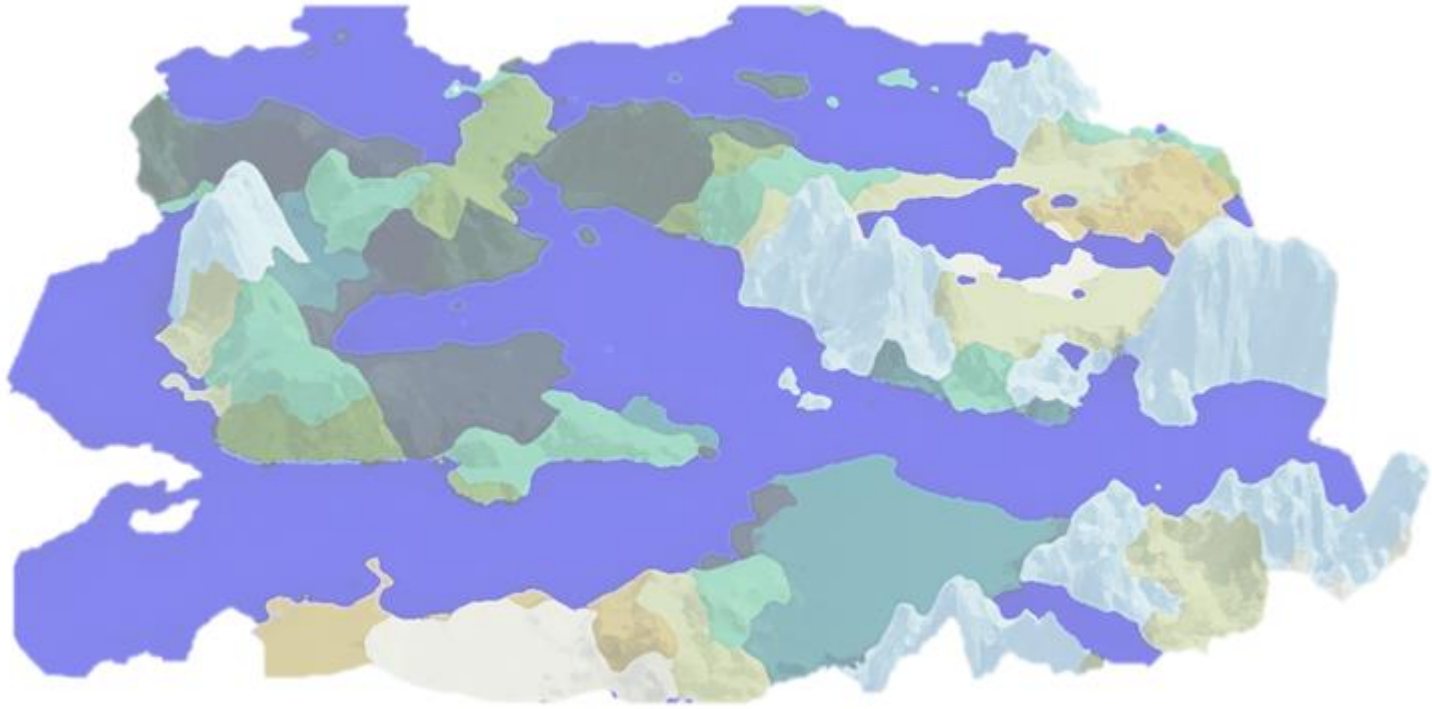
e.g. a living room

Generate Render

SceneDreamer: Unbounded 3D Scene Generation

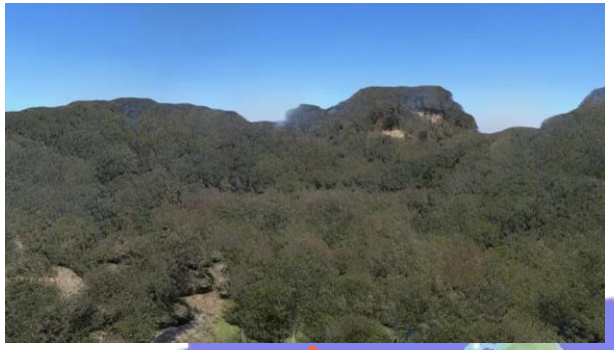


In-the-wild 2D Image Collections



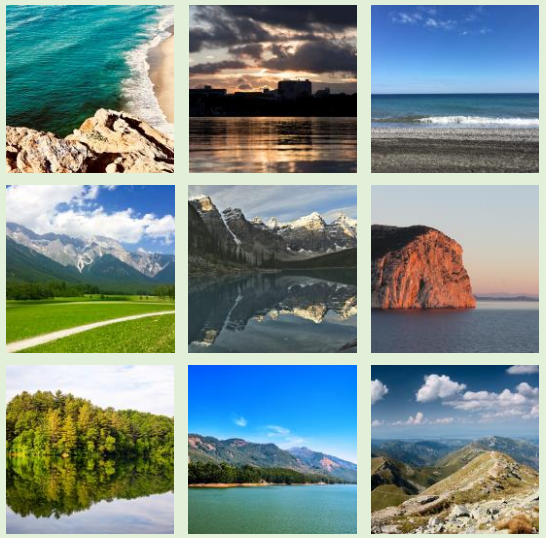
Photorealistic Unbounded 3D Scenes

SceneDreamer: Unbounded 3D Scene Generation

Multi-view consistent

Well-defined geometry



In-the-wild
Image Collections

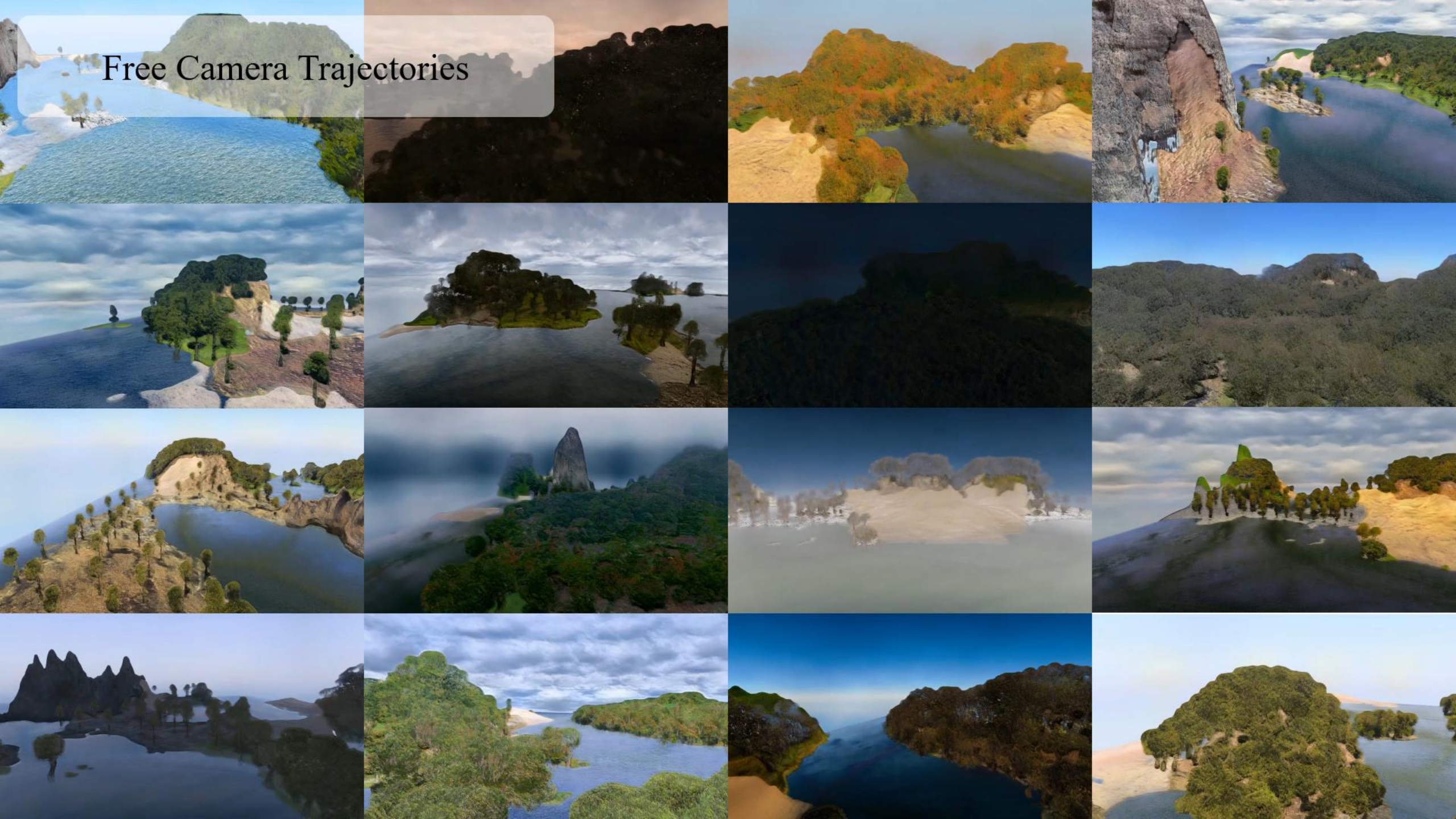
Photorealistic
Unbounded 3D Scenes

Diverse scenes and styles

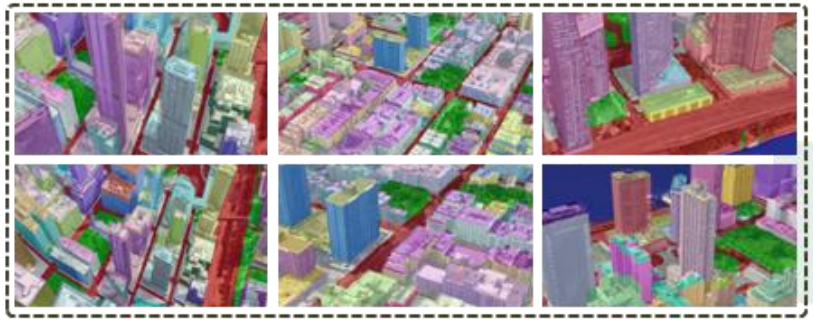


Infinite 3D World!

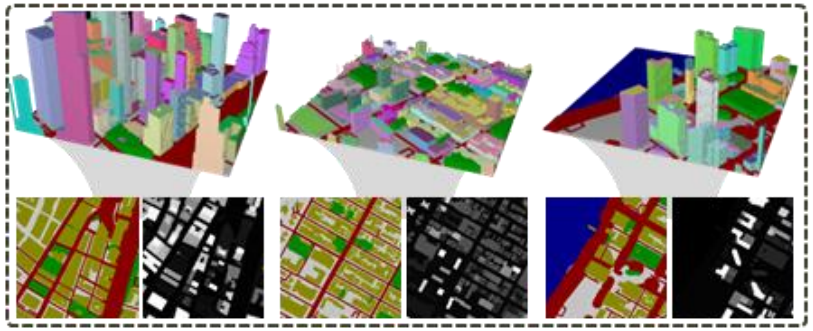

Generate with Different Styles


Free Camera Trajectories

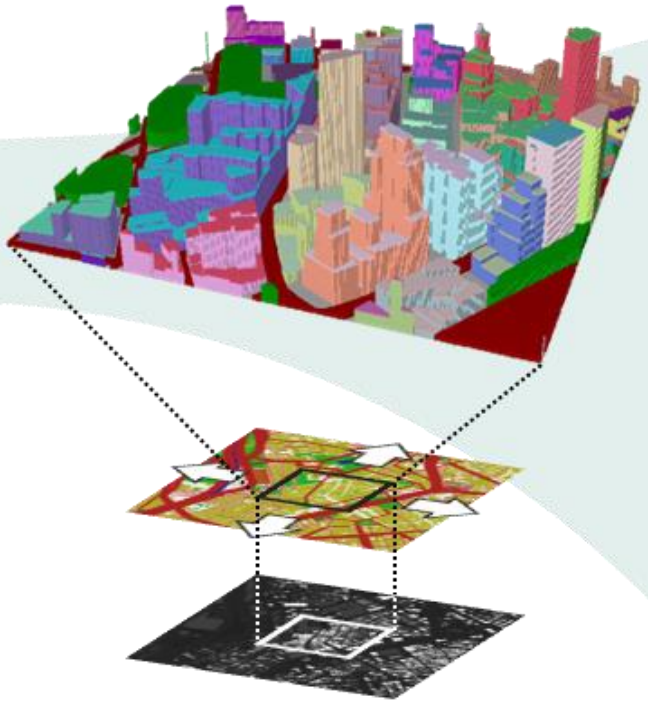
CityDreamer: Unbounded 3D City Generation



(a) GoogleEarth Dataset: Real-world City Appearance



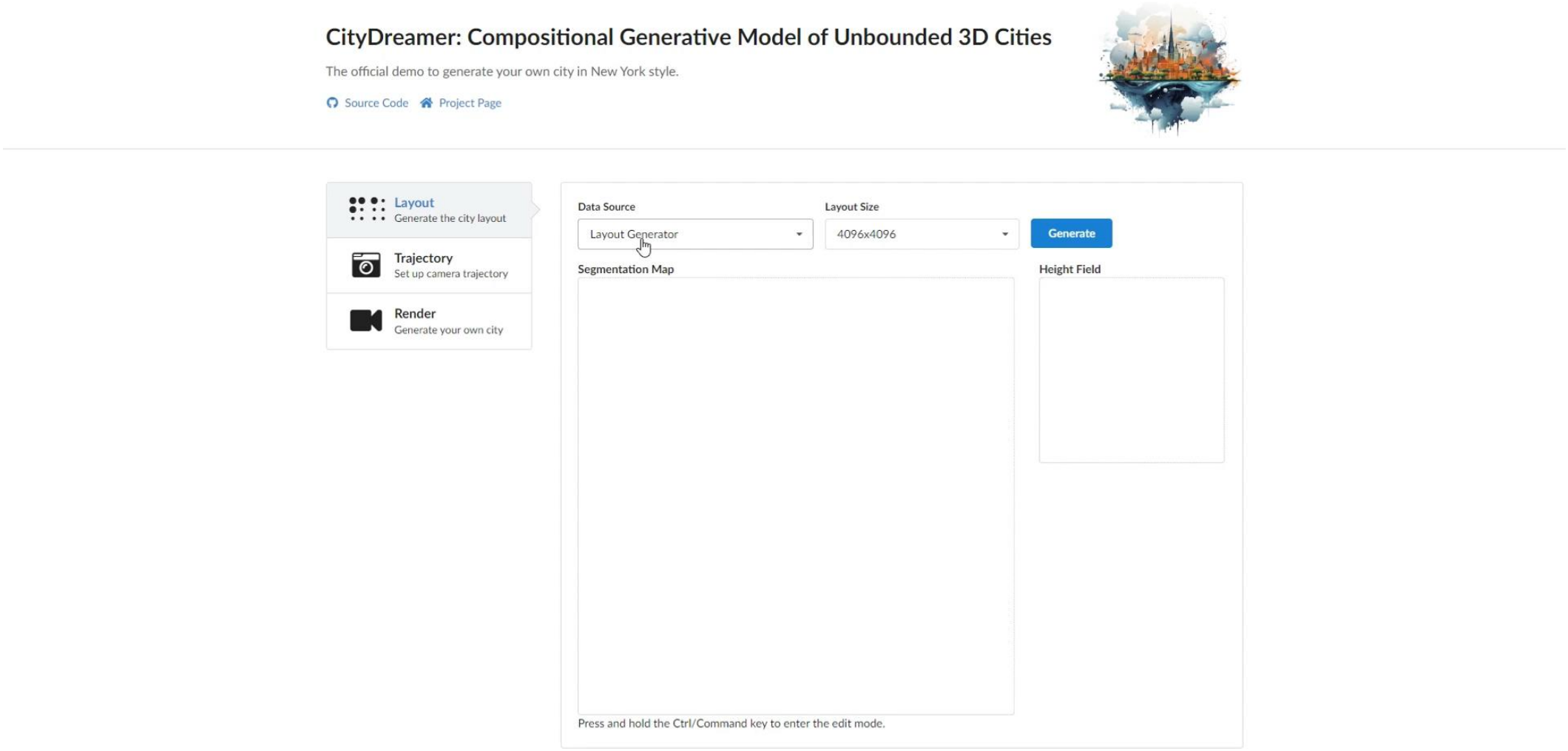
(b) OSM Dataset: Real-world City Layout



(c) Unbounded City Layout Generation

(d) CityDreamer Generated 3D Cities

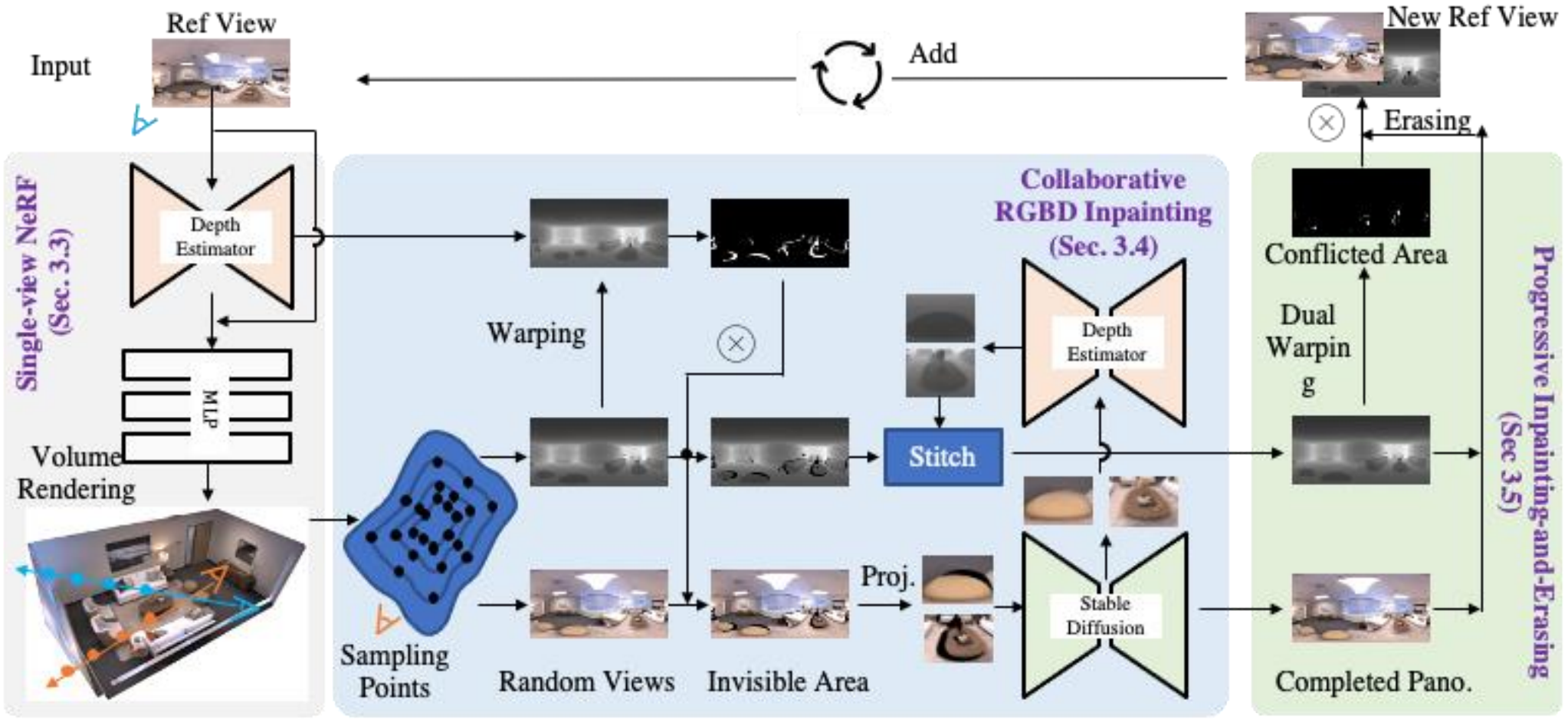
CityDreamer Demo



Online available at https://huggingface.co/spaces/hzxie/city-dreamer

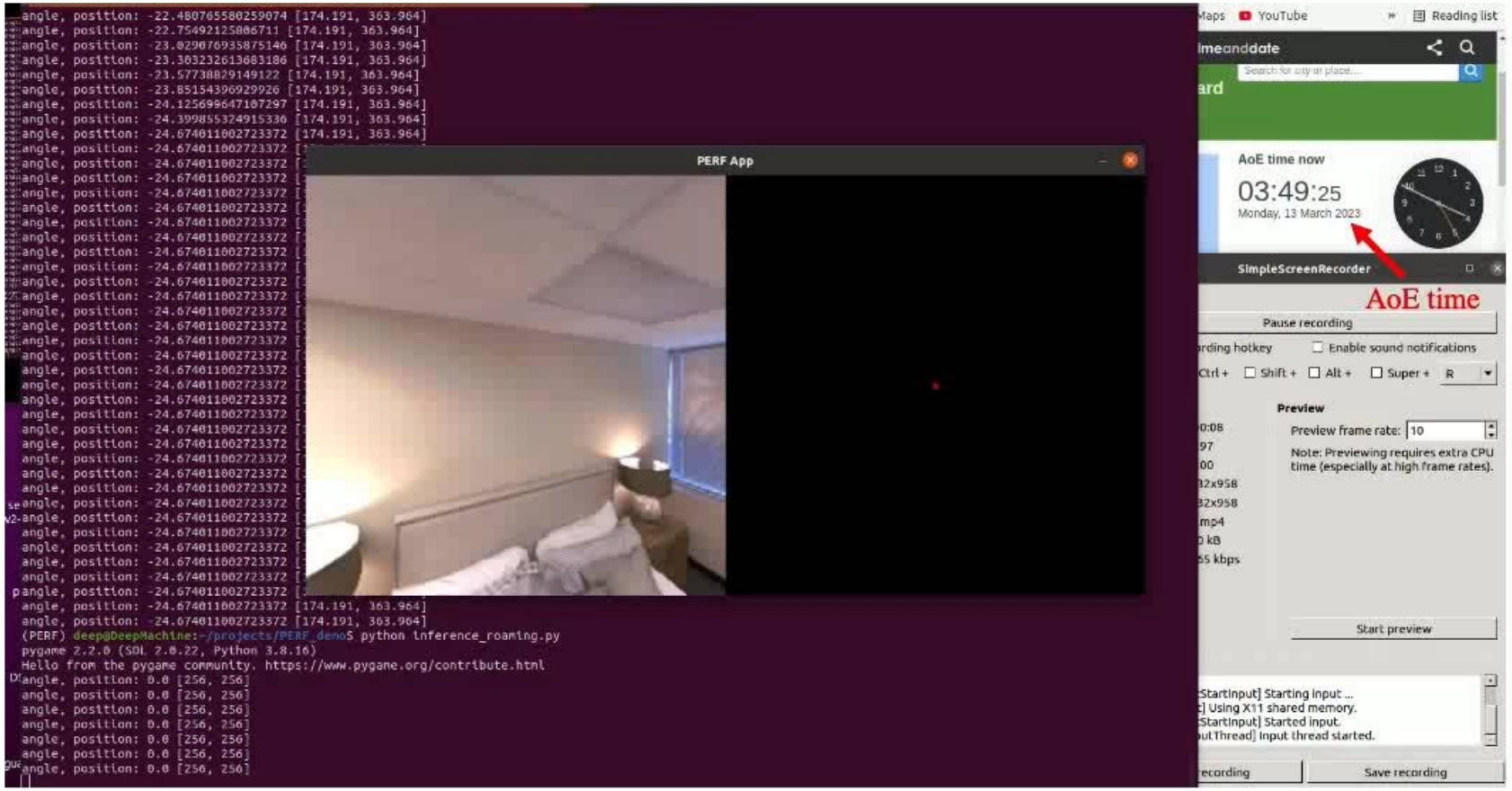
GaussianCity: Real-Time Rendering

PERF: Panoramic Neural Radiance Field



PERF: Panoramic Neural Radiance Field

PERF: Panoramic Neural Radiance Field

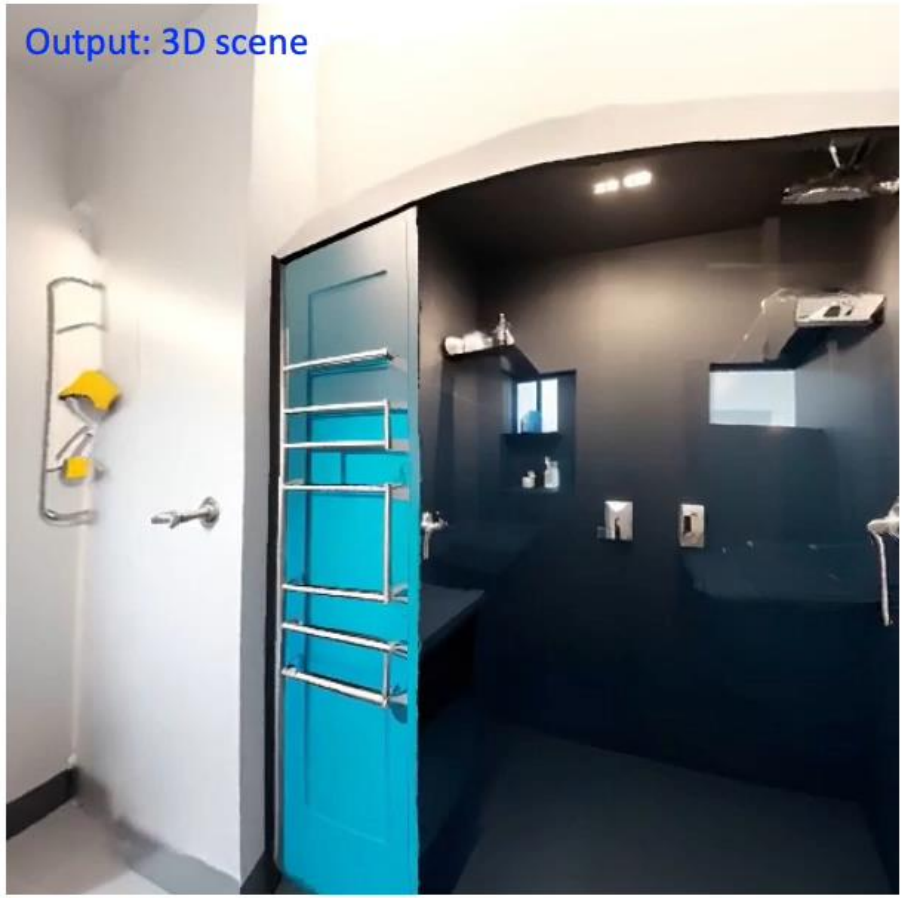


Online rendering (**NOT** saved images)

Controlled by W, A, S, D keys

PERF: Panoramic Neural Radiance Field

Input: text [A bathroom]



Input: text [A Chinese kitchen]

Text to 3D Scene = Text to 2D Panorama + PERF

PERF: Panoramic Neural Radiance Field

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Input: text [A large bedroom]

Input: text [A living room with one TV]

Text to 3D Scene = Text to 2D Panorama + PERF

PERF: Panoramic Neural Radiance Field

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

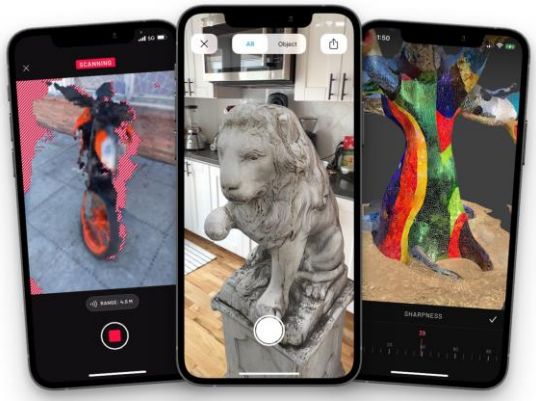
Input: text [A kitchen]

Input: text [A large bedroom with colorful beds]

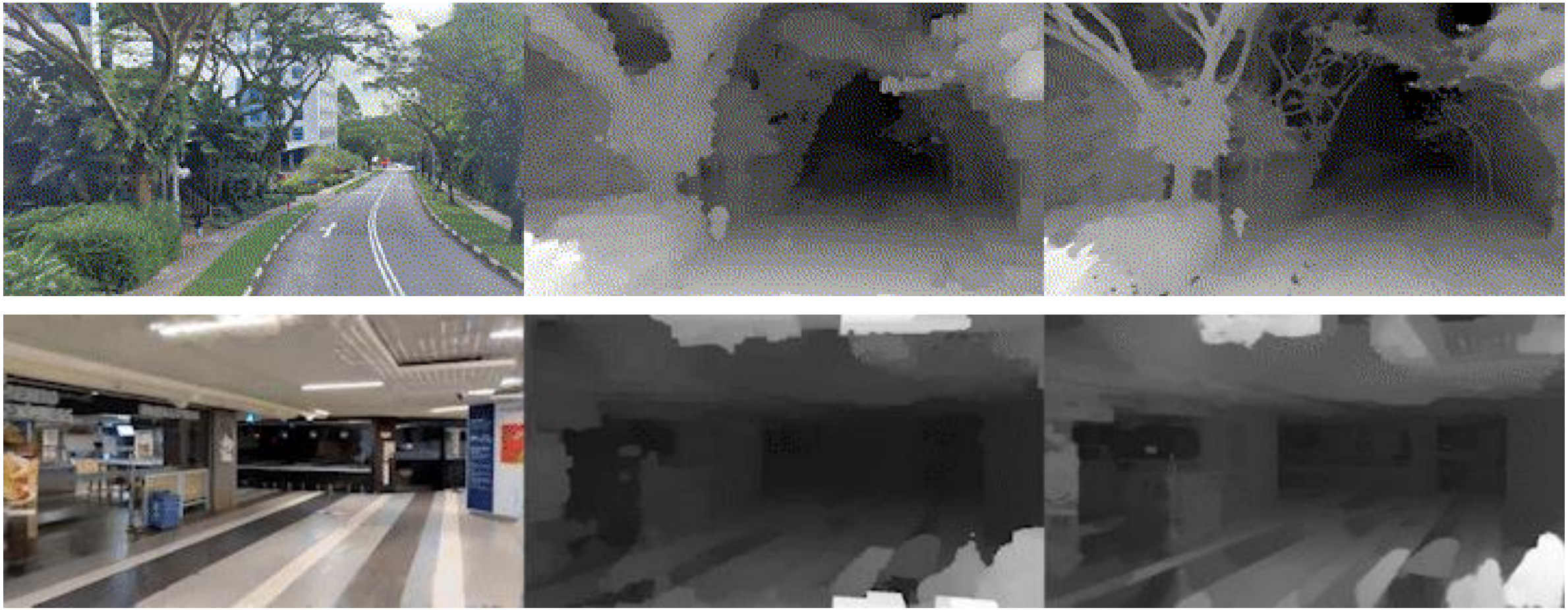
Text to 3D Scene = Text to 2D Panorama + PERF

F2NeRF: Mobile 3D Scene Reconstruction

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE



F2NeRF: Mobile 3D Scene Reconstruction



Visual AIGC

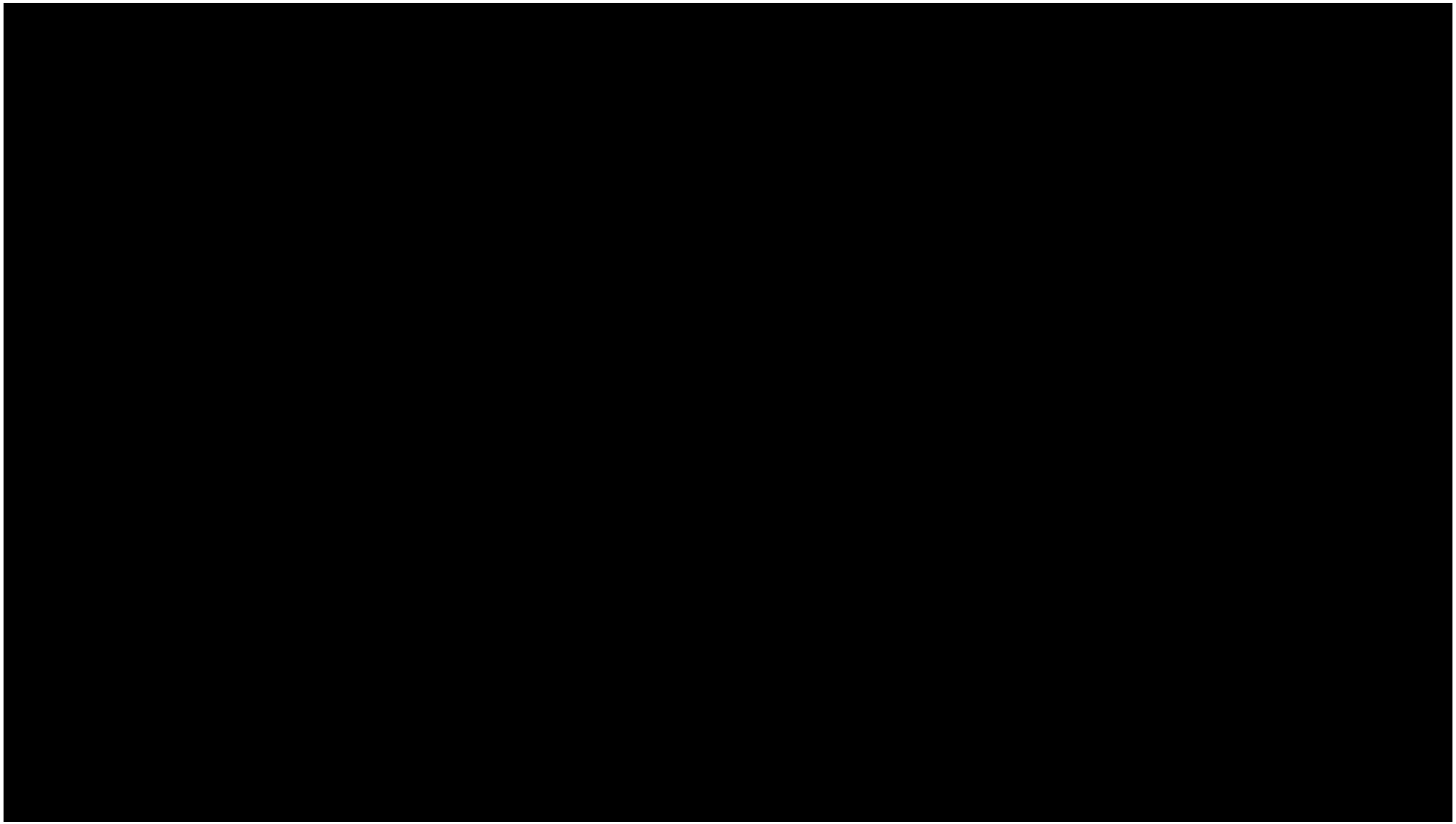
Avatar

Foundation Model

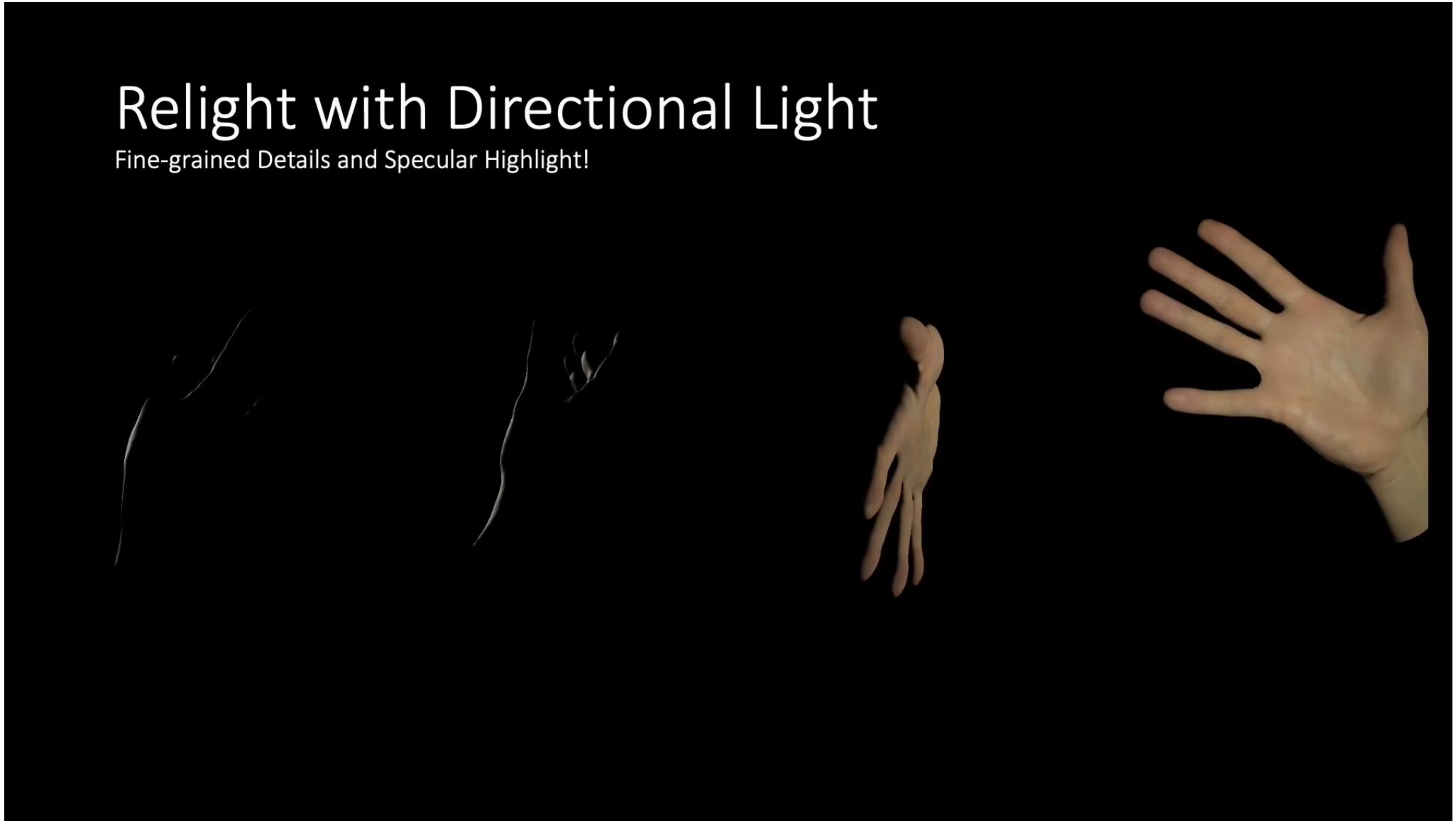
Object

Scene

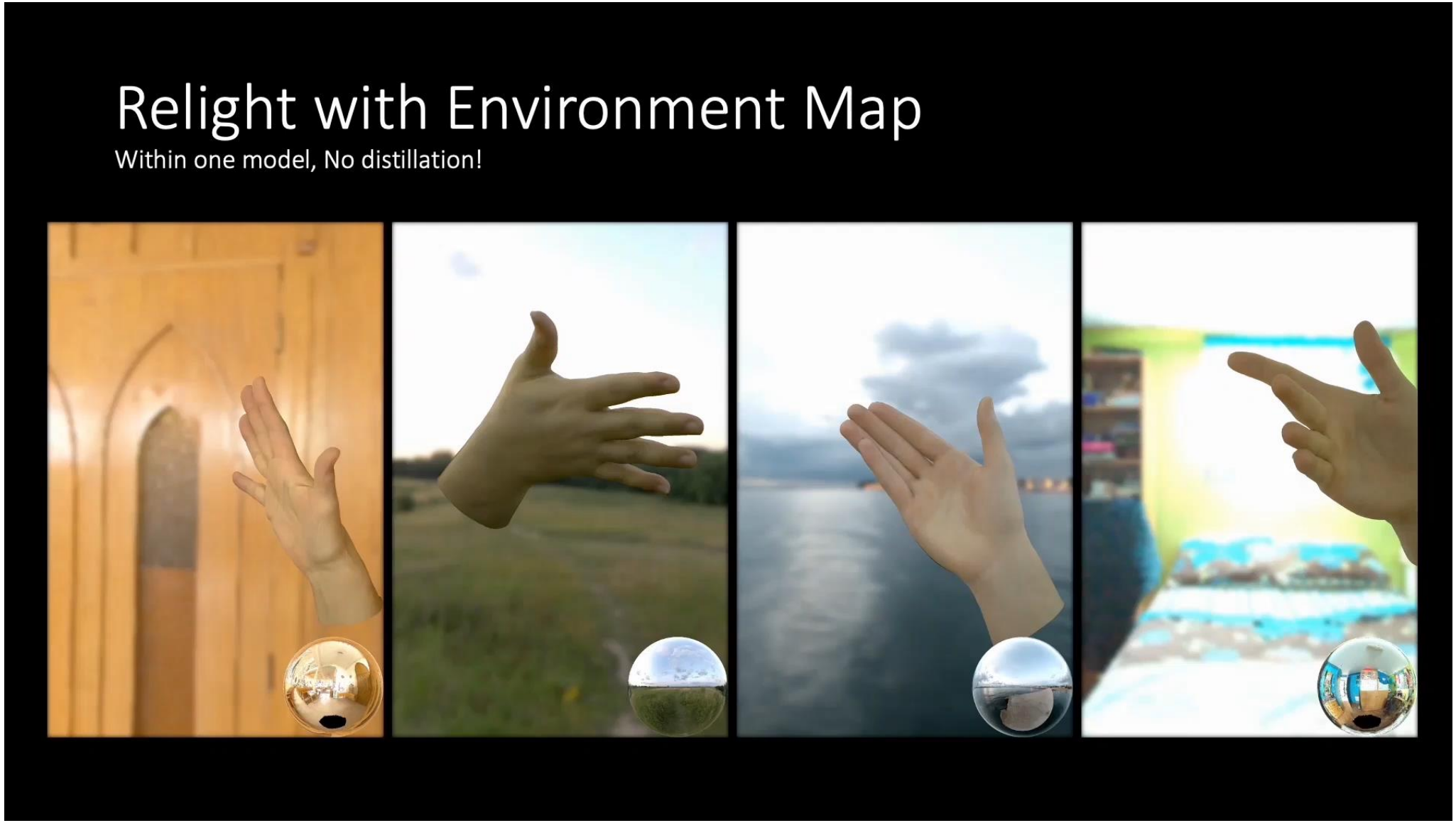
URHand: Universal Model for Relightable Hand



URHand: High-fidelity Details



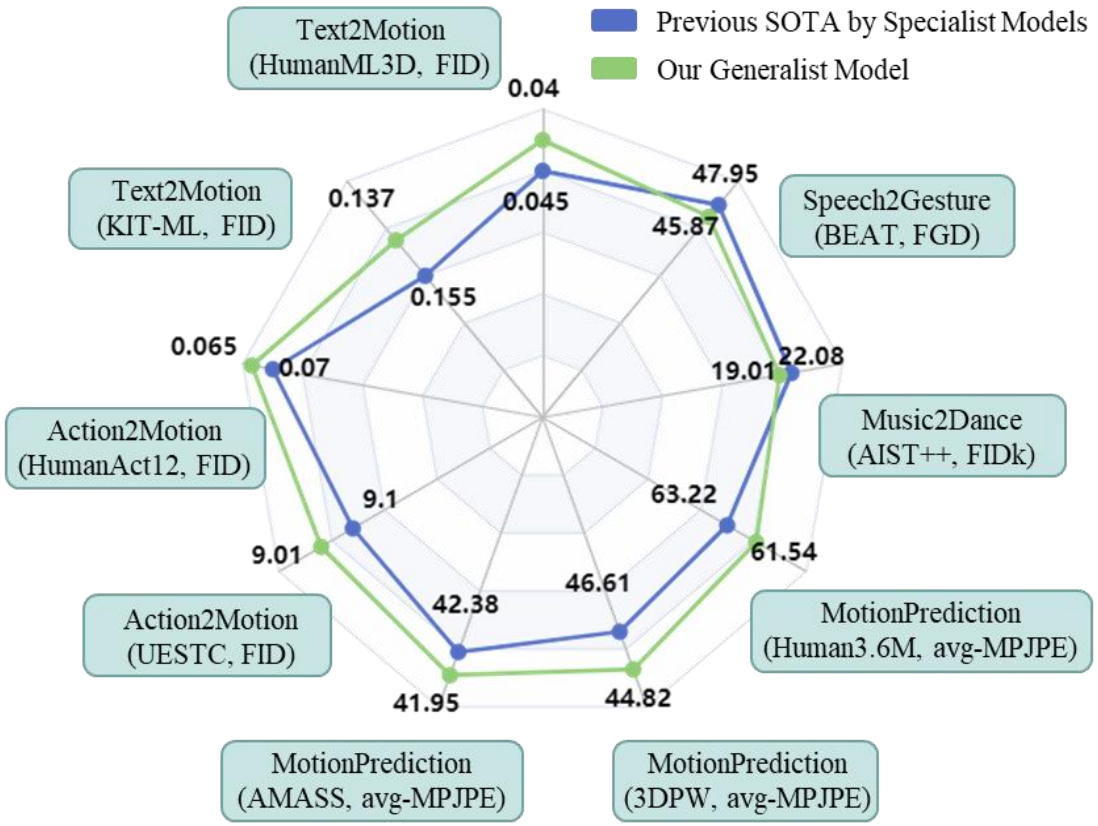
URHand: Generalize to Arbitrary Lighting and ID

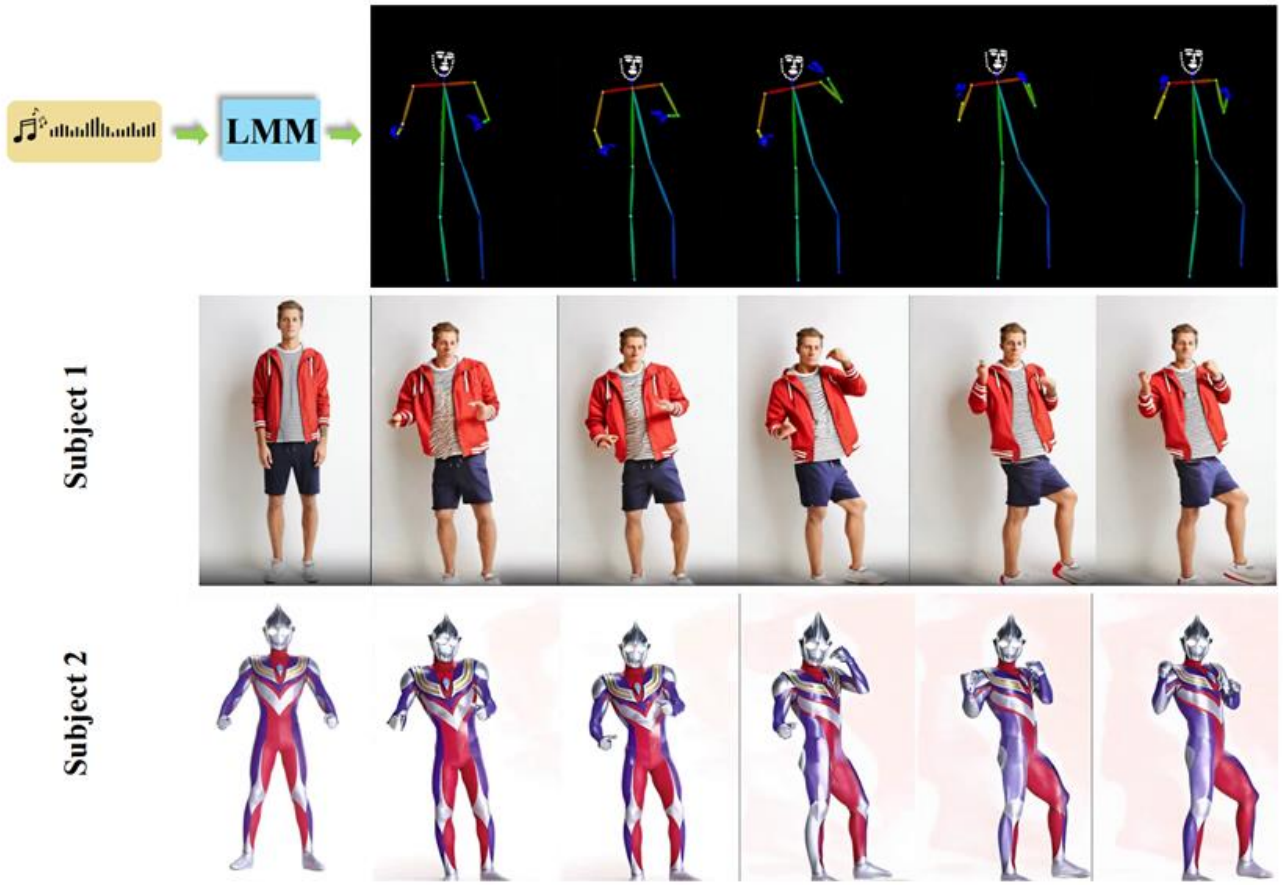


URHand: Personal Relightable Hand from iPhone

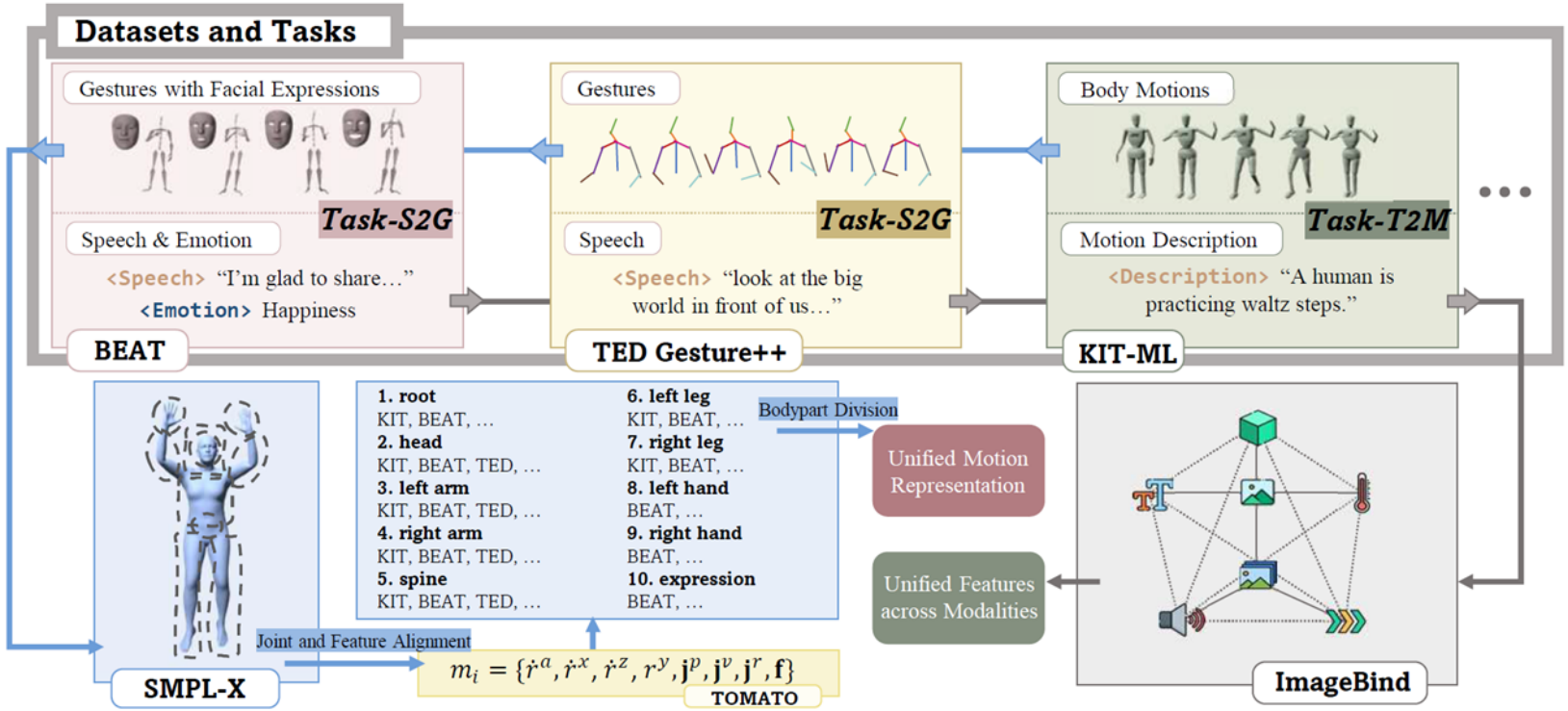
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

LMM | Large Motion Model



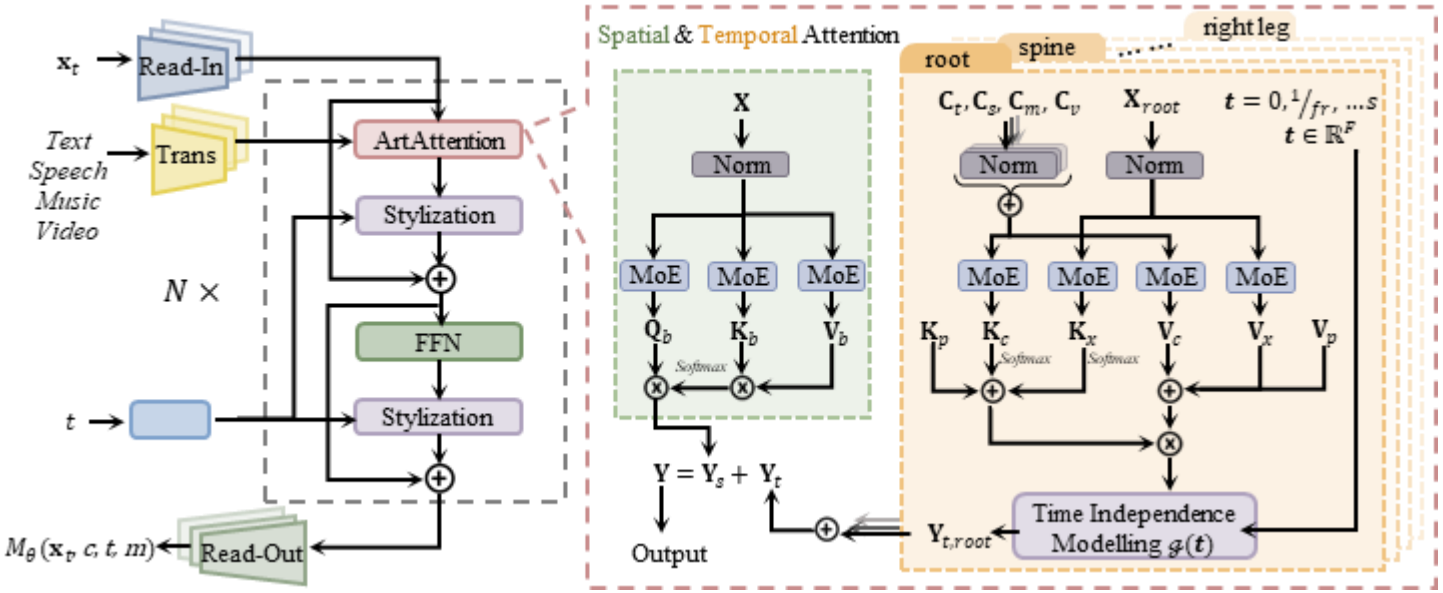


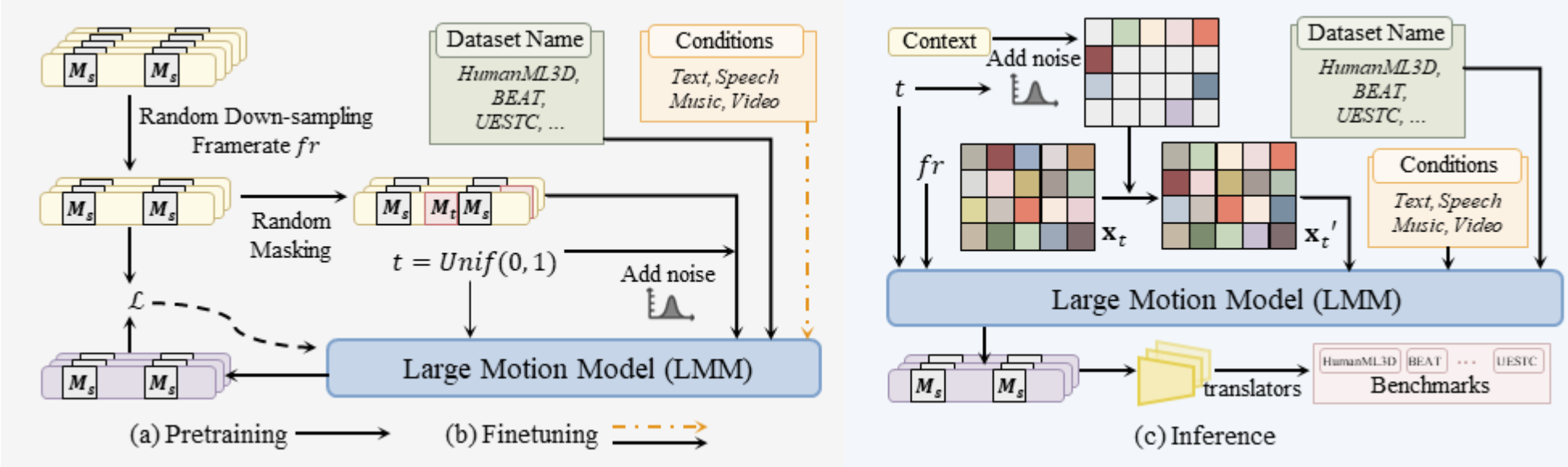
LMM | MotionVerse

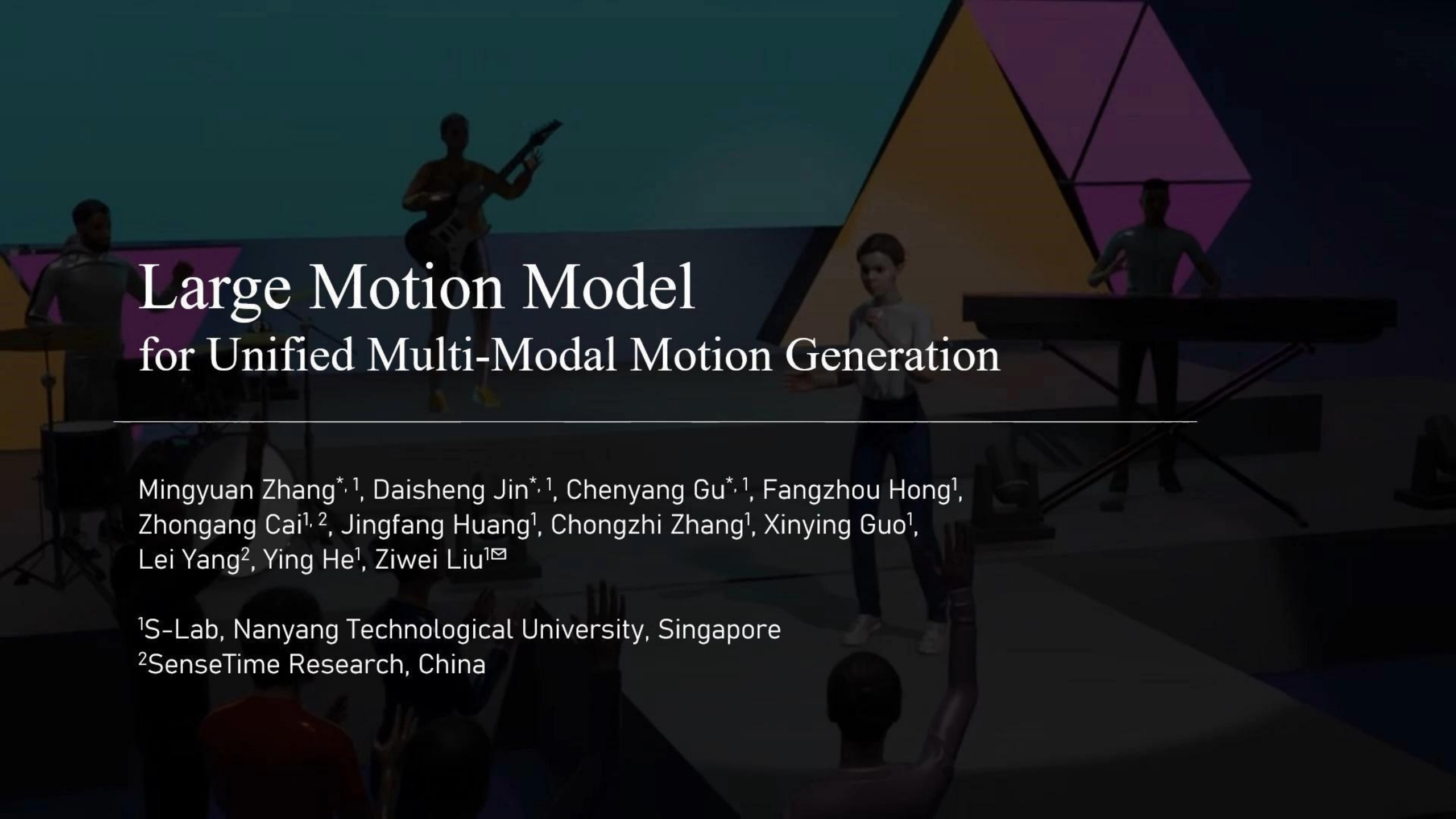


Dataset	#Seq	#Frames	Repr	Condition
HumanML3D [33]	14614	2M	H3D	Text
KIT-ML [109]	2485	245K	H3D	Text
Motion-X [78]	50863	9M	SMPLX	Text
BABEL [110]	5123	7M	SMPLX	Text
UESTC [48]	25600	10M	SMPL	Action
HumanAct12 [35]	1191	90K	Kpt3D	Action
NTU-RGB-D 120 [86, 119]	139656	10M	Kpt3D	Action
AMASS [95]	14244	20M	SMPLX	-
3DPW [98]	81	140K	SMPL	Video
Human3.6M [47]	210	530K	Kpt3D	Video
TED-Gesture++ [153]	34491	10M	Kpt3D	Speech
TED-Expressive [87]	27221	8M	Kpt3D	Speech
Speech2Gesture-3D [60]	1047	1M	Kpt3D	Speech
BEAT [84]	1639	18M	Kpt3D	Speech
AIST++ [69]	1408	1M	SMPL	Music
MPI-INF-3DHP [99]	16	1M	Kpt3D	Video
Total	320K	100M	-	-

LMM | Architecture and Training Scheme







Large Motion Model
for Unified Multi-Modal Motion Generation

Mingyuan Zhang[*,1], Daisheng Jin[*,1], Chenyang Gu[*,1], Fangzhou Hong[1], Zhongang Cai[1,2], Jingfang Huang[1], Chongzhi Zhang[1], Xinying Guo[1], Lei Yang[2], Ying He[1], Ziwei Liu[1]✉

[1]S-Lab, Nanyang Technological University, Singapore
[2]SenseTime Research, China

SMPLer-X | Motion Capture Foundation Model

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.

SMPLer-X | Data & Model Scaling

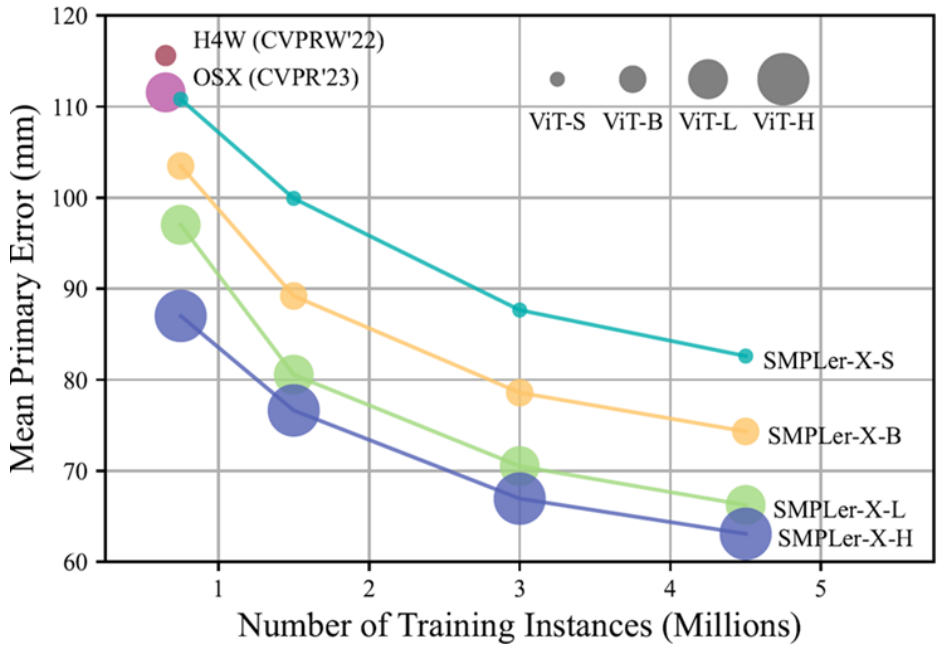


Table 2: **Foundation Models.** We study the scaling law of the amount of data and the model sizes. The metrics are MPJPE for 3DPW, and PVE for other evaluation benchmarks. Foundation models are named "SMPLer-X-MN", where M indicates the size of ViT backbone (S, B, L, H), N is the number of datasets used in the training. FPS: inference speed (frames per second) on a V100 GPU. MPE: mean primary error. AGORA uses the validation set, and EgoBody uses the EgoSet.

#Datasets	#Inst.	Model	#Param.	FPS	AGORA [50]	EgoBody [68]	UBody [39]	3DPW [58]	EHF [51]	MPE
5	0.75M	SMPLer-X-S5	32M	36.2	119.0	114.2	110.1	110.2	100.5	110.8
10	1.5M	SMPLer-X-S10	32M	36.2	116.0	88.6	107.7	97.4	89.9	99.9
20	3.0M	SMPLer-X-S20	32M	36.2	109.2	84.3	70.7	87.5	86.6	87.7
32	4.5M	SMPLer-X-S32	32M	36.2	105.2	82.5	68.1	83.2	74.1	82.6
5	0.75M	SMPLer-X-B5	103M	33.1	102.7	108.1	105.8	104.8	96.1	103.5
10	1.5M	SMPLer-X-B10	103M	33.1	97.8	76.4	107.3	89.9	74.7	89.2
20	3.0M	SMPLer-X-B20	103M	33.1	95.6	75.5	65.3	83.5	73.0	78.6
32	4.5M	SMPLer-X-B32	103M	33.1	88.0	72.7	63.3	80.3	67.3	74.3
5	0.75M	SMPLer-X-L5	327M	24.4	88.3	98.7	110.8	97.8	89.5	97.0
10	1.5M	SMPLer-X-L10	327M	24.4	82.6	69.7	104.0	82.5	64.0	80.6
20	3.0M	SMPLer-X-L20	327M	24.4	80.7	66.6	61.5	78.3	65.4	70.5
32	4.5M	SMPLer-X-L32	327M	24.4	74.2	62.2	57.3	75.2	62.4	66.2
5	0.75M	SMPLer-X-H5	662M	17.5	89.0	87.4	102.1	88.3	68.3	87.0
10	1.5M	SMPLer-X-H10	662M	17.5	81.4	65.7	100.7	78.7	**56.6**	76.6
20	3.0M	SMPLer-X-H20	662M	17.5	77.5	63.5	59.9	**74.4**	59.4	67.0
32	4.5M	SMPLer-X-H32	662M	17.5	**69.5**	**59.5**	**54.5**	75.0	56.8	**63.1**

SMPLer-X | SoTA on Sevem Benchmarks

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

BEDLAM [smplx4]	179.5	132.2	177.5	131.4	131.0	96.5	25.8	38.8/39.0	129.6	95.9	27.8	36.6/36.7
Hand4Whole-finetuned [smplx7]	144.1	96.0	141.1	92.7	135.5	90.2	41.6	46.3/48.1	132.6	87.1	46.1	44.3/46.2
BEDLAM-finetuned [smplx5]	142.2	102.1	141.0	101.8	103.8	74.5	**23.1**	**31.7/33.2**	102.9	74.3	**24.7**	**29.9/31.3**
PyMAF-X [smplx8]	141.2	94.4	140.0	93.5	125.7	84.0	35.0	44.6/45.6	124.6	83.2	37.9	42.5/43.7
OSX [smplx10]	130.6	85.3	127.6	83.3	122.8	80.2	36.2	45.4/46.1	119.9	78.3	37.9	43.0/43.9
HybrIK-X [smplx9]	120.5	73.7	115.7	72.3	112.1	68.5	37.0	46.7/47.0	107.6	67.2	38.5	41.2/41.4
SMPLer-X [smplx11]	107.2	68.3	104.1	66.3	99.7	63.5	29.9	39.1/39.5	96.8	61.7	31.4	36.7/37.2
AiOS (0.3 score) [smplx13]	103.0	63.5	100.8	62.6	98.9	61.0	27.7	42.5/43.4	96.8	60.1	29.2	40.1/40.9
SMPLer-X (AiOS) [smplx14]	102.4	63.8	99.5	62.1	98.3	61.2	30.3	40.4/40.7	95.5	59.6	31.7	37.9/38.2
AiOS (0.5 score) [smplx12]	**97.8**	**61.3**	**96.0**	**60.7**	**91.9**	**57.6**	24.6	38.7/39.6	**90.2**	**57.1**	25.7	36.4/37.3

Table 6: **UBody.** † denotes the methods that are finetuned on the UBody training set. ∗ denotes the methods that are trained on UBody training set only.

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
PIXIE [15]	61.7	12.2	4.2	168.4	55.6	45.2
Hand4Whole [46]	44.8	8.9	2.8	104.1	45.7	27.0
OSX [39]	42.4	10.8	2.4	92.4	47.7	24.9
OSX [39]†	42.2	**8.6**	**2.0**	81.9	41.5	**21.2**
SMPLer-X-B1∗	38.5	10.8	3.0	64.8	45.4	22.3
SMPLer-X-L20	33.2	10.6	2.8	61.5	43.3	23.1
SMPLer-X-L32	**30.9**	10.2	2.7	**57.3**	**39.2**	21.6
SMPLer-X-L-20†	31.9	10.3	2.8	57.4	40.2	21.6

Table 7: **EgoBody-EgoSet.** † denotes the methods that are finetuned on the EgoBody-EgoSet training set. ∗ denotes the methods that are trained on EgoBody-EgoSet training set only.

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
Hand4Whole [46]	58.8	**9.7**	3.7	121.9	50.0	42.5
OSX [39]	54.6	11.6	3.7	115.7	50.6	41.1
OSX [39]†	45.3	10.0	3.0	82.3	46.8	35.2
SMPLer-X-B1∗	56.1	10.7	3.5	87.2	49.4	34.9
SMPLer-X-L20	38.9	9.9	3.0	66.6	42.7	31.8
SMPLer-X-L32	**36.3**	9.8	**2.9**	**62.2**	**41.4**	**30.7**
SMPLer-X-L20†	37.8	9.9	**2.9**	63.6	42.5	30.8

Table 8: **ARCTIC.** † and ∗ denote the methods that are finetuned on the ARCTIC training set and trained on the ARCTIC training set only, respectively.

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
Hand4Whole [46]	63.4	18.1	4.0	136.8	54.8	59.2
OSX [39]	56.9	**17.5**	3.9	102.6	56.5	44.6
OSX [39]†	33.0	18.8	3.3	58.4	39.4	30.4
SMPLer-X-B1∗	45.2	18.9	3.4	66.6	42.5	34.0
SMPLer-X-L10	46.9	18.1	**2.3**	76.9	50.8	33.2
SMPLer-X-L32	**29.4**	18.9	2.7	**48.6**	**38.8**	**26.8**
SMPLer-X-L10†	33.1	19.0	2.7	54.9	40.1	27.3

Table 9: **DNA-Rendering-HiRes**. † and ∗ are finetuned on the DNA-Rendering-HiRes training set and trained on the DNA-Rendering-HiRes training set only, respectively.

	PA-PVE↓ (*mm*)			PVE↓ (*mm*)		
Method	All	Hands	Face	All	Hands	Face
Hand4Whole [46]	62.8	11.0	4.2	111.4	56.4	52.6
OSX [39]	59.9	10.6	4.3	105.7	55.0	52.5
OSX [39]†	43.5	7.5	3.5	67.1	43.3	38.2
SMPLer-X-B1∗	45.6	7.5	3.4	63.2	40.7	34.2
SMPLer-X-L20	44.4	11.1	4.5	77.7	47.5	43.2
SMPLer-X-L32	**35.8**	**7.2**	**3.2**	**54.4**	**36.7**	**34.0**
SMPLer-X-L20†	37.9	7.3	3.4	56.5	38.4	34.9

SMPLer-X | Benchmarking Datasets

Table 1: **Benchmarking EHPS datasets.** For each dataset, we train a model on its training set and evaluate its performance on the *val* set of AGORA and *testing* sets of UBody, EgoBody (EgoSet), 3DPW, and EHF. Datasets are then ranked by mean primary error (MPE). Top-1 values are bolded, and the rest of Top-5 are underlined. #Inst.: number of instances used in training. ITW: in-the-wild. EFT [27], NeuralAnnot (NeA) [47] and UP3D [34] produce pseudo labels.

						AGORA [50]		UBody [39]		EgoBody [68]		3DPW [58]		EHF [51]		
Dataset	#Inst.	Scene	Real/Synthetic	SMPL	SMPL-X	PVE↓	★	PVE↓	★	PVE↓	★	MPJPE↓	★	PVE↓	★	MPE↓
BEDLAM [5]	951.1K	ITW	Syn	-	Yes	164.7	4	132.5	8	109.1	2	**98.1**	1	**81.1**	1	**117.1**
SynBody [63]	633.5K	ITW	Syn	-	Yes	166.7	5	144.6	11	136.6	4	106.5	5	112.9	5	133.5
InstaVariety [29]	2184.8K	ITW	Real	NeA	-	195.0	9	125.4	4	140.1	9	100.6	3	110.8	4	134.3
GTA-Human II [3]	1802.2K	ITW	Syn	-	Yes	161.9	3	143.7	10	139.2	8	103.4	4	126.0	12	134.8
MSCOCO [40]	149.8K	ITW	Real	EFT	NeA	191.6	8	107.2	2	139.0	7	121.2	10	116.3	7	135.0
EgoBody-MVSet [68]	845.9K	Indoor	Real	Yes	Yes	190.9	7	191.4	18	127.0	3	99.2	2	101.8	2	142.1
AGORA [50]	106.7K	ITW	Syn	Yes	Yes	**124.8**	1	128.4	6	138.4	6	131.1	12	164.6	24	145.4
Egobody-EgoSet [68]	90.1K	Indoor	Real	Yes	Yes	207.1	15	126.8	5	**103.1**	1	134.4	18	121.4	10	147.5
RICH [22]	243.4K	ITW	Real	-	Yes	195.6	10	168.1	15	137.9	5	115.5	8	127.5	13	148.9
MPII [2]	28.9K	ITW	Real	EFT	NeA	202.1	11	123.9	3	155.5	15	131.9	14	140.8	16	150.8
MuCo-3DHP [45]	465.3K	ITW	Real	Yes	-	187.7	6	185.4	17	146.4	12	119.4	9	134.7	15	154.7
PROX [19]	88.5K	Indoor	Real	-	Yes	204.1	13	180.3	16	151.8	13	132.5	17	122.5	11	158.2
UBody [39]	683.3K	ITW	Real	-	Yes	207.0	14	**78.7**	1	145.6	11	149.4	23	132.1	14	158.5
SPEC [32]	72.0K	ITW	Syn	Yes	-	161.5	2	146.1	12	154.8	14	139.7	21	197.8	27	160.0
CrowdPose [36]	28.5K	ITW	Real	NeA	-	207.1	16	129.8	7	156.9	16	156.3	25	154.5	22	160.9
MPI-INF-3DHP [44]	939.8K	ITW	Real	NeA	NeA	221.5	20	166.7	14	142.7	10	131.6	13	155.5	23	163.6
HumanSC3D [47]	288.4K	Studio	Real	-	Yes	215.2	18	237.8	22	167.3	17	113.0	7	107.1	3	168.1
PoseTrack [1]	28.5K	ITW	Real	EFT	-	218.1	19	161.0	13	180.8	21	150.2	24	149.9	21	172.0
BEHAVE [3]	44.4K	Indoor	Real	Yes	-	208.3	17	205.8	20	175.8	19	132.0	15	145.0	18	173.4
CHI3D [16]	252.4K	Studio	Real	-	Yes	203.3	12	264.7	25	175.7	18	122.6	11	121.0	9	177.5
Human3.6M [23]	312.2K	Studio	Real	Yes	NeA	226.0	21	276.1	26	200.6	24	112.3	6	120.8	8	187.2
DNA-R-HiRes [9]	998.1K	Studio	Real	-	Yes	230.0	22	278.2	27	179.2	20	134.5	19	149.7	20	194.3
3DPW [58]	22.7K	ITW	Real	Yes	NeA	234.0	23	259.3	23	192.6	23	140.6	22	142.9	17	207.2
ARCTIC [14]	1539.1K	Studio	Real	-	Yes	308.5	29	200.7	19	186.4	22	202.5	26	182.5	25	216.1
DNA-R [9]	3992.0K	Studio	Real	-	Yes	274.7	26	341.5	30	214.4	27	138.4	20	115.5	6	216.9
UP3D [34]	7.1K	ITW	Real	UP3D	-	257.5	24	224.1	21	216.6	28	211.5	27	194.8	26	220.9
Talkshow [64]	3326.9K	Indoor	Real	-	Yes	286.4	27	133.2	9	203.6	25	291.3	29	201.9	28	223.3
FIT3D [13]	1779.3K	Studio	Real	-	Yes	329.7	30	404.0	31	213.8	26	132.1	16	148.1	19	245.5
MTP [48]	3.2K	ITW	Real	Yes	Yes	272.7	25	284.9	28	273.2	29	265.2	28	244.6	29	268.1
OCHuman [69]	2.5K	ITW	Real	EFT	-	307.1	28	263.3	24	279.3	30	293.4	30	281.7	30	285.0
LSPET [25]	2.9K	ITW	Real	EFT	-	365.7	31	292.6	29	340.1	31	339.8	31	316.3	31	330.9
SSP3D [55]	311	ITW	Real	Yes	-	549.8	32	522.4	32	548.1	32	439.0	32	539.5	32	519.8

SMPLer-X | Systematic Study on Data

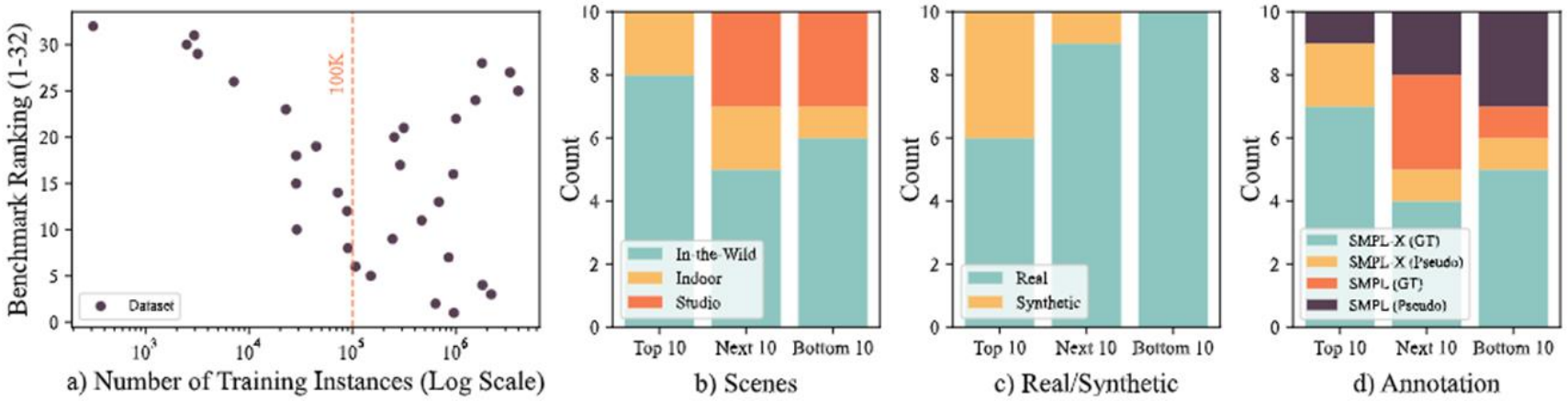


Figure 3: **Analysis on dataset attributes.** We study the impact of a) the number of training instances, b) scenes, c) real or synthetic appearance, and d) annotation type, on dataset ranking in Table 1.

Training Instances

Scenes

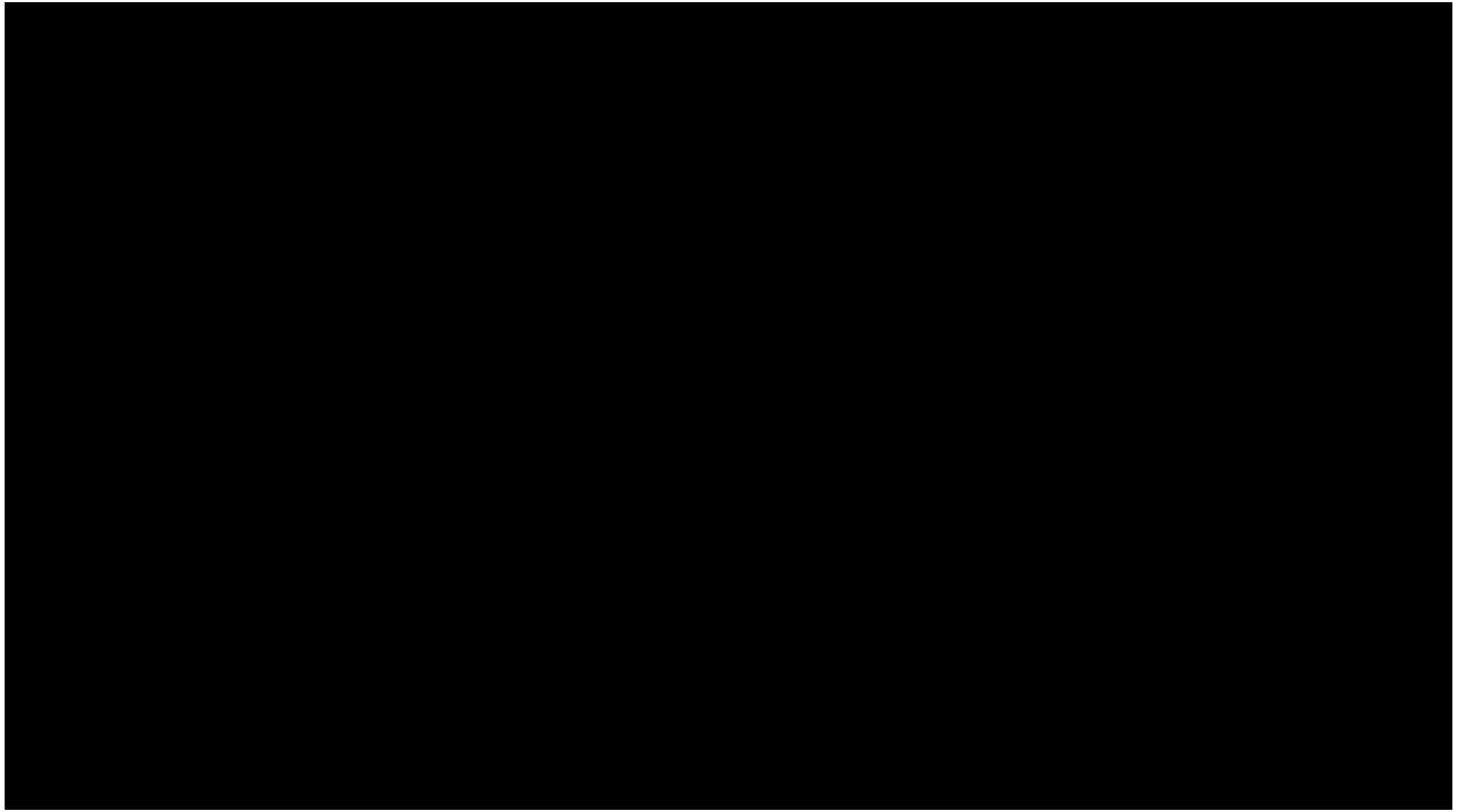
Real/ Synthetic

Annotation

SMPLer-X | Demo

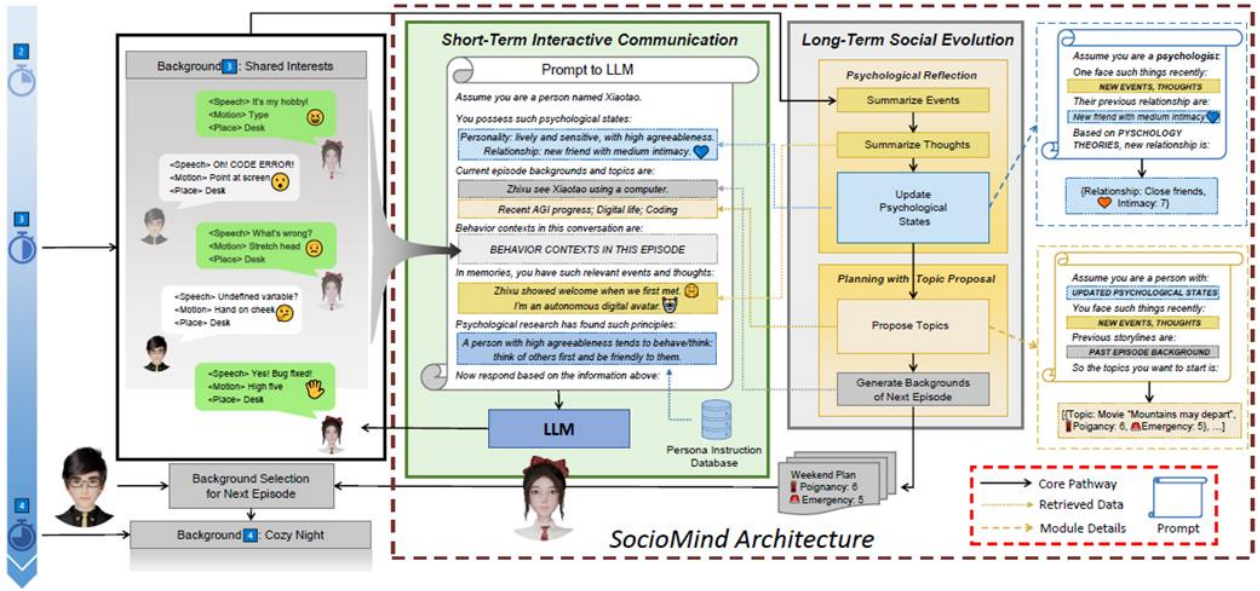
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Digital Life Project (DLP)

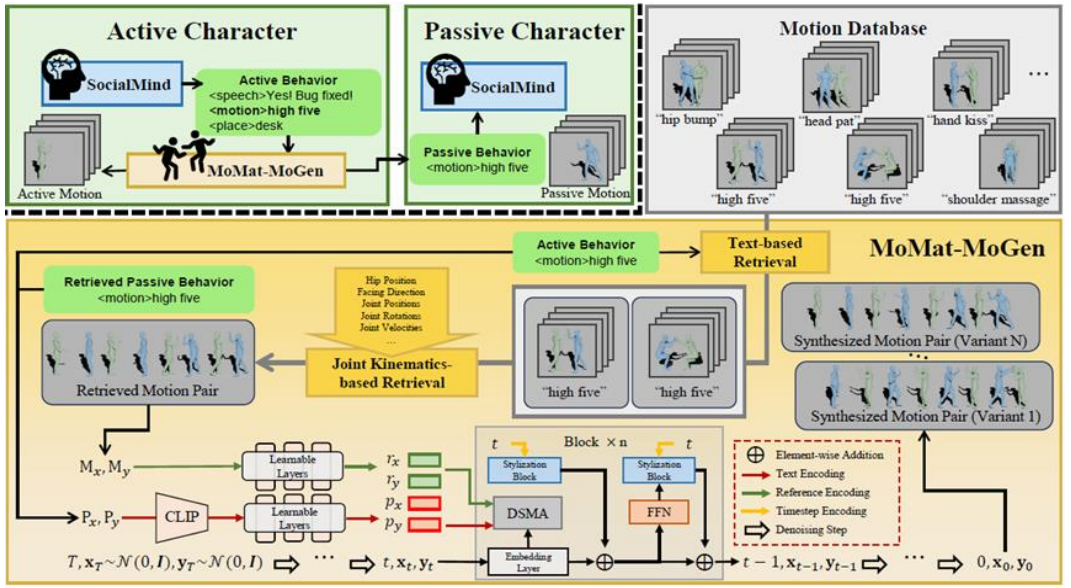


Body & Facial Animation

Digital Brain & Body

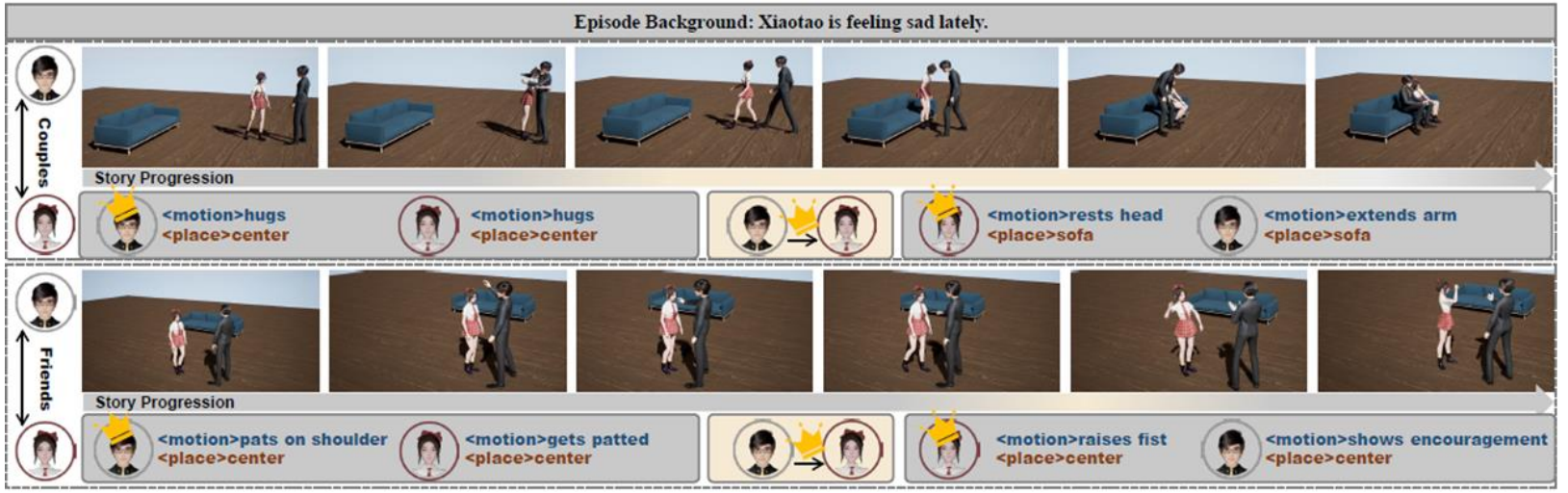


SocioMind: LLM-empowered Simulation of Human Psychology



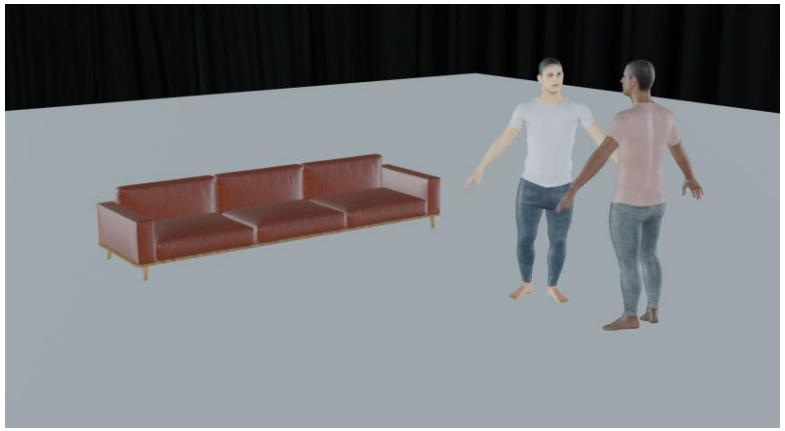
MoMat-MoGen: High-quality Interactive Motion Synthesis

How relationships affect behaviors?



3D Characters with Social Intelligence

"Friends"



"Couples"

“First Meet”

<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

“Music Lovers”

<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

“Shared Interests”

<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

“Cozy Night”

<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

Visual AIGC

Avatar

Thank You!

Object

Scene