Visual AIGC with Foundation Models

Ziwei Liu 刘子纬

Nanyang Technological University

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

2023

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

AI-Generated Content

Movie

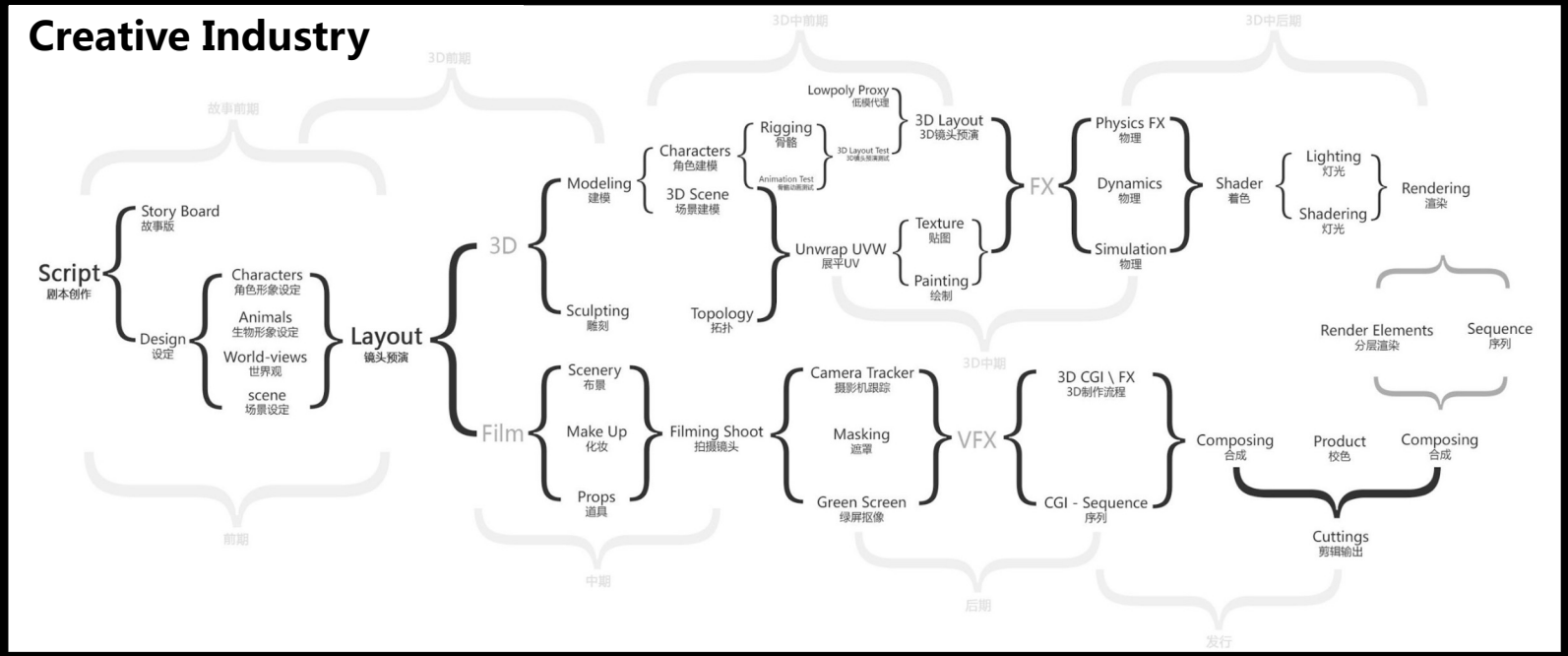
Game

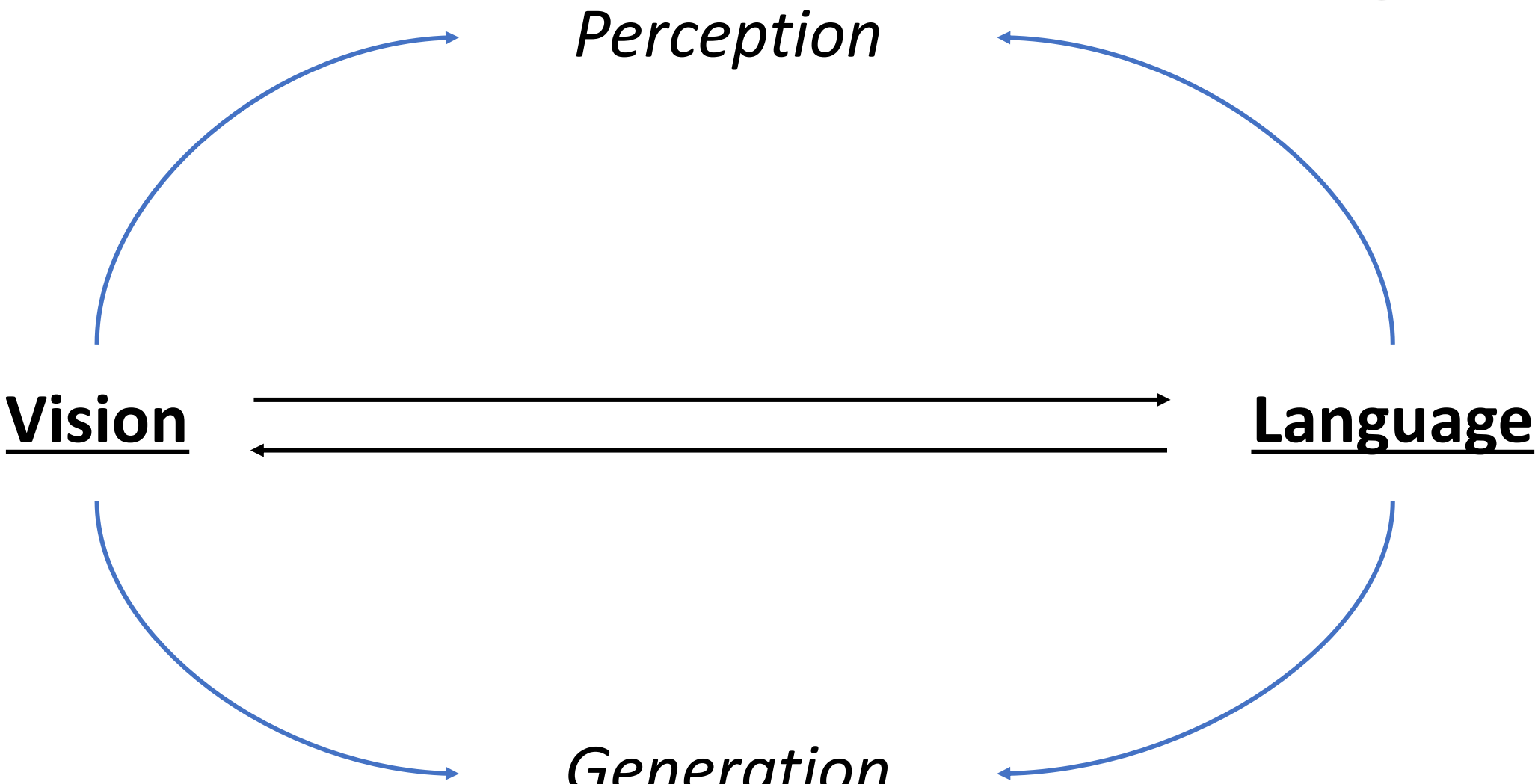
Anime

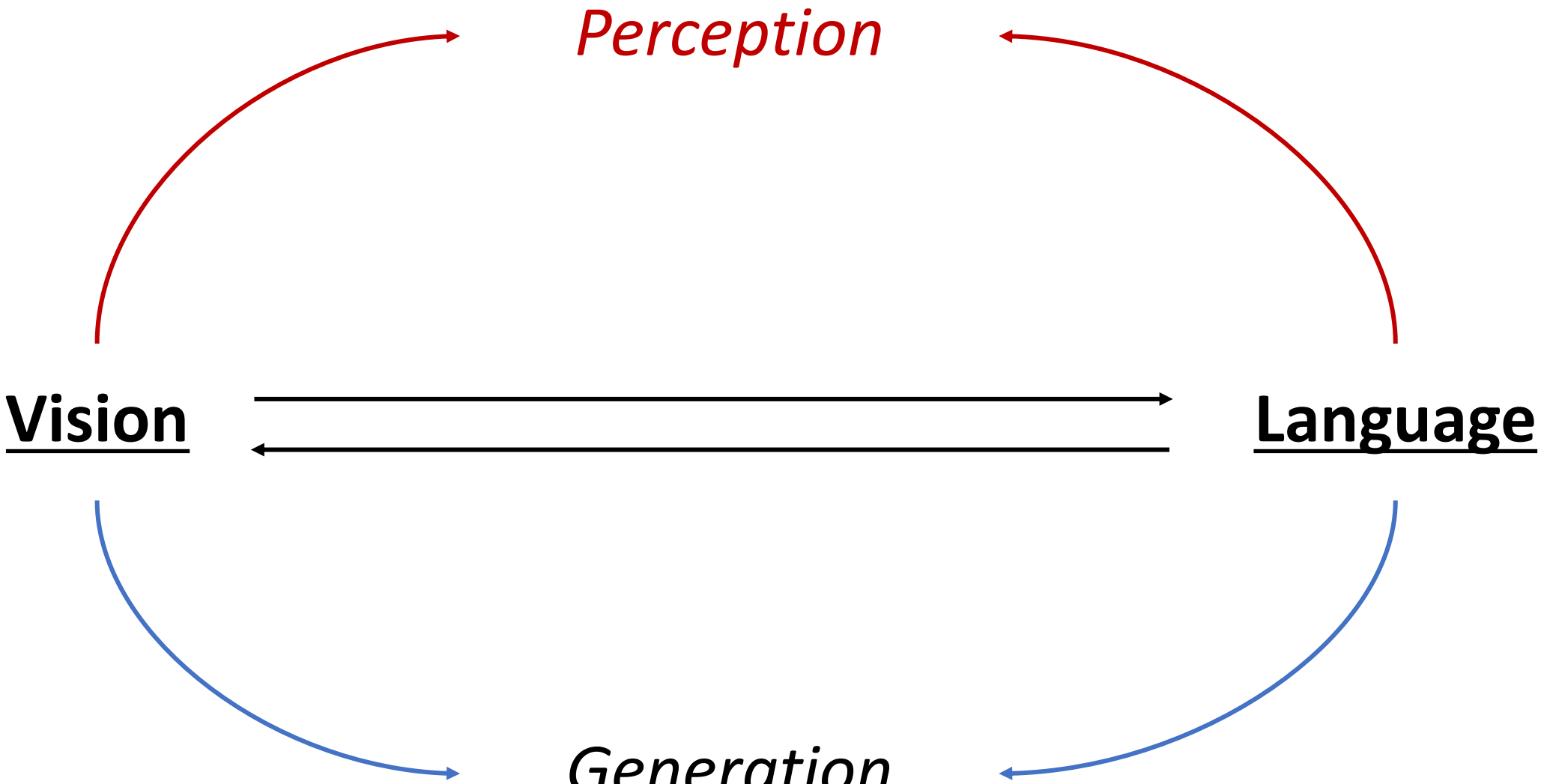
VTuber

Virtual Beings

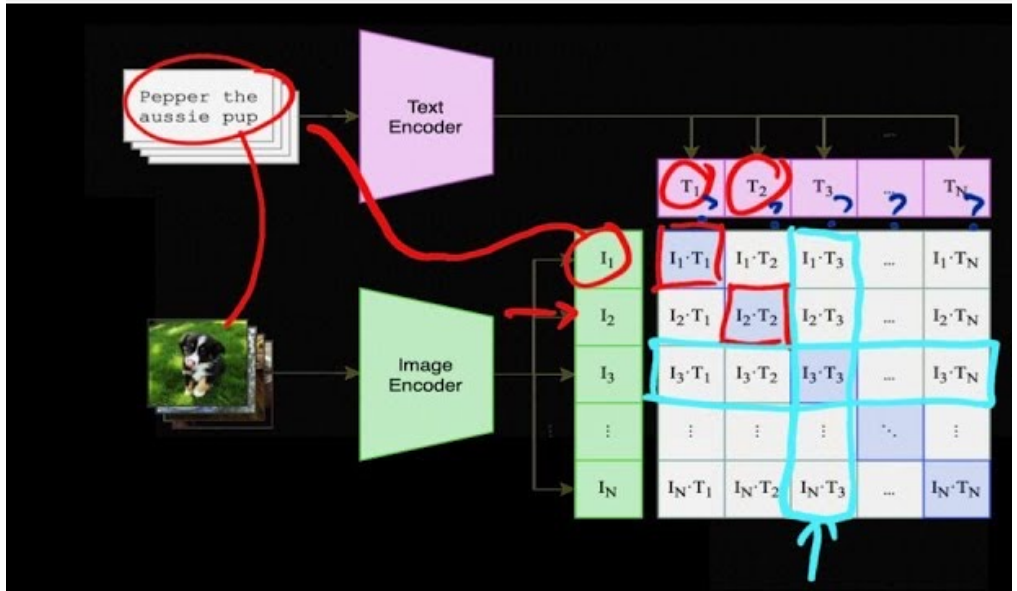
Creative Industry




Perception
Vision
Language
Generation


Perception
Vision
Language
Generation

CLIP & GPT-4



User What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

CoOp & CoCoOp

Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

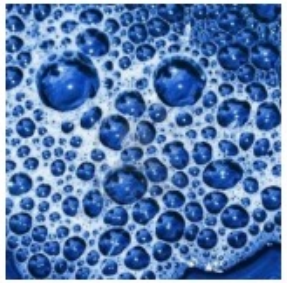
(a)

Flowers102

Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**94.51**

(b)

Describable Textures (DTD)



Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**63.58**

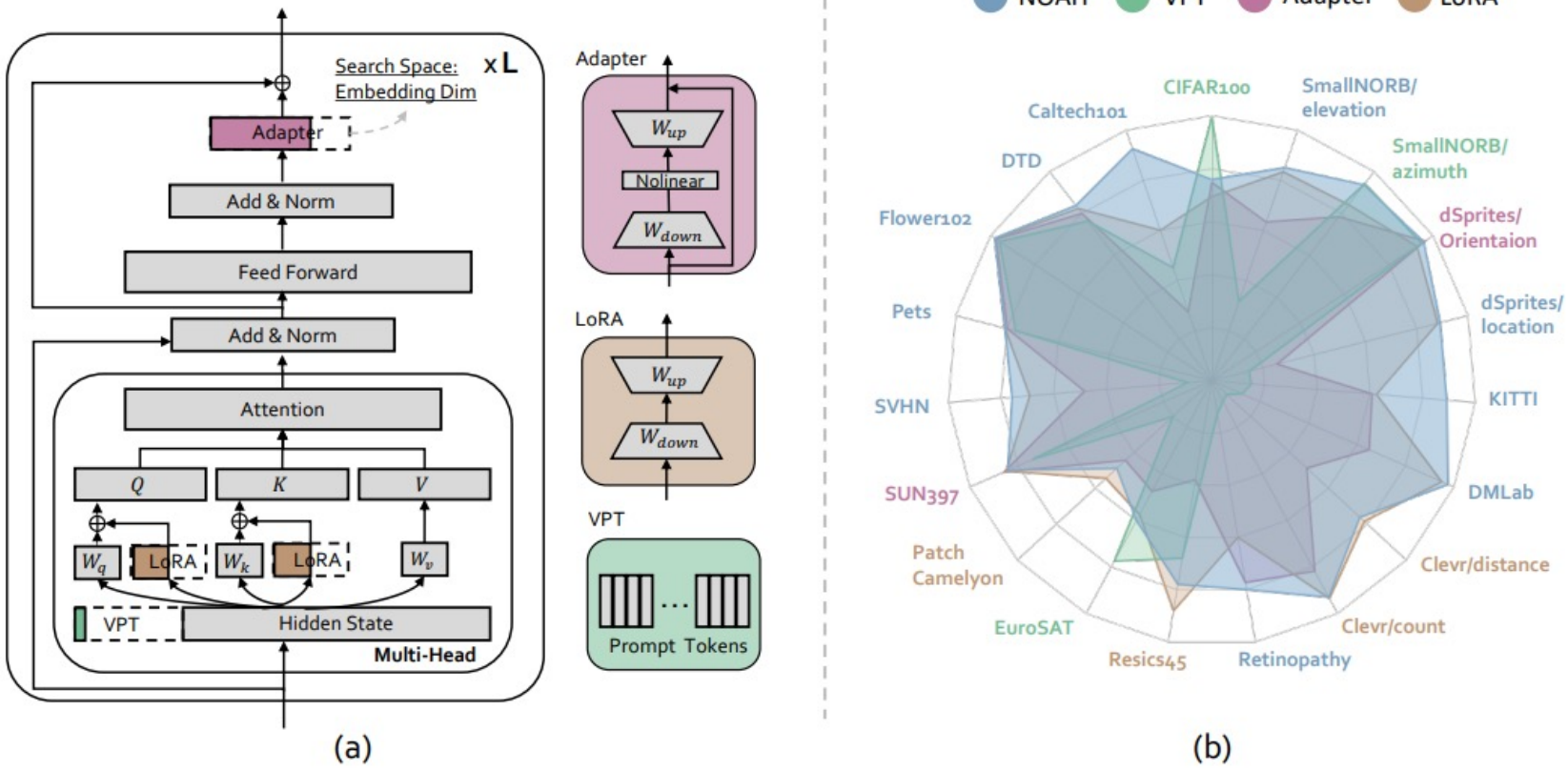
(c)

EuroSAT

Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**83.53**

(d)

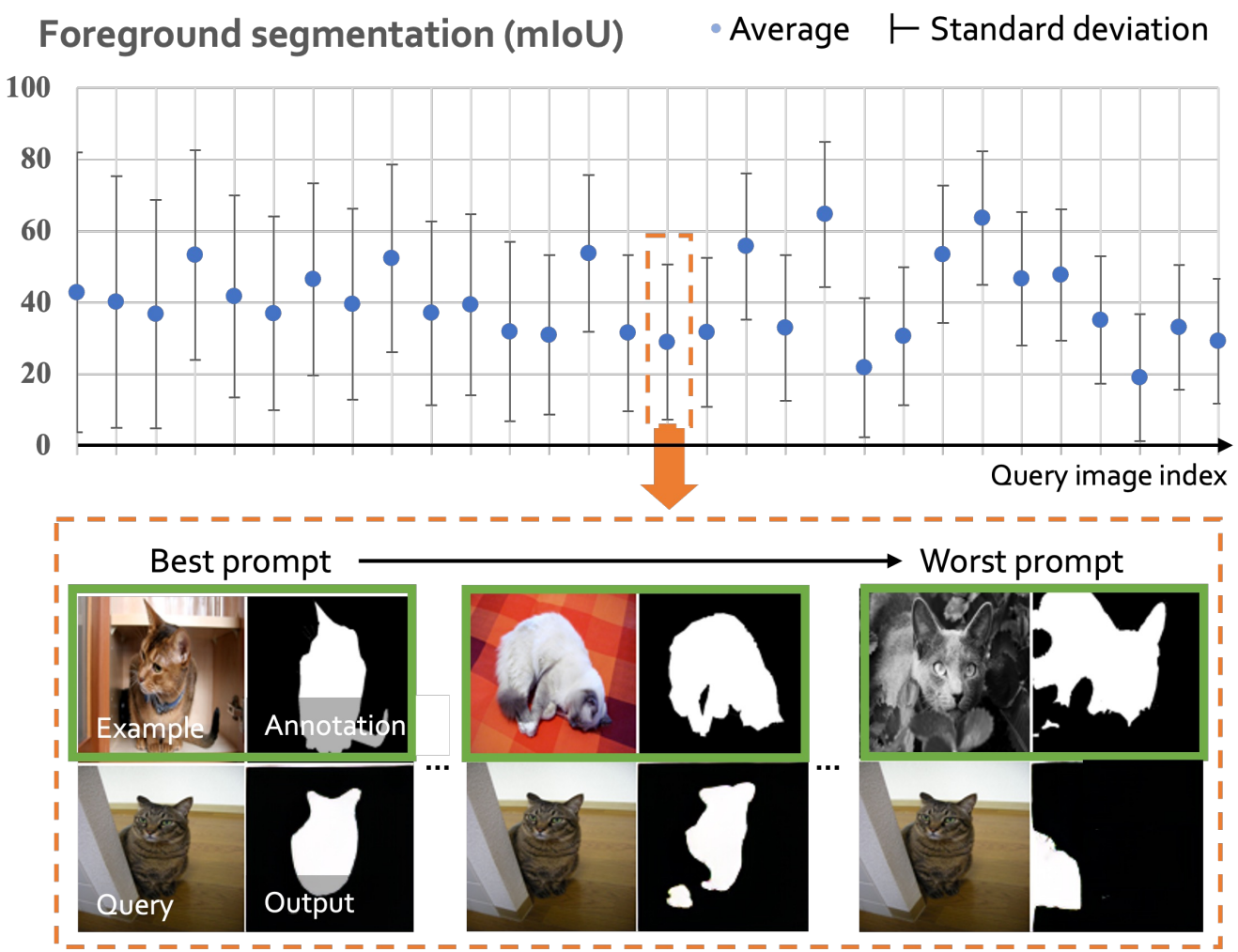
NOAH


Search Space:
Embedding Dim
x L
Adapter
Add & Norm
Feed Forward
Add & Norm
Attention
Q
K
V
LoRA
LoRA
VPT
Hidden State
Multi-Head
Adapter
Nolinear
LoRA
VPT
Prompt Tokens
(a)
NOAH
VPT
Adapter
LoRA
CIFAR100
SmallNORB/ elevation
SmallNORB/ azimuth
dSprites/ Orientaion
dSprites/ location
KITTI
DMLab
Clevr/distance
Clevr/count
Retinopathy
Resics45
EuroSAT
Patch Camelyon
SUN397
SVHN
Pets
Flower102
DTD
Caltech101
(b)

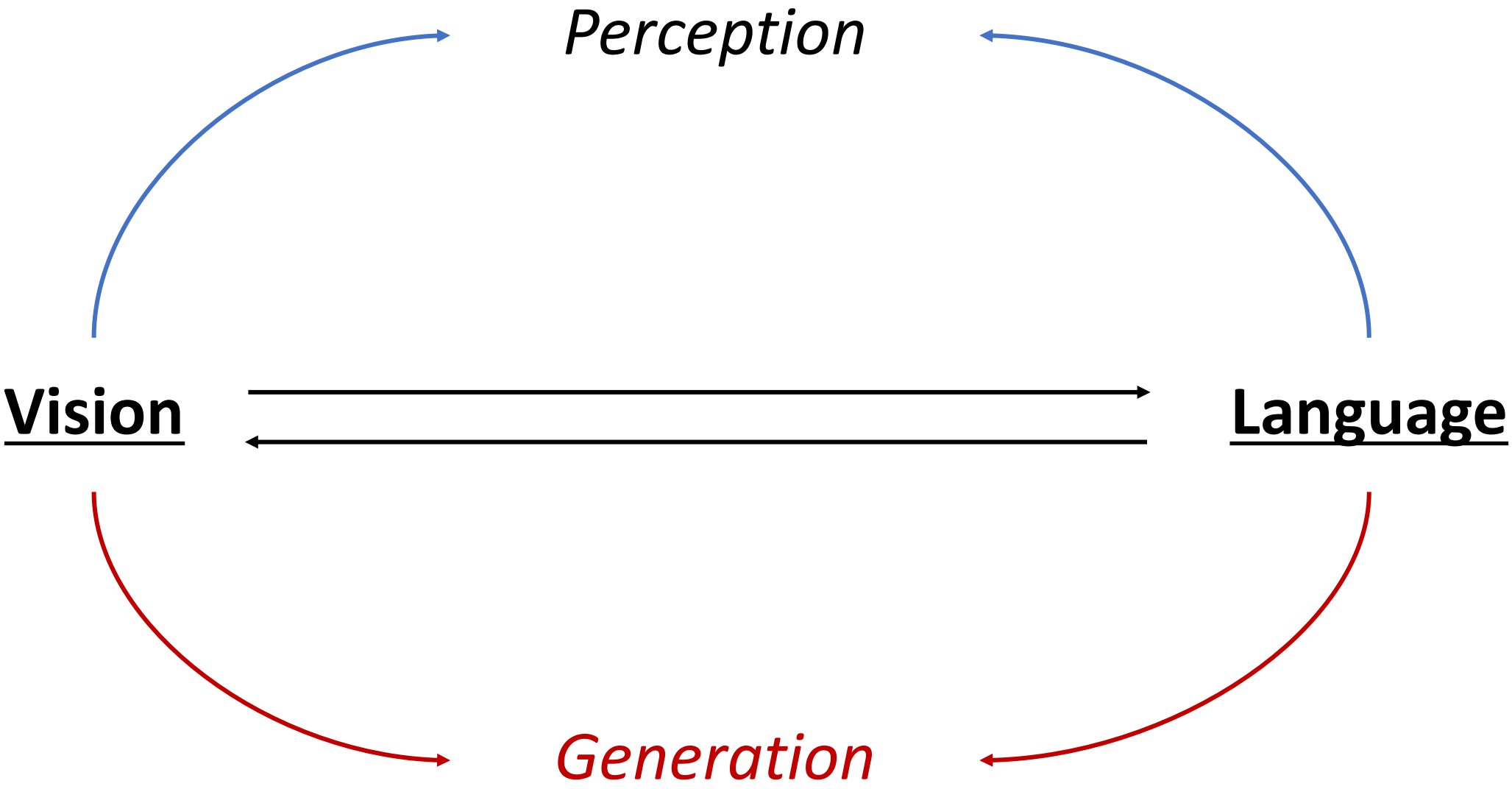
Bamboo

Method	Data	Annotation	Model	Paradigm	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	IN1K	AVG↑
SwAV [10]	IN1K	1.2M	RN50	Self.	92.5	76.6	76.4	88.0	93.0	65.5	60.5	78.1	91.0	56.0	66.9	76.8
DINO [11]	IN1K	1.2M	RN50	Self.	93.7	79.2	77.2	89.2	96.2	66.0	68.3	77.6	92.3	63.1	83.3	79.8
SWSL [71]	IG-1B	1B	RN50	Semi.	94.7	79.5	79.1	94.4	94.6	67.8	65.9	77.8	96.1	58.4	81.2	80.9
WSL [46]	IG-1B	1B	RX101	Weak.	95.0	78.2	83.5	**95.5**	90.8	67.9	72.3	75.3	93.3	53.9	83.3	81.0
CLIP [53]	WIT	400M	RN50	Lang.	88.7	70.3	86.4	88.2	96.1	73.3	78.3	76.4	89.6	49.1	73.3	79.1
CLIP [53]	WIT	400M	B/16	Lang.	96.2	83.1	92.8	93.1	98.1	78.4	86.7	79.2	94.7	59.5	80.2	85.6
BiT [37]	IN1K	1.2M	RN50	Sup.	91.7	74.8	72.5	92.3	92.0	61.1	53.5	72.4	91.2	52.5	75.2	73.6
BiT [37]	IN22K	14M	RN50	Sup.	94.9	82.2	83.3	91.5	99.4	69.9	59.0	77.3	93.9	55.6	76.7	80.3
RN50	Bamboo	69M	RN50	Sup.	93.9	81.2	85.3	92.0	99.4	72.2	91.1	76.5	93.2	84.0	77.2	86.0 (+5.1)
B/16	Bamboo	69M	B/16	Sup.	**98.2**	**90.2**	**92.9**	95.1	**99.8**	**79.0**	**93.3**	**81.2**	**97.0**	**88.1**	**83.6**	**91.8** (+6.2)

Visual In-Context Learning



(a) Visual in-context learning is sensitive to prompt selection


Perception
Vision
Language
Generation

DALLE2 & Stable Diffusion

Visual AIGC

Avatar

Foundation Model

Object

Scene

Visual AIGC

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Avatar

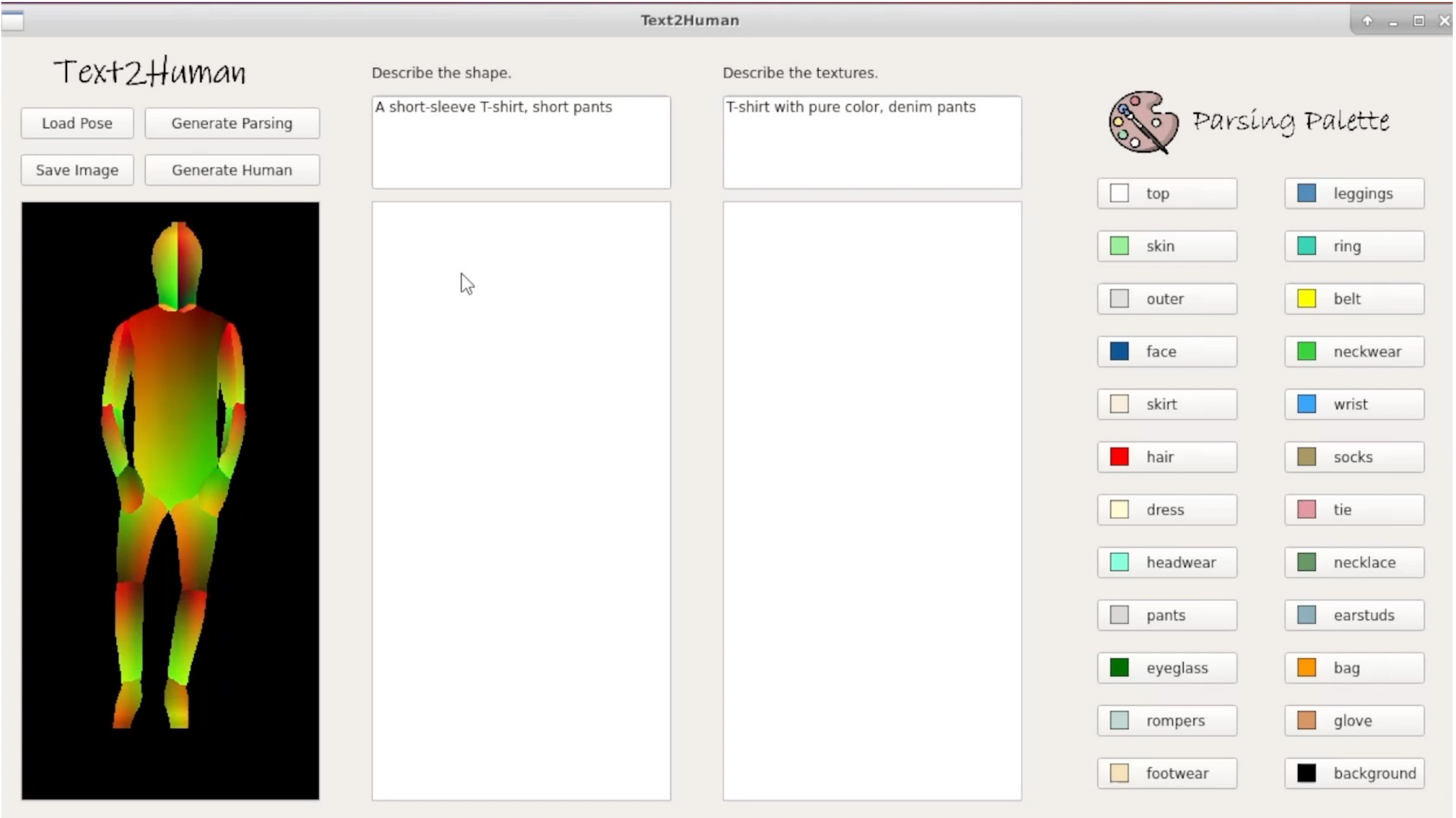
Foundation Model

Object

Scene

StyleGAN-Human: 2D Human Generation

Text2Human: Text-to-2D Human



Text2Performer: Text-to-2D Human Video

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

The dress the person wears has medium sleeves and it is of short length. The texture of it is pure color.

The lady moves to the left.

She is turning right from the front to the side.

She is turning right from the side to the back.

She turns right from the back to the side.

She turns right from the side to the front.

She moves to the right.

EVA3D: 3D Human Generation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

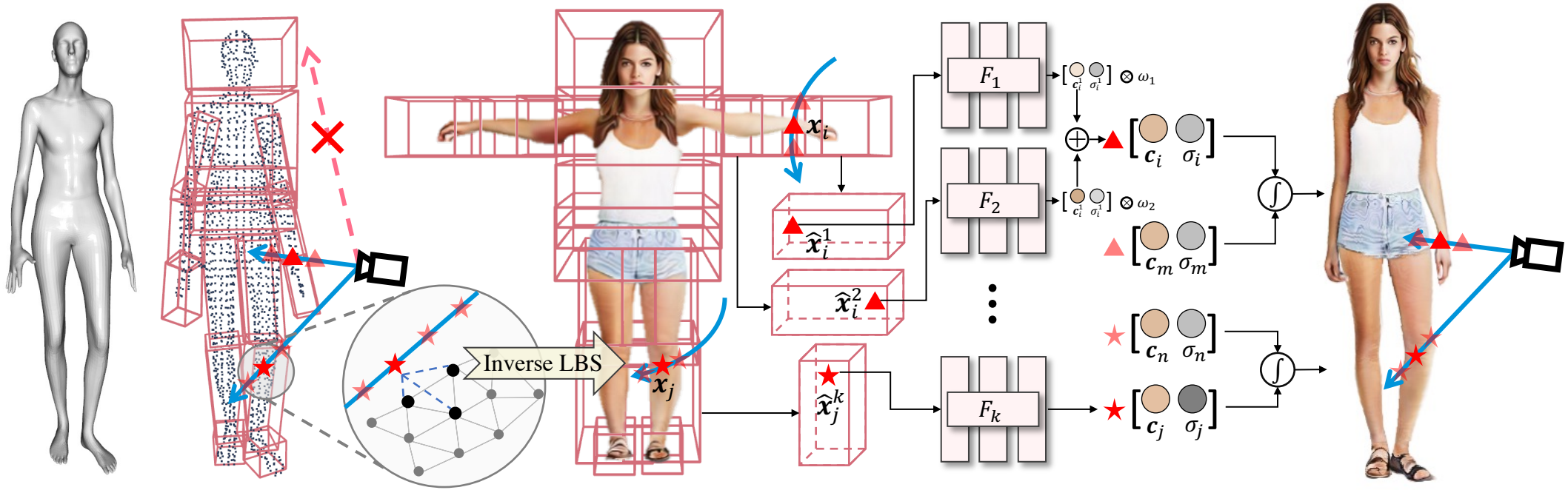
- Learn 3D generation from 2D image collections

Static **?** ***Articulated***

EVA3D: 3D Human Generation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

- Compositional Human NeRF



a) Sample Rays & Points at Observation Space SMPL(β, θ)

b) Inverse LBS to Canonical Space

c) Transform to Local Bounding Box & Query Subnetworks

d) Integrate Queried Results Along Rays

EVA3D: 3D Human Generation

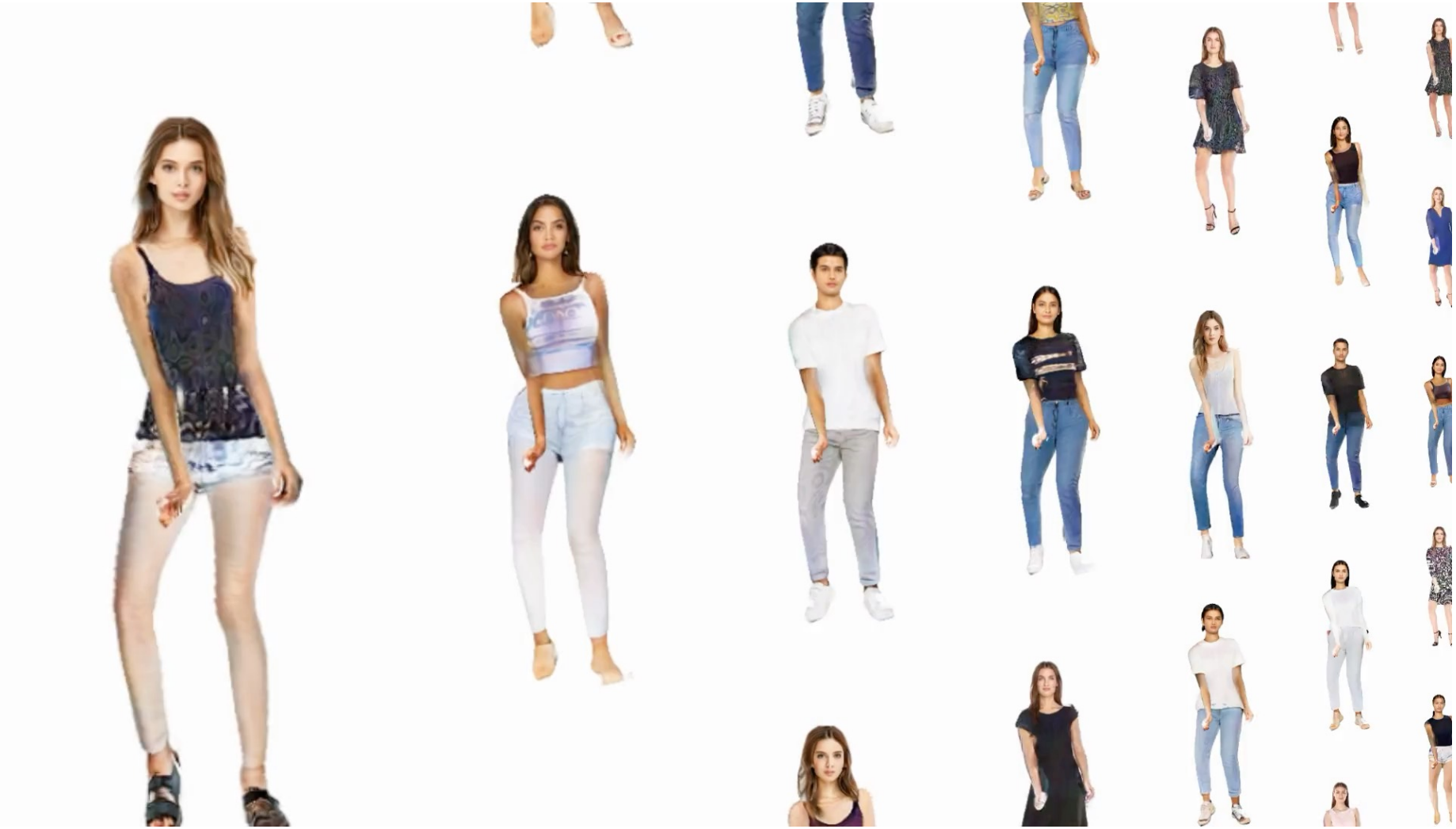
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

- Qualitative Results

EVA3D: 3D Human Generation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

- Explicit Pose/ Shape Control



AvatarCLIP: Text-to-3D Avatar

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

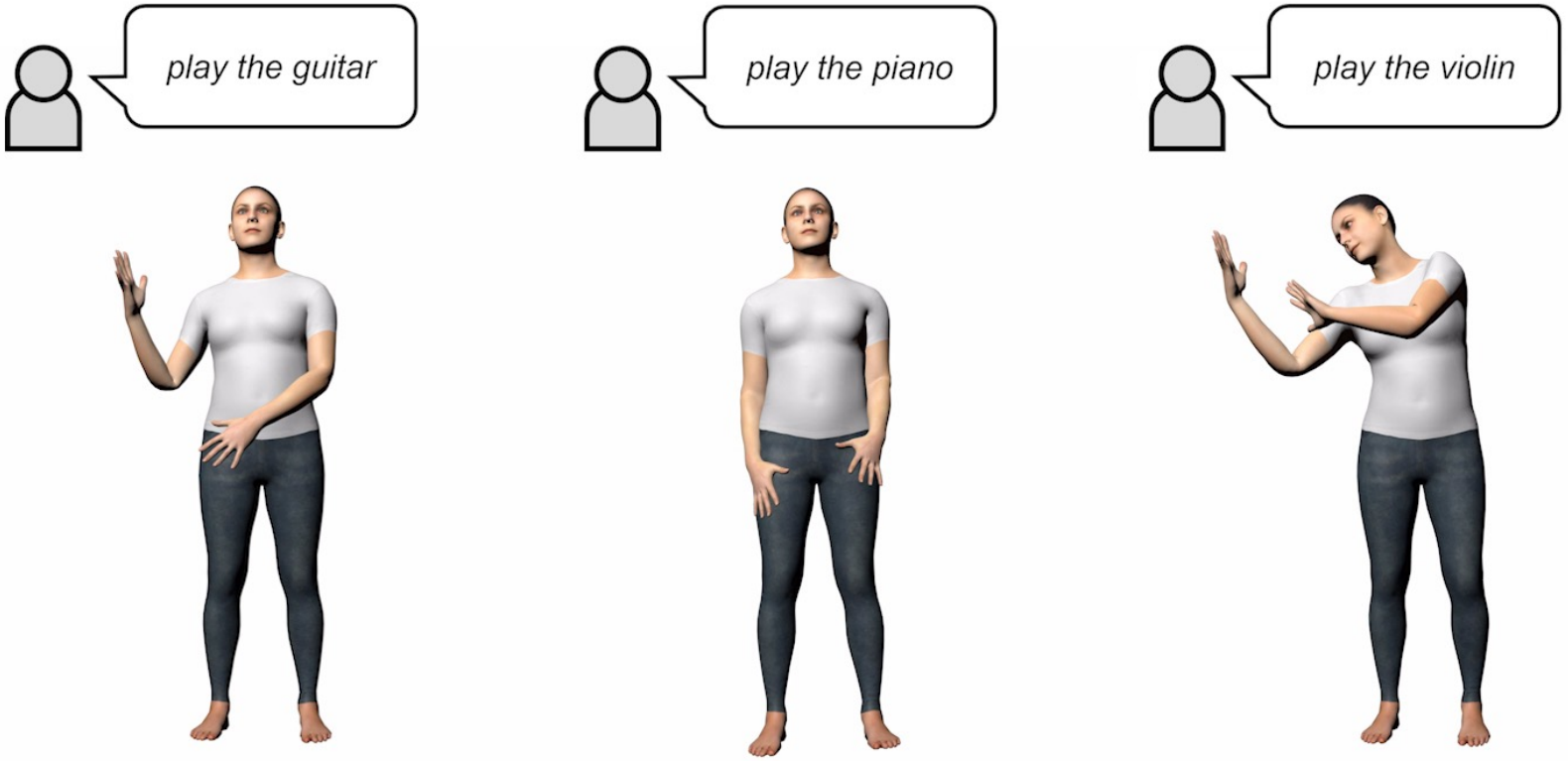
I want to generate a tall and fat Iron Man that is running.

I would like to generate a skinny ninja that is raising arms.

I want to generate a tall and skinny female soldier that is arguing.

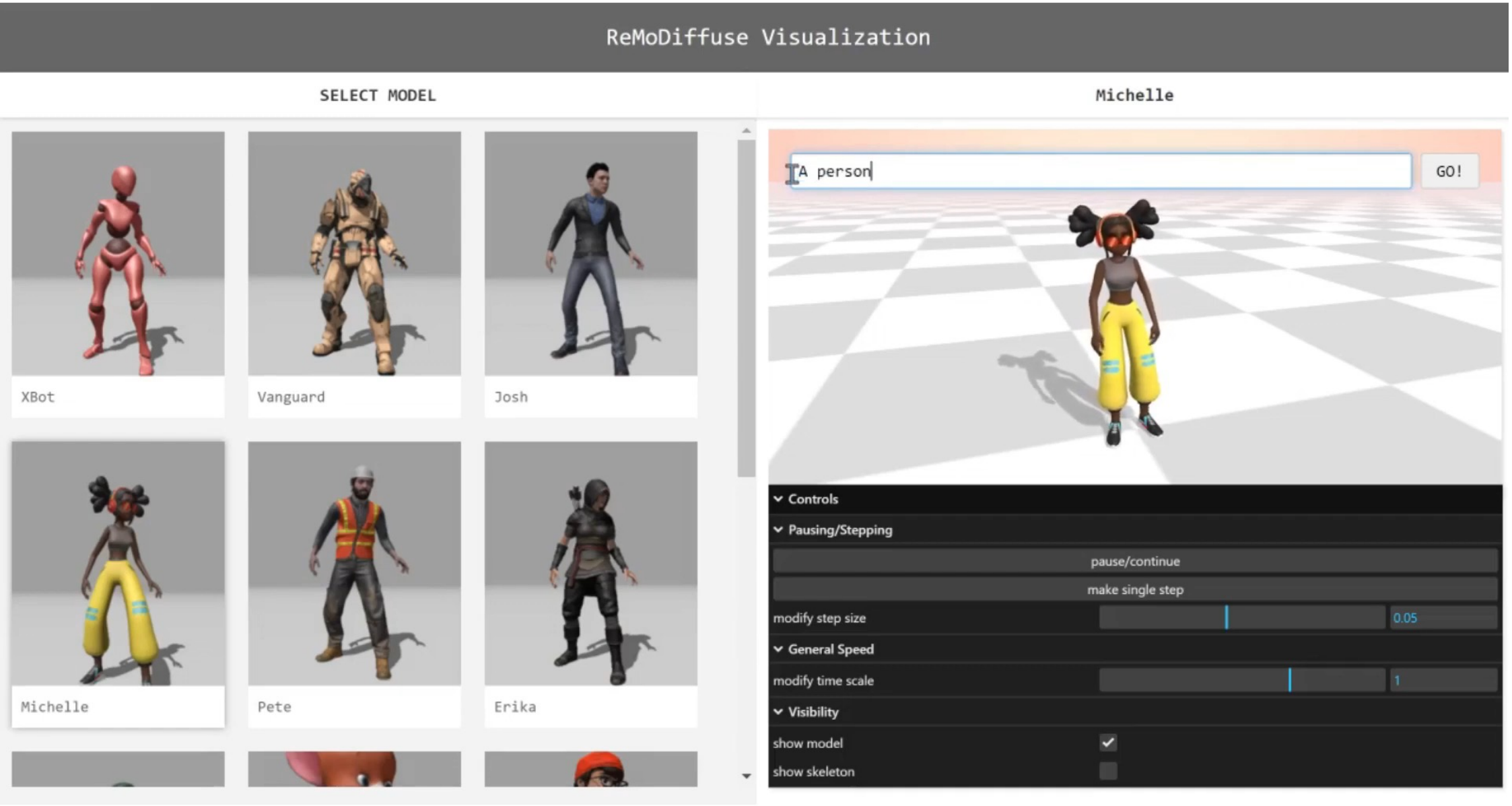
I want to generate an overweight sumo wrestler that is sitting.

MotionDiffuse: Text-to-3D Human Video



ReMoDiffuse: Text-to-3D Human Video

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE



Visual AIGC

Avatar

Foundation Model

Object

Scene

OmniObject3D: Text-to-3D Object

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: Text-to-3D Object

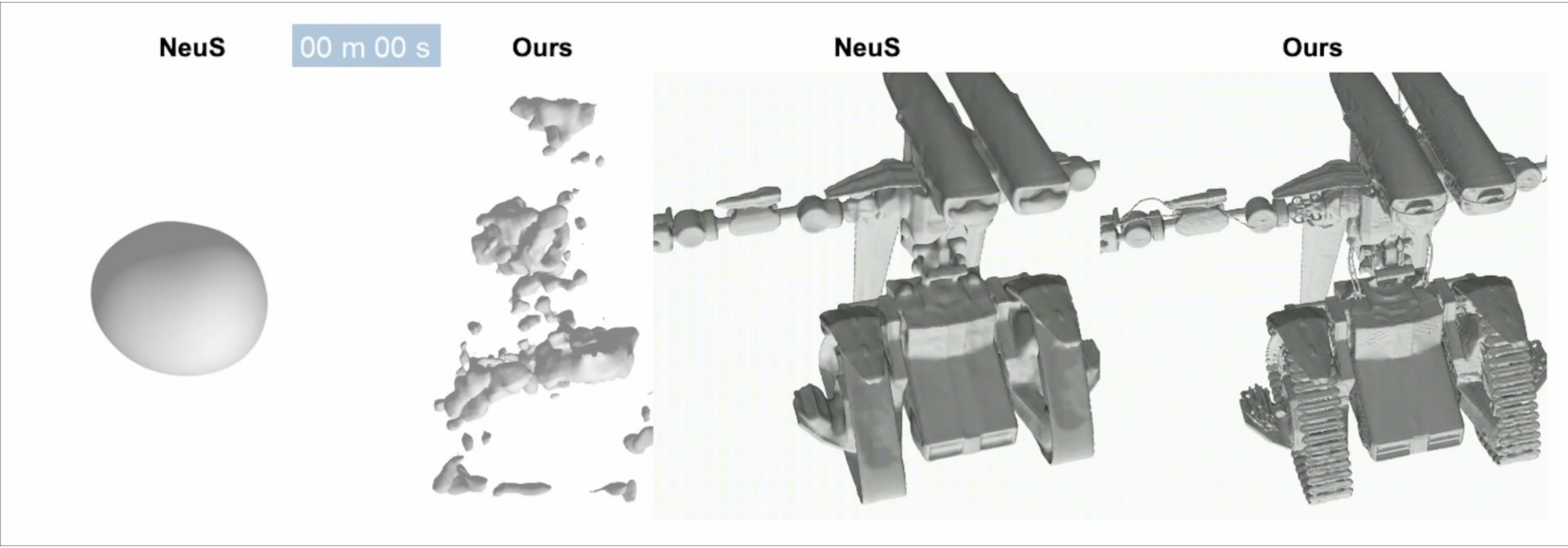
I want to generate a toy dinosaur.

I want to generate a music box.

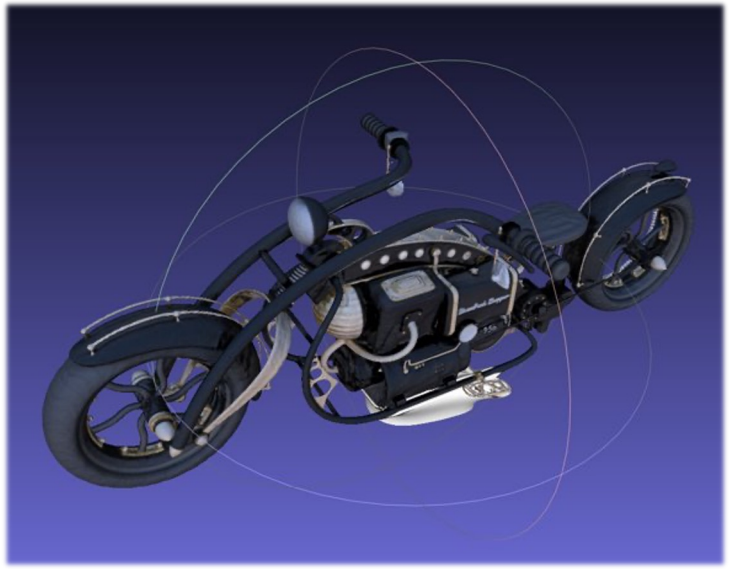
I want to generate a plaster statue.

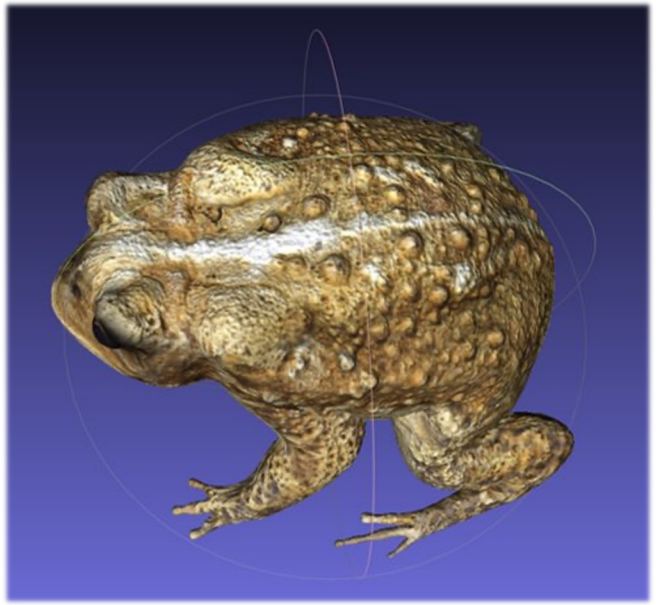
Voxurf: Fast 3D Object Reconstruction

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE



Voxurf: Fast 3D Object Reconstruction




Visual AIGC

Avatar

Foundation Model

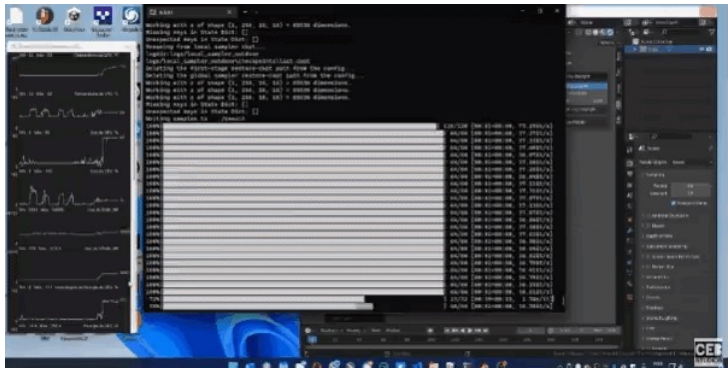
Object

Scene

Text2Light: Text-to-3D Environment

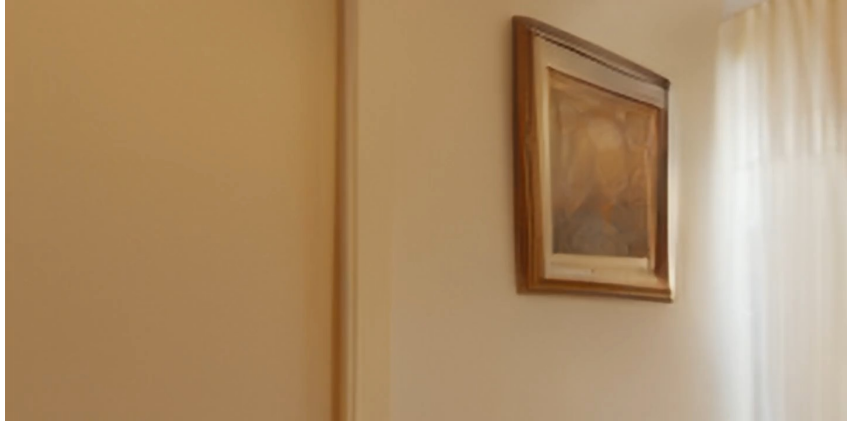
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

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

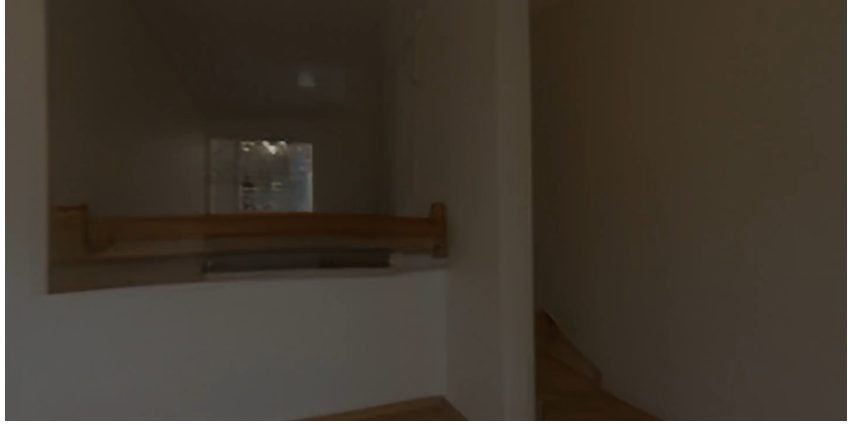


4K+ Resolution with High Dynamic Range

“white bed linen with white pillow”

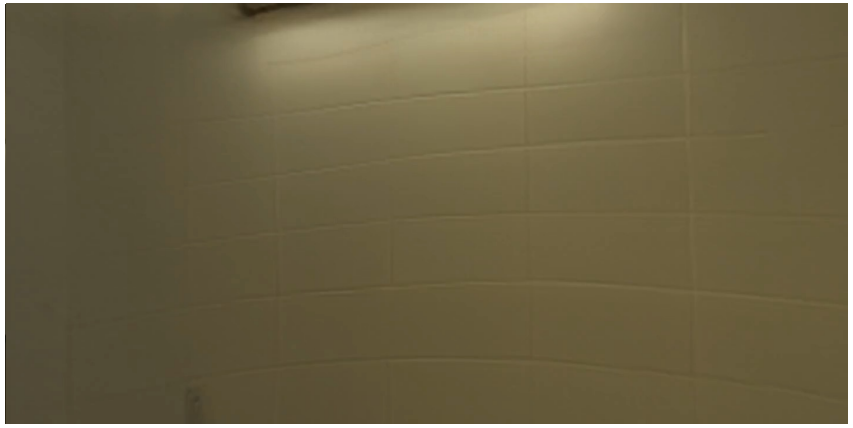


“brown wooden floor with white wall”

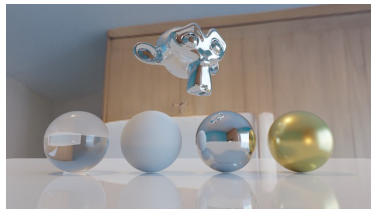


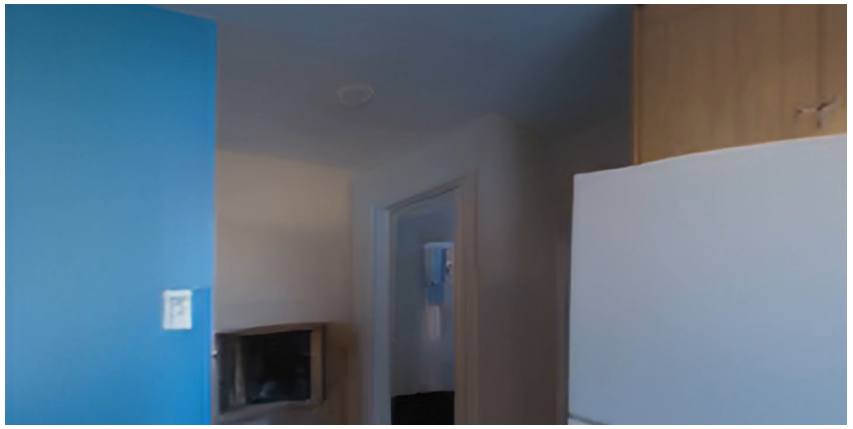
“gray concrete pathway with wall signages”

“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

Text2Light

Own Your Reality with Any Sentences

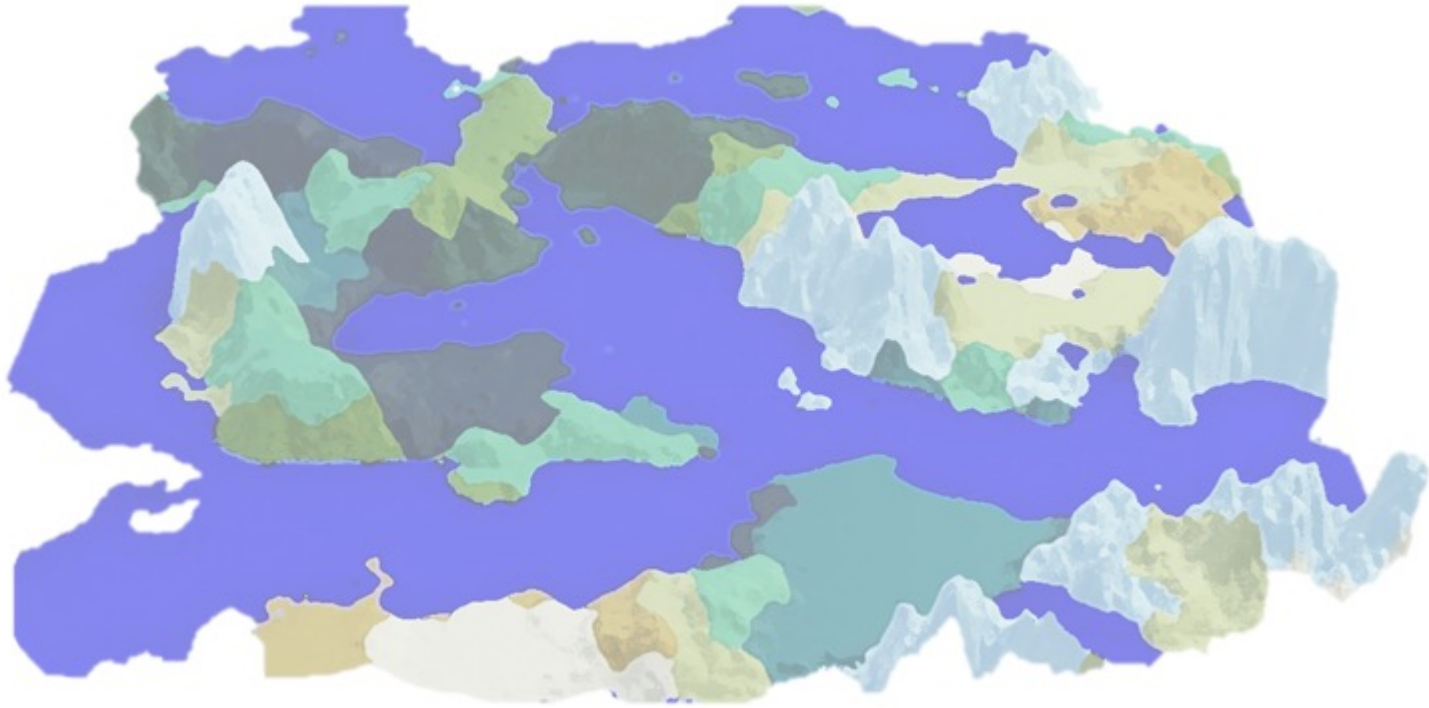
Describe Your Scene

e.g. a living room

Generate Render

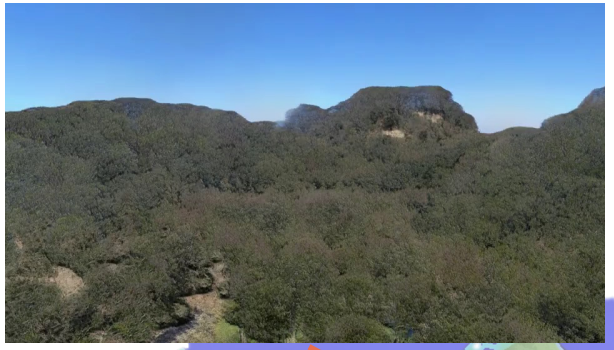
SceneDreamer: Unbounded 3D Scene Generation

In-the-wild 2D
Image Collections



Photorealistic
Unbounded 3D Scenes

SceneDreamer: Unbounded 3D Scene Generation

Multi-view consistent

Well-defined geometry

Diverse scenes and styles

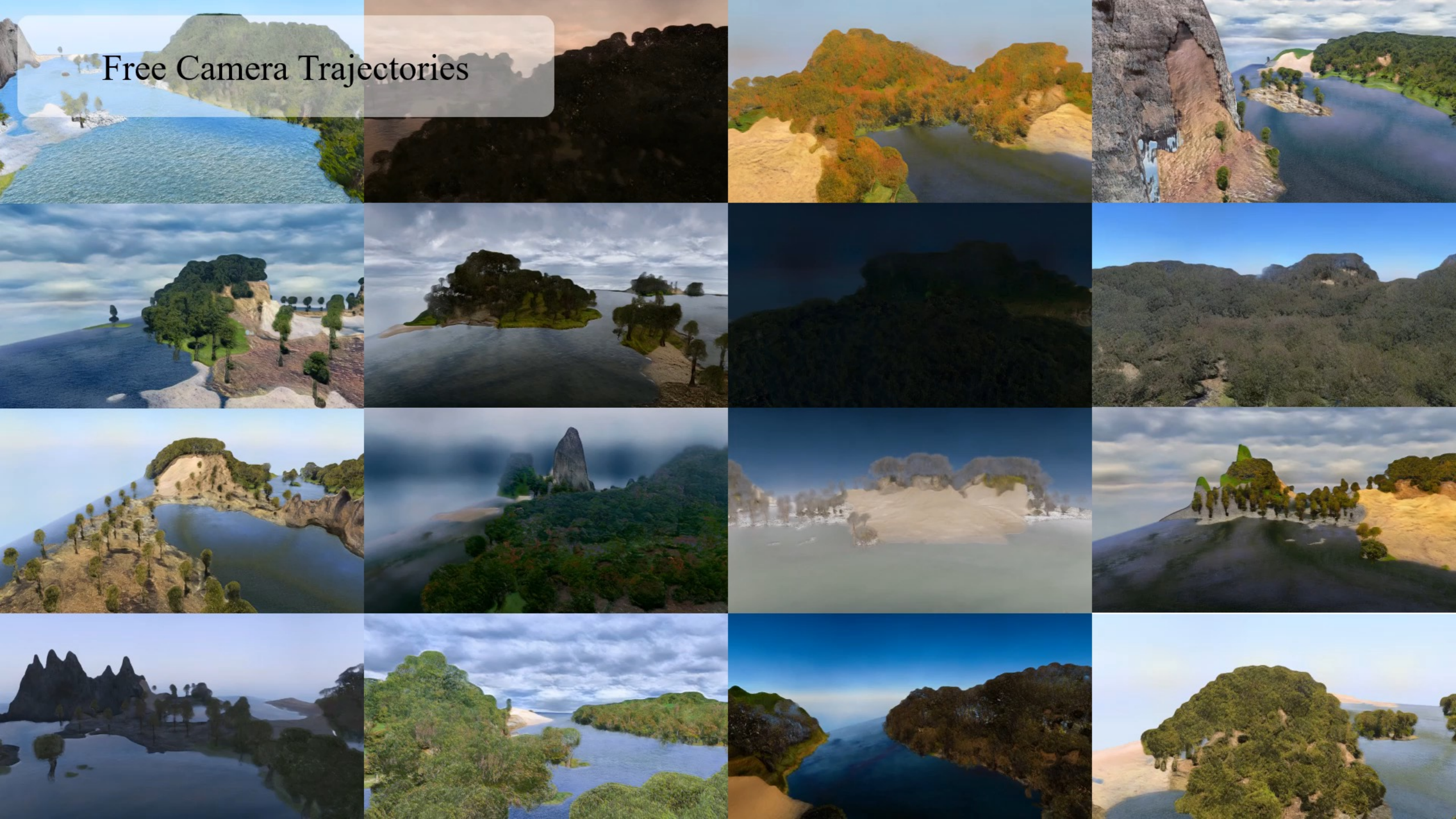
In-the-wild
Image Collections

Photorealistic
Unbounded 3D Scenes



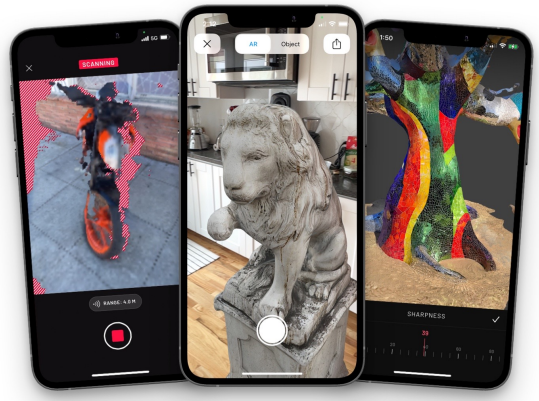
Infinite 3D World!


Generate with Different Styles

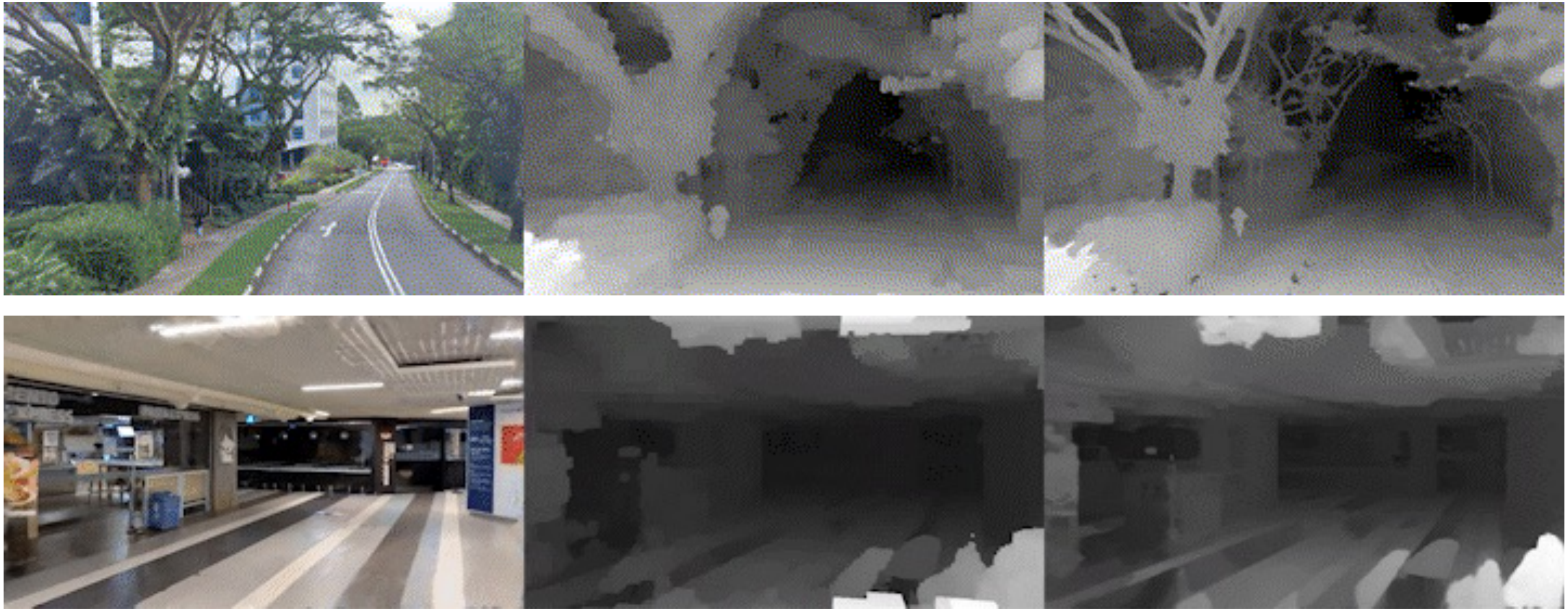

Free Camera Trajectories

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

Relighting4D: Relightable 3D Human

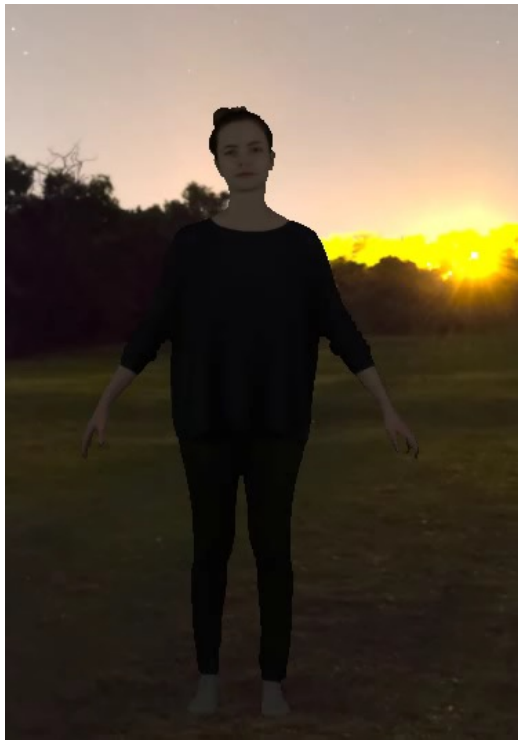
Prior works

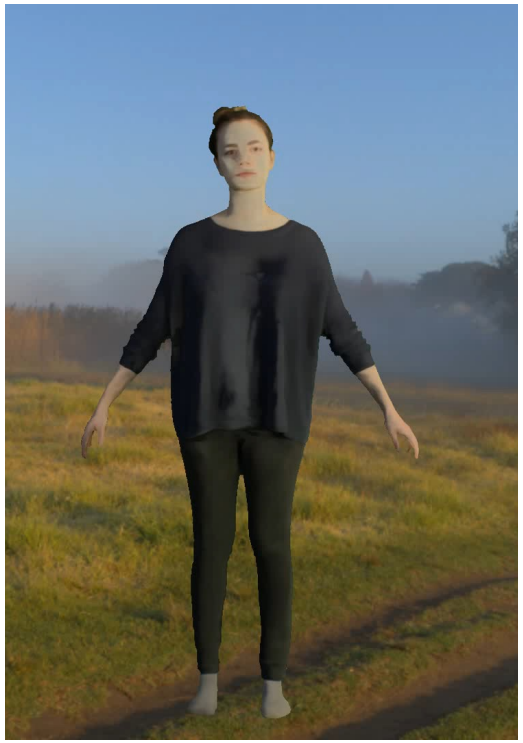
Synthetic dataset

Light Stage data

Relighting4D uses **only** videos to relight dynamic human actors from free viewpoints

Relighting4D: Relightable 3D Human



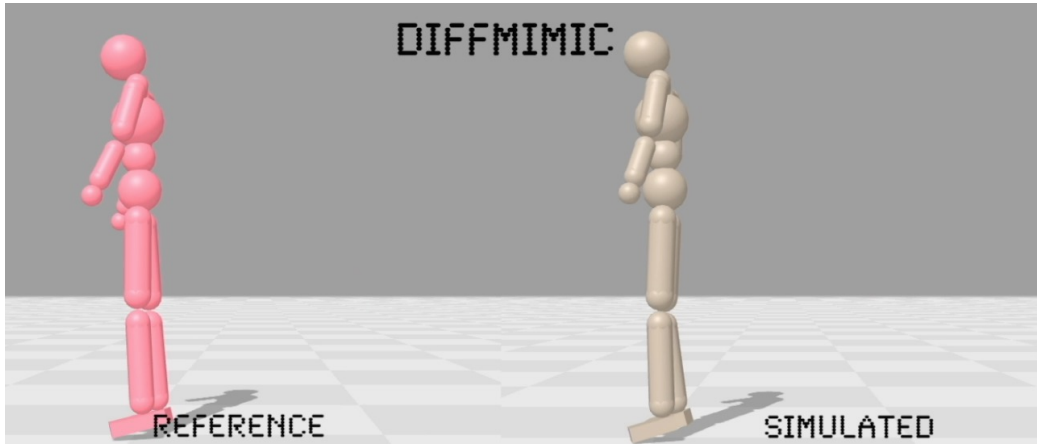


Video of human

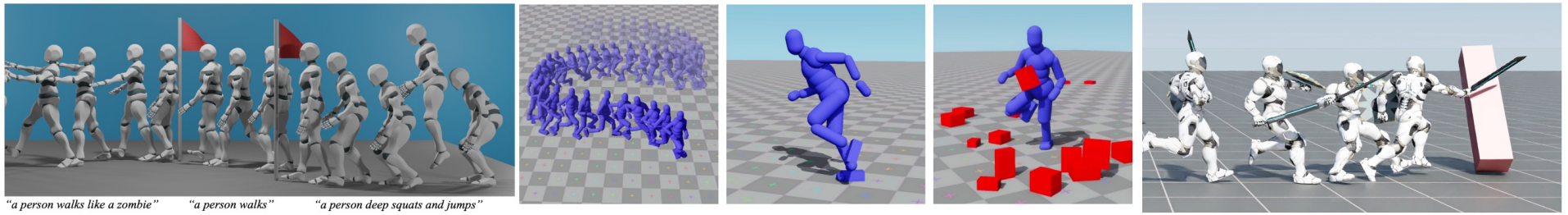
Relight with different illuminations and free viewpoints

DiffMimic: Physically-Simulated Character

- Motion mimicking: let a **physically-simulated** character imitate a reference motion.



- A fundamental task for downstream animation applications.

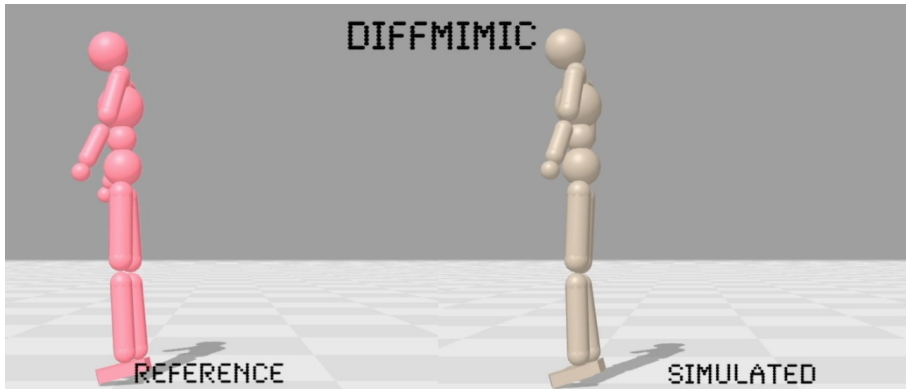


Language-Conditioned Control | Responsive Control | Skill Composition

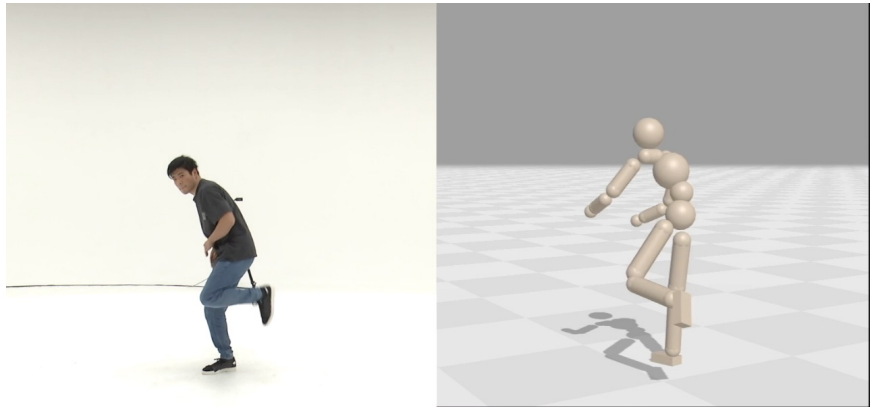
DiffMimic: Physically-Simulated Character

Motion	T_{cycle}(s)	DeepMimic	Spacetime Bound	Ours w/ RSI
Back-Flip	1.75	31.18	41.20 +32.1%	3.82 -87.7%
Cartwheel	2.72	30.45	17.35 -43.0%	4.72 -84.5%
Walk	1.25	23.80	4.08 -79.5%	1.55 -93.5%
Run	0.80	19.31	4.11 -78.7%	1.41 -92.7%
Jump	1.77	25.65	41.63 +77.8%	2.12 -91.7%
Dance	1.62	24.59	10.00 -59.3%	2.19 -91.1%

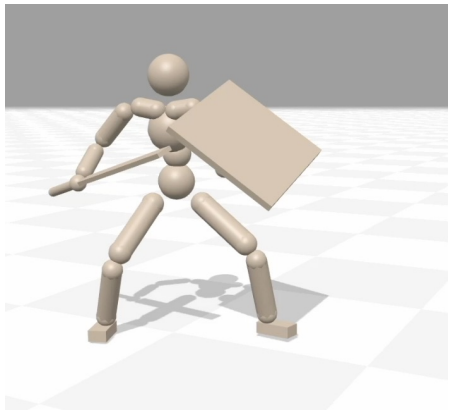
a) ~10x better sample efficiency compared to DeepMimic



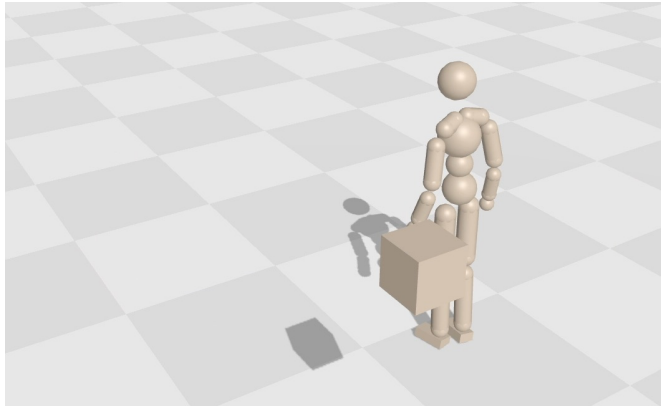
b) Learning backflip in 5 minutes



c) Scalable



d) General



e) Robust

Visual AIGC

Avatar

Foundation Model

Object

Scene

ReVersion: Object Relation Generation

Input

Exemplar Images

"vegetable is contained inside basket"

Output

Relation Prompt

<R>

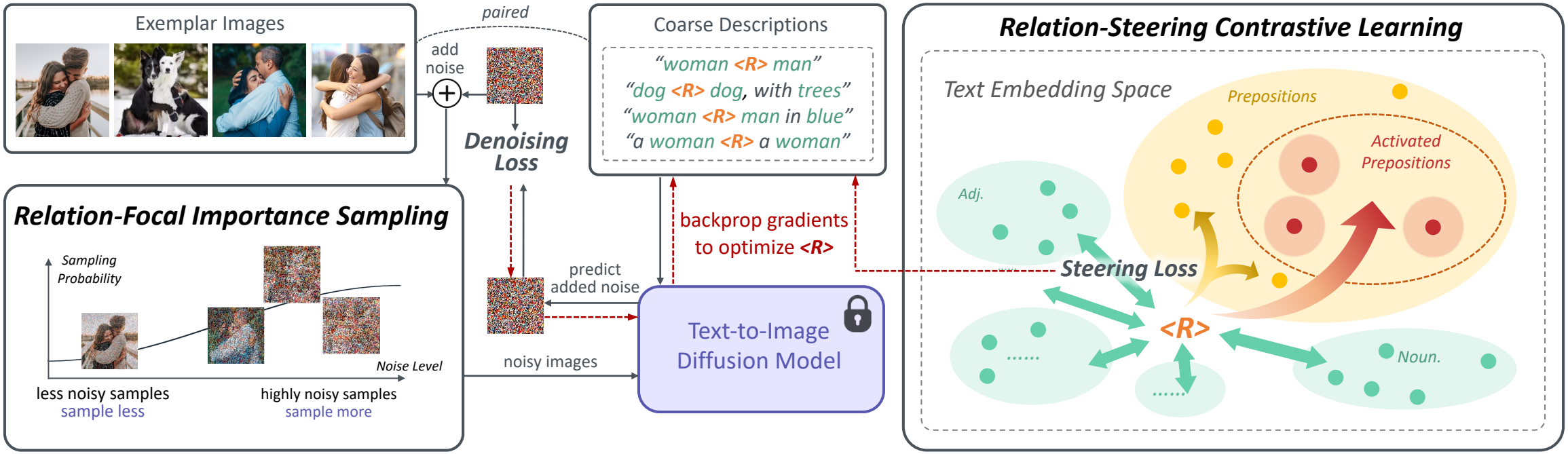
represent the co-existing relation in exemplar images

Application

Relation-Specific Text-to-Image Synthesis

"Sculpture <R> museum"

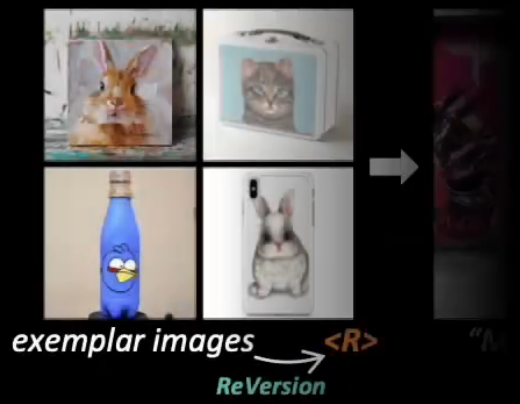
ReVersion: Object Relation Generation



ReVersion: Object Relation Generation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Visual Results: ReVersion



Visual AIGC

Avatar

Thank You!

Object

Scene