

Multi-Modal Generative AI 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

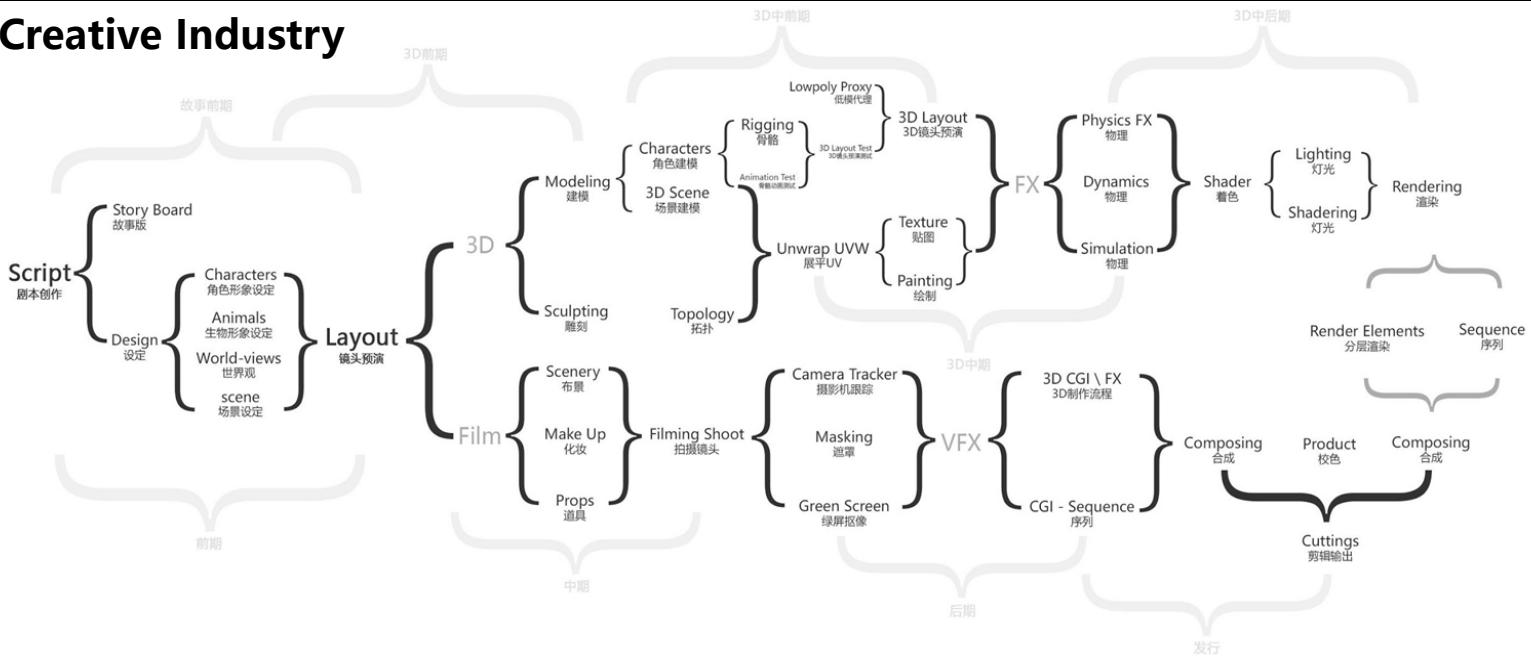


VTuber



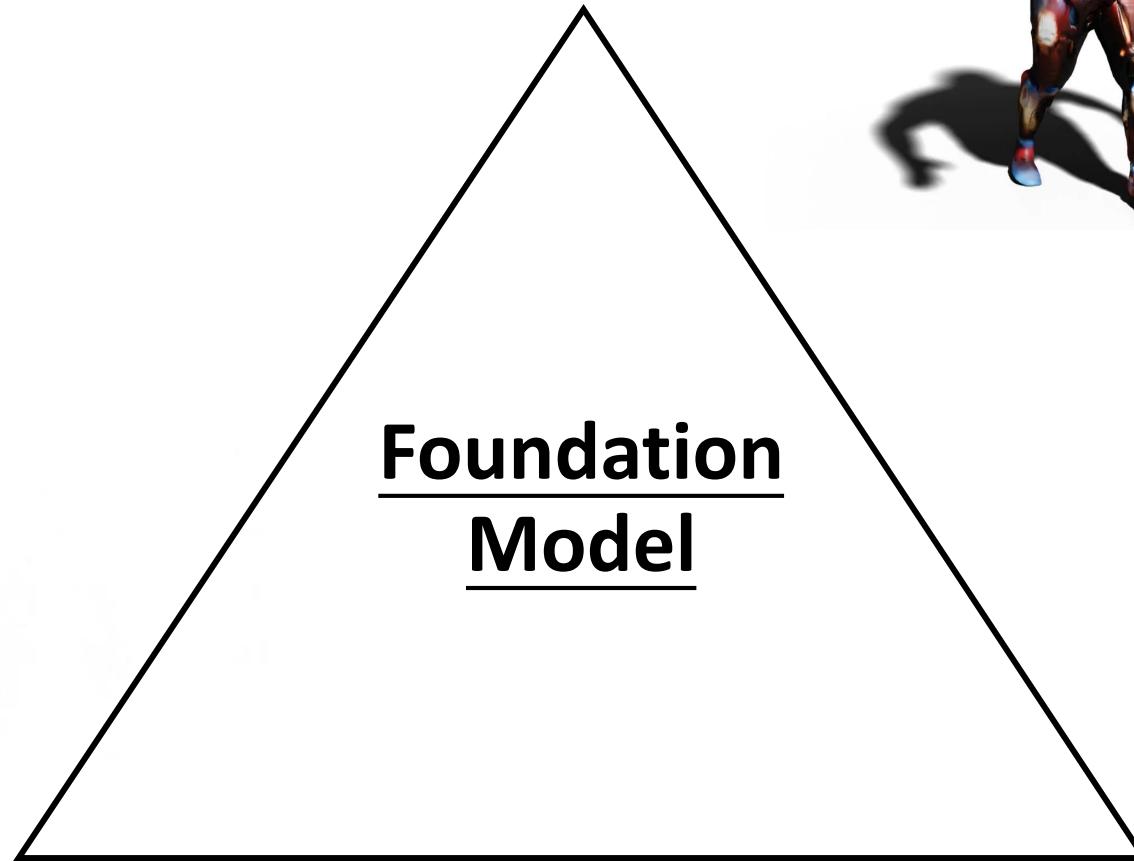
Virtual Beings

Creative Industry





Object



Avatar



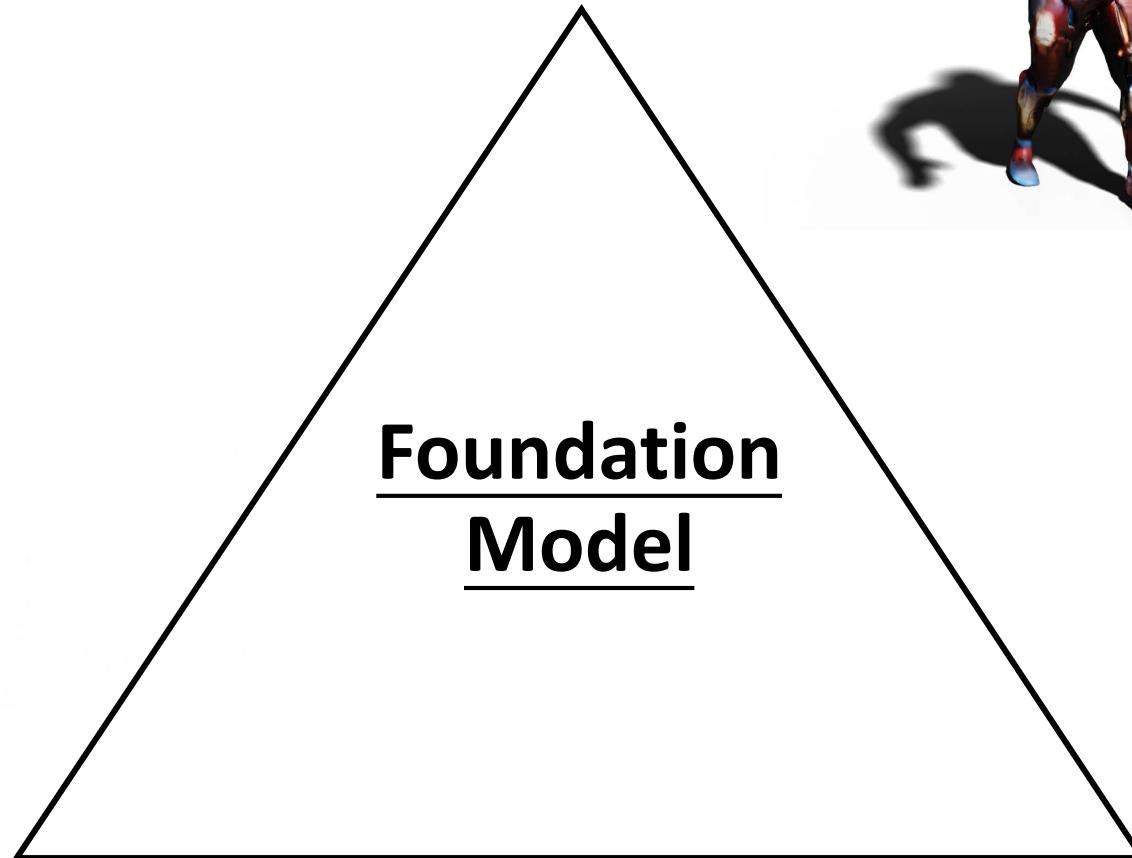
Scene





Object

Avatar



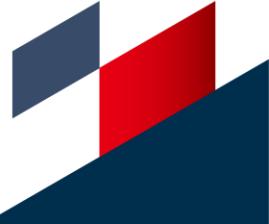
Scene



StyleGAN-Human: 2D Human Generation



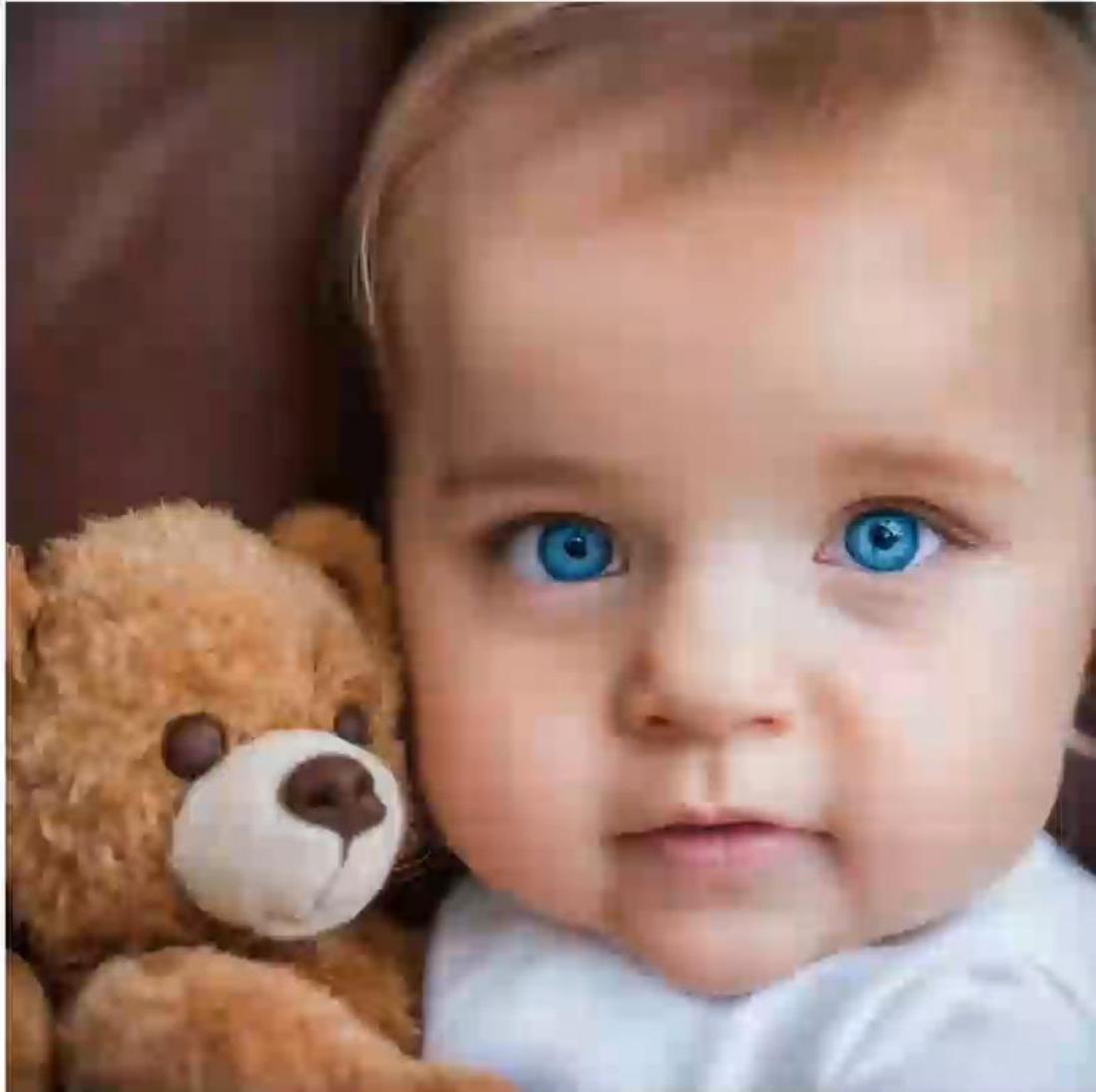
S-LAB
FOR ADVANCED
INTELLIGENCE



HyperHuman: 2D Human Generation



S-LAB
FOR ADVANCED
INTELLIGENCE



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.



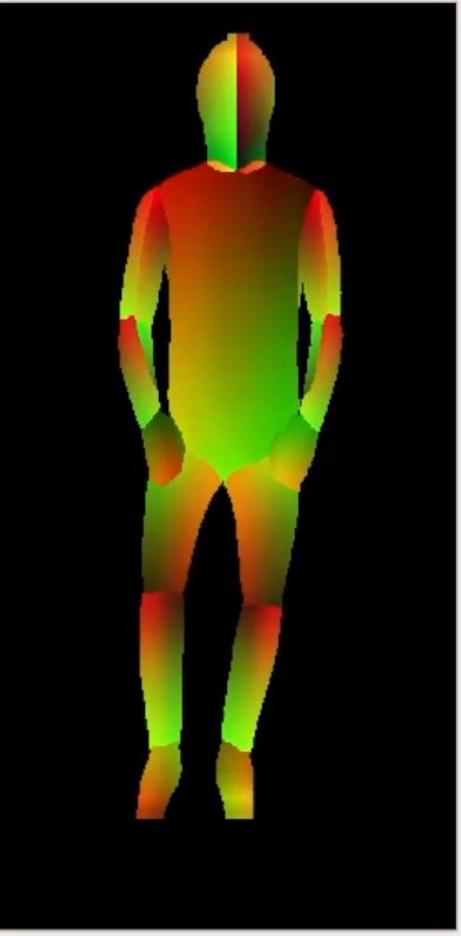
Text2Human: Text-to-2D Human

Text2Human

Text2Human

Load Pose Generate Parsing

Save Image Generate Human



Describe the shape.
A short-sleeve T-shirt, short pants

Describe the textures.
T-shirt with pure color, denim pants

Parsing Palette

<input type="checkbox"/> top	<input checked="" type="checkbox"/> leggings
<input checked="" type="checkbox"/> skin	<input checked="" type="checkbox"/> ring
<input type="checkbox"/> outer	<input checked="" type="checkbox"/> belt
<input checked="" type="checkbox"/> face	<input checked="" type="checkbox"/> neckwear
<input type="checkbox"/> skirt	<input checked="" type="checkbox"/> wrist
<input checked="" type="checkbox"/> hair	<input checked="" type="checkbox"/> socks
<input type="checkbox"/> dress	<input checked="" type="checkbox"/> tie
<input checked="" type="checkbox"/> headwear	<input checked="" type="checkbox"/> necklace
<input type="checkbox"/> pants	<input checked="" type="checkbox"/> earstuds
<input checked="" type="checkbox"/> eyeglass	<input checked="" type="checkbox"/> bag
<input type="checkbox"/> rompers	<input checked="" type="checkbox"/> glove
<input type="checkbox"/> footwear	<input checked="" type="checkbox"/> background

Text2Performer: Text-to-2D Human Video



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.

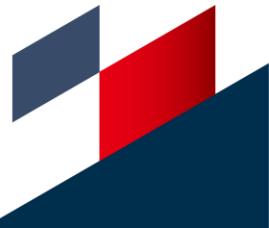
She is turning right from the front to the side.

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

- Learn 3D generation from 2D image collections



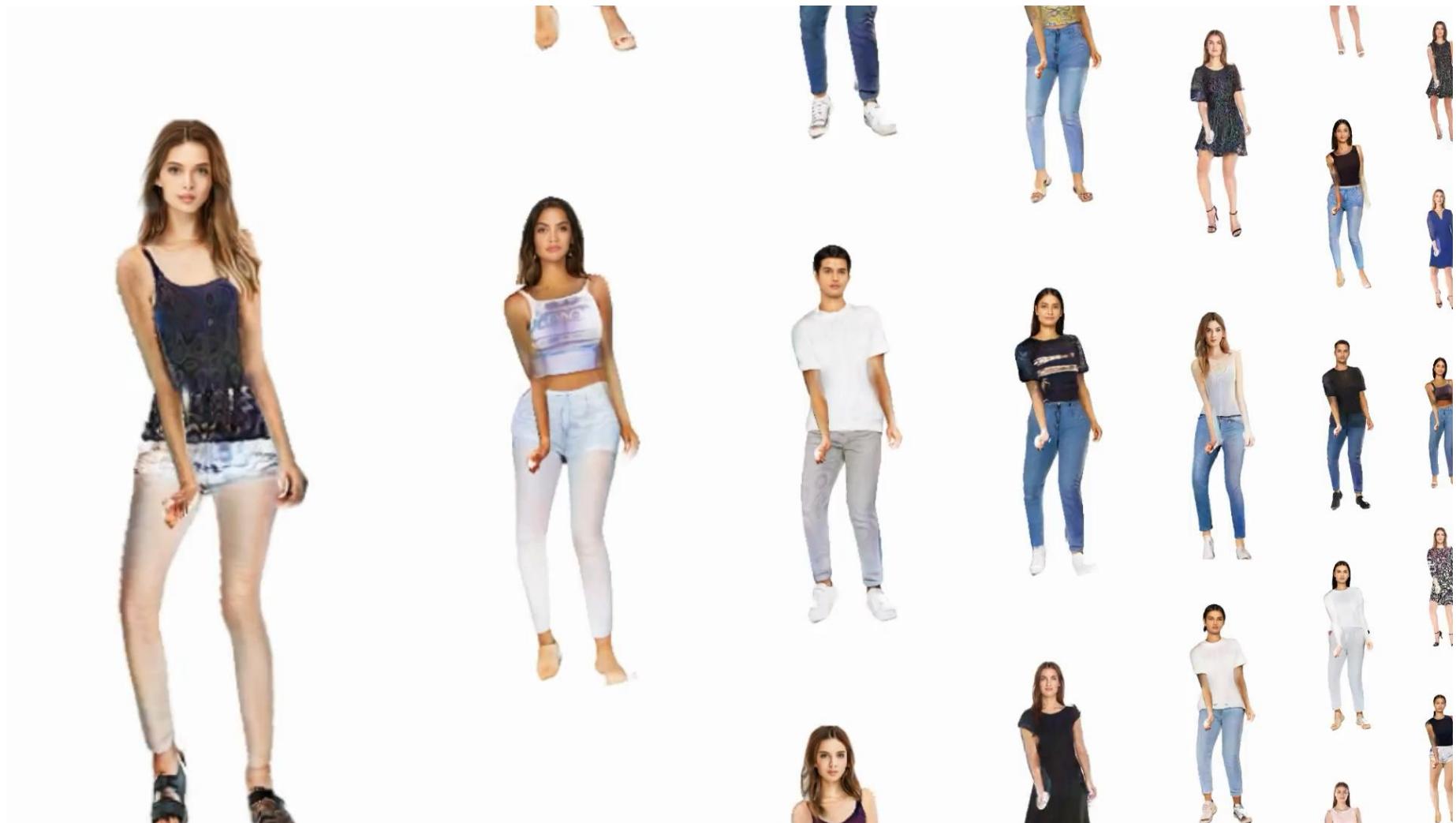
?
Static → *Articulated*



EVA3D: 3D Human Generation



S-LAB
FOR ADVANCED
INTELLIGENCE



AvatarCLIP: Text-to-3D Avatar



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



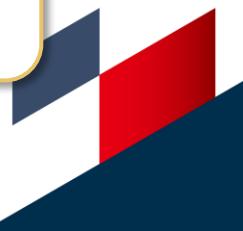
I would like to generate a skinny ninja that is raising arms.



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



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



AvatarCLIP: Text-to-3D Avatar



S-LAB
FOR ADVANCED
INTELLIGENCE

60 FPS (1-60)

Generate "A very skinny ninja that is shooting back arrows"

AvatarCLIP

Create Your Own Avatar
with Natural Languages!



Describe the Shape



Generate

Next Step

Renderer Controller

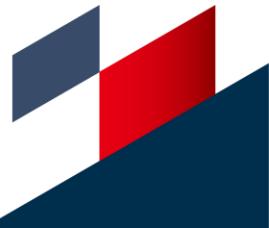
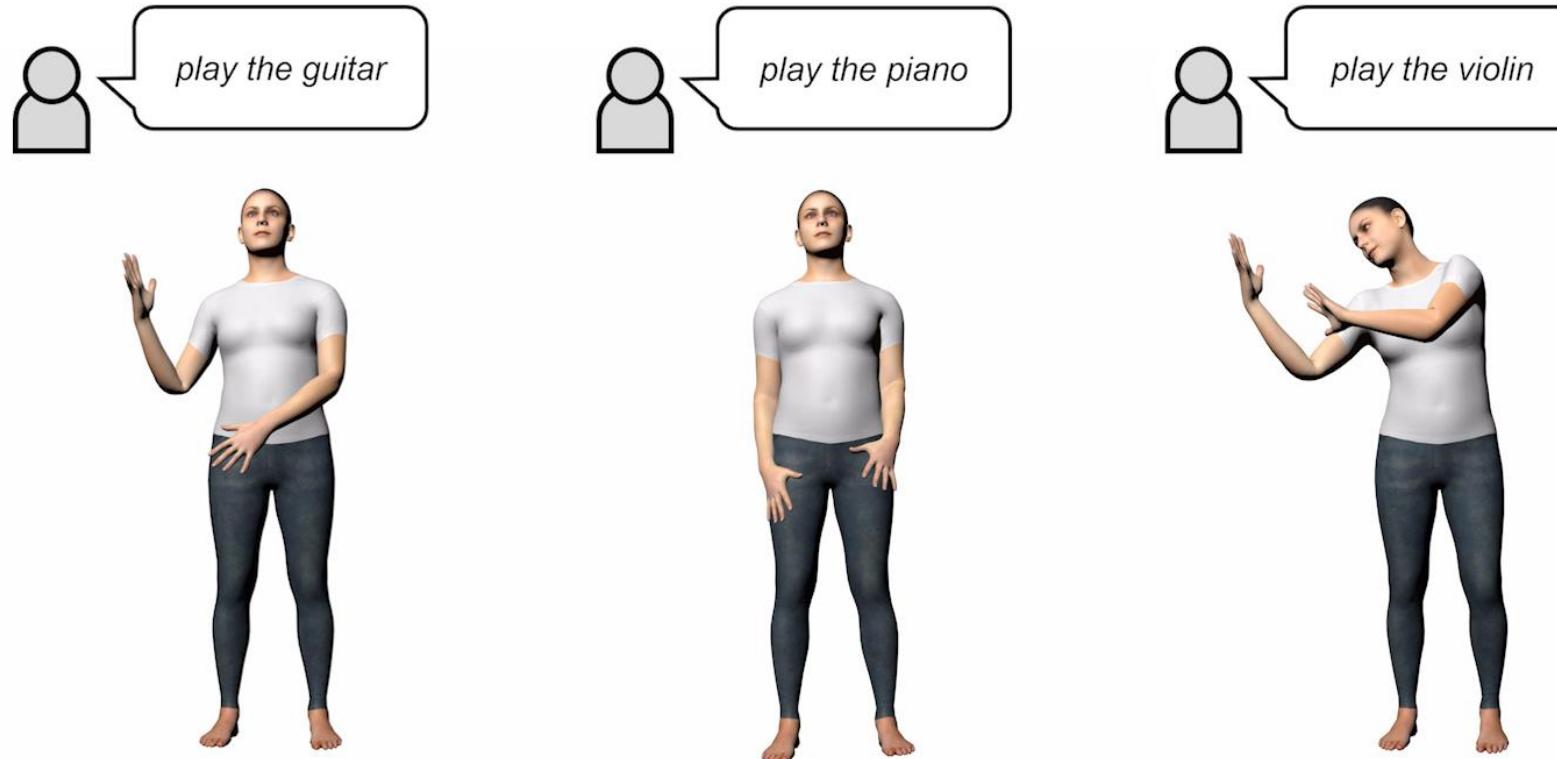
- Vertex Color
- Wireframe
- Normal



MotionDiffuse: Text-to-3D Human Video



S-LAB
FOR ADVANCED
INTELLIGENCE



ReMoDiffuse: Text-to-3D Human Video



S-LAB
FOR ADVANCED
INTELLIGENCE

ReMoDiffuse Visualization

SELECT MODEL



XBot



Vanguard



Josh



Michelle



Pete



Erika

Michelle

A person



The visualization shows a 3D character named Michelle with dark skin, black hair in pigtails, wearing a grey tank top and yellow pants, walking on a checkered floor. Her shadow is cast on the floor behind her.

Controls

Pausing/Stepping

pause/continue

make single step

modify step size

General Speed

modify time scale

Visibility

show model

show skeleton

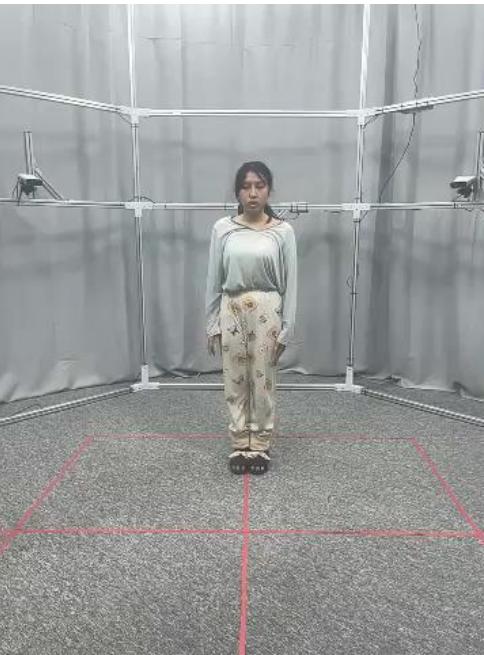
HuMMan Dataset



S-LAB
FOR ADVANCED
INTELLIGENCE



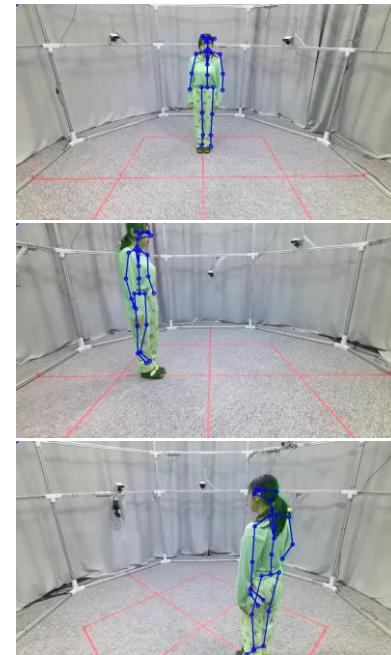
Artec Eva



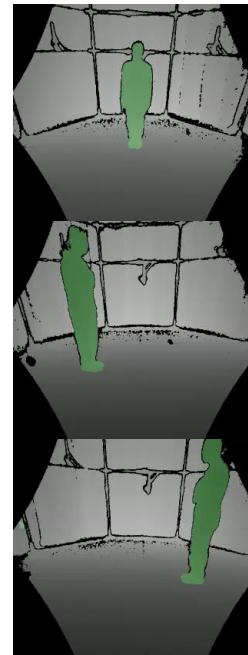
iPhone RGB



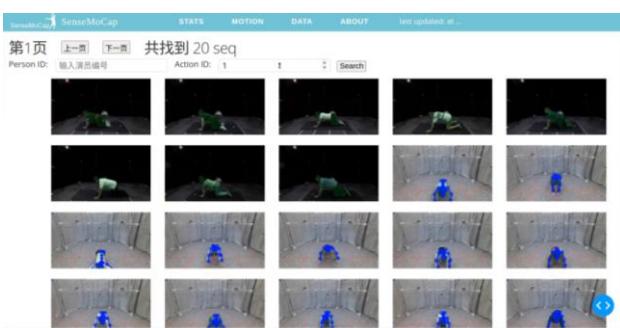
iPhone Depth



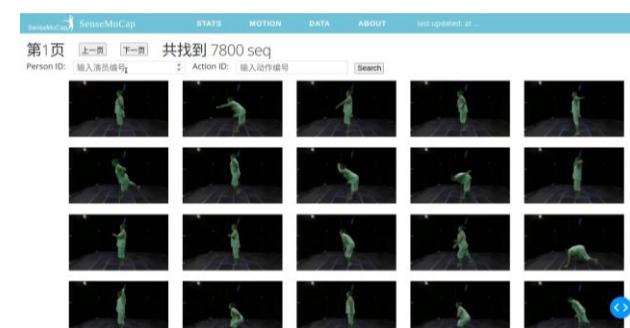
Kinect RGB



Kinect Depth



Search by Action



Search by Actor

0.1mm
Accuracy

11
Cameras

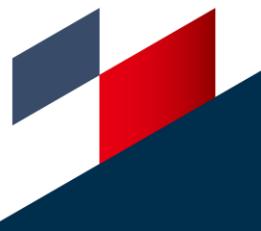
1G
Data / Sec

6
Actor / Day

SMPLer-X: MoCap Anybody



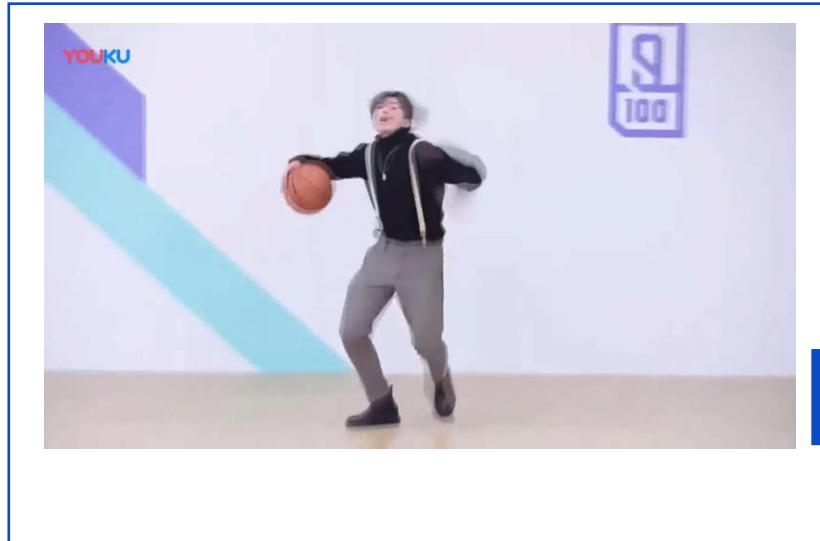
S-LAB
FOR ADVANCED
INTELLIGENCE



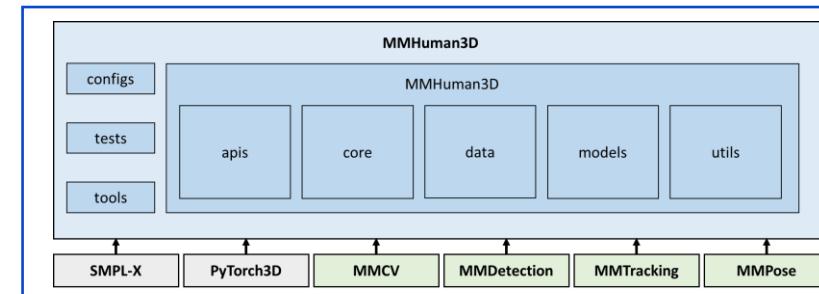
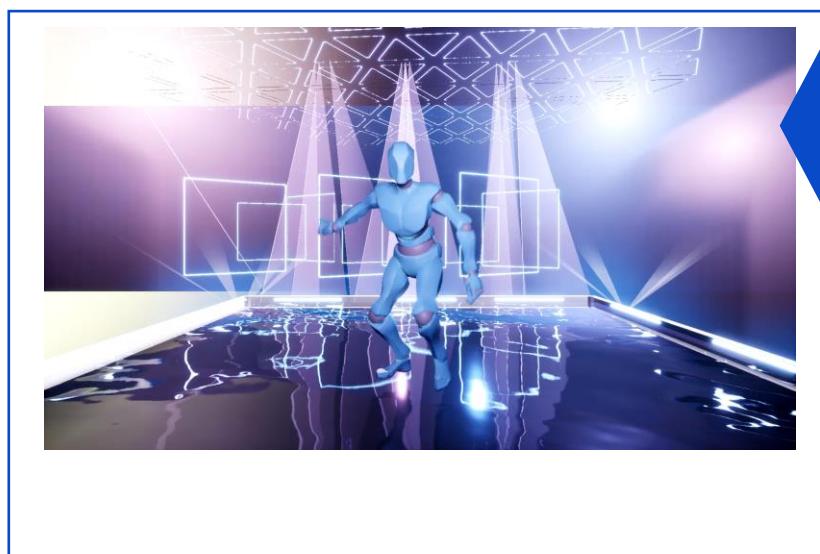
MMHuman3D Software



S-LAB
FOR ADVANCED
INTELLIGENCE



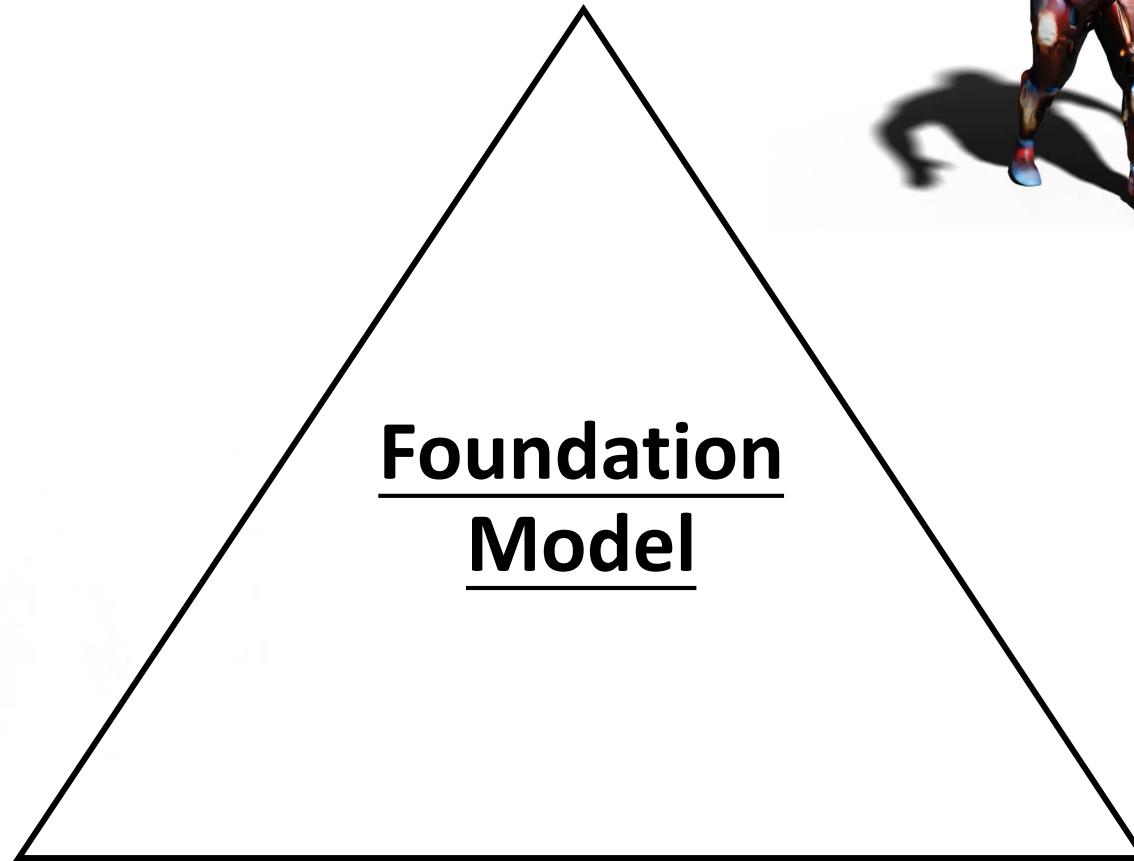
The screenshot shows the 'Mocap' software interface. At the top, it says 'Simple Task' and '10.10.30.159:8765/mocap'. Below that is a large button labeled '上传原始视频' (Upload original video) with the instruction '拖动要上传的视频到此处，或 点击这里 打开上传窗口' (Drag the video to be uploaded here, or click here to open the upload window). Below this is a section labeled '执行' (Execute) with a blue button '开始执行mocap任务' (Start executing mocap task). On the left, there's a dropdown menu showing '选择视频' (Select video) with '2021-10-27 19:58:24 |> SingDanceRapBasketball.mp4' selected. At the bottom, there's a '显示执行' (Show execution) button.



3D Animation Production:
3 days -> 30 min



Object



Avatar



Scene



OmniObject3D: 3D Object Datasets

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

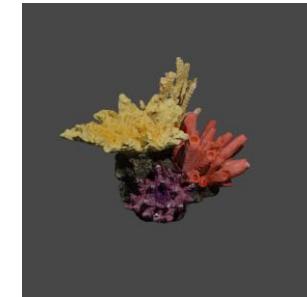
- ✓ 6k high-quality 3D models
- ✓ 190 categories
- ✓ 4 modalities: textured mesh, point cloud, real-captured video, synthetic multi-view images.
- ✓ Many down-stream tasks

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
DTU	2014	✓	✓		124	NA
GSO	2021	✓	✓		1k	17
AKB-48	2022	✓	✓		2k	48
Ours	2022	✓	✓	✓	6k	190

Real-world
3D scans



OmniObject3D: 3D Object Datasets



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

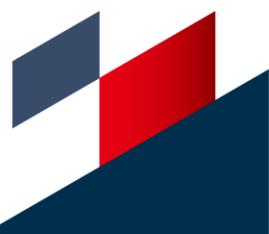
Entertainment, decoration.



DiffTF: Large-Vocabulary 3D Diffusion Model



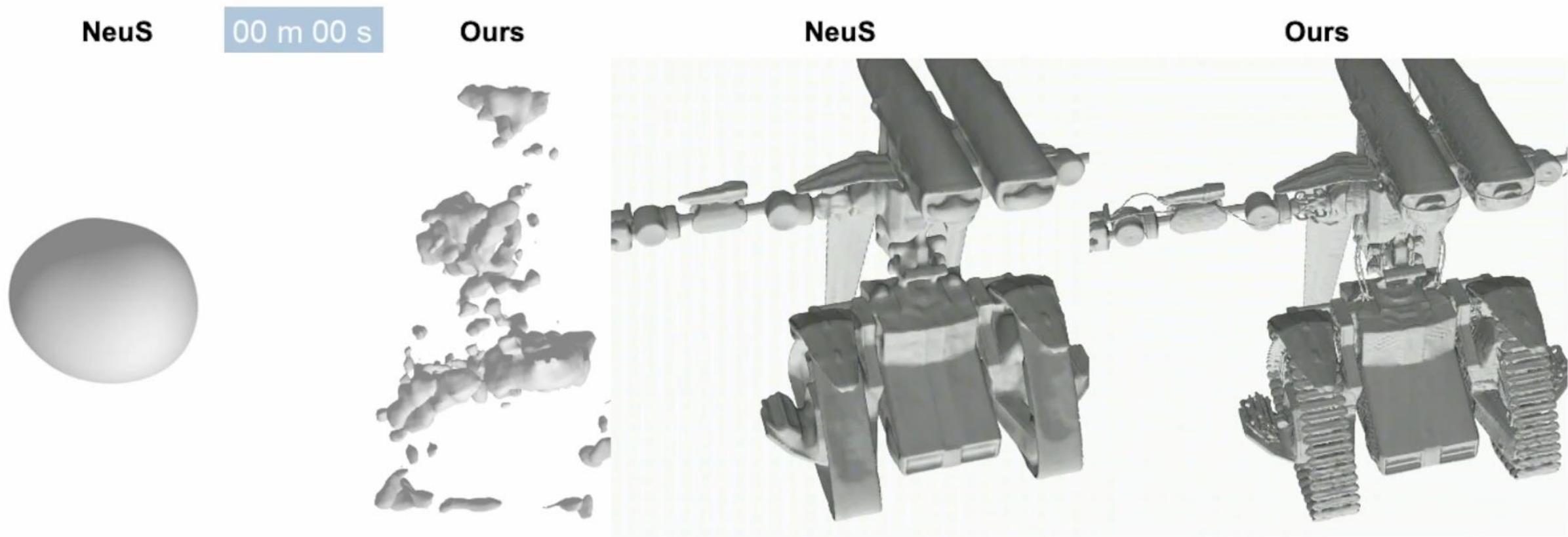
S-LAB
FOR ADVANCED
INTELLIGENCE



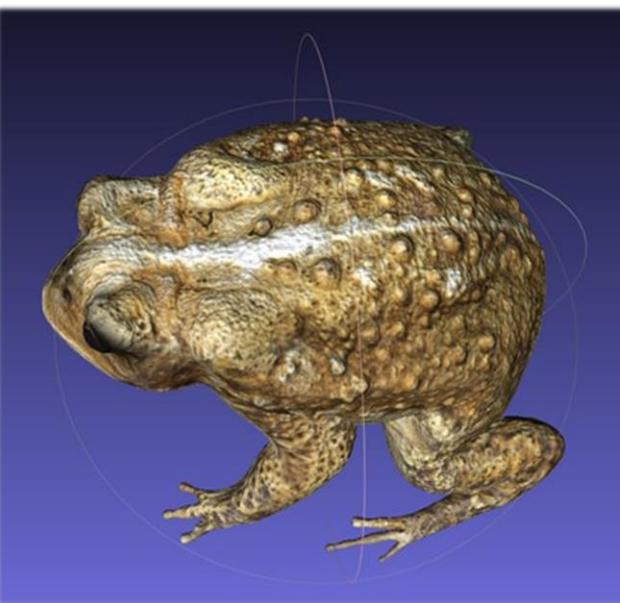
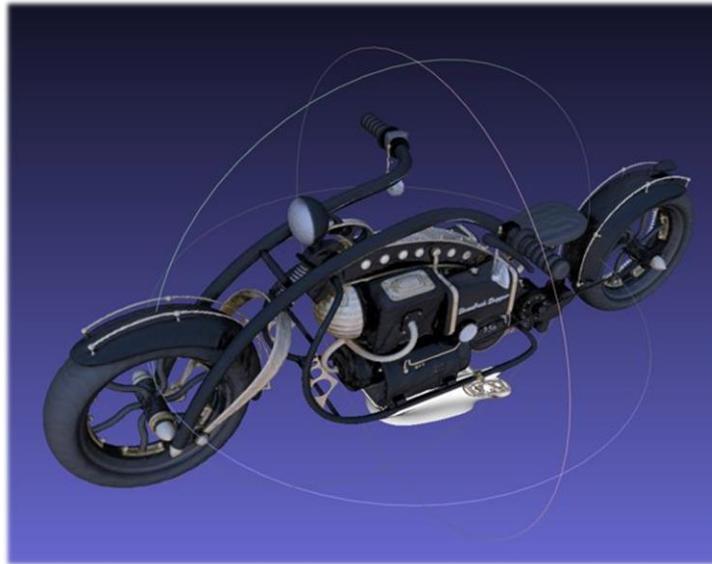
Voxurf: Fast 3D Object Reconstruction



S-LAB
FOR ADVANCED
INTELLIGENCE



Voxurf: Fast 3D Object Reconstruction



DreamGaussian: Efficient 3D Generation

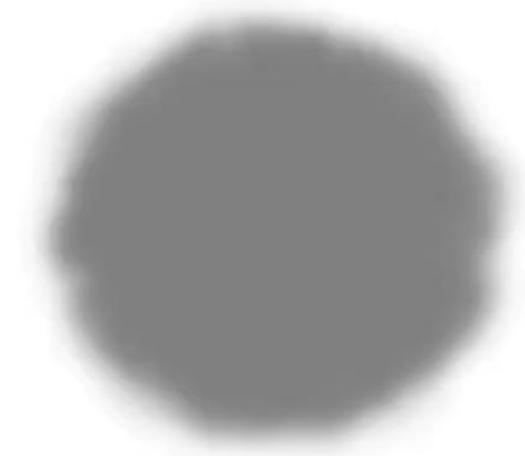


S-LAB
FOR ADVANCED
INTELLIGENCE



Zero-1-to-3 (NeRF)

00:00
Minute Second

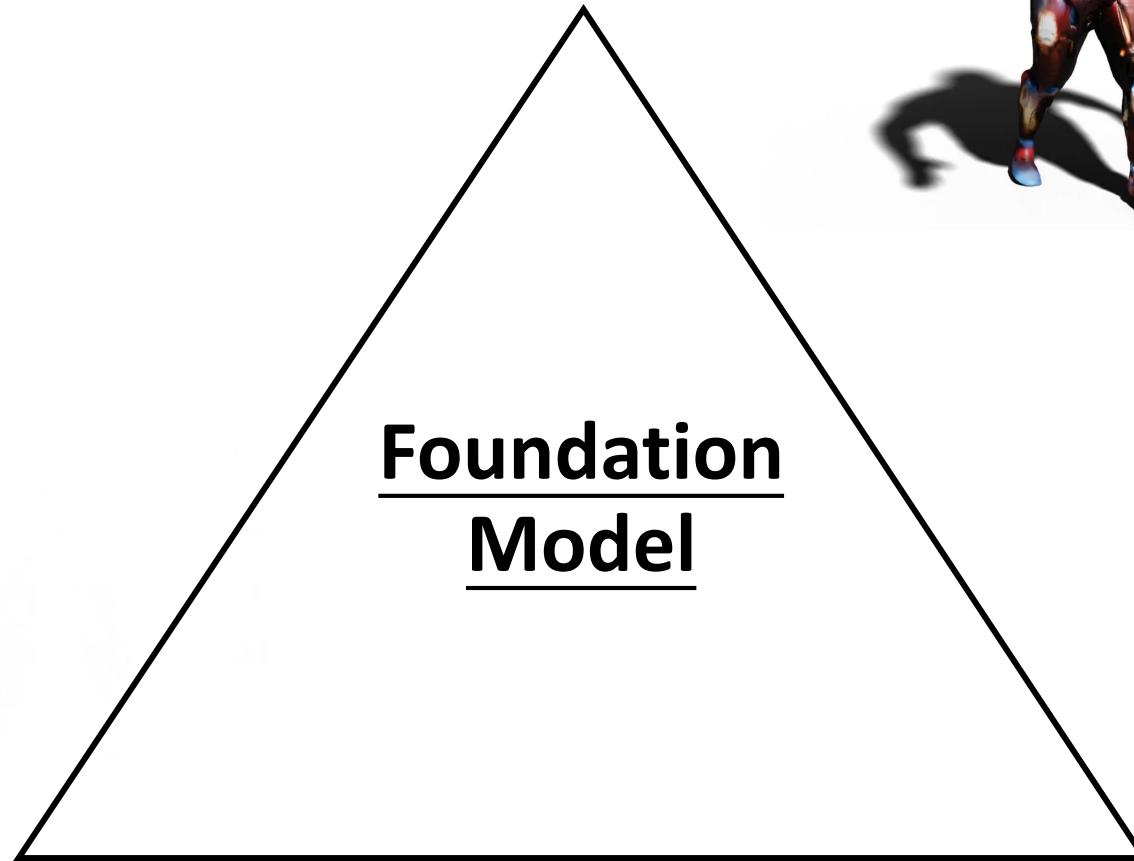


Ours (Gaussian Splatting)





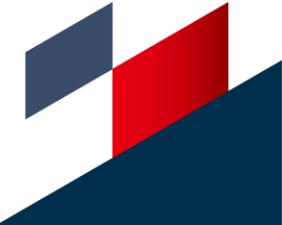
Object



Avatar



Scene

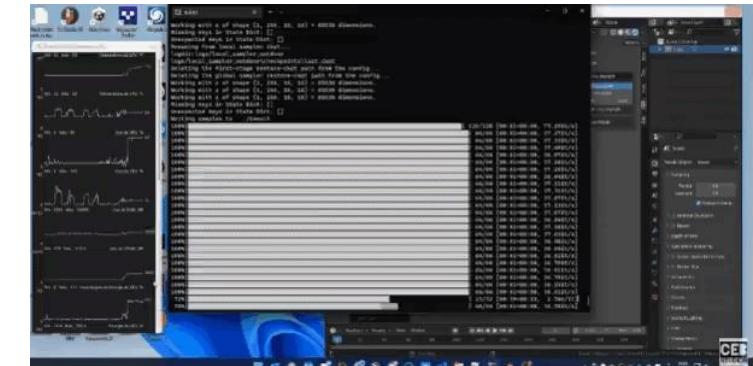


Text2Light: Text-to-3D Environment



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”



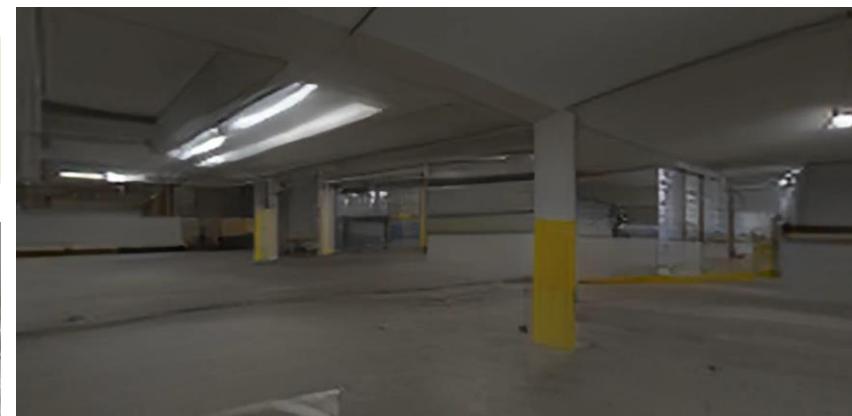
“gray concrete
pathway with
wall signages”



“blue and
brown wooden
counter”



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



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

Text2Light: Text-to-3D Environment



S-LAB
FOR ADVANCED
INTELLIGENCE

Text2Light
Own Your Reality
with Any Sentences

Describe Your Scene

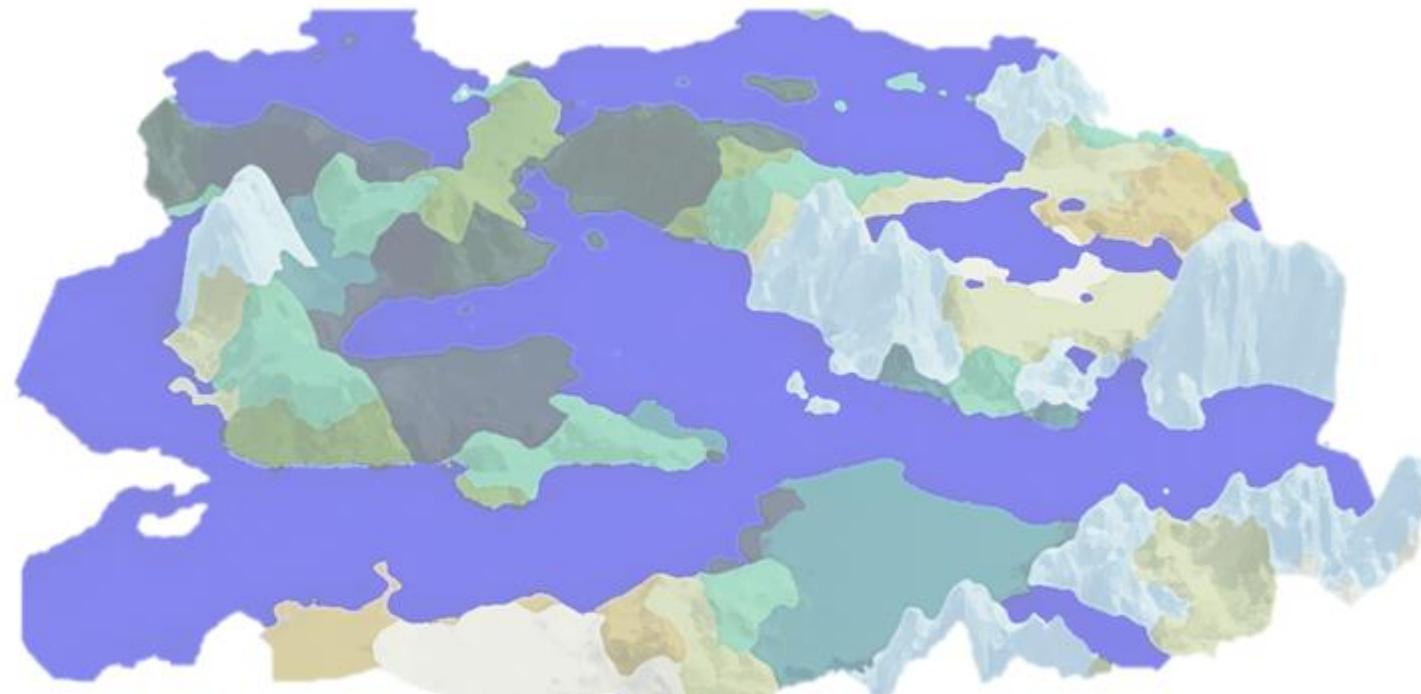
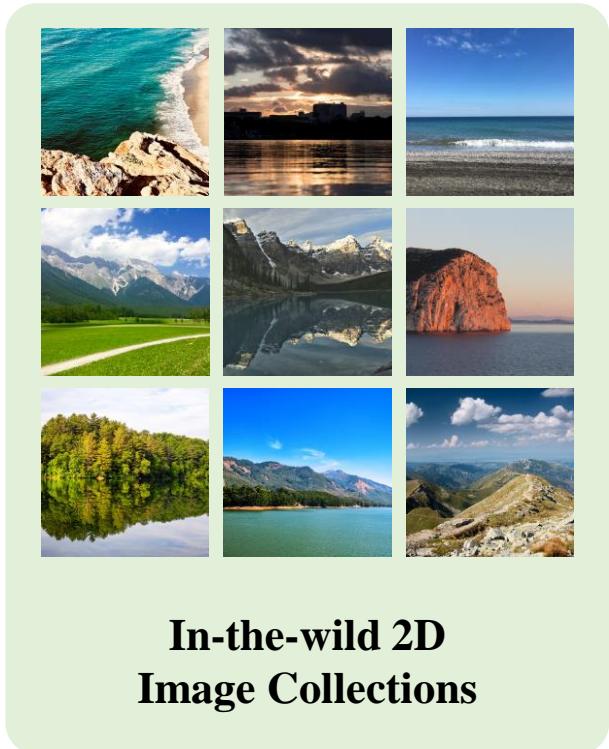
e.g. a living room

Generate

Render

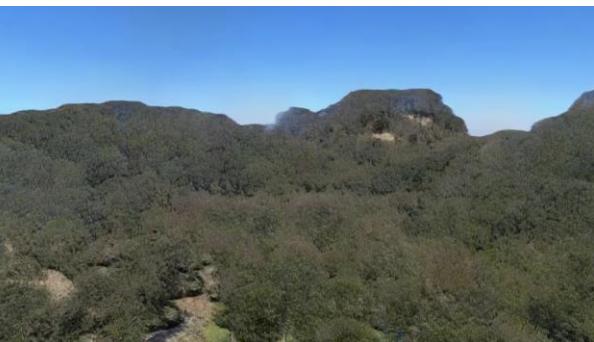


SceneDreamer: Unbounded 3D Scene Generation

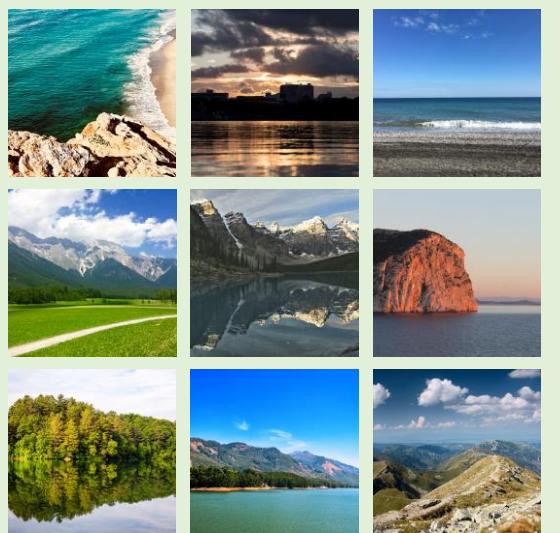


Photorealistic
Unbounded 3D Scenes

SceneDreamer: Unbounded 3D Scene Generation



Multi-view consistent



In-the-wild
Image Collections

Well-defined geometry

Diverse scenes and styles



Photorealistic
Unbounded 3D Scenes



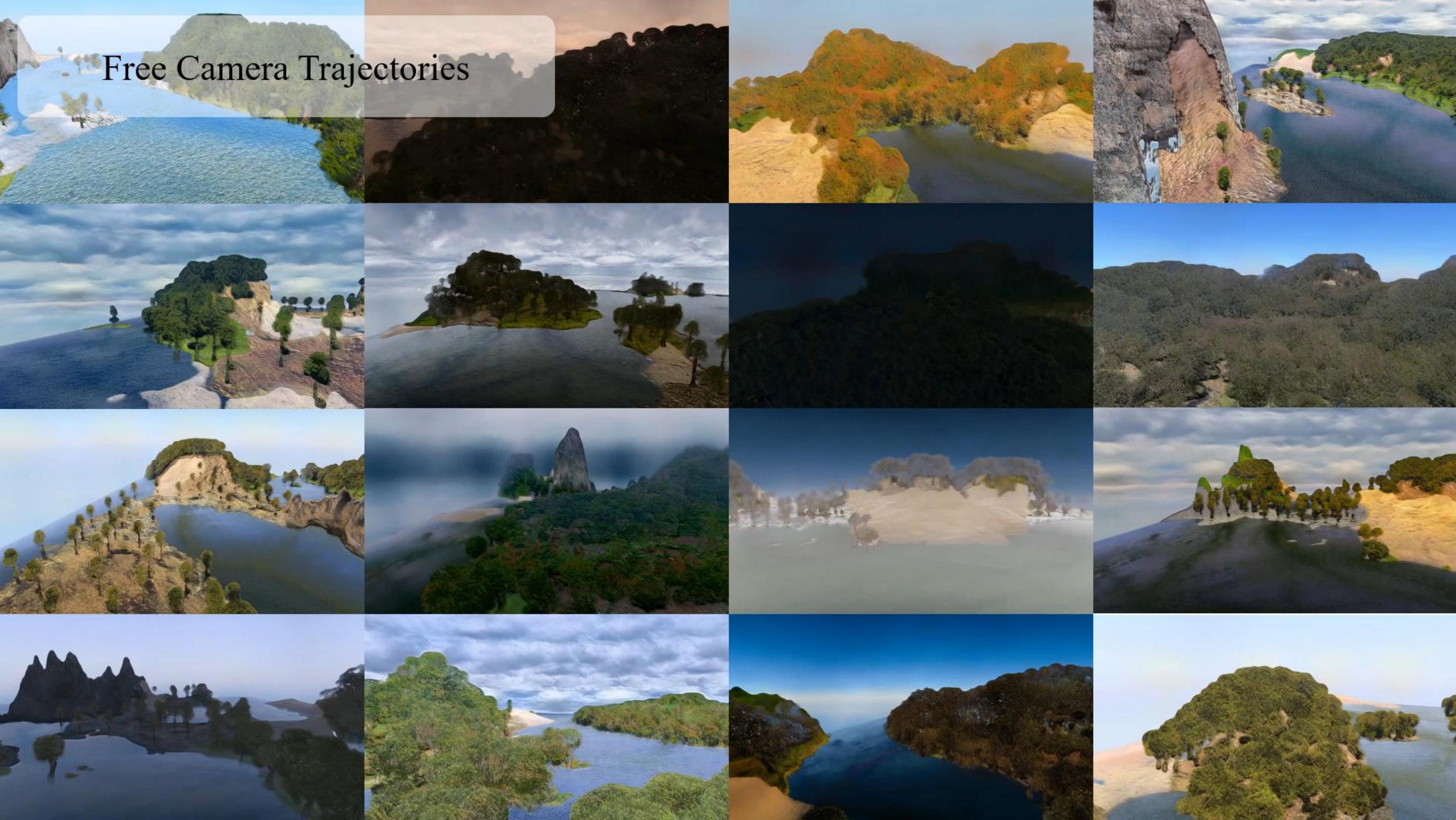
Infinite 3D World!



Generate with Different Styles



Free Camera Trajectories



CityDreamer: Unbounded 3D City Generation



S-LAB
FOR ADVANCED
INTELLIGENCE

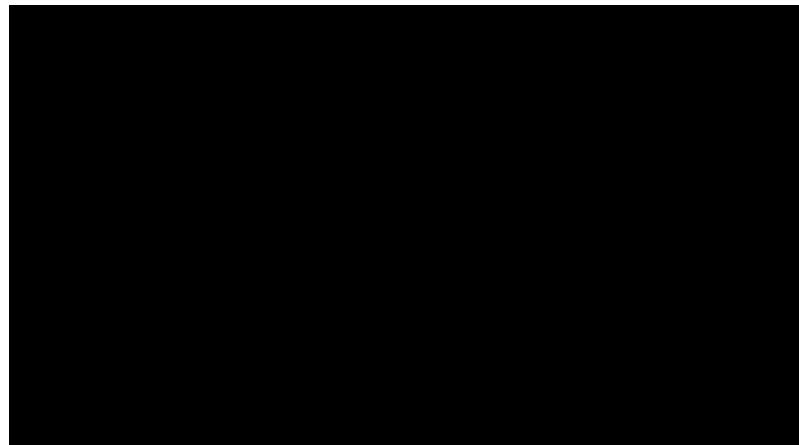
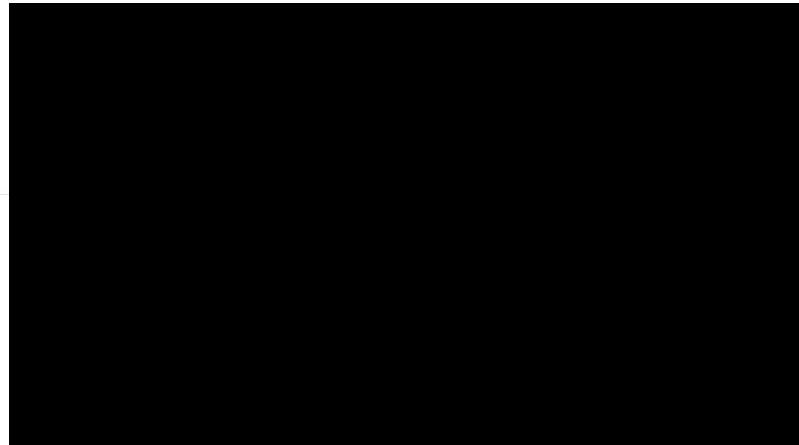
CityDreamer: Compositional Generative Model of Unbounded 3D Cities

The official demo to generate your own city in New York style.

[Source Code](#) [Project Page](#)



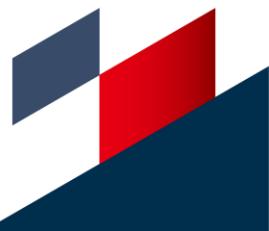
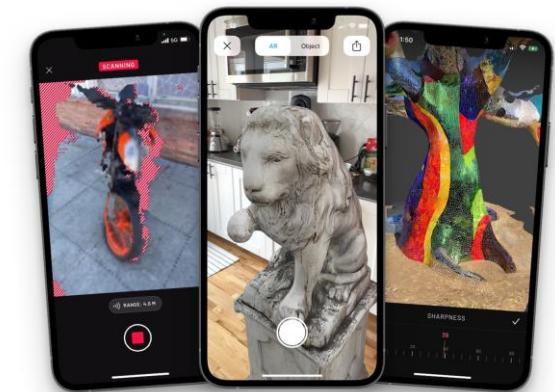
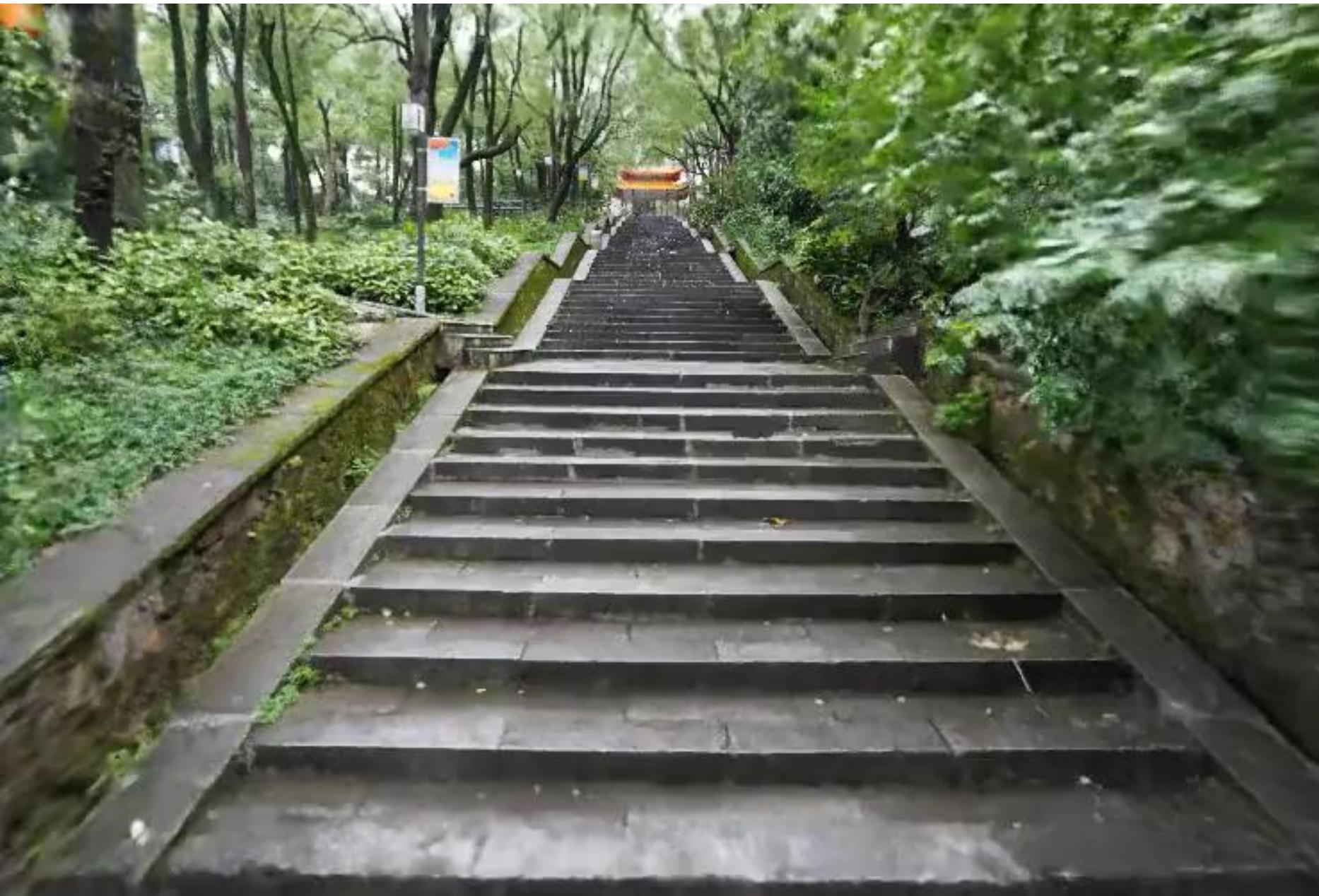
The screenshot shows the CityDreamer web interface. On the left, a sidebar menu has three items: "Layout" (selected), "Trajectory" (camera icon), and "Render" (video camera icon). The main area is titled "Layout" and contains controls for "Data Source" (set to "Layout Generator") and "Layout Size" (set to "4096x4096"). A large blue "Generate" button is prominent. Below these controls are two empty rectangular boxes labeled "Segmentation Map" and "Height Field". At the bottom of the main area, a note says "Press and hold the Ctrl/Command key to enter the edit mode."



F2NeRF: Mobile 3D Scene Reconstruction



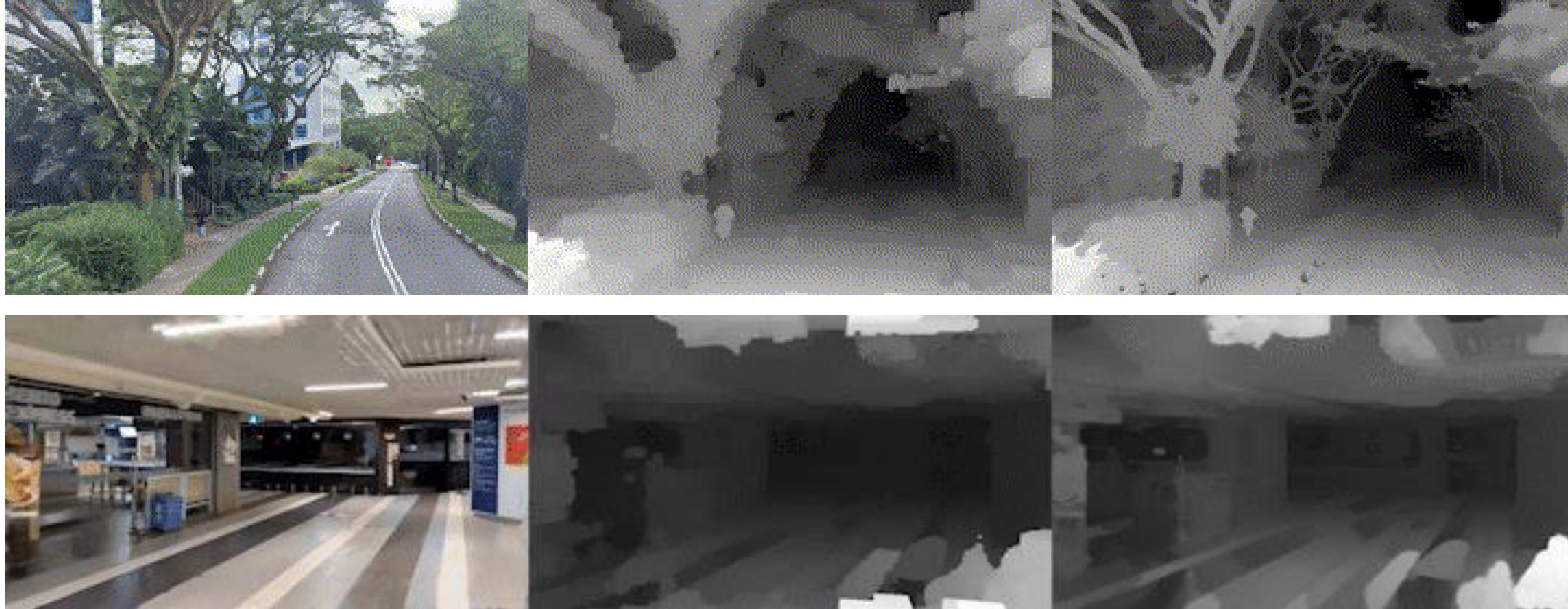
S-LAB
FOR ADVANCED
INTELLIGENCE



F2NeRF: Mobile 3D Scene Reconstruction

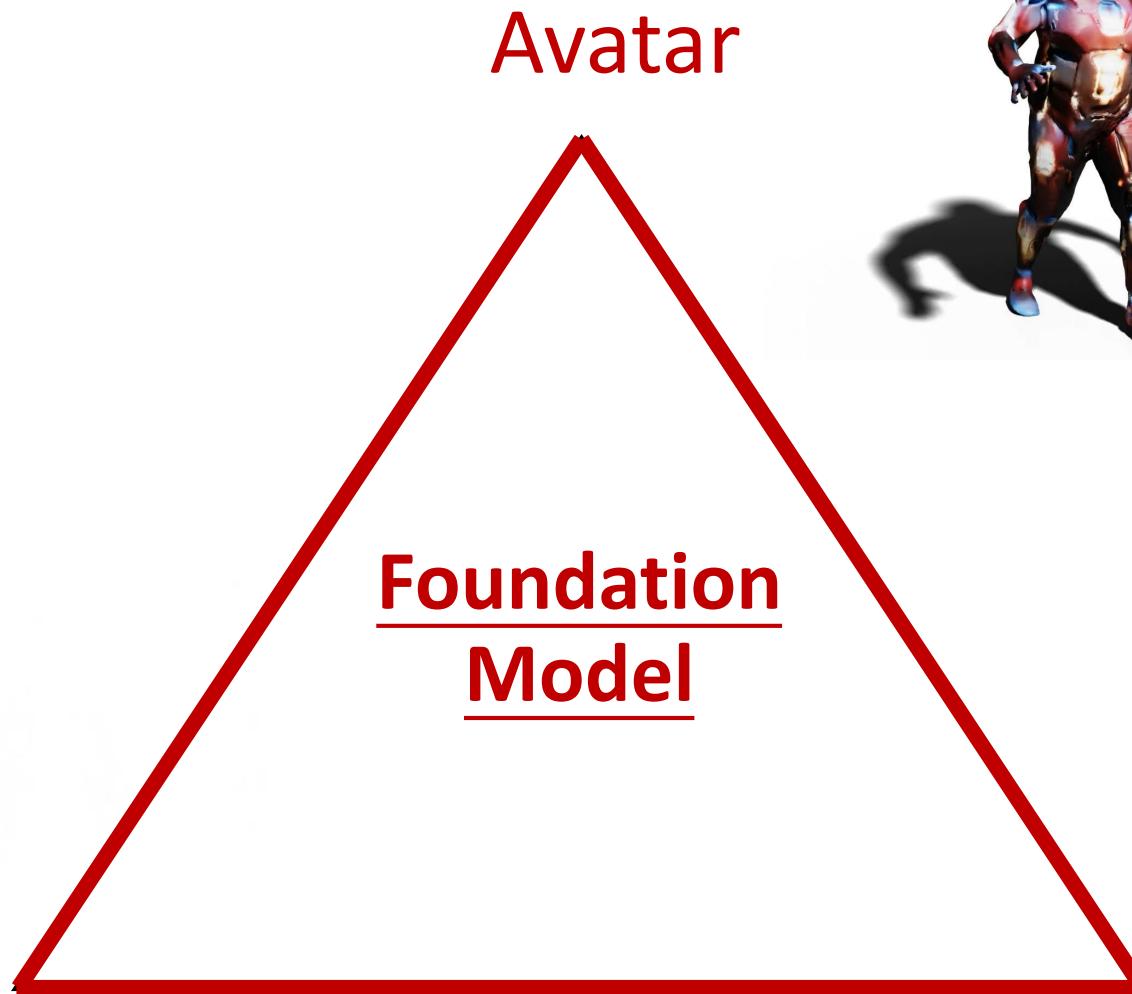


S-LAB
FOR ADVANCED
INTELLIGENCE





Object



Scene



FreeU: Free Lunch in Diffusion U-Net

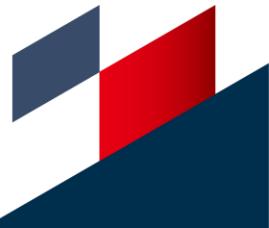


S-LAB
FOR ADVANCED
INTELLIGENCE

FreeU: Free Lunch in Diffusion U-Net

Chenyang Si, Ziqi Huang, Yuming Jiang, Ziwei Liu

S-Lab, Nanyang Technological University



LaVie: Text-to-Video Foundation Model



Cinematic shot of Van Gogh's selfie, Van Gogh style.



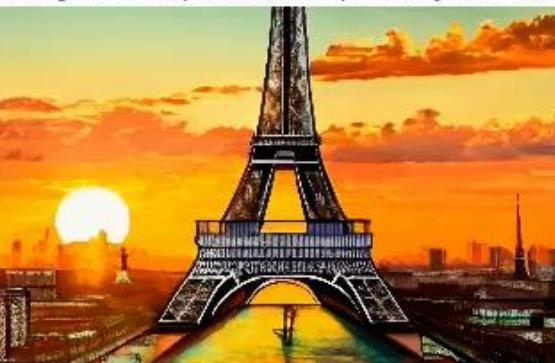
A corgi's head depicted as an explosion of a nebula.



Yoda playing guitar on the stage.



The bund Shanghai, oil painting.



The Eiffel Tower at sunset, trending on artstation, oil painting.



A fantasy landscape, trending on artstation, 4k high resolution.



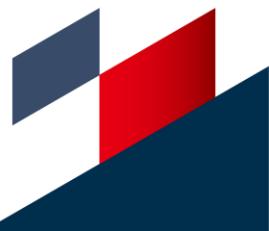
A super cool giant robot in Cyberpunk city, artstation.



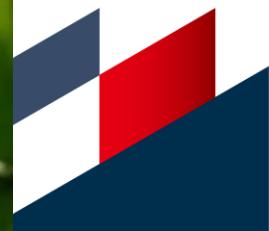
A Mars rover moving on Mars.



A space shuttle launching into orbit, with flames and smoke billowing out from the engine.



Adventure of a Panda



SEINE: Image-to-Video Foundation Model



S-LAB
FOR ADVANCED
INTELLIGENCE



Image2Video



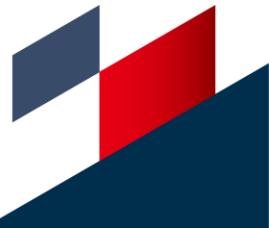
Transition

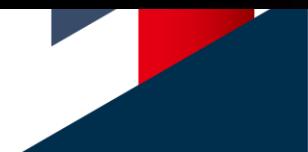


Image2Video



Transition

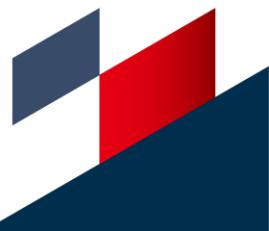




HyperDreamer: Image-to-3D Foundation Model



S-LAB
FOR ADVANCED
INTELLIGENCE





Object

Avatar



Thank You!



Scene

