NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
S-LAB FOR ADVANCED INTELLIGENCE

3DTopia: Foundation Ecosystem for 3D Generative Models

Ziwei Liu（刘子纬）

https://liuziwei7.github.io/
Nanyang Technological University

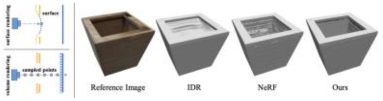
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Learning 3D from Multi-View Supervision

Efficiency

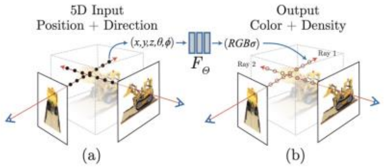
[Instant-NGP, Müller et al. 2022]

Surface



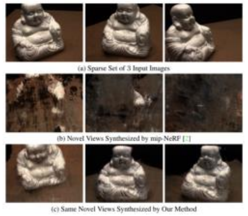
[Neus, Wang et al. 2021]

Original NeRF



[NeRF, Mildenhall et al. 2020]

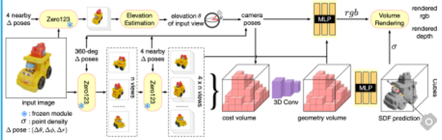
Regularization



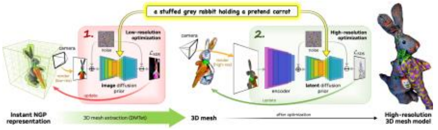
[RegNeRF, Niemeyer et al., 2022]

[FreeNeRF, Yang et al., 2023]

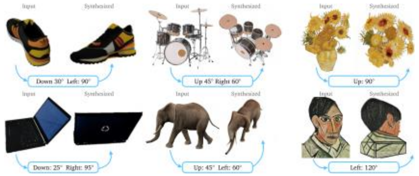
Diffusion priors



[One-2-3-45, Liu et al. 2023]



[Magic3D, Lin et al. 2022]

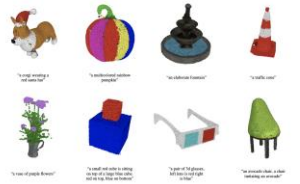


[Zero-123, Liu et al. 2023]

[DreamFusion, Poole et al. 2022]

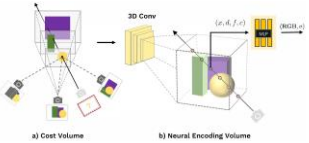
Foundation models

[Shap-E, Jun et al. 2023]

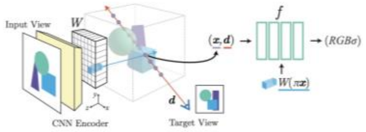


[Point-E, Nichol et al. 2022]

Generalizable NeRFs



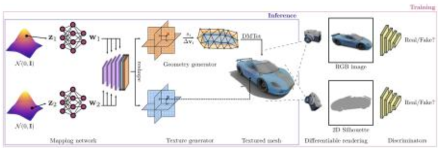
[MVSNeRF, Chen et al. 2021]



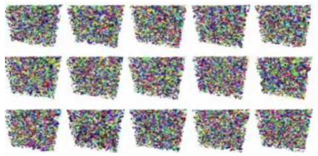
[pixelNeRF, Yu et al. 2020]

3D generative models

[AutoSDF, Mittal et al. 2022]



[GET3D, Gao et al. 2022]

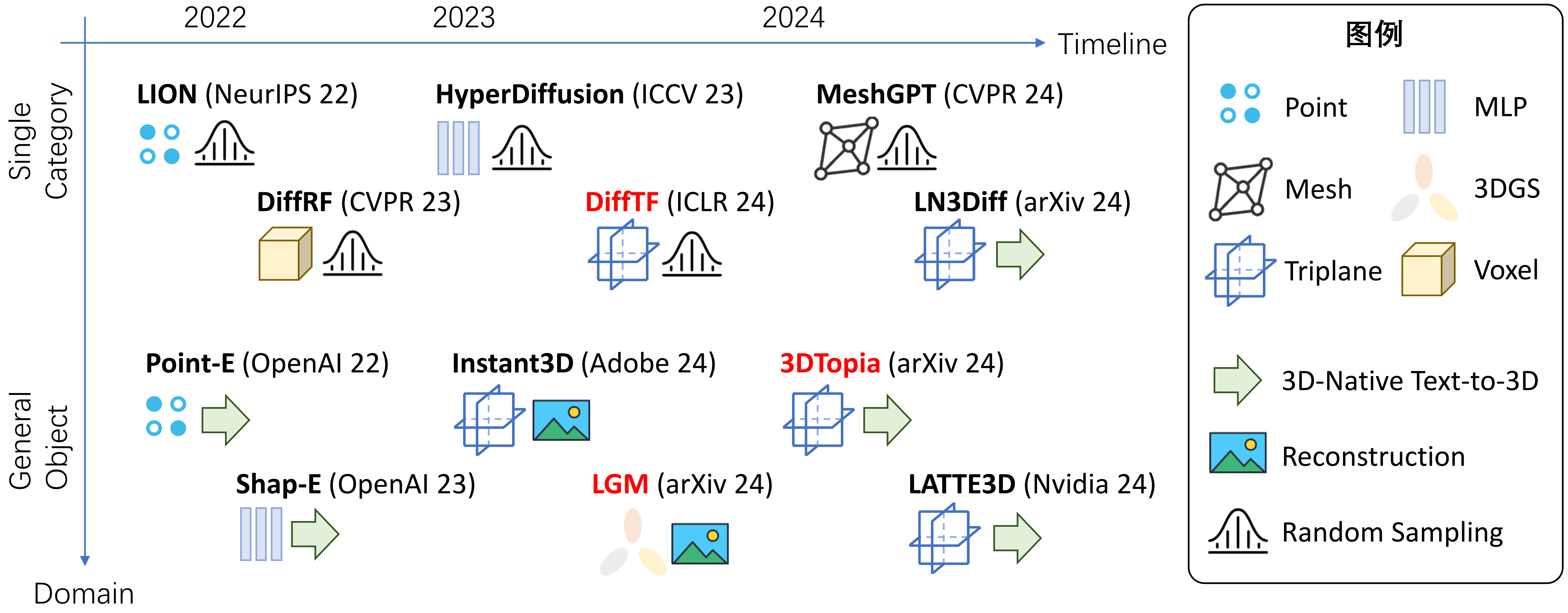


[DiffRF, Müller et al. 2023]

Reconstruction ⟵⟶ **Generation**

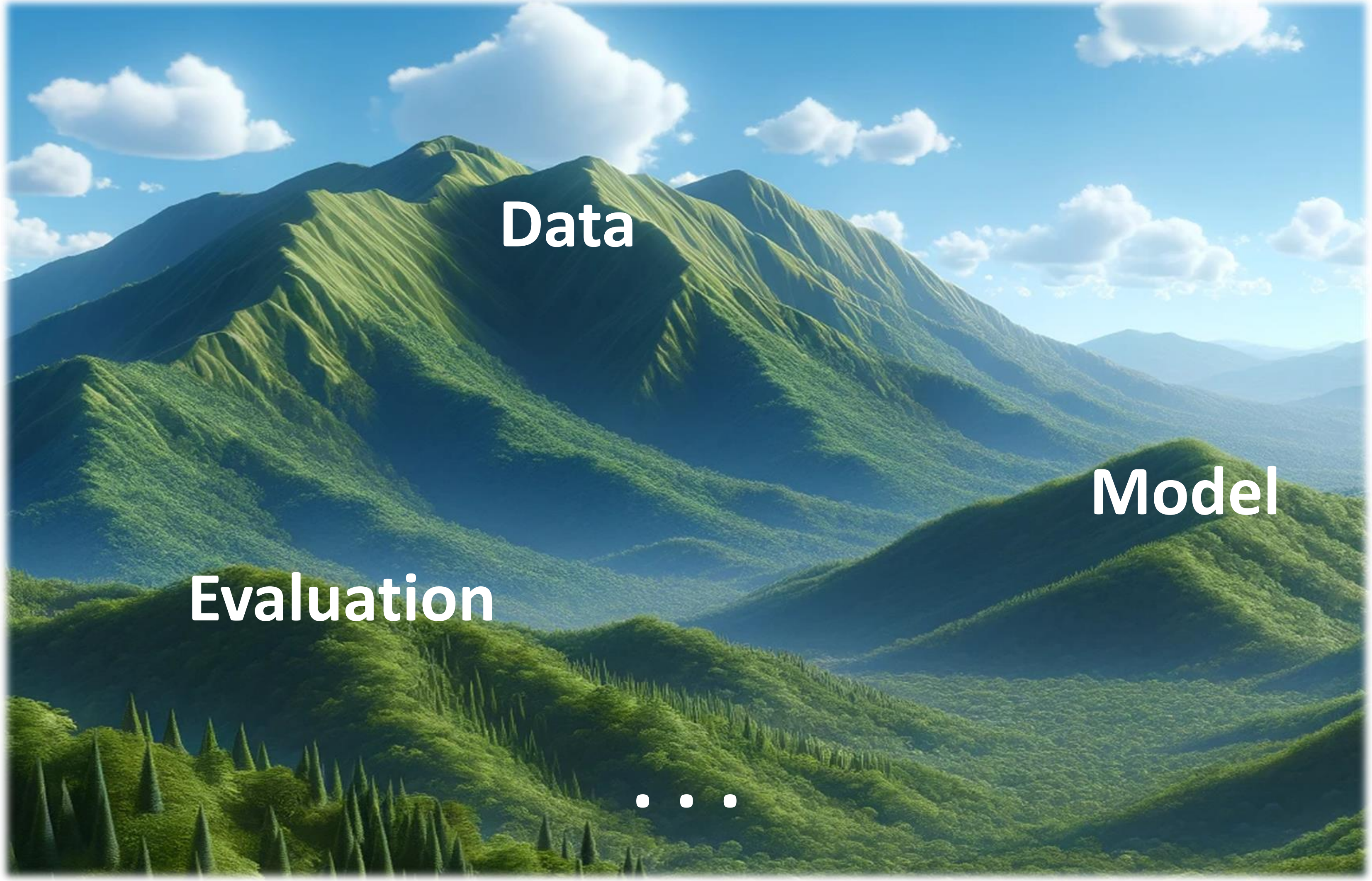
Dense Views | **Sparse Views** | **Single View** | · · · · | **Category/Text/Image Condition**

3D Generative Models



Foundation Ecosystem

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE



Overview

1. 3D Object Dataset
2. 3D Generative Models
3. 4D Generative Models
4. 3D Generation Evaluation

3D Object Dataset

OmniObject3D

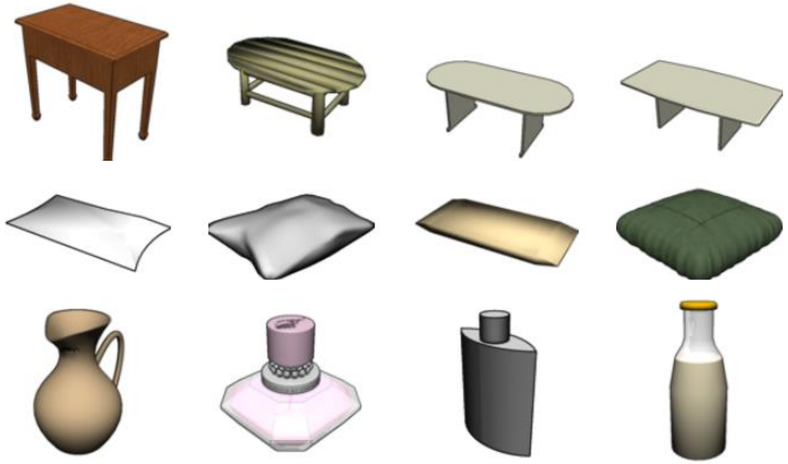
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation

[CVPR 2023 Best Paper Candidate] (0.51%, 12/2359)

OmniObject3D: Background & motivation

Synthetic data



ShapeNet
large in scale
low quality
not realistic

Multi-view images

CO3D
large in scale
No 3D GT

Real-world 3D scans

Google scanned objects
high quality
real-world scans
household objects

OmniObject3D

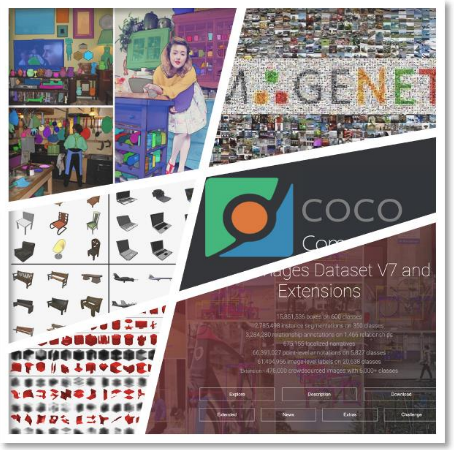
large-vocabulary
high quality
real-world scans

OmniObject3D: Statistics

We are still collecting data, and the size of OmniObject3D is still growing.

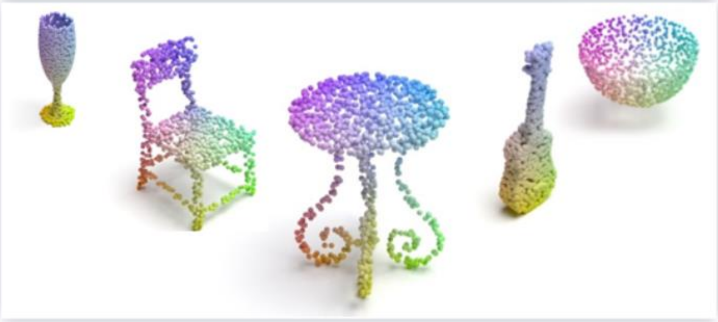
A new version with 18k real scanned 3D objects will be released as soon as possible.

	Dataset	Year	Real	Full 3D	Video	Num Objs	Num Cats
Synthetic data	ShapeNet	2015		√		51k	55
	ModelNet	2014		√		12k	40
online assets with a variety of data types	Objaverse	2023		√		818k	21k
	3D-Future	2020		√		16k	34
	ABO	2021		√		8k	63
	Toys4K	2021		√		4k	105
Multi-view image	CO3D V1/V2	2021	√		√	19k/40k	50
	MVImgNet	2023	√		√	219k	238
Real-world 3D scans	DTU	2014	√	√		124	NA
	GSO	2021	√	√		1k	17
	AKB-48	2022	√	√		2k	48
	Ours	2022	√	√	√	**6k**	**190**

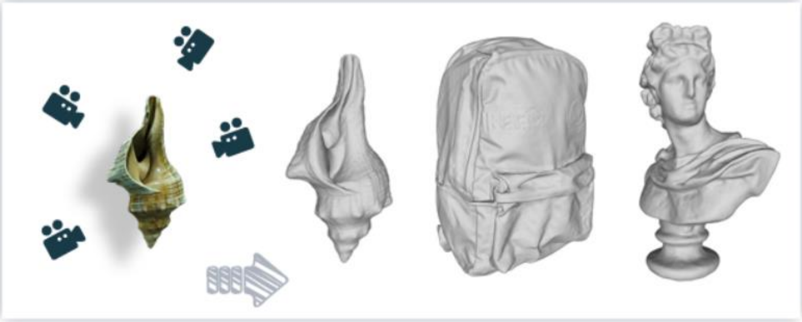


OmniObject3D: Applications

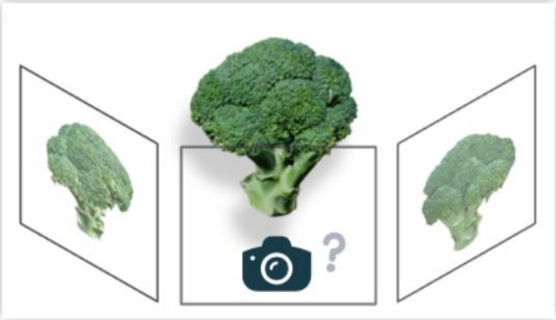
Perception



Surface Reconstruction



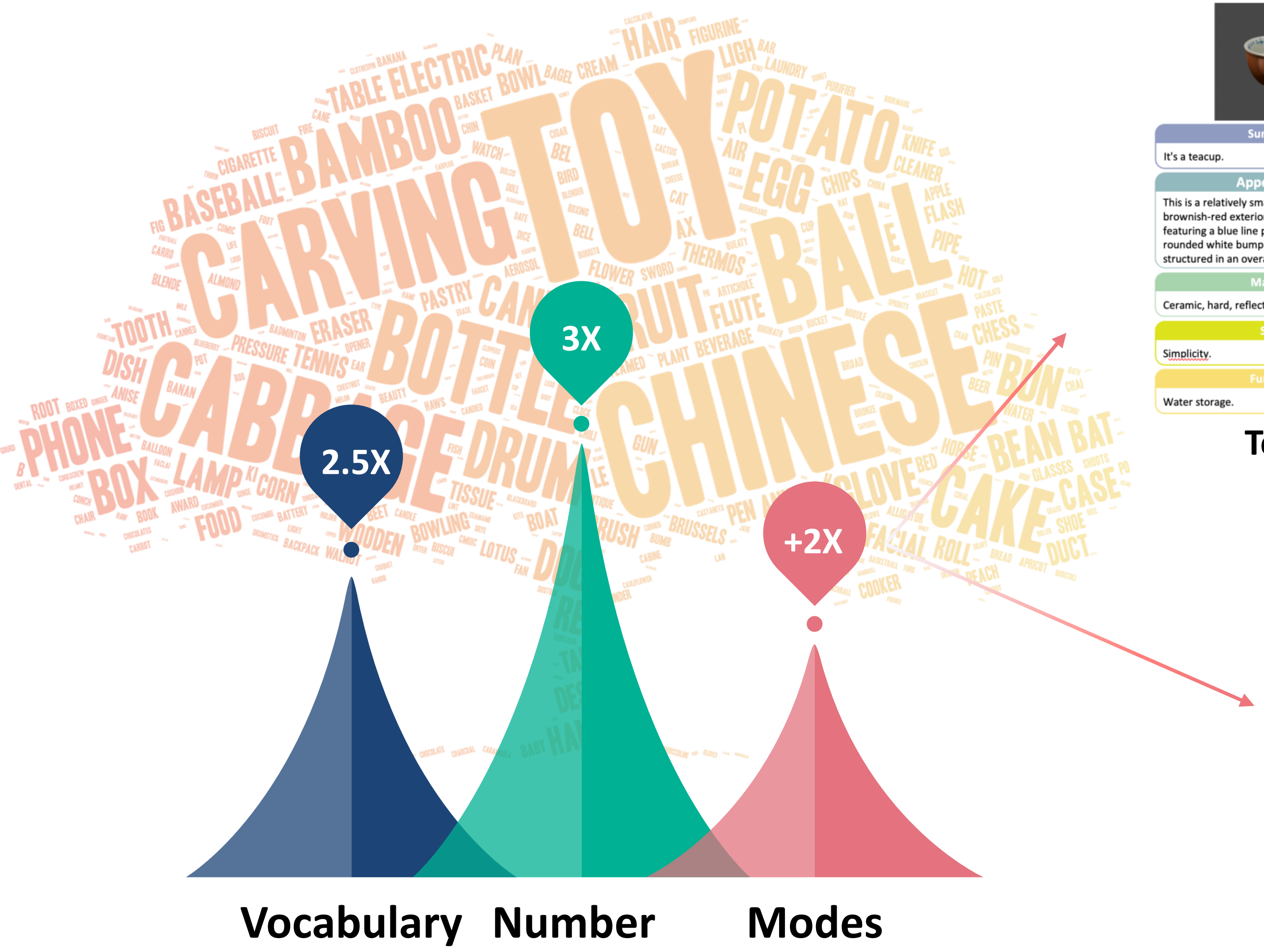
Novel View Synthesis

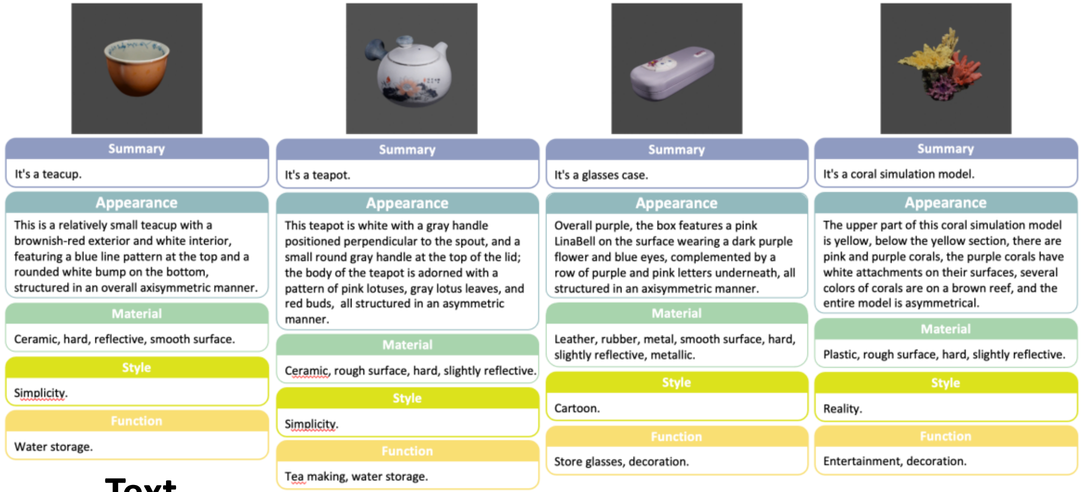


Generation

OmniObject3D V2

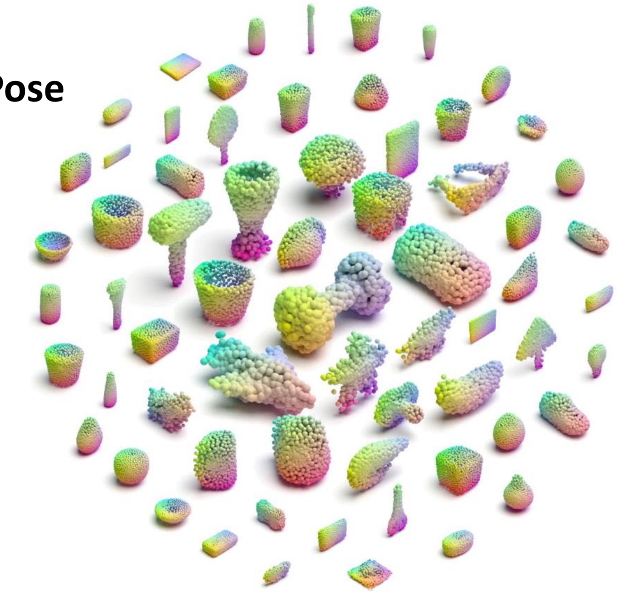
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE





Text

Pose



OmniObject3D V2

Summary

It's a teacup.

Appendance

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

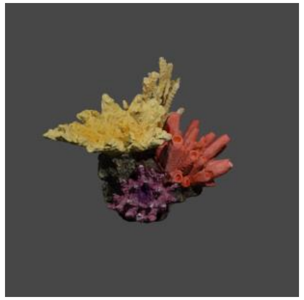
Leather, rubber, metal, smooth surface, hard, slightly reflective, metallic.

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.

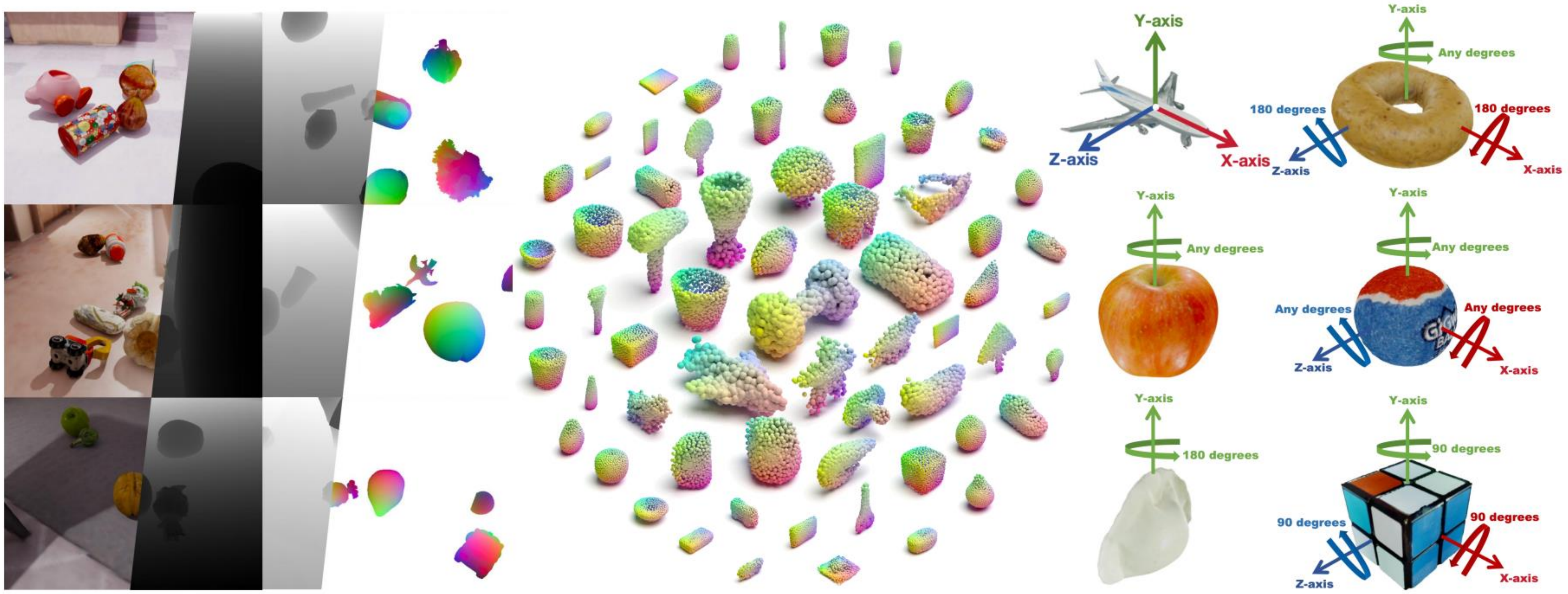
Style

Reality.

Function

Entertainment, decoration.

Omni6D: Large-Vocabulary 6D Pose Estimation



(a) Omni6D Dataset (b) Shape Priors (c) Symmetry Annotation

3D Generative Models

3DTopia: Motivation & Overview

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

3DTopia: Large Text-to-3D Generation Model with Hybrid Diffusion Prior

Motivation:

- 3D Diffusion Model (**3D Prior**):
  - ✓ Fast Sampling Speed (a few seconds),
  - ☐ but Low Generation Quality

- Score Distillation Sampling (**2D Prior**):
  - ✓ High Generation Quality,
  - ☐ but Slow Generation Speed (a few hours)

➢ 3DTopia (**Hybrid Diffusion Prior**)

- ✓ Fast Sampling Speed (3min)

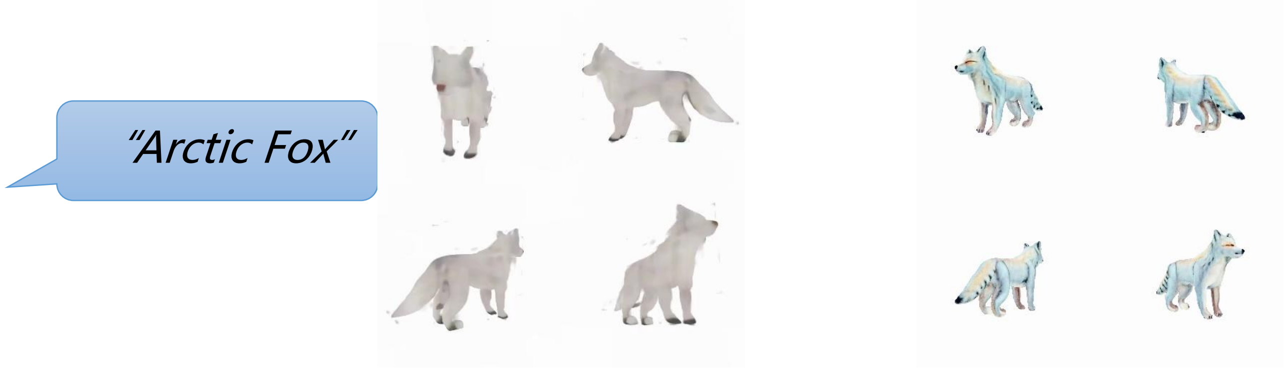
  Feed forward inference of first stage diffusion model

- ✓ High Generation Quality

  Second stage SDS ensures high quality texture

First Stage
Fast 3D Diffusion Sampling (30s)

Second Stage
Rapid SDS Finetuning (2.5min)



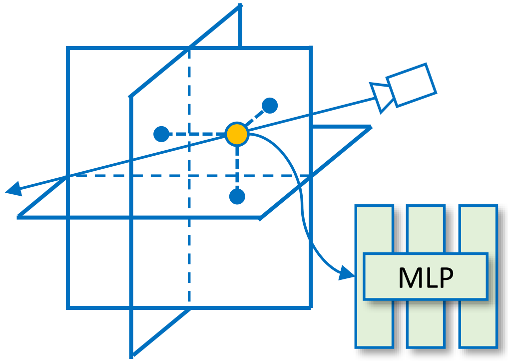
3DTopia: Dataset Preparation

Large 3D Dataset

Multi-View Images Rendering

3D Captioning

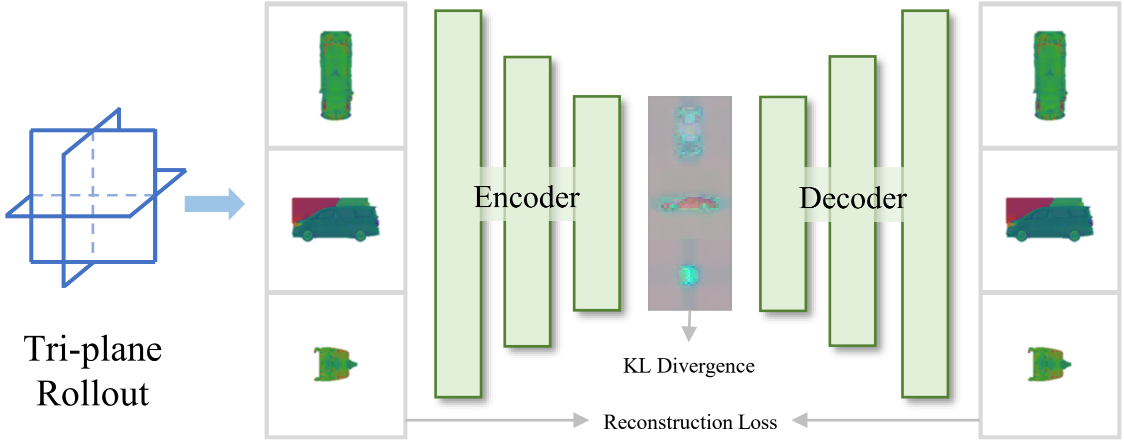
Data Cleaning



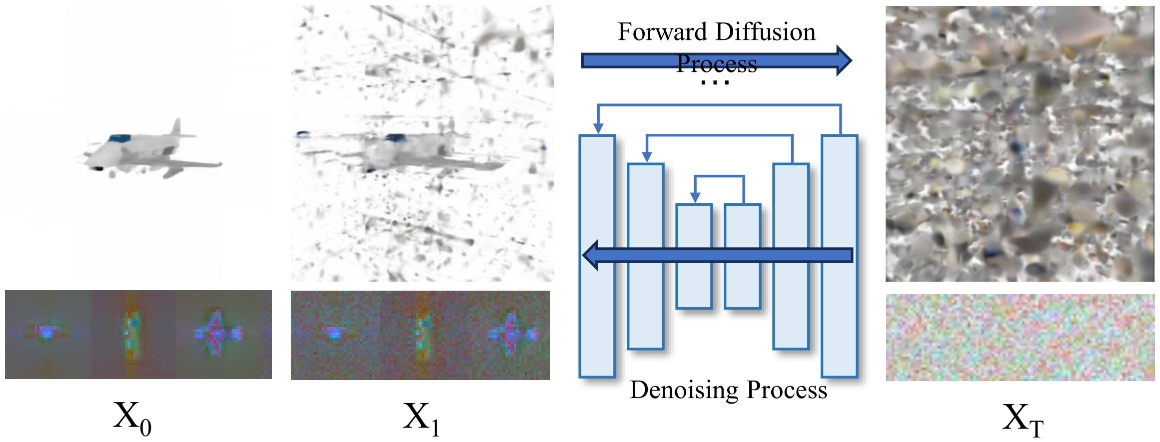
Tri-plane Fitting

3DTopia: Method

Stage I: Tri-plane Latent Diffusion Model



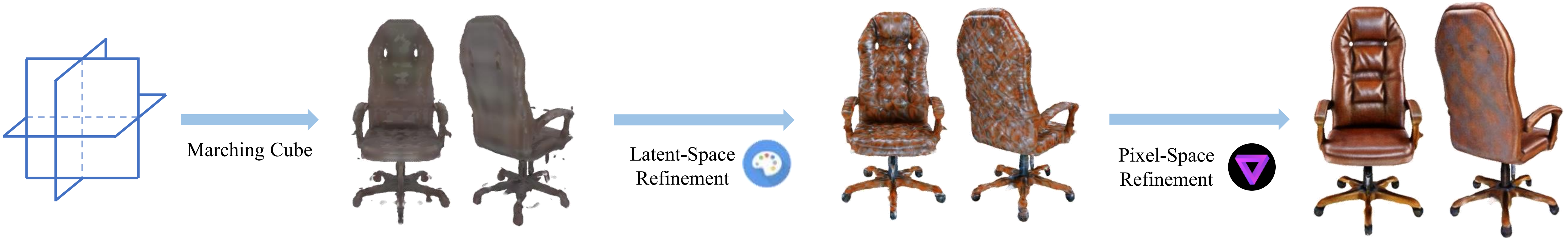
Tri-plane VAE



Latent Diffusion Model

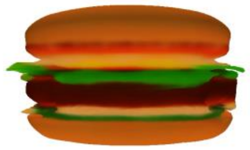
3DTopia: Method

Stage II: SDS-Based Refinement

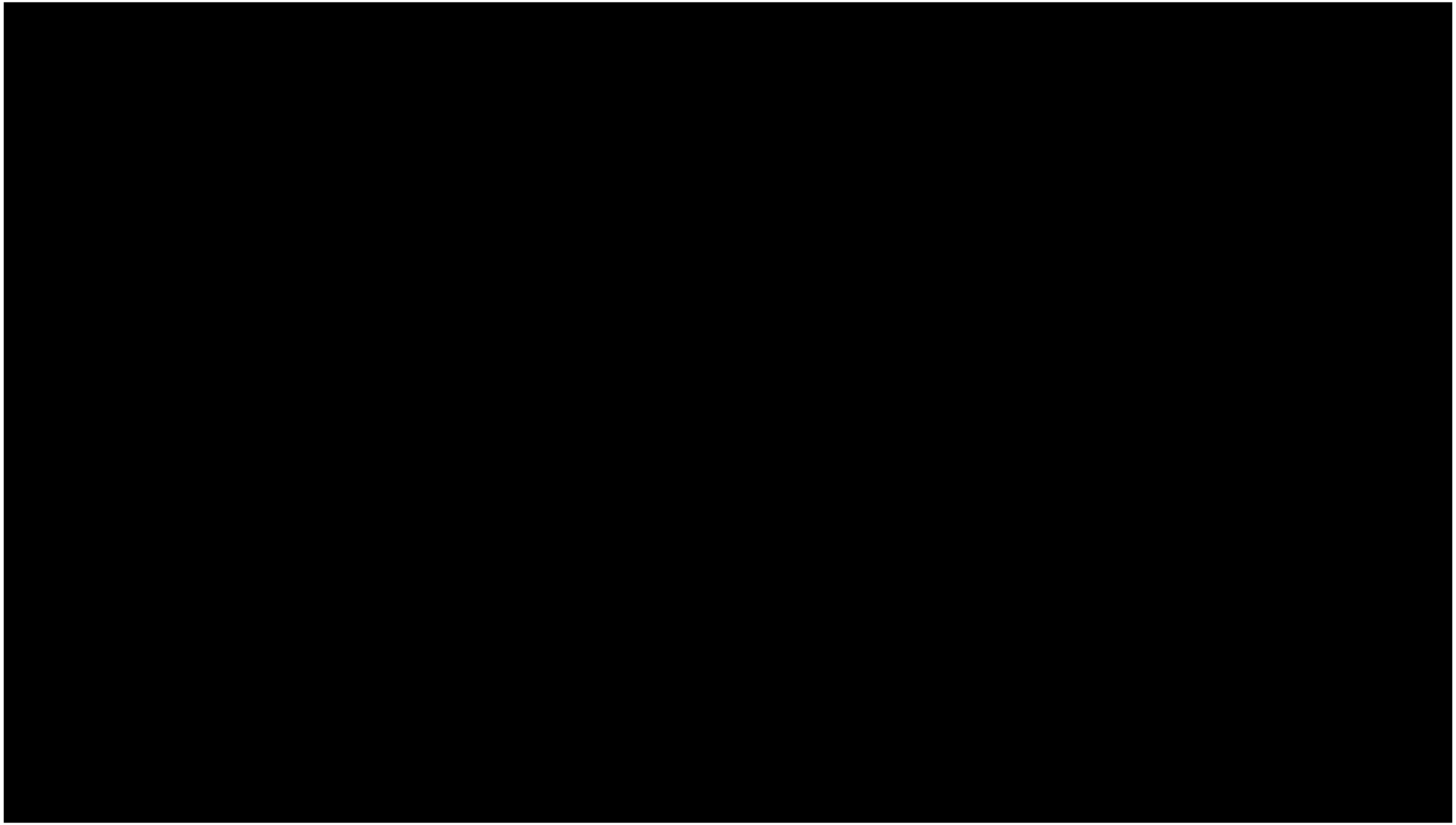


Two-Step SDS-Based Texture Refinement

3DTopia: Results

	A hamburger	Arctic Fox	Astronaut Suit and Helmet	Bean Bag Chair	Sailboat
Point-E					
Shap-E					
3DTopia					

3DTopia: Demo Video



ThemeStation: Motivation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

ThemeStation: Generating Theme-Aware 3D Assets from Few Exemplars
[SIGGRAPH 2024]

Motivation:

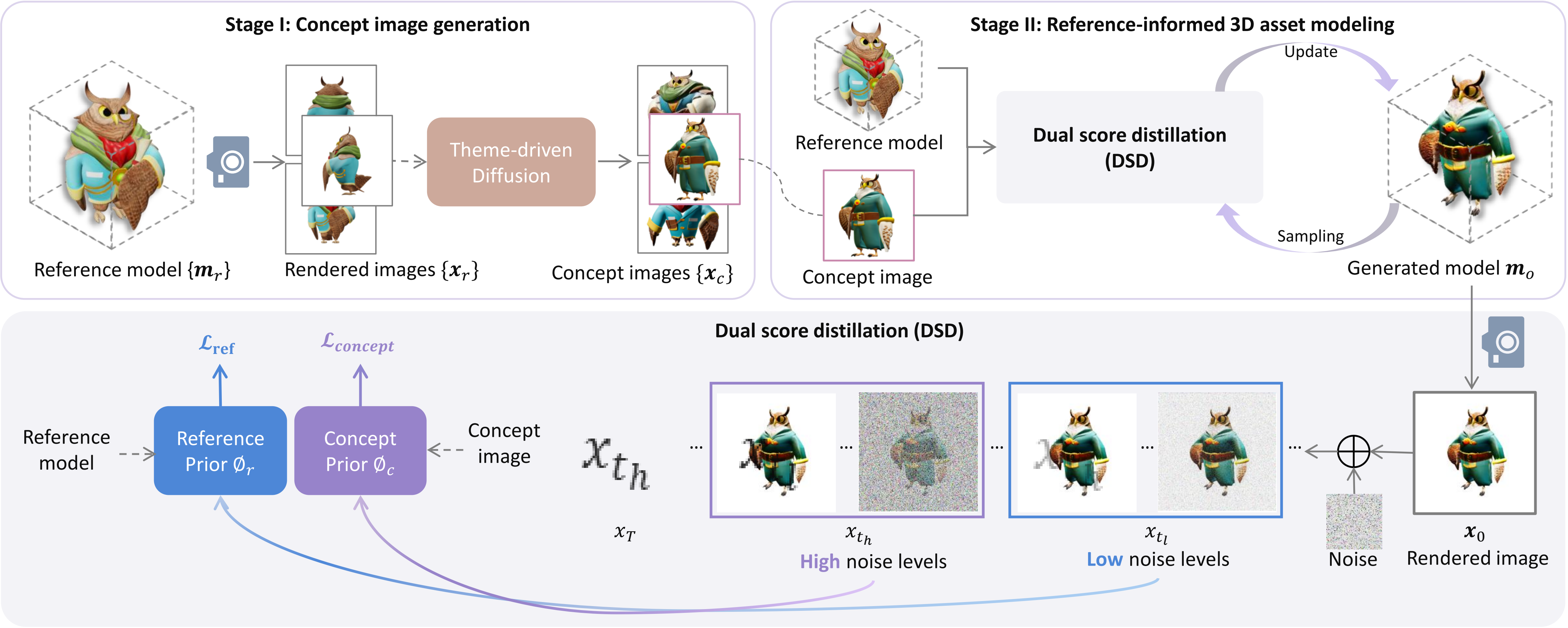
Generate 3D assets from a few 3D exemplars

ThemeStation can synthesize **customized 3D** assets based on **given few exemplars**.

Reference models

Generated 3D galleries

ThemeStation: Method Overview

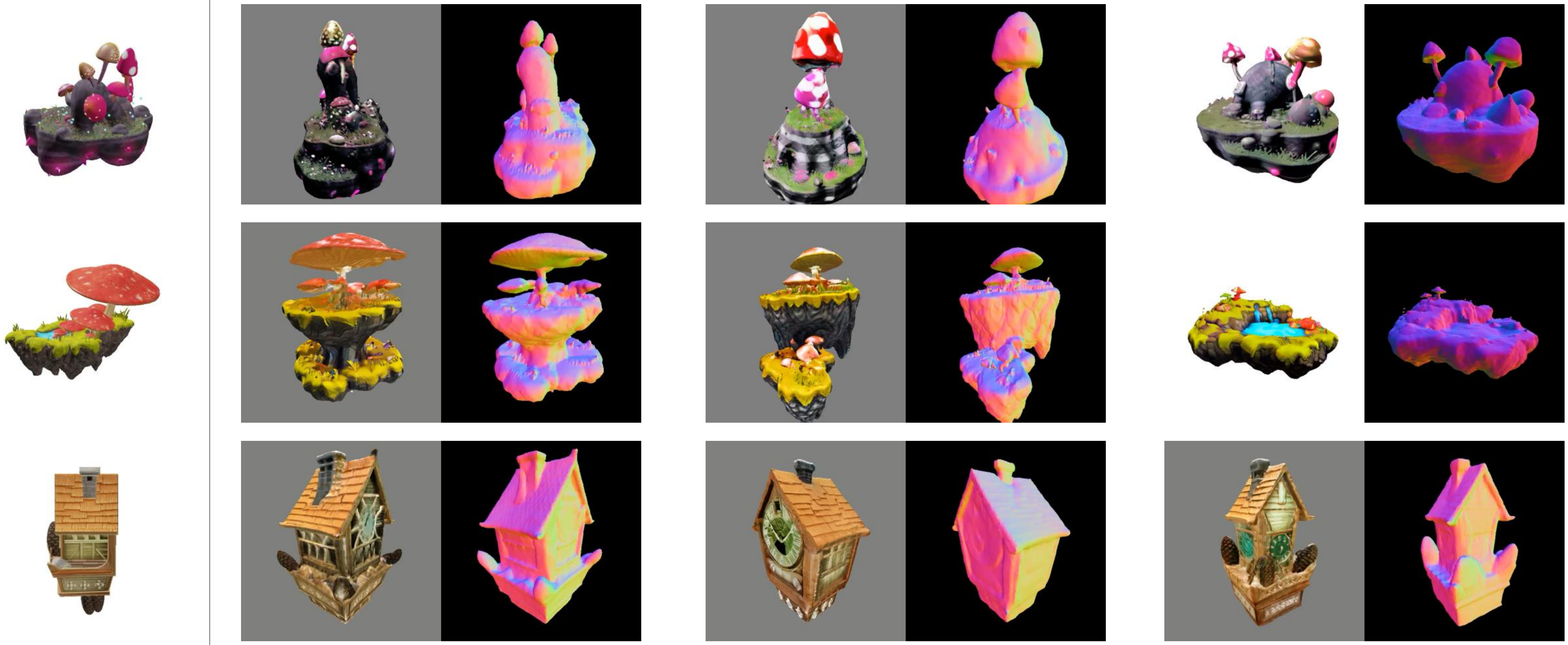


ThemeStation: a few 3D exemplar based generation

Reference models

Generated 3D galleries

ThemeStation: one 3D exemplar based generation



Reference models

Generated 3D galleries

HyperDreamer: Single image to 3D

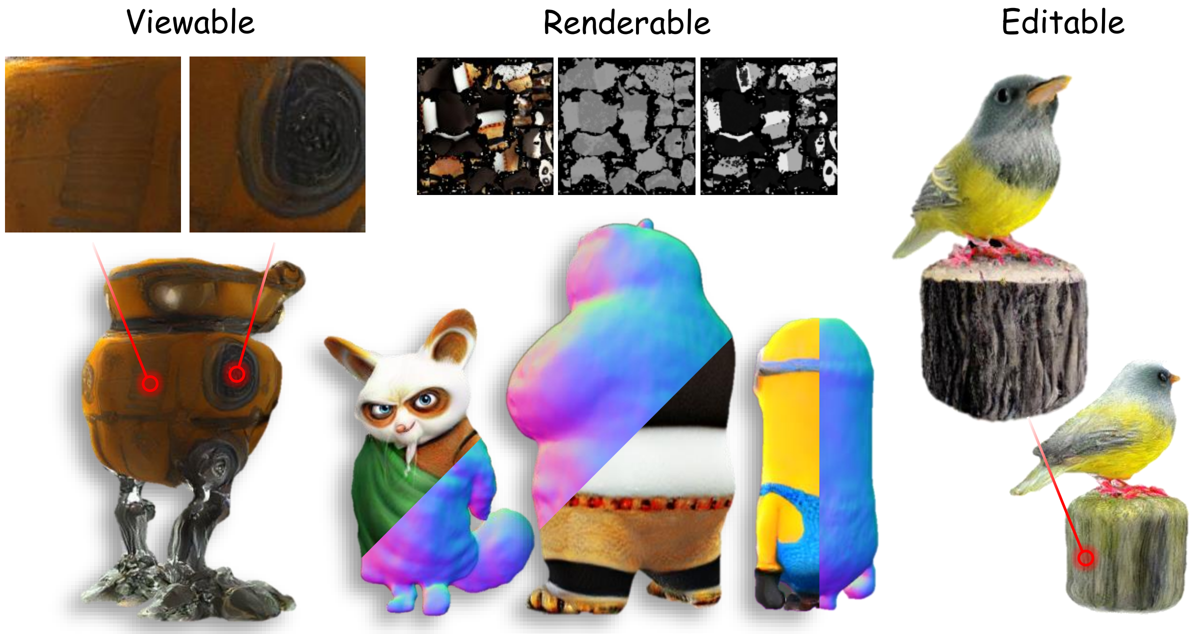
HyperDreamer: Hyper-Realistic 3D Content Generation and Editing from a Single Image

[SIGGRAPH Asia 2023]

Motivation:

To improve the texture resolution; add material modelling; support interactive texture editing.

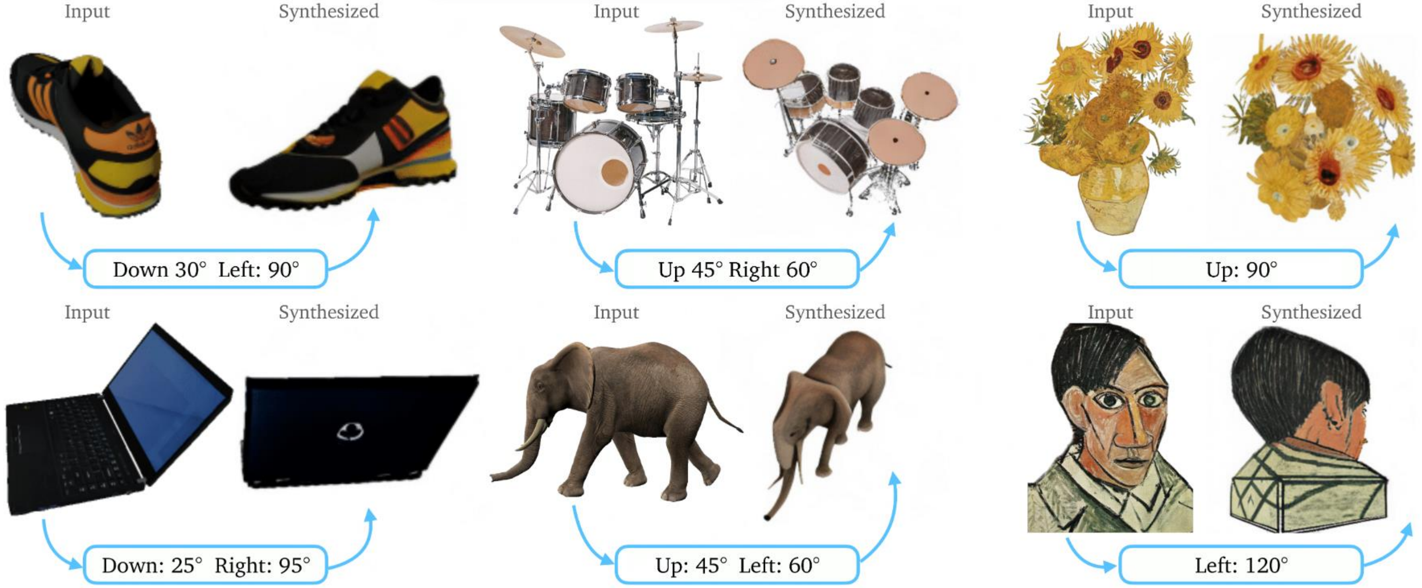
Reference images



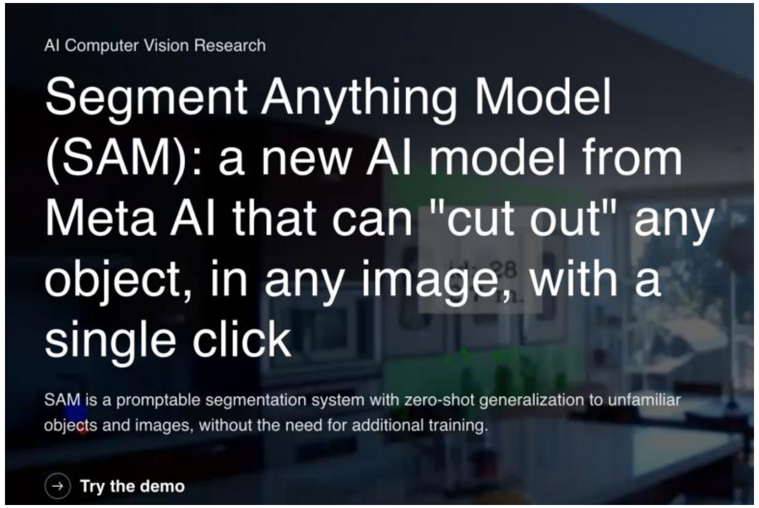
3D generation and editing

HyperDreamer: Priors to use

[Zero123, Liu et al., 2023] ***2D Diffusion Prior***

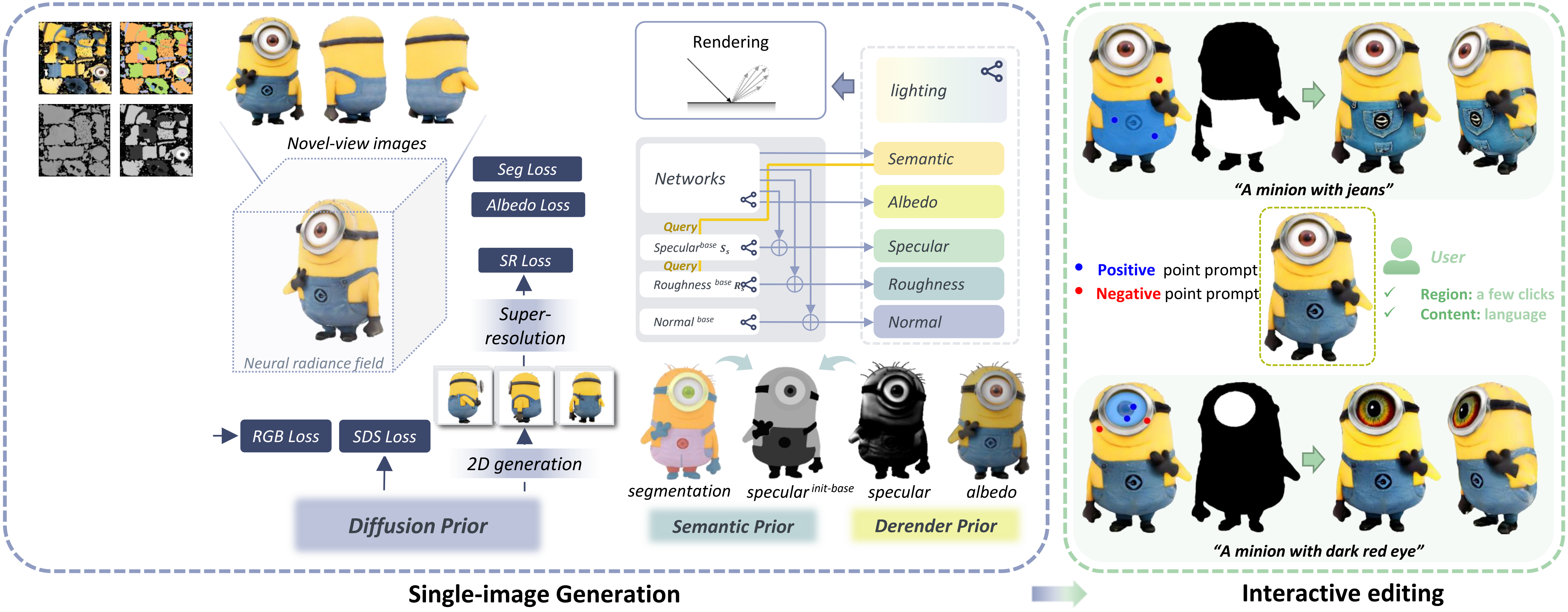


[SAM, Kirillov et al., 2023] ***Semantic Prior***



[De-rendering, Wimbauer et al., 2022] ***Material Prior***

HyperDreamer: Method

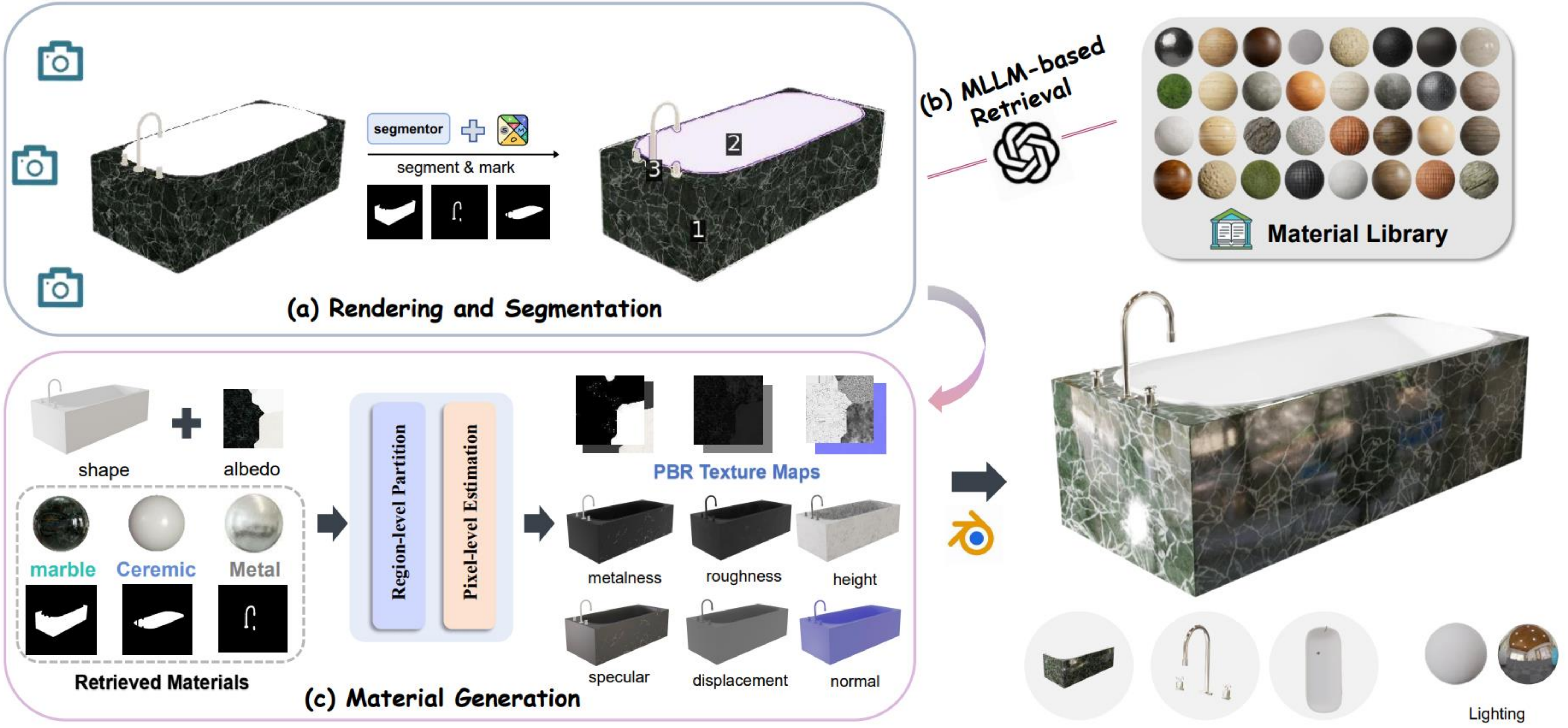


HyperDreamer: Demo

Make-it-Real: Reliable material inference and modelling

Motivation: Leverage Multimodal Large Language Model to provide priors for this highly ill-posed problem.

Make-it-Real: Method



Make-it-Real: Demo

DreamGaussian: Motivation & Overview

DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation

[ICLR 2024 Oral] (3.75%, 86/2296)

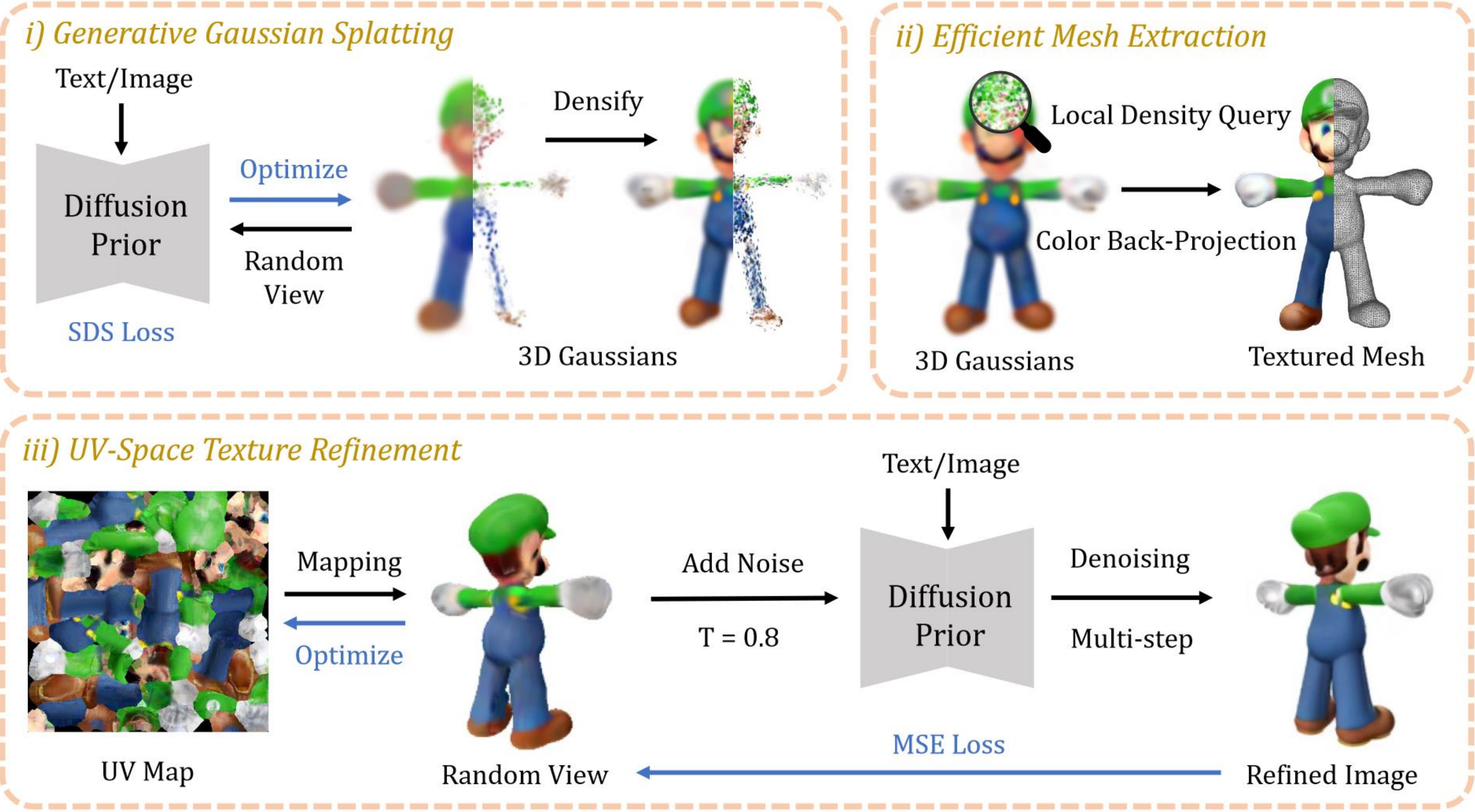
Motivation:

Current obstacles for practical 3D generation:

Success rate and **Generation speed**.

3D representation matters:

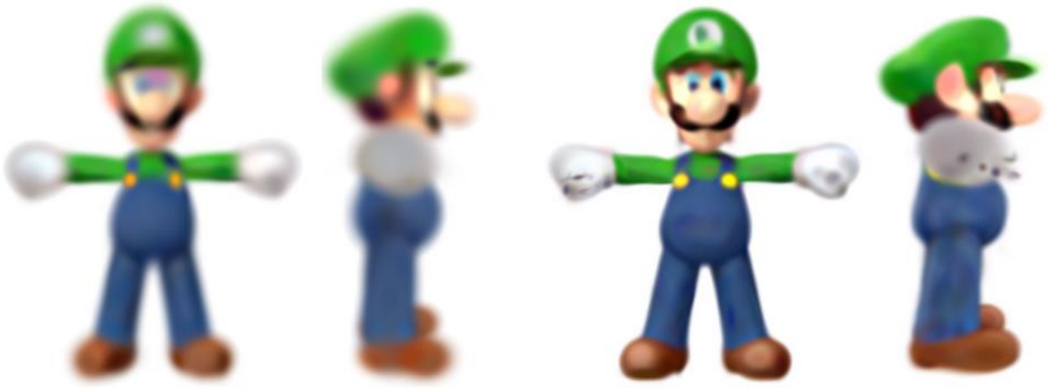
- Gaussian Splatting: **grow** from a small number of Gaussians
- Polygonal Mesh: hard to be optimized from scratch, but good as a second stage.



Two-stage optimization

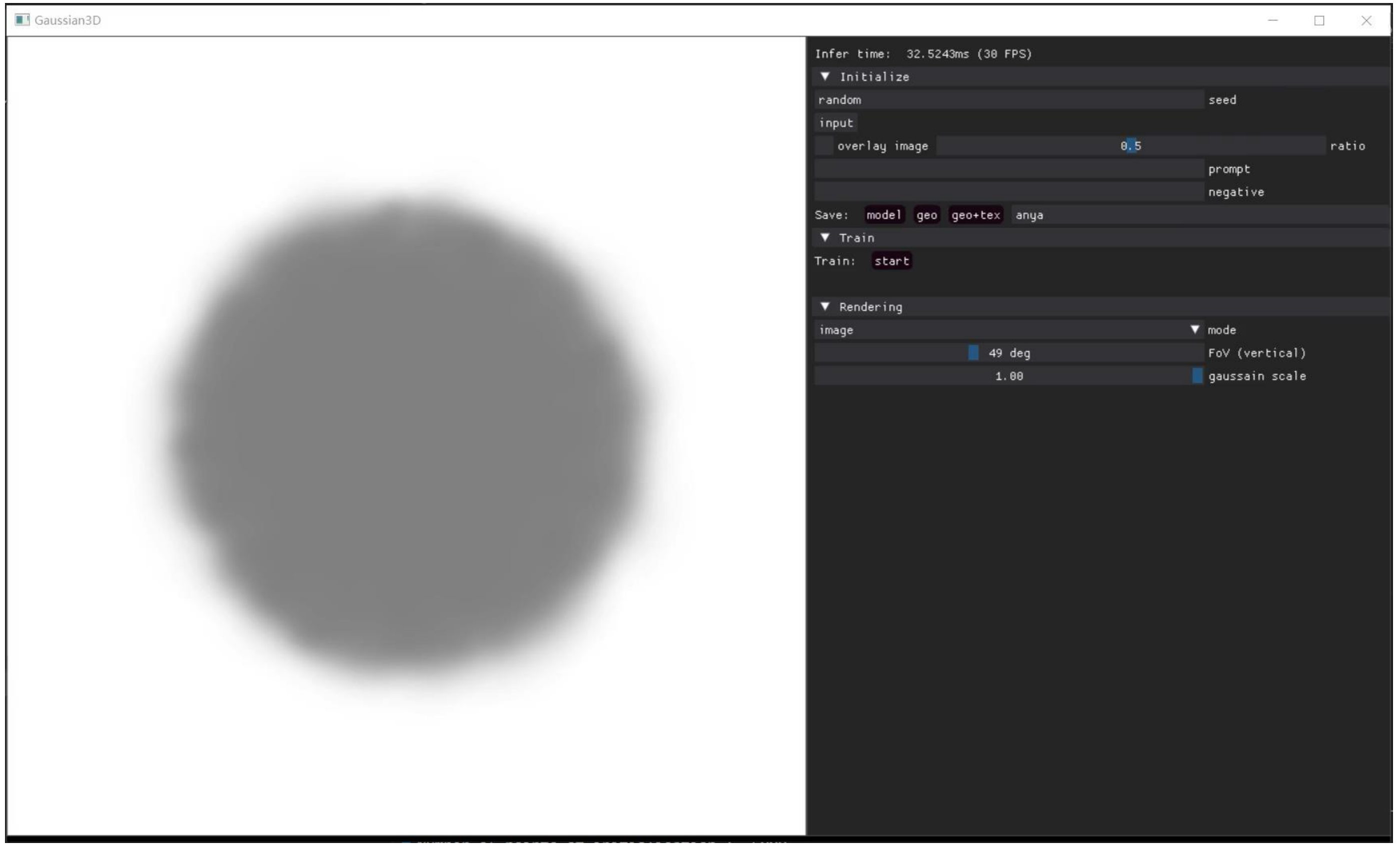
DreamGaussian: Method

Densification is important for details



w/o Densification
5000 Gaussians

Stage 1 Full Setting
~10000 Gaussians

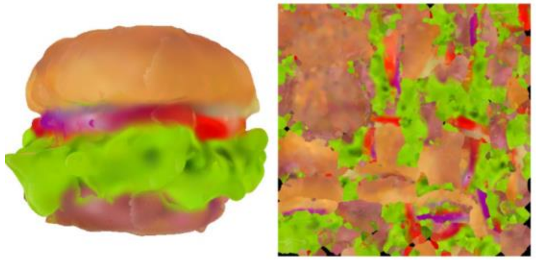


We start from a small number (e.g., 5000) of Gaussians so the shape can initialize very fast.

During training, we gradually add more Gaussians to generate details.

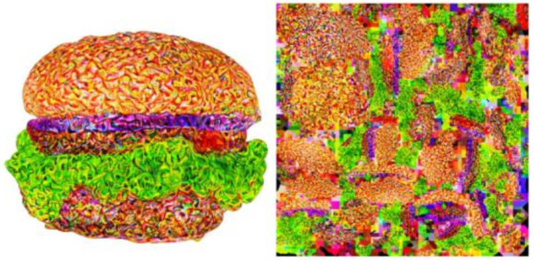
DreamGaussian: Method

Optimize the UV-space mesh texture



Stage 1

We adopt a mesh stage to optimize the UV-space texture.



Stage 2 (SDS)

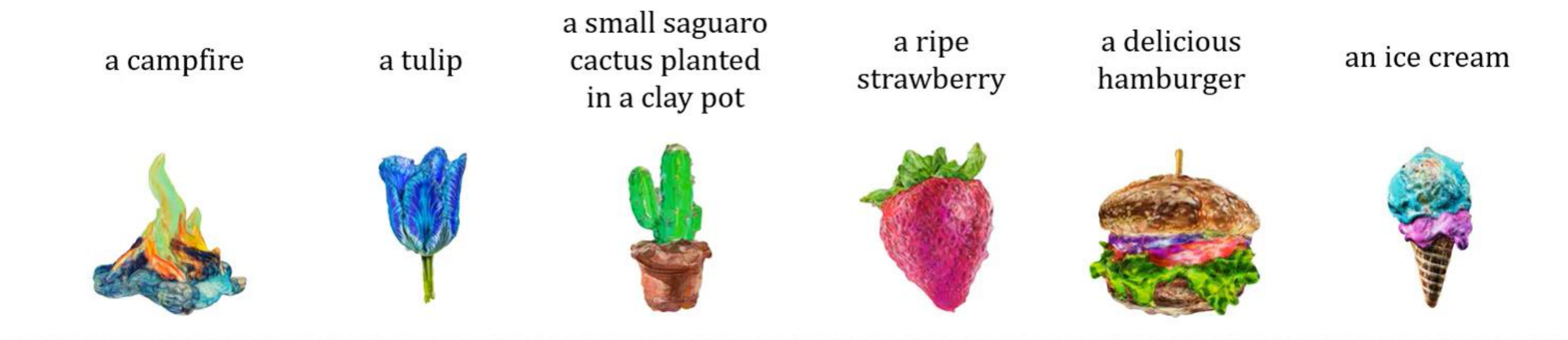
SDS is ambiguous/stochastic at the denoising direction

Stage 2 (MSE)

We learn from SDEdit (image-to-image) to enhance the details by using multi-step optimization.

DreamGaussian: Text-to-3D Results

Text-to-3D

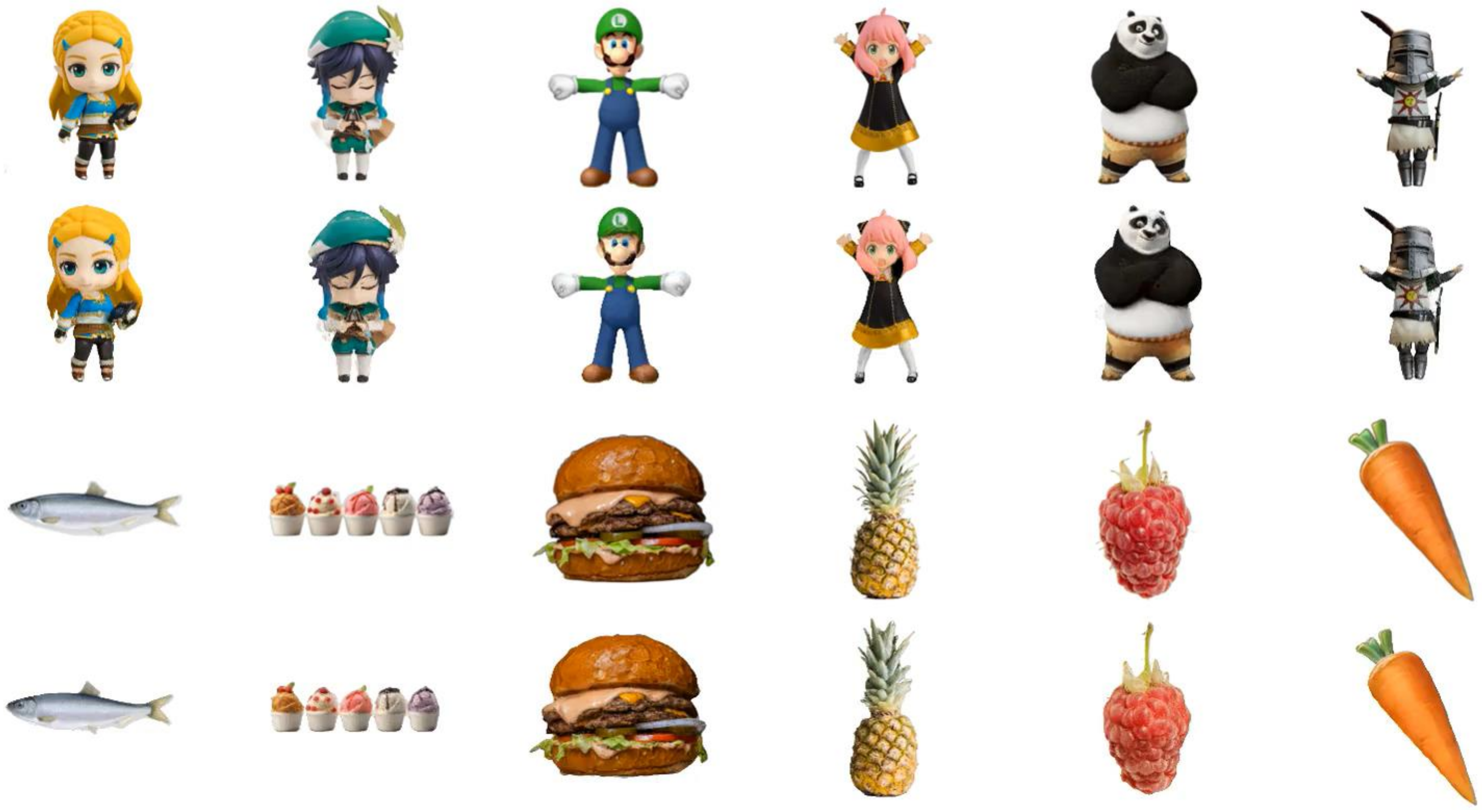


Both stages take about 2.5 minutes to converge (due to a larger resolution of SD).

Very terrible Janus Problem, maybe due to the fast convergence.

DreamGaussian: Image-to-3D Results

Image-to-3D



Both stages take about 1 minute to converge.

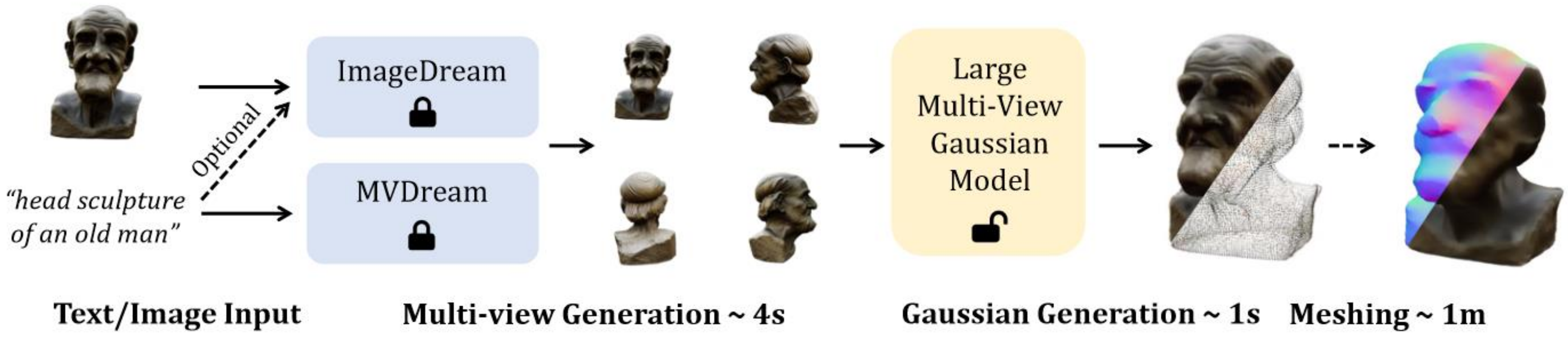
Back-view is still blurry compared to front-view.

LGM: Motivation & Overview

LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation

Motivation:

- ✓ LRM (32 x 32 triplane-based, limited resolution) + Gaussian Splatting for higher fidelity
- ✓ Multi-view settings could achieve 3D generation with higher quality



LGM: Method

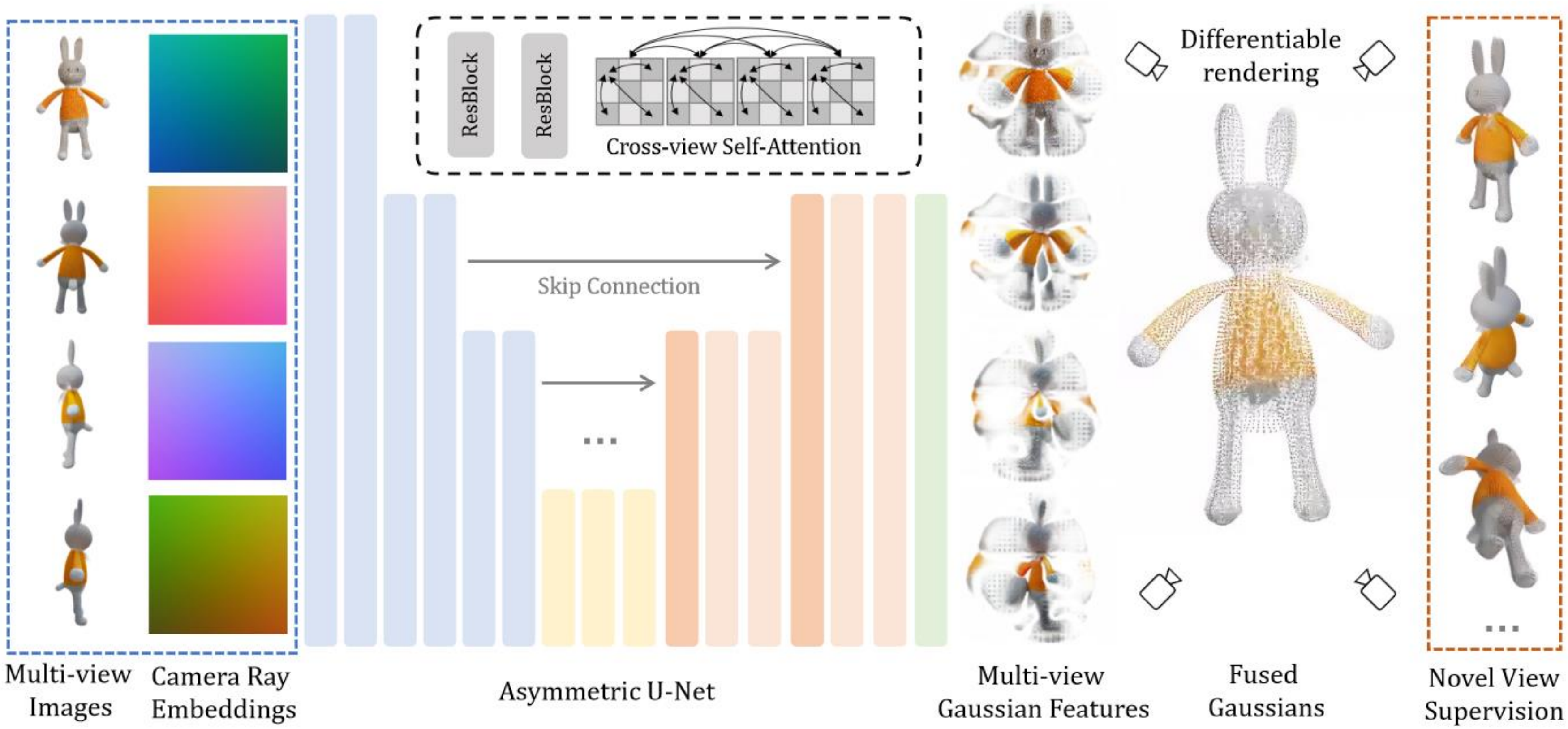
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

1. Fused multi-view Gaussian features

   Cross-View Self-Attention

2. Asymmetric U-Net

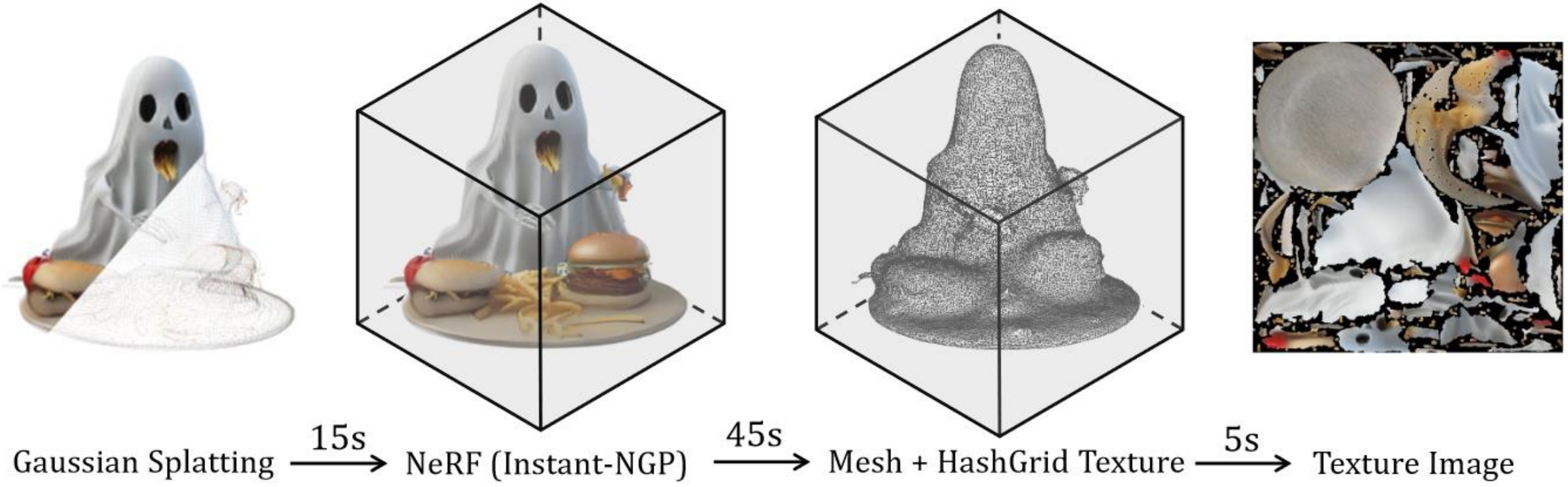
   high-resolution in training (e.g., 512x512)



LGM: Method

3. Meshing

Better than "density field + marching cube"



LGM: Image-to-3D

LGM: Text-to-3D

"motorcycle"

"mech suit"

"ghost lantern"

"furry fox head"

"dresser"

"swivel chair"

"astronaut"

"mushroom house"

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

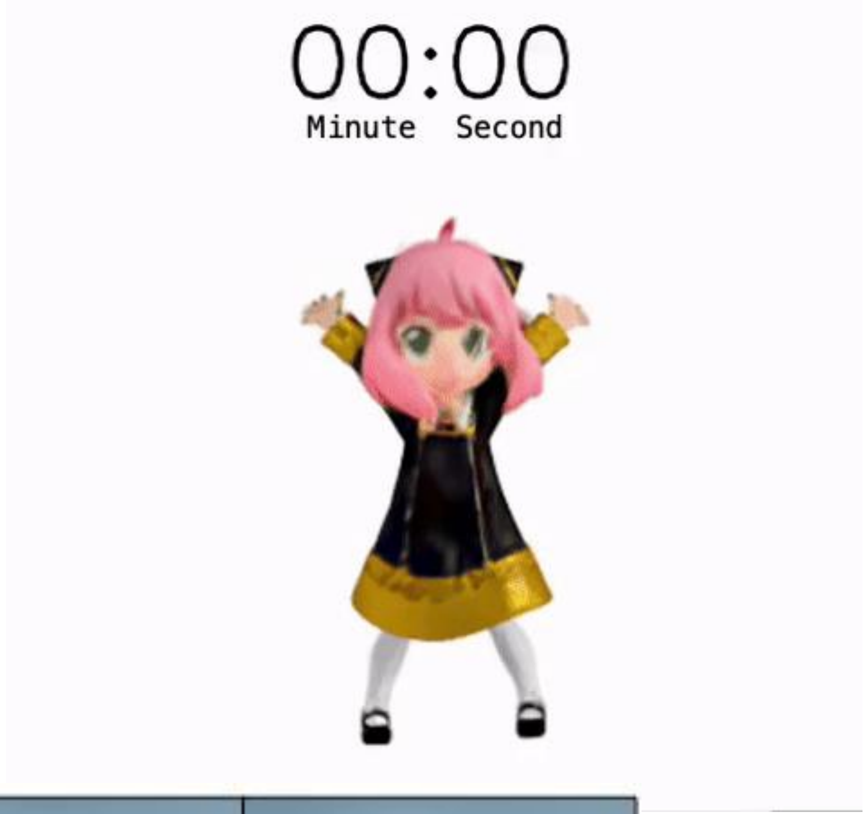
4D Generative Models

DG4D：Motivation & Overview

DreamGaussian4D: Generative 4d gaussian splatting

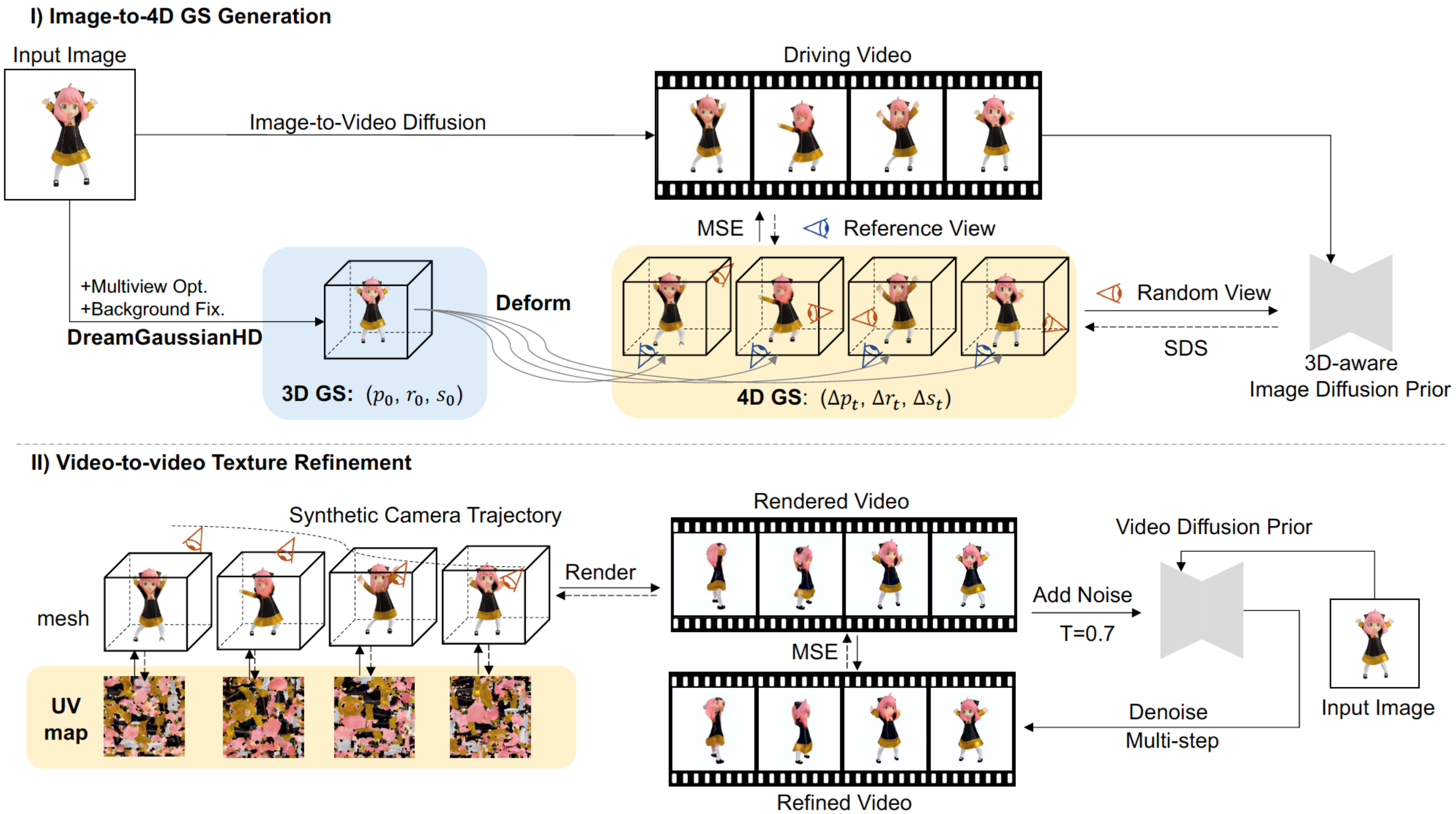
Motivation:

- ✓ **3D Prior** + **Video Prior** for diverse and high-quality 4D generation
- ✓ 3D Gaussian Kernels could be well-suited for dynamic 4D
- ✓ Mesh Sequences could be well-optimized by using video priors
- ✓ The whole 4D generation could be resolved in **5 min**

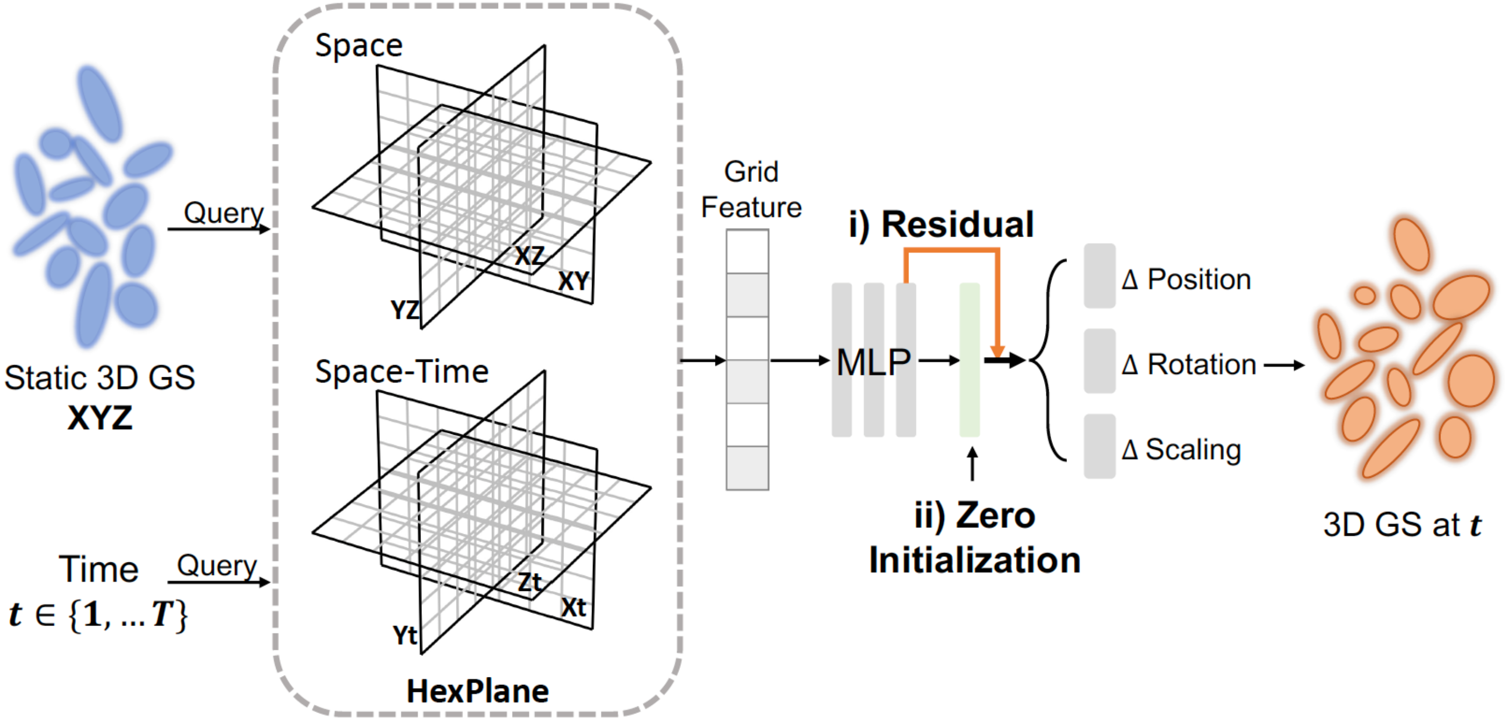


Composited Scene

DG4D：Method Overview



DG4D：Motion Representation



1. Initialization: Static 3D Gaussian

   Image-to-3D Generation based on 3DGS

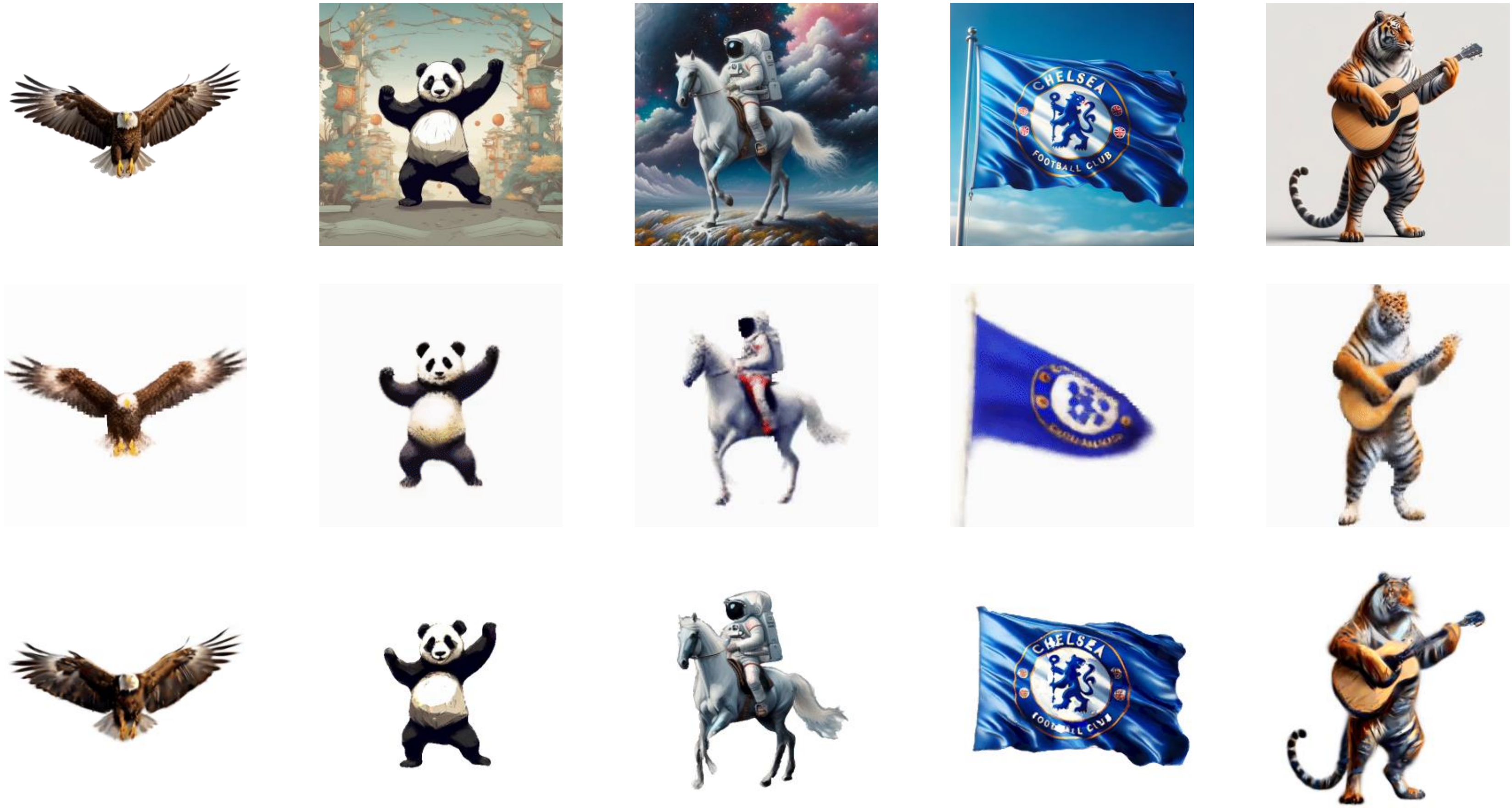
2. Dynamics: HexPlane Optimization

   Deformation Field for 4D Generation

3. Refinement: Video Prior

   Video-to-Video Refinement

DG4D：Results



	Gen. Time
Input Image	
Animate124	>> 8 hours
DG4D (Ours)	<5mins

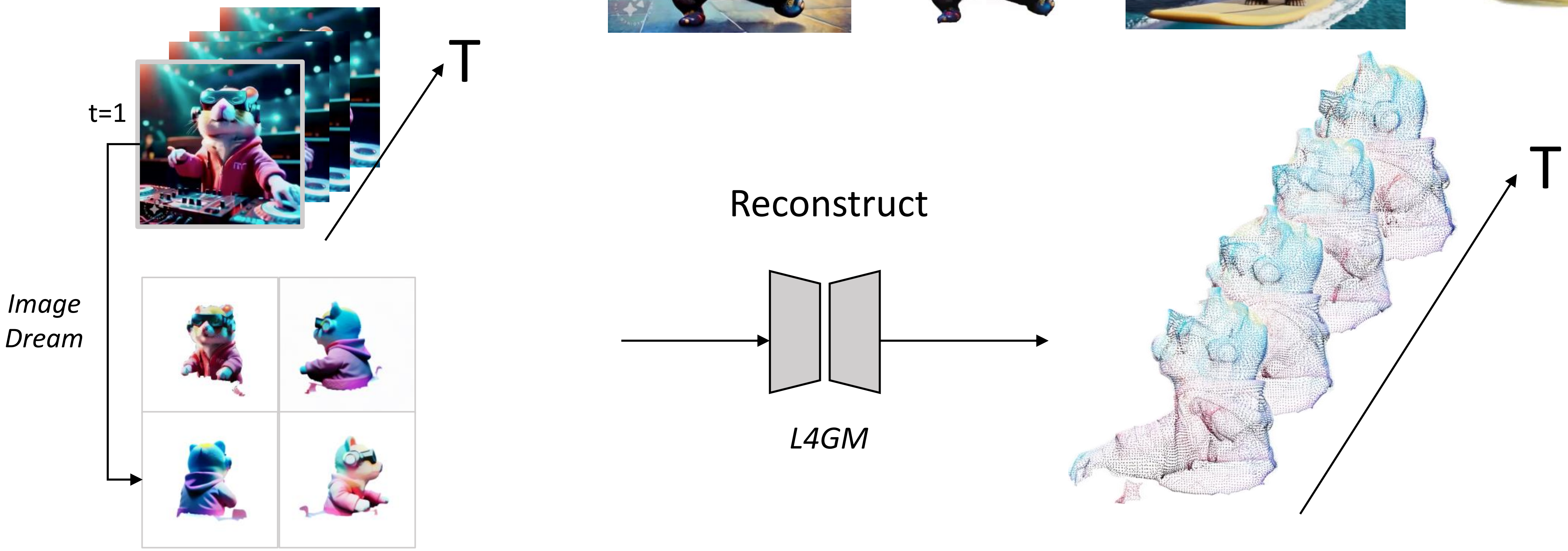
L4GM: Motivation & Overview

Motivation:

Feedforward 3D Generation + **Video Prior** for diverse and high-quality 4D generation

L4GM: Large 4D Gaussian Reconstruction Model

Input 1: video frames



Input 2: multiview images at t=1

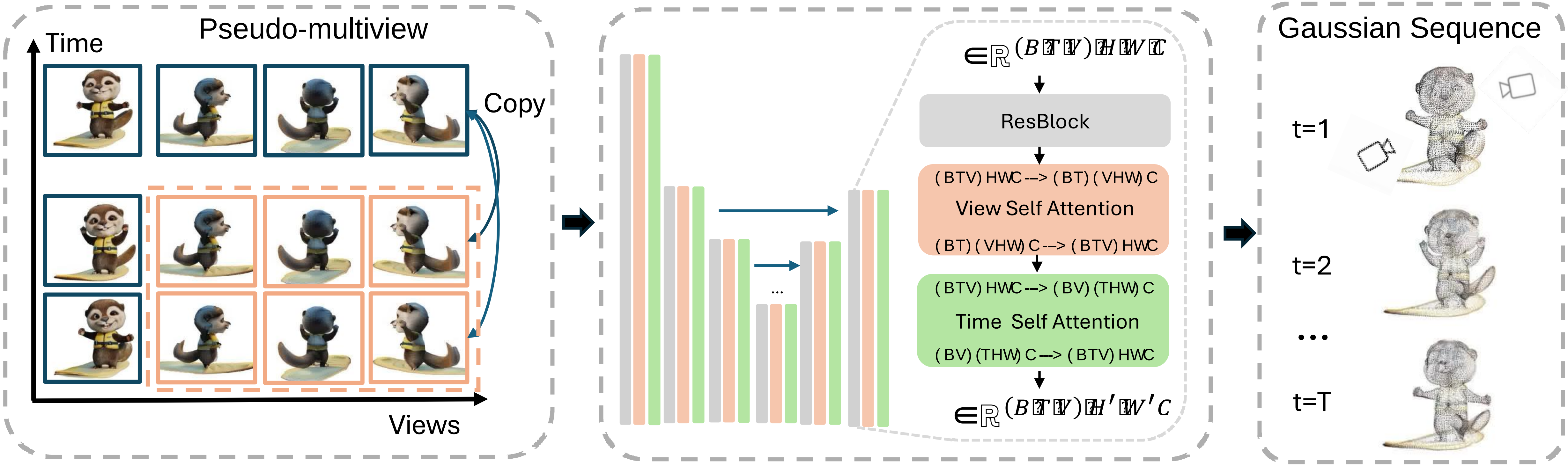
Output: a sequence of 3DGS

L4GM: Objaverse-4D dataset

Animated objects in *Objaverse-1.0*

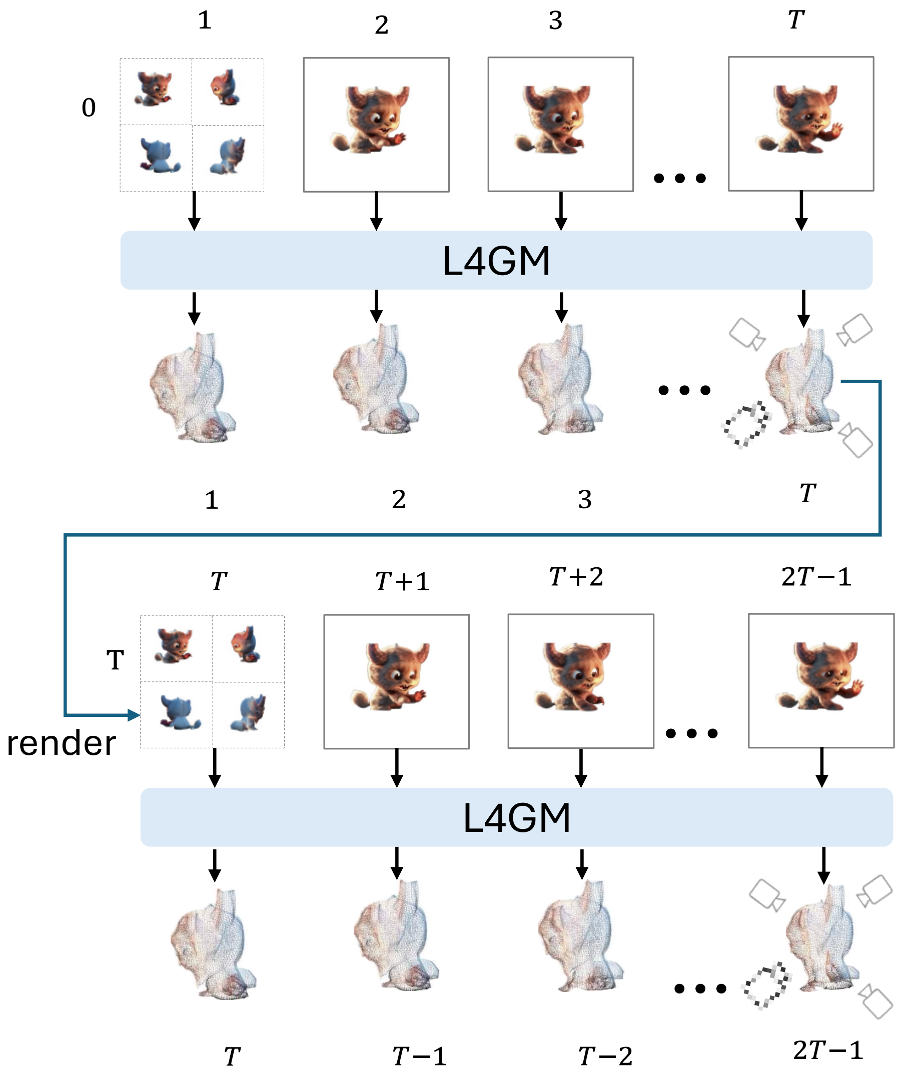
- 44K objects (5%) / 110K animations / 500K 3D frames
- After filtering, rendered in 48 views: 300M images

L4GM: Method



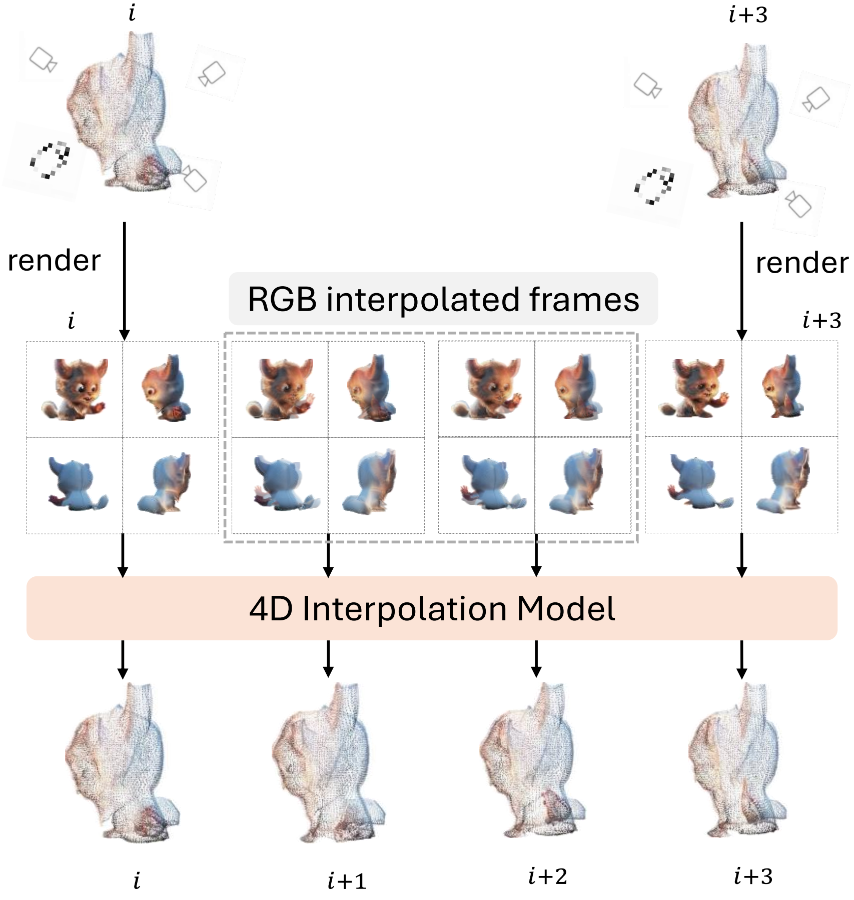
- Fully exploit the LGM pretrain
- Simple design for better scalaibility

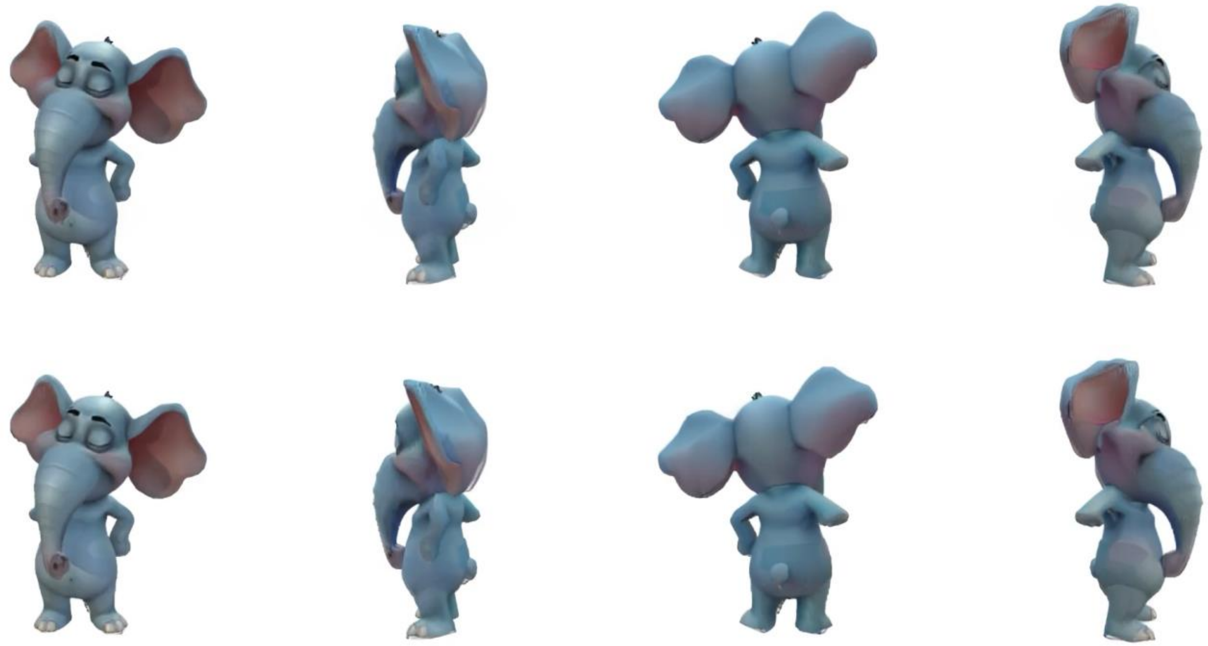
L4GM: Autoregressive reconstruction for long videos



00:00

L4GM: 4D Interpolation model





L4GM: 4D Results

Input Video

L4GM (ours)

Runtime: 7.0 seconds

DG4D

Runtime: ~10 minutes

STAG4D

Runtime: ~2 hrs

3D Generation Evaluation

GPT4Eval: Motivation & Overview

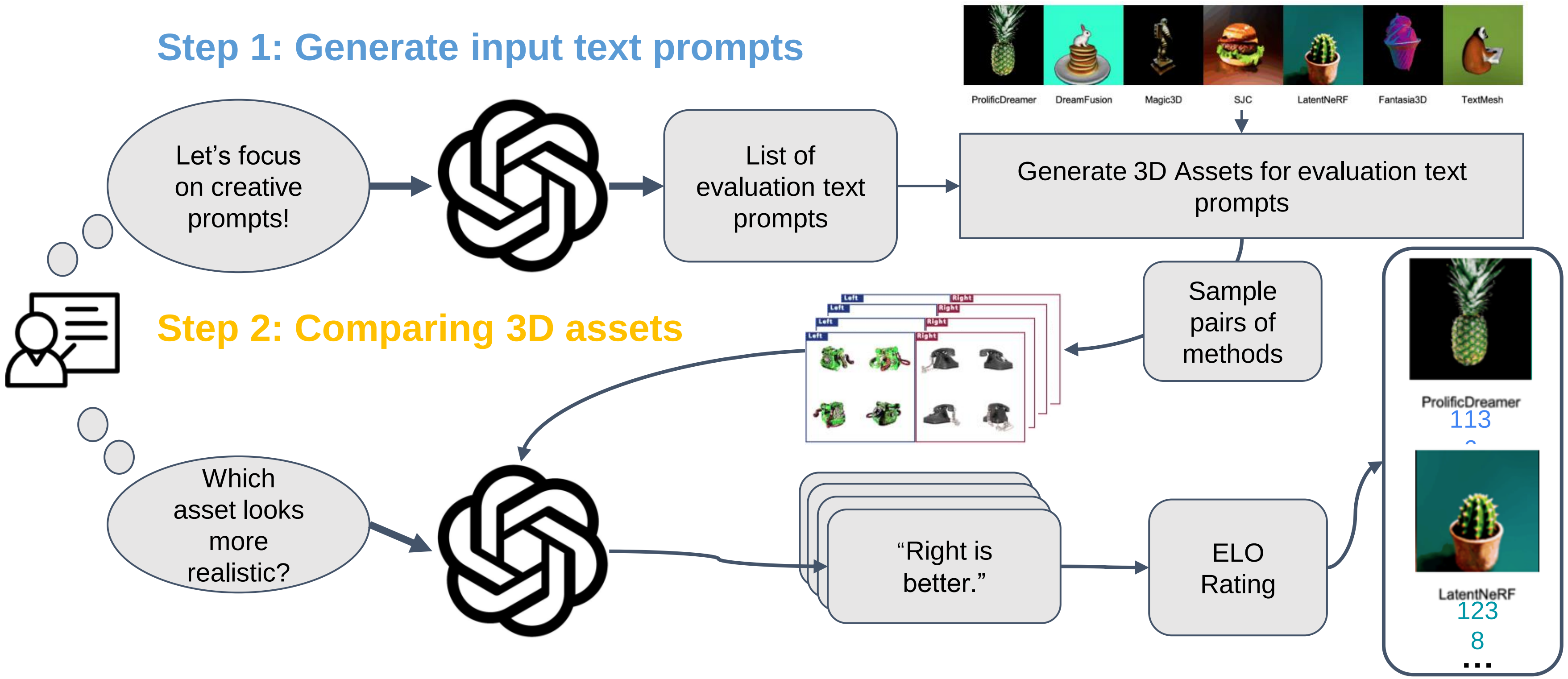
GPT4Eval: GPT-4V(ision) is a Human-Aligned Evaluator for Text-to-3D Generation
[CVPR 2024]

Text-to-3D has gained increasing attention!

But **evaluation metric** has lagged behind...

How do we achieve it?

Use GPT-4V!



GPT4Eval: Input-text prompt generation pipeline

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Subject

Living Beings, Plants, Buildings and Structures, Vehicles, Furniture, Electronics, Household Items, Clothing and Accessories, Abstract Objects, Natural Elements, Food and Beverages

Appearance

- **Color:** Specific color, patterns, gradients, ...
- **Materials:** Wood, metal, glass, fabric, stone, ...
- **Textures:** Smooth, rough, furry, scaly, ...
- **Finish:** Glossy, matte, translucent, opaque, ...
- **Size:** Small, medium, large, specific dimensions, ...
- **State:** New, old, worn, pristine, ...

Geometry

- **Volume:** hollow, solid, porous, or layered, ...
- **Symmetry:** symmetrical, asymmetrical, or radially symmetrical, ...
- **Contours:** smooth, jagged, irregular, or undulating, ...
- **Internal Structures:** empty, compartmentalized, or multi-layered, ...
- **Shape:** cone, cylinder, sphere, ...

Status

- **Static:** Still, motionless, ...
- **Dynamic:** Moving, changing, ...
- **Emotional State:** Happy, sad, angry, ...
- **Physical State:** Broken, intact, in use, ...
- **Interaction:** Interacting with another object or environment, ...

Scene

- **Environment:** Indoor, outdoor, urban, rural, natural, fantastical, ...
- **Context:** Part of a larger scene, event, or story, ...
- **Lighting:** Day, night, artificial, natural, shadows, highlights, ...
- **Weather:** Sunny, rainy, cloudy, stormy, ...
- **Scale:** The relative size of the object in the scene, ...

Style

- **Aesthetic:** Minimalistic, ornate, modern, vintage, ...
- **Cultural:** Asian, African, Western, Middle Eastern, ...
- **Emotional:** Cheerful, gloomy, energetic, calm, ...
- **Functional:** Practical, decorative, symbolic, ...
- **Conceptual:** Abstract, realistic, surrealistic, impressionistic, ...

GPT4Eval: How to formulate the input?

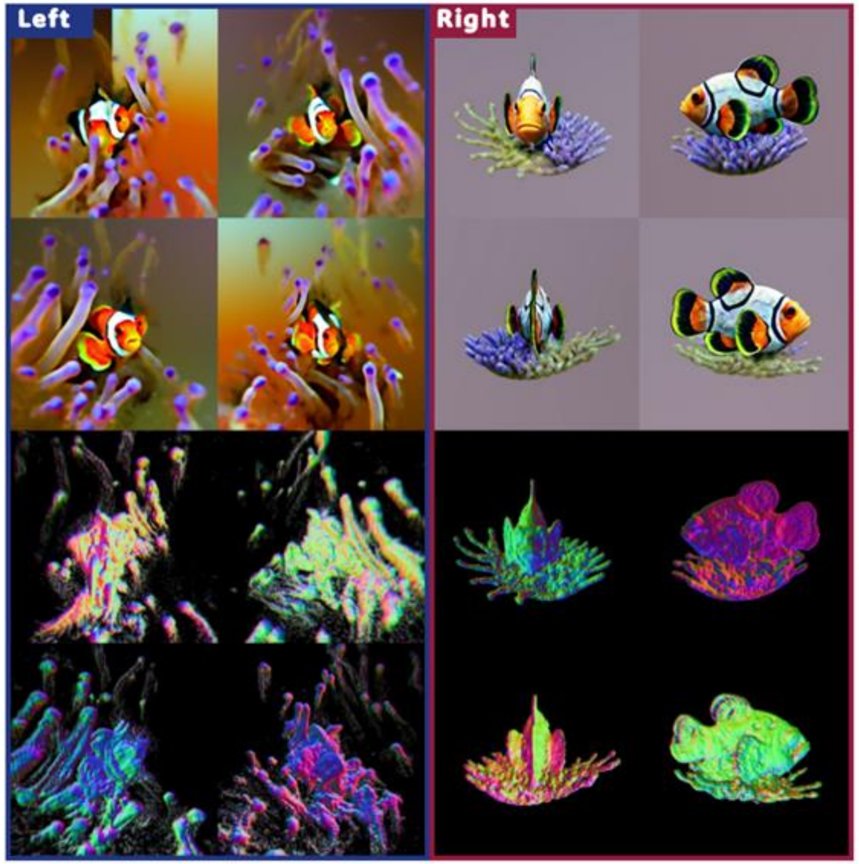
1. **Multiple views** provide a full-view 3D perception

GPT-4V Caption: Intricately detailed steampunk apparatus, primarily of mechanical design nature, appearing three-dimensional. With worn metallic and glassy texture. Showcasing a central clock face and multiple gauges, and accentuated by pipes, gears, and levers. Crafted mainly from aged bronze and accented with glass and wood. Intended for time display and possible atmospheric measurements, and is static. Exhibiting a Victorian steampunk style, set in an industrial workshop environment with a nostalgic and inventive mood & atmosphere.

GPT-4V Caption: Detailed potted plant on a rugged terrain, primarily of organic and naturalistic structure, appearing full and lifelike. Showcasing a vibrant green plant with yellow flowers and accompanied by smaller pink blossoms, and accentuated by a scattering of pebbles and rocks. Crafted mainly from digital textures mimicking natural materials and accented with subtle shading. Intended for environmental visualization and is static. Exhibiting a contemporary and natural aesthetic, set in an outdoor-like setting with a serene and peaceful atmosphere.

2. **Paired comparison** alleviates the ambiguity inherent in an absolute score

GPT4Eval: Example outputs



"Clownfish peeking out from sea anemone tendrils."

Text-Asset Alignment: The right model shows ... **without any tendrils obstructing it**. The left one ... and seems **more consistent with the "peeking out" aspect**.

3D Plausibility: The left model's fish appears **distorted and blended with the anemone tendrils**, while the right model depicts both the fish and anemone as **distinct and solid entities**, being more plausible in the real world.

Texture-Geometry Coherency: The left object is less compelling due to the **less integrated positioning of the fish**. The right object shows a strong correspondence between the geometry and the texture; **the anemone tendrils and clownfish stripes align well across both the RGB and normal maps**.

Texture Details: The texture on the left has **more blur and less sharpness**. The right clownfish presents with **sharper, clearer textures and distinct patterns**.

Geometry Details: When observing the local geometry of the normal maps, the right object exhibits **sharper details and more defined structures**. It is particularly evident within the tendrils of the anemone and the body of the clownfish, where **individual scales and tentacle textures appear more pronounced**.

Final answer: left right right right right

GPT4Eval: Elo System

Problem: **Comparison is noisy.** Same match can lead to different results.

E.g. {(A > B), (A > C), (B > C), (B > A), (A > B), …}

Our solution: use **Elo system** from chess

Elo score's assumption:

- each player has a **unique score** quantifying the performance
- When two players compete, we draw a number from Gaussian whose mean equals their score; the player who got a larger number wins.

$$\Pr(\text{“}i\text{ beats }j\text{”}) = \left(1 + 10^{(\sigma_j - \sigma_i)/400}\right)^{-1}$$

GPT4Eval: Experiments

Our metric is human-alignment across criteria

Methods	Alignment	Plausibility	Color-Geo	Texture	Geometry	Average
PickScore [34]	0.667	<u>0.484</u>	0.458	0.510	0.588	0.562
CLIP-S [23]	0.718	0.282	0.487	0.641	0.667	0.568
CLIP-E [23]	<u>0.813</u>	0.426	**0.581**	0.529	0.658	0.628
Aesthetic-S [58]	0.795	0.410	<u>0.564</u>	0.769	<u>0.744</u>	<u>0.671</u>
Aesthetic-E [58]	0.684	0.297	0.555	<u>0.813</u>	0.684	0.611
Ours	**0.821**	**0.641**	<u>0.564</u>	**0.821**	**0.795**	**0.710**

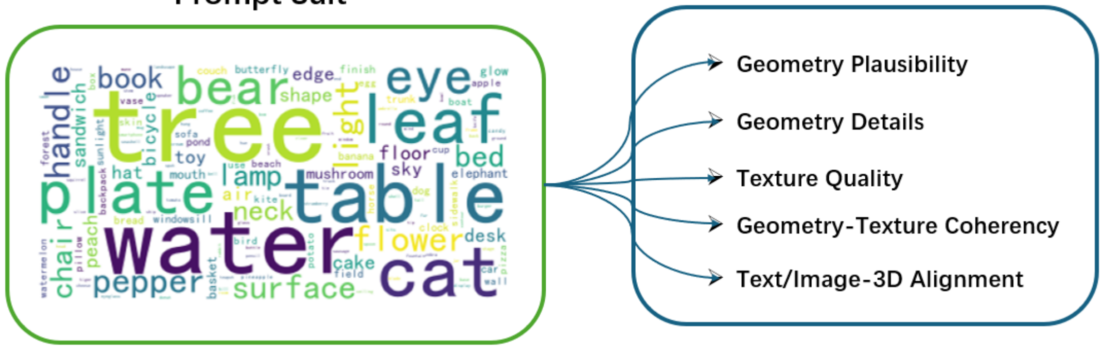
Kendall's Tau between metric predicted ranking and expert predicted ranking (higher is better).

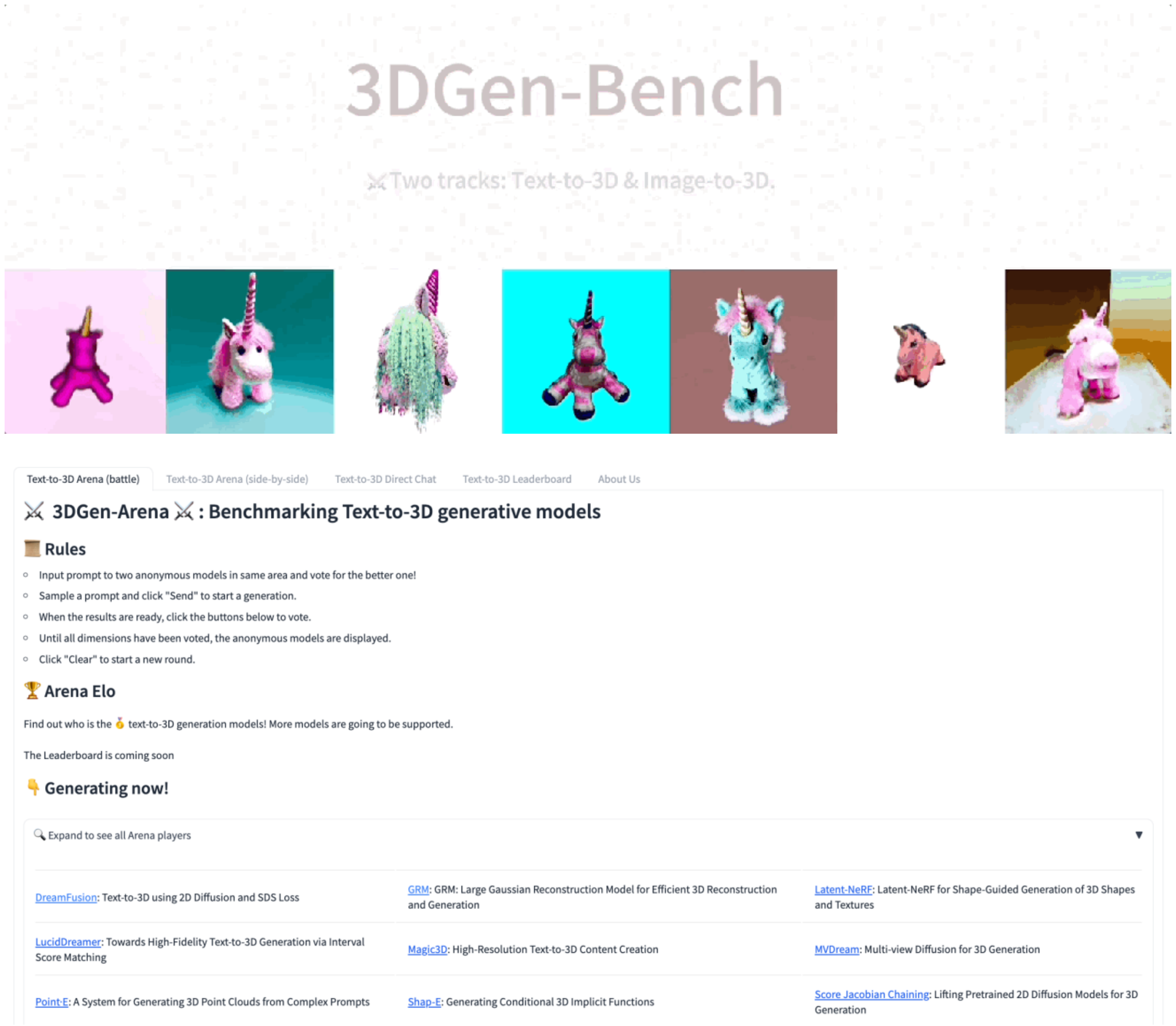
Our metrics **reach top-2 alignment in all six criteria** while existing automatic metrics are usually good at one or two.

3DGen-Bench: 3D Generation Human Preference Arena

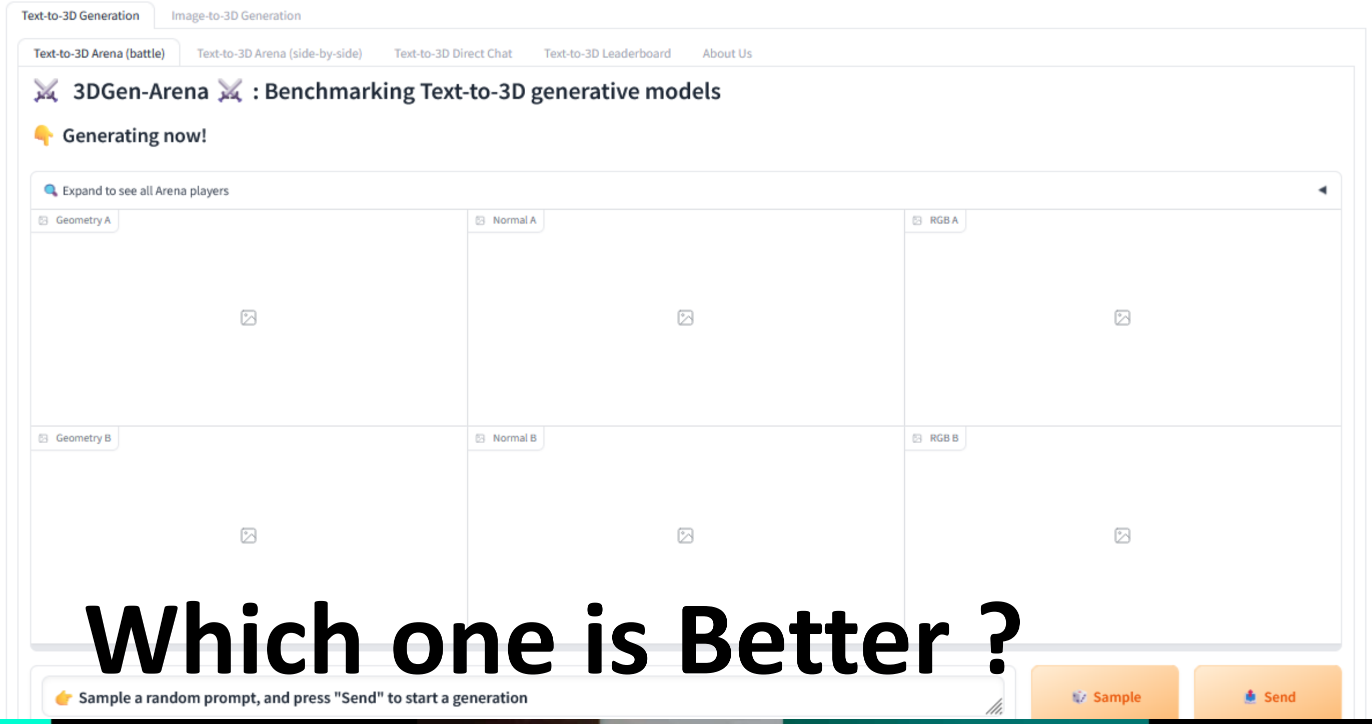
- **2** Tracks
- **1k+** prompts
- **10k+** generated models
- **60k** paired data

Prompt Suit





3DGen-Bench: 3D Generation Human Preference Arena



Which one is Better ?

ProlificDreamer | Dreamfusion | Magic3D | SJC | LatentNeRF | Fantasia3D | MVDream

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Thank you!