Microsoft

Structured 3D Latents
for Scalable and Versatile 3D Generation

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang,
Bowen Zhang, Dong Chen, Xin Tong, Jiaolong Yang

Motivation

Large image generation models have enabled ready-to-use tools that exert a profound impact on today's digital industry

… +

⋮

Motivation

Unlike 2D images which is typically represented by pixel grids, **3D data has diverse representations**.

Surface Representations (Occupancy, SDF etc.)

- ***Pros:*** Good shape and geometry; Usable in current industry.
- ***Cons:*** Falter in appearance detail

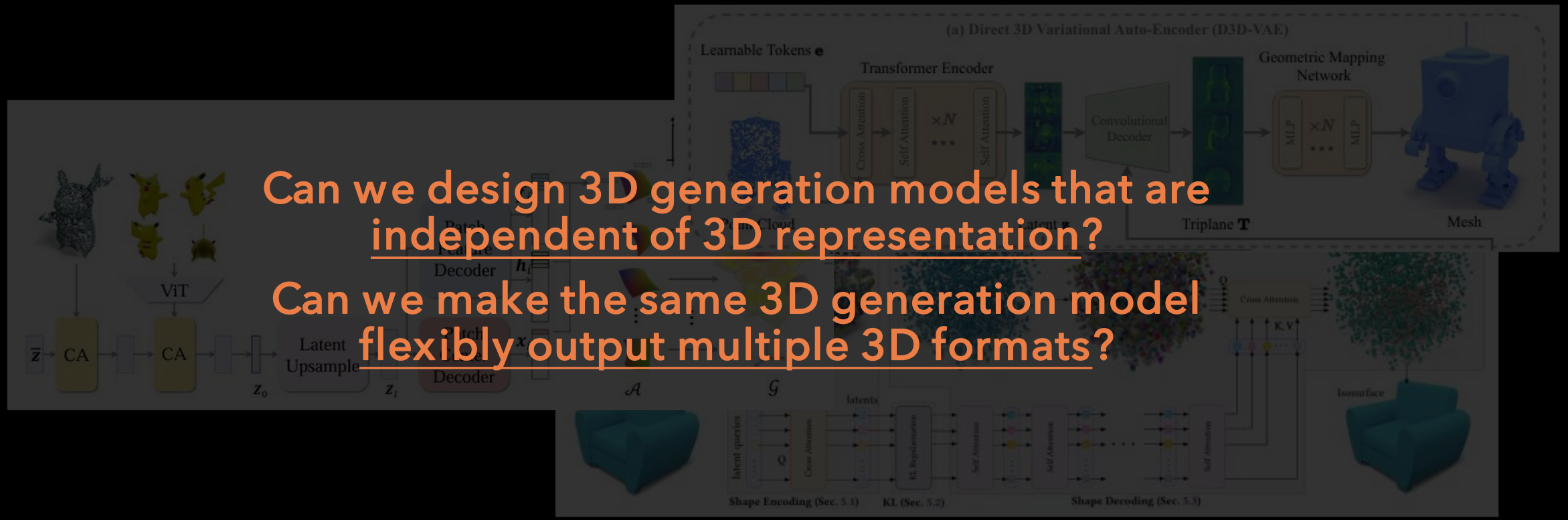
Volumetric Representations (Radiance Fields, 3DGS etc.)

- ***Pros:*** High-quality appearances (thin structures, translucent…)
- ***Cons:*** Struggle with plausible geometry extraction

*Image taken from *OctFusion & GaussianCube*

Motivation

The unique characteristics of different representations may result in different designs for different representations.



Can we design 3D generation models that are independent of 3D representation?

Can we make the same 3D generation model flexibly output multiple 3D formats?

*Image taken from *3DShape2VecSet*, *AtlasGS* and *Direct3D*

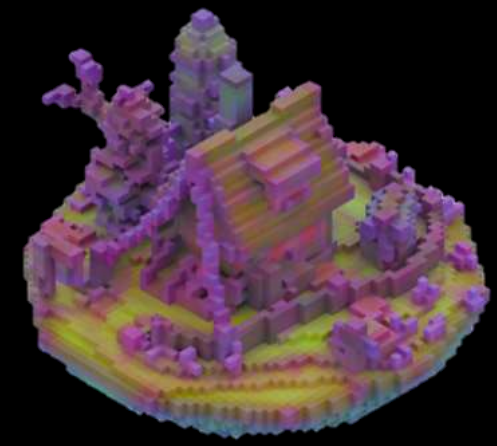
SLAT: Structured Latents

Microsoft

Explicit 3D grid structure with sparsity

We encode the 3D asset as latents within active voxels from regular 3D grid that intersected with its surfaces. The benefits are three-fold:

- **Versatility**: Supporting different 3D representations
- **Locality**: Easy to model; Easy to edit
- **Efficiency**: Fast runtime and improved fidelity



Structured Latents

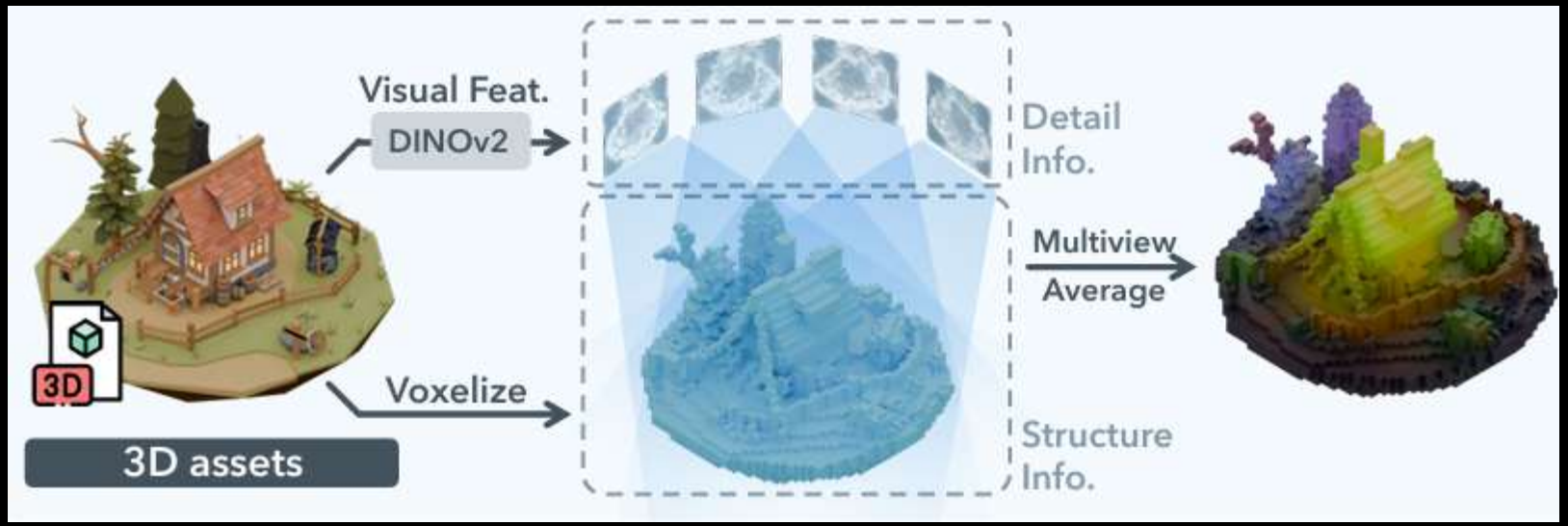
(64^3 grid)

SLAT: Structured Latents

Microsoft

Integrating powerful vision foundation models

Marry sparse structure with powerful **visual representations** to encapsulate both comprehensive geometry and appearance –– bypassing the need for a dedicated 3D encoder and eliminating the costly pre-fitting process.

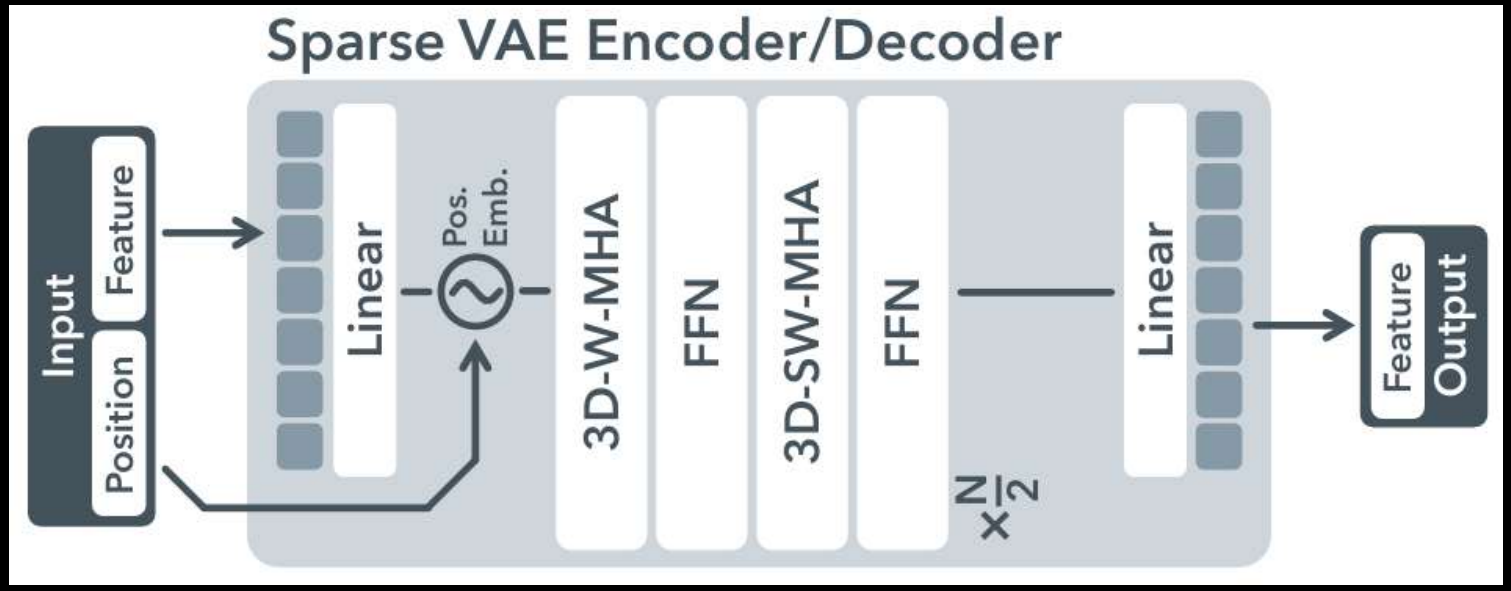


SLAT: Structured Latents

Microsoft

Encoder and Decoder(s):

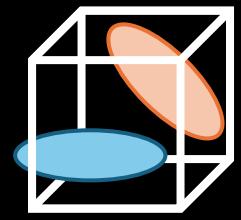
- Same overall architecture with minor differences in output heads
- Sparse Transformer with Shifted Window Attention



SLAT: Structured Latents

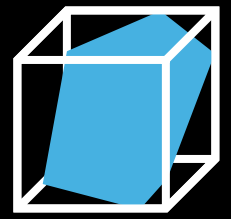
Different decoders:

- **3D Gaussians**
  Regress multiple 3D Gaussians for each active voxel



- **Radiance Fields**
  Represented as CP-decomposed vectors (trivecs)

- **Meshes**
  Appending two sparse con. upsamplers in the end
  Output SDF (Scalar Field) and interpolation params in FlexiCubes



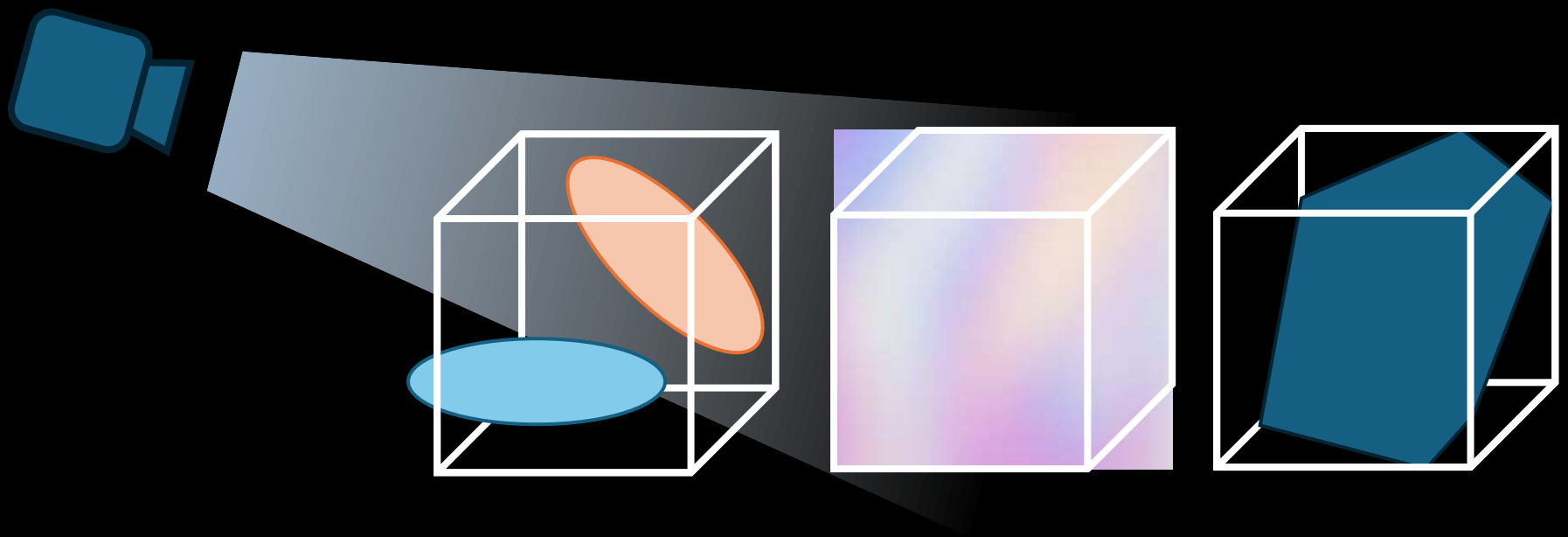
- Others ….

SLAT: Structured Latents

Microsoft

Training: Efficient image-space supervisions

- For 3DGS and RF: color losses (L1, SSIM, LPIPS) on rendered images
- For mesh: both color and geometry (normal & depth) losses on rendered views

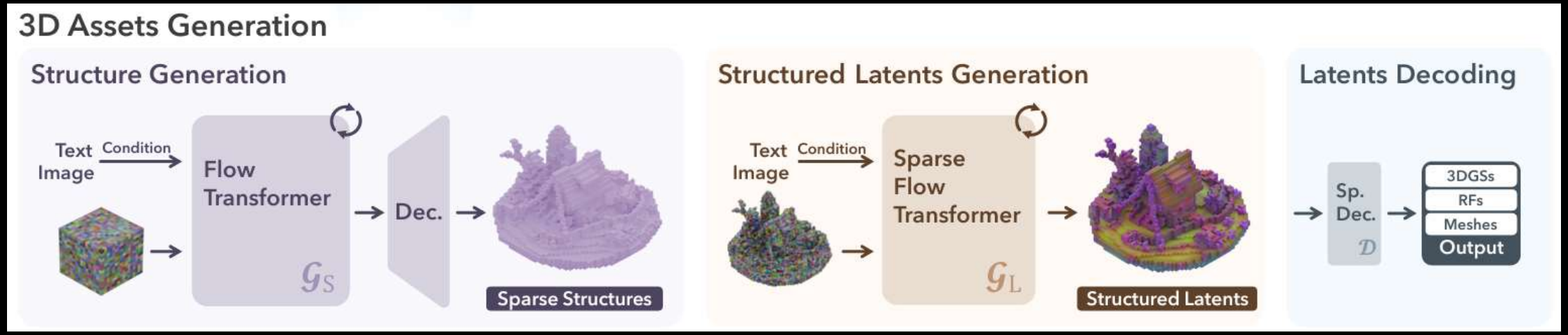


TRELLIS: Generative models with SLAT

Microsoft

Two-stage generation pipeline

- First sparse structures, then latents for non-empty cell

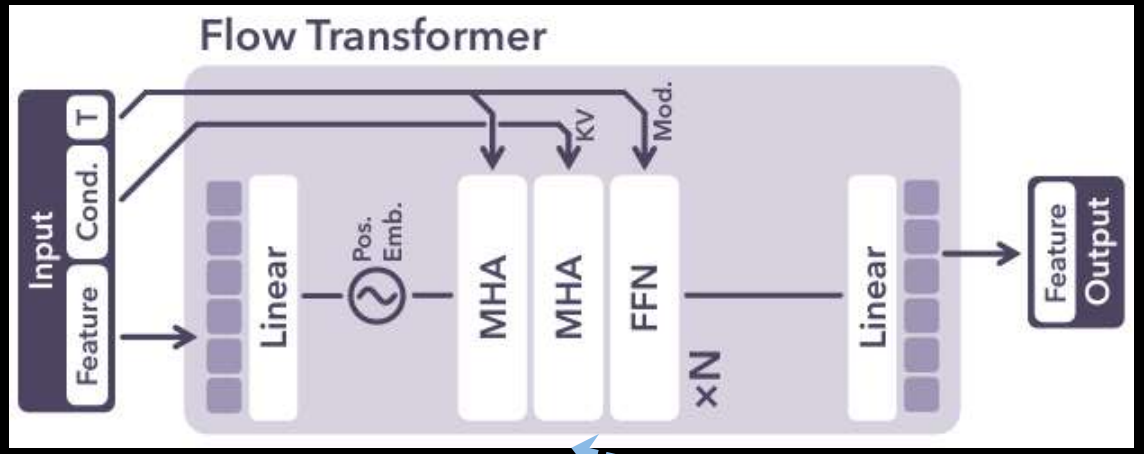


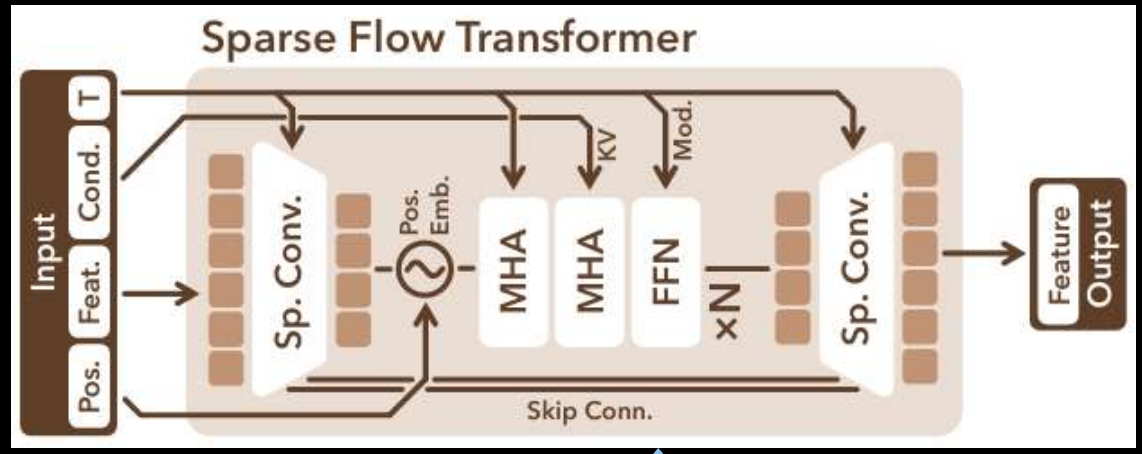
TRELLIS: Generative models with SLAT

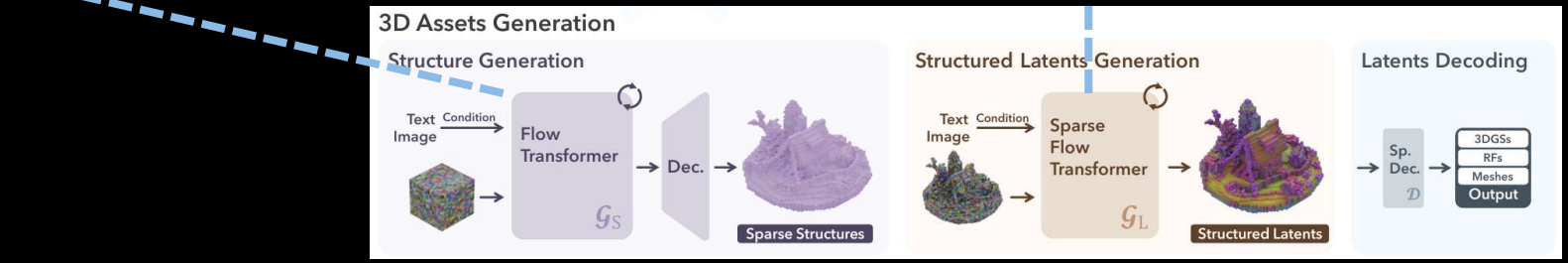
Microsoft

Standard DiT, with sparsity handling

- Positional Encoding for spatial information
- Sparse Conv packing for efficiency







Dataset

We collected and filtered 3D assets from various open-sourced datasets, resulting in ~500K high quality 3D assets used for training.

Source	Aesthetic Score Threshold	Filtered Size

Da

We

rip

Please describe the details of the 3D object, the detailed description will be used for a text to 3d model to generate this 3D object. Please provide details of the shape, color of each part, avoid imagination and solve it step by step. You do not need to describe the background.

<raw_captions>: The 3D object appears to be a stylized, medieval knight character standing in a T-pose, typically used in character modeling for animation and game design. Here is a detailed description of the character's features:

1. **Helmet**:
   - **Shape**: Bucket-style helmet with a flat top and a slight flare at the bottom.
   - **Color**: Silver/gray with signs of wear and tear.

2. **Upper Body**:
   - **Chest Armor**: Breastplate that covers the torso.
   - **Colors**:
     - Primary: Red fabric with a blue and green emblem at the center.
     - Secondary: Silver shoulder pads on both sides.
   - **Additional Details**: Textured chainmail visible on the arms.

3. **Arms**:
   - **Upper Arms**: Covered in a combination of chainmail and fabric, resembling a tunic underneath the armor.
   - **Forearms**: Encased in silver bracers that match the helmet and shoulder pads.
   - **Hand Position**: Both hands are extended outward in a T-pose.

4. **Lower Body**:
   - **Tunic**:
     - **Shape**: Red fabric tunic extending from the waist to the mid-thigh.
     - **Color**: Same as the upper tunic with a visible blue vertical stripe on the sides.
   - **Leg Armor**:
     - **Shape**: Iron greaves over the calves.
     - **Color**: The same silver/gray as the helmet and bracers.
   - **Footwear**:
     - **Shape**: Traditional medieval-style boots with straps and buckles.
     - **Color**: Dark brown.

5. **Belt**:
   - **Shape**: Simple, broad belt wrapped around the waist.
   - **Color**: Brown with a silver buckle.

These details should help in using a text-to-3D model to accurately recreate the knight character. The object is stylized, likely for a game or animation, with exaggerated features and simplified textures.

This is a hard problem. Carefully summarize in ONE caption aiming for **no more than 40 words** based on the following captions (possibly incorrect) by people describing the 3D object. The caption will be used for a text to 3D model to generate this 3D object. Ensure the summary is concise and captures the essential information without including any additional commentary or unnecessary details. Please avoid hallucination. Raw captions: <raw_captions>. Please directly output the detailed caption without extra text. Provide the caption in a simple, plain text format with no markdown, bullet points, or special formatting.

<detailed_captions>: Stylized medieval knight in T-pose with a worn silver bucket helmet, red tunic and chest armor with blue and green emblem, textured chainmail, silver shoulder pads, bracers and greaves, dark brown medieval boots, and a brown waist belt.

This is a hard problem. Carefully summarize the given caption describing the 3D object. The caption will be used for a text to 3D model to generate this 3D object. Ensure the summary is concise and captures the essential information without including any additional commentary or unnecessary details. Please avoid hallucination. Please give me TEN versions of the caption. ONE caption per line. The captions should be more and more brief. The first caption should be the most detailed and the last caption should be the most brief. The first caption should be about 12 words long and the last caption should describe ALL major objects in the scene AS SIMPLY AS POSSIBLE for at most 5 words. Please directly output the caption without extra text. Provide the caption in a simple, plain text format with no markdown, bullet points, numbering, or special formatting. Raw captions: <detailed_captions>

Medieval knight in T-pose, silver armor, red tunic, blue-green emblem, chainmail, and brown boots.
Medieval knight, silver gear, red tunic, blue-green emblem, chainmail.
Knight in T-pose, silver armor, red tunic, emblem, chainmail.
Knight with silver armor, red tunic, and chainmail.
Knight in T-pose, silver gear, and chainmail.
Medieval knight, silver armor, and tunic.
Knight with silver armor and chainmail.
Silver-armored knight with emblem.
Silver-armored medieval knight.
Armored knight.

Dataset & captions are released; see our GitHub repo.

Results | Text to 3D

Rustic lantern with a flickering flame.
(乡村风格的灯笼，火焰闪烁)

Carved wooden chess piece. (queen)
(木制雕花棋子(皇后))

The train carriage has a classic, vintage design with a dark, rounded roof, teal exterior, detailed windows, and red wheels.
(火车车厢采用经典、复古的设计，配有深色圆形车顶、蓝绿色外观、精致的窗户和红色车轮)

A space colony with domed habitats and connecting tunnels.
(一个拥有圆顶栖息地和连接隧道的太空殖民地)

Ceramic mug with a crack.
(有裂纹的陶瓷杯)

A stylized, cartoonish rocket with a red dome top and black antenna, teal cylindrical middle section with red bands and black connectors.

Metallic dog-like robot with articulated legs and futuristic design elements.

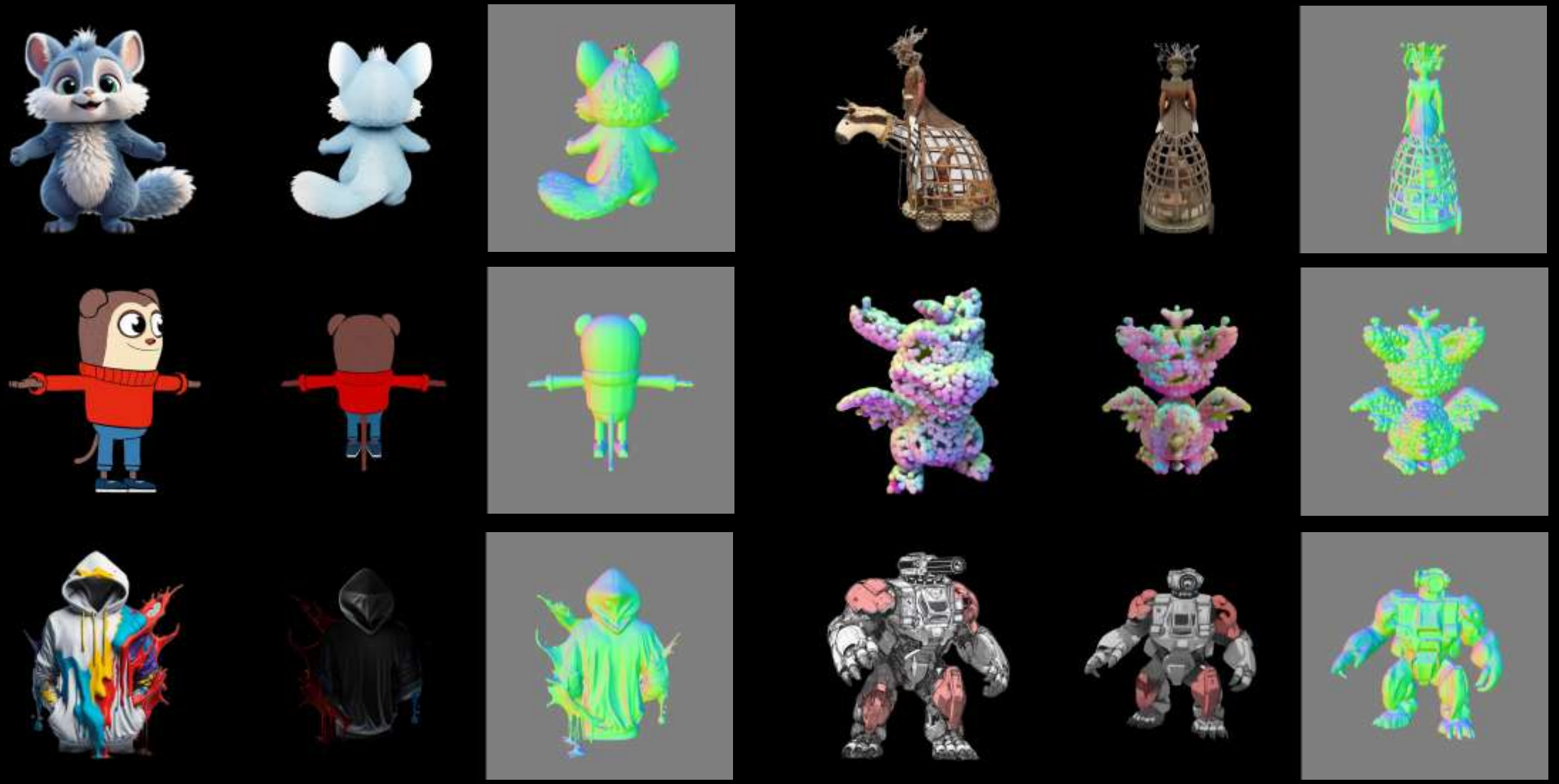
Futuristic robotic arm on a table.

Portable transistor radio, dark cover, speaker grille, brand logo on front.

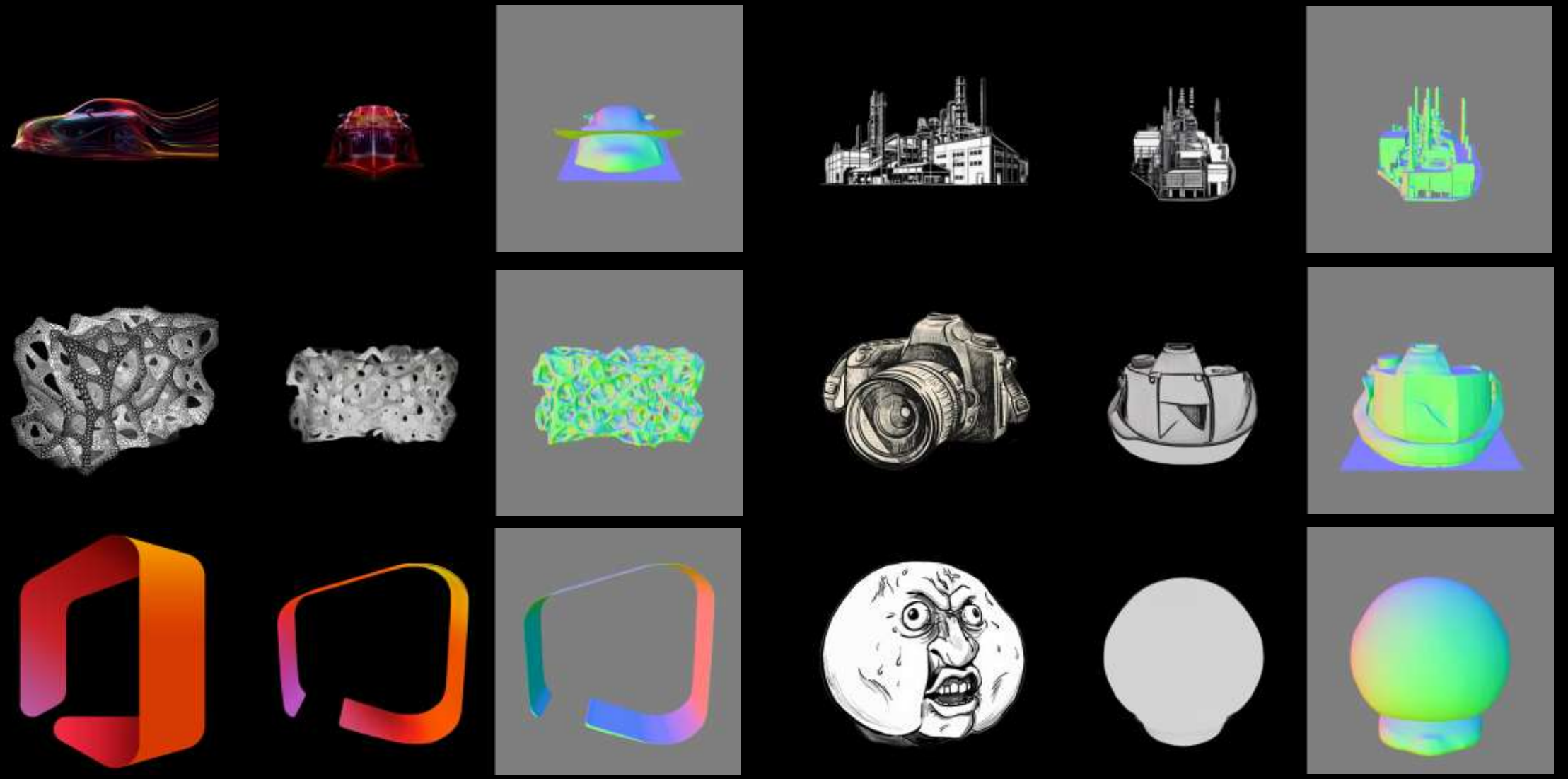
Bronze owl sculpture perched on a branch.

Results | Image to 3D

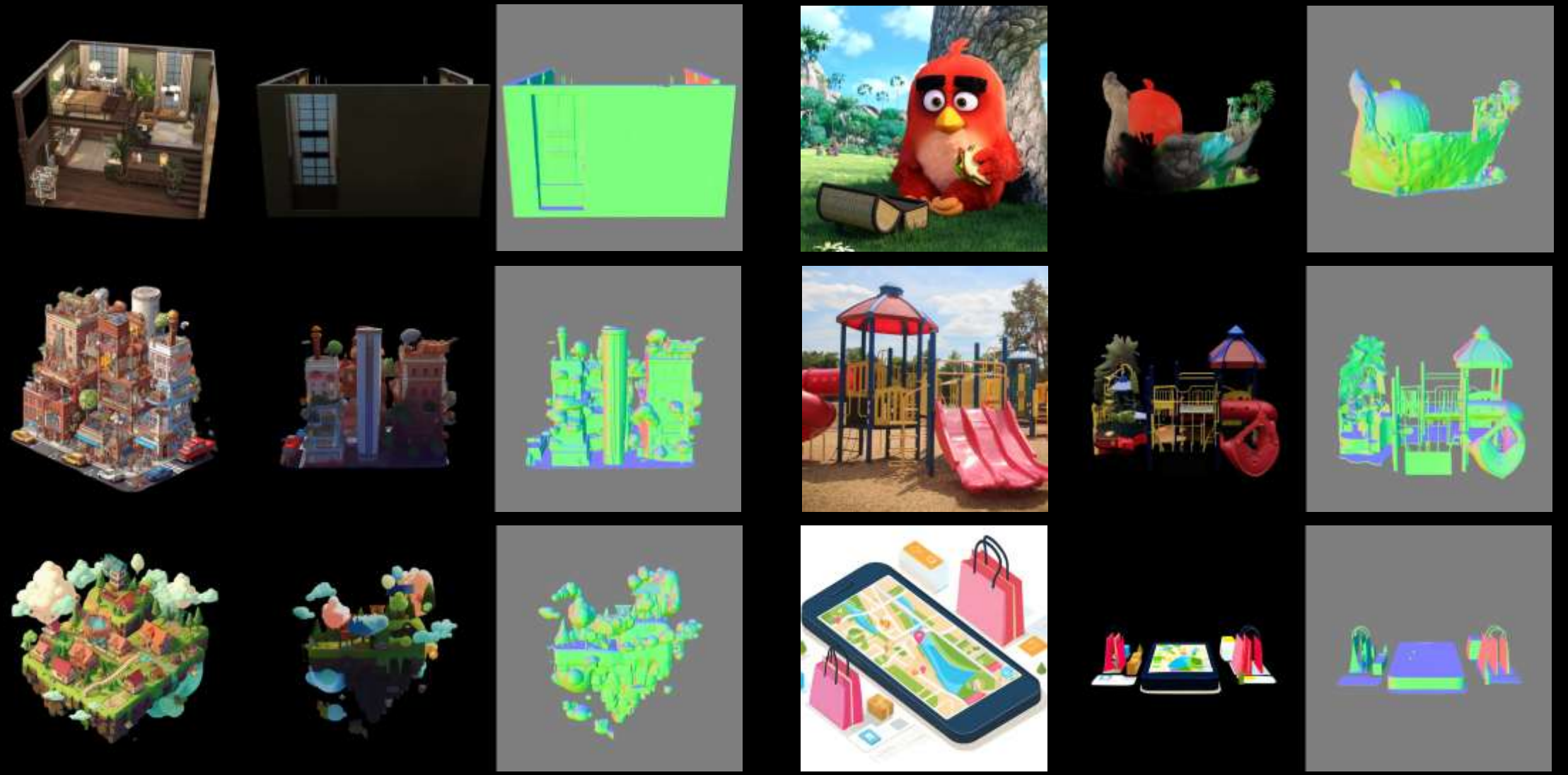
Results | Generalization Ability



Results | Generalization Ability

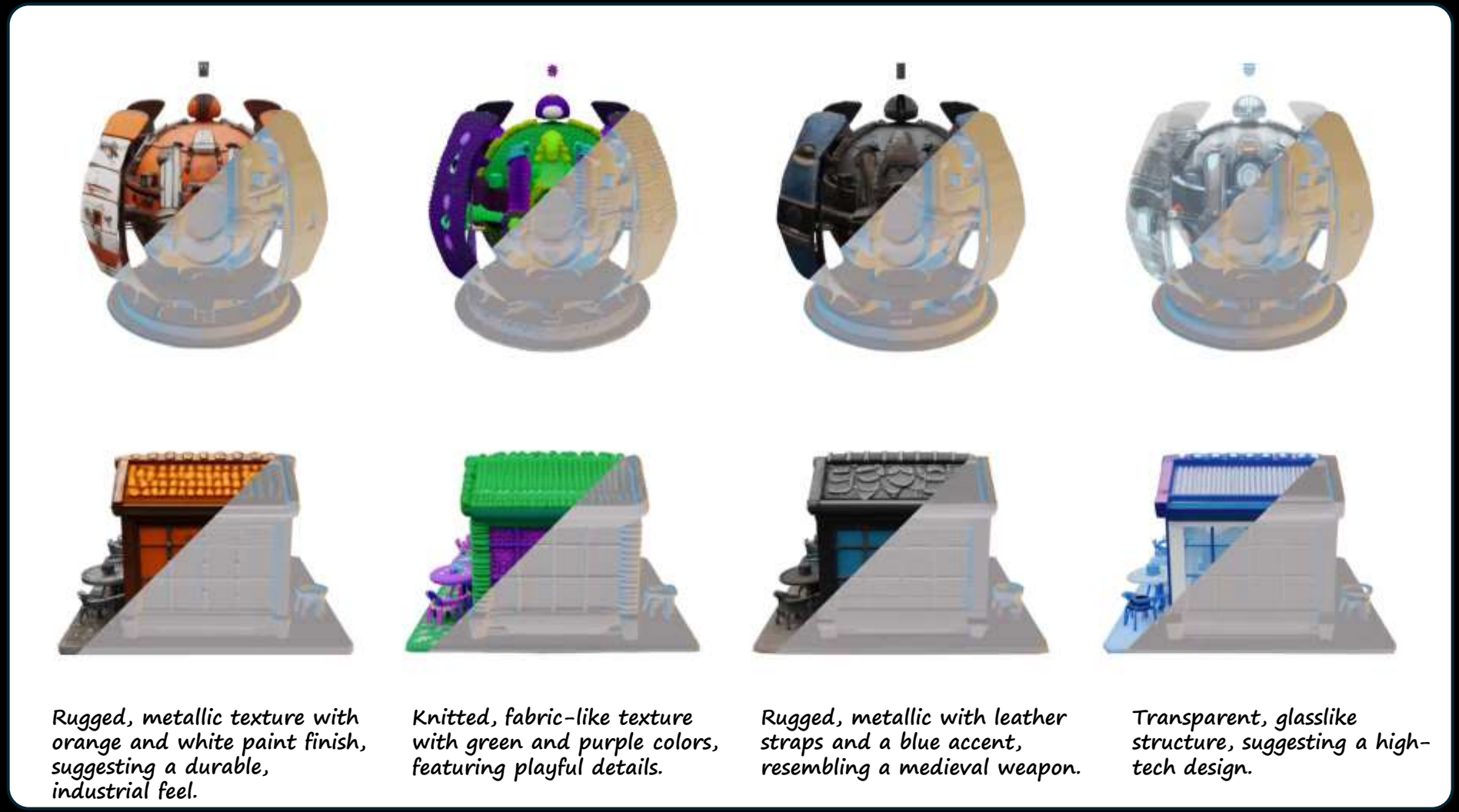


Results | Generalization Ability



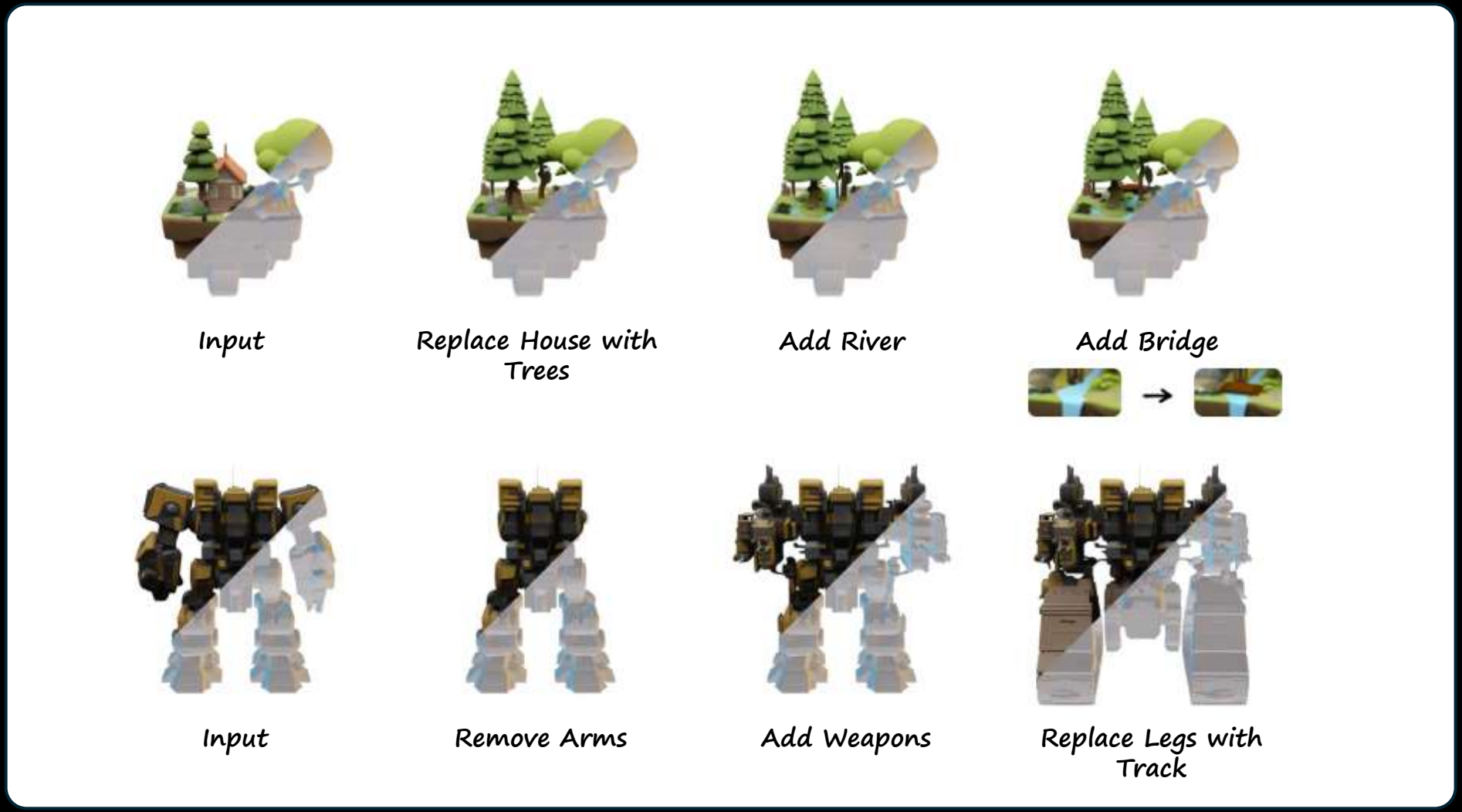
Applications | Asset Variants

Keep the first stage sparse structure fixed and generate varied latents:



Applications | Local Editing

Applying tuning-free inpainting algorithm Repaint to both stage



Results | Method Comparison

Table 1. Reconstruction fidelity of different latent representations. (†: evaluated using albedo color; ‡: evaluated via Radiance Fields)

Method	Appearance		Geometry			
	PSNR↑	LPIPS↓	CD↓	F-score↑	PSNR-N↑	LPIPS-N↓
LN3Diff	26.44	0.076	0.0299	0.9649	27.10	0.094
3DTopia-XL	25.34†	0.074†	0.0128	0.9939	31.87	0.080
CLAY	–	–	0.0124	0.9976	35.35	0.035
Ours	**32.74**/32.19‡	**0.025**/0.029‡	**0.0083**	**0.9999**	**36.11**	**0.024**

Figure 6. User study for text/image-to-3D generation.

Table 2. Quantitative comparisons using Toys4k

Method	Text-to-3D						Image-to-3D					
	CLIP↑	FD_{incep}↓	KD_{incep}↓	FD_{dinov2}↓	KD_{dinov2}↓	FD_{point}↓	CLIP↑	FD_{incep}↓	KD_{incep}↓	FD_{dinov2}↓	KD_{dinov2}↓	FD_{point}↓
Shap-E	25.04	37.93	0.78	497.17	49.96	6.58	82.11	34.72	0.87	465.74	62.72	8.20
LGM	24.83	36.18	0.77	507.47	61.89	24.73	83.97	26.31	0.48	322.71	38.27	15.90
InstantMesh	25.56	36.73	0.62	478.92	49.77	10.79	84.43	20.22	0.30	264.36	25.99	9.63
3DTopia-XL	22.48†	53.46†	1.39†	756.37†	87.40†	13.72	78.45†	37.68†	1.20†	437.37†	53.24†	18.21
Ln3Diff	18.69	71.79	2.85	976.40	154.18	19.40	82.74	26.61	0.68	357.93	50.72	7.86
GaussianCube	24.91	27.35	0.30	460.07	39.01	29.95	–	–	–	–	–	–
Ours L	26.60	20.54	**0.08**	238.60	4.24	5.24	**85.77**	**9.35**	**0.02**	**67.21**	**0.72**	**2.03**
Ours XL	**26.70**	**20.48**	**0.08**	**237.48**	**4.10**	**5.21**	–	–	–	–	–	–

Results | Ablation Study

Table 3. Ablation study on the size of SLAT.

Resolution	Channel	PSNR↑	LPIPS↓
32	16	31.64	0.0297
32	32	31.80	0.0289
32	64	31.85	0.0283
64	8	**32.74**	**0.0250**

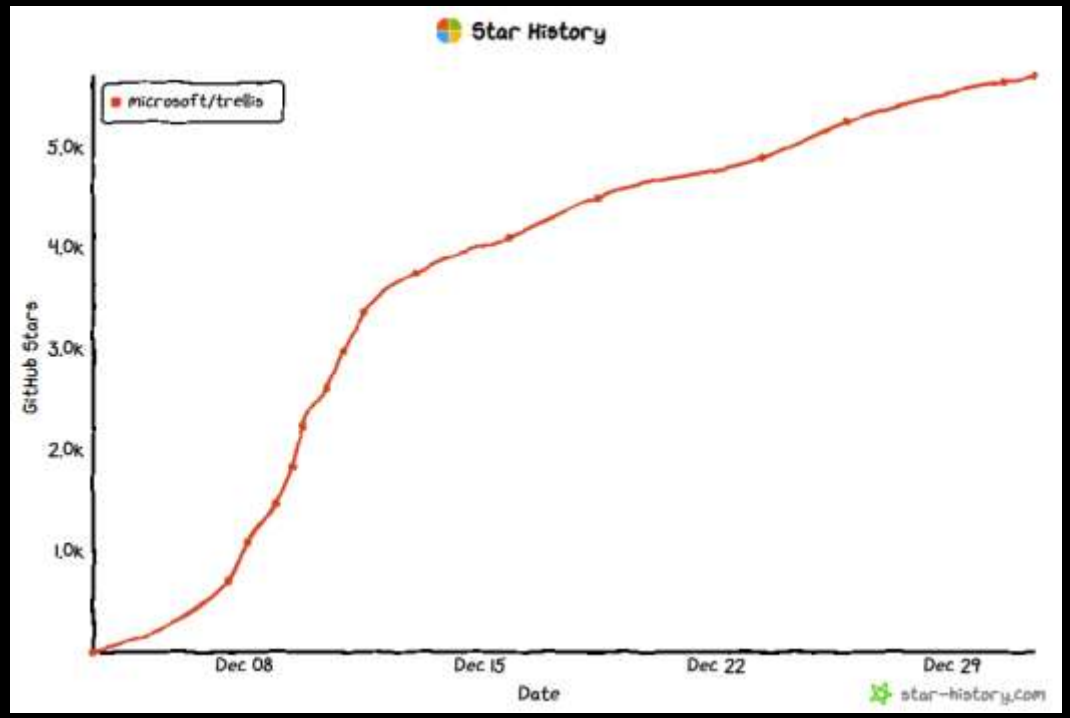
Table 5. Ablation study on model size.

Method	Training set		Toys4k	
	CLIP↑	FD_{dinov2}↓	CLIP↑	FD_{dinov2}↓
B	25.41	121.45	26.47	265.26
L	25.62	99.92	26.60	238.60
XL	**25.71**	**93.96**	**26.70**	**237.48**

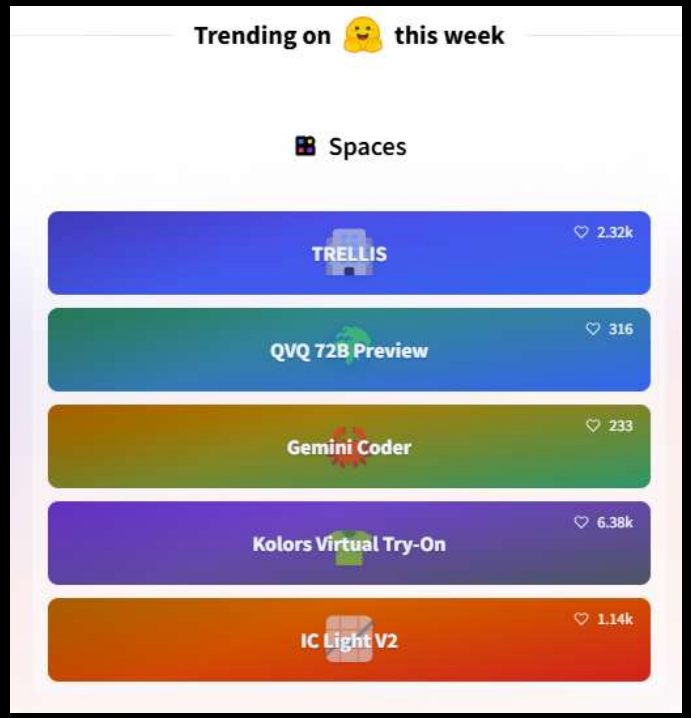
Table 4. Ablation study on different generation paradigms.

	Method	Training set		Toys4k	
		CLIP↑	FD_{dinov2}↓	CLIP↑	FD_{dinov2}↓
Stage 1	Diffusion	25.09	132.71	25.86	295.90
	Rectified flow	**25.40**	**113.42**	**26.37**	**269.56**
Stage 2	Diffusion	25.58	100.88	26.45	244.08
	Rectified flow	**25.65**	**95.97**	**26.61**	**240.20**

Trending



GitHub: 5700+ Stars in 30 days



Hugging Face: Most Popular Space

Trending

Extensive community engagement

4D Generation

Summary

Structured Latent representation (SLAT) for comprehensive information encoding and versatile decoding

Two-stage Rectified Flow Transformers for SLAT generation

We hope our model can:

- *Serve as powerful 3D generation foundations and unlock new possibilities for the 3D vision community.*
- *Shed some light on 3D-representation-agnostic asset modeling, in contrast to the field's relentless pursuit of and adaptation to new representations.*

Thanks

For more information, visit:
https://trellis3d.github.io