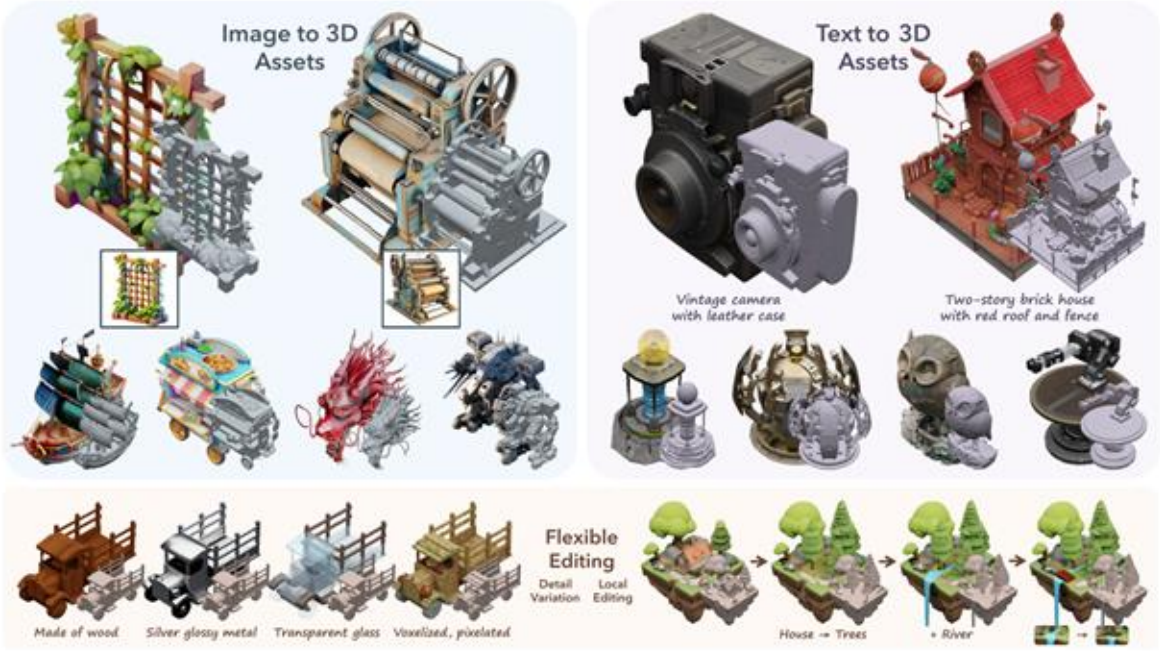
TRELLIS

Structured 3D Latents for Scalable and Versatile 3D Generation

Xiang et al., Microsoft Research

Blurb

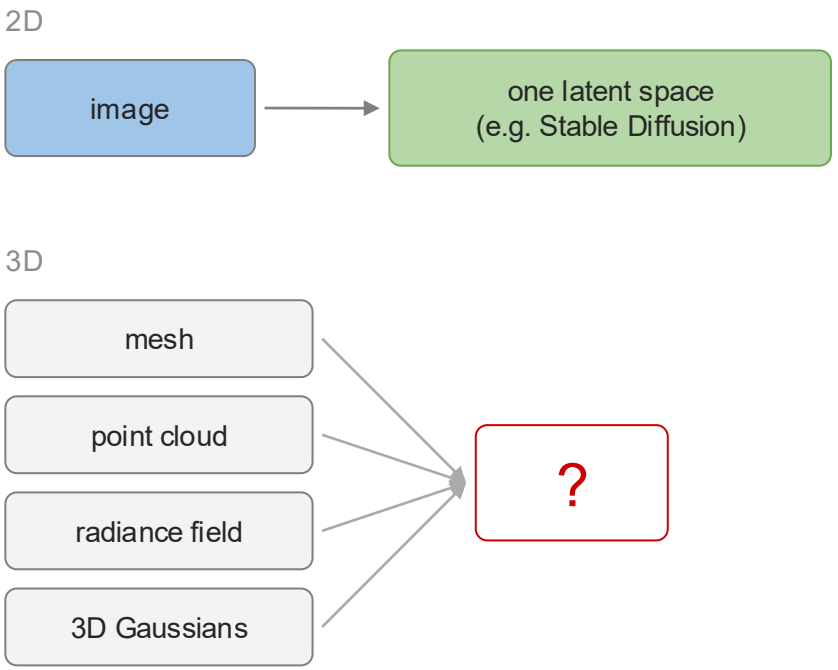
“A unified Structured LATent (SLAT) representation which allows decoding to different output formats, such as Radiance Fields, 3D Gaussians, and meshes”



Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

The Problem

- Task: text or image in, 3D asset out
- 3D quality lags far behind 2D image generation
- Meshes: good geometry, weak appearance
- Gaussians / NeRFs: great appearance, poor geometry
- **Why care:**
  - 2D generators are already everyday tools, 3D isn't there yet
- **Why it's hard:**
  - No shared latent, so no shared architecture

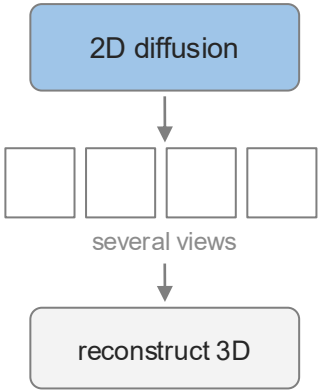


Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

Design Space

Lift from 2D

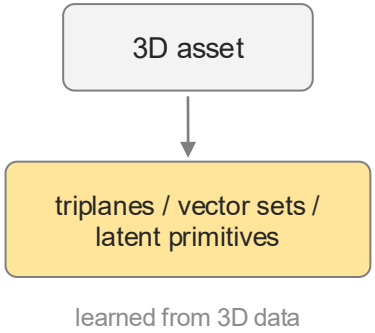
e.g. InstantMesh, LGM (also DreamFusion)



Strong 2D priors
but generated views disagree

Native 3D Latents

e.g. Shap-E, LN3Diff, 3DTopia-XL, CLAY

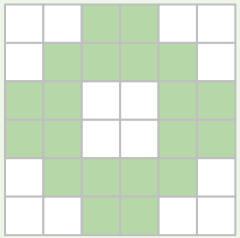


learned from 3D data

Consistent 3D, but shape only,
weak appearance, or costly pre-fitting

Structured Sparse Latent

TRELLIS



latents only on surface voxels

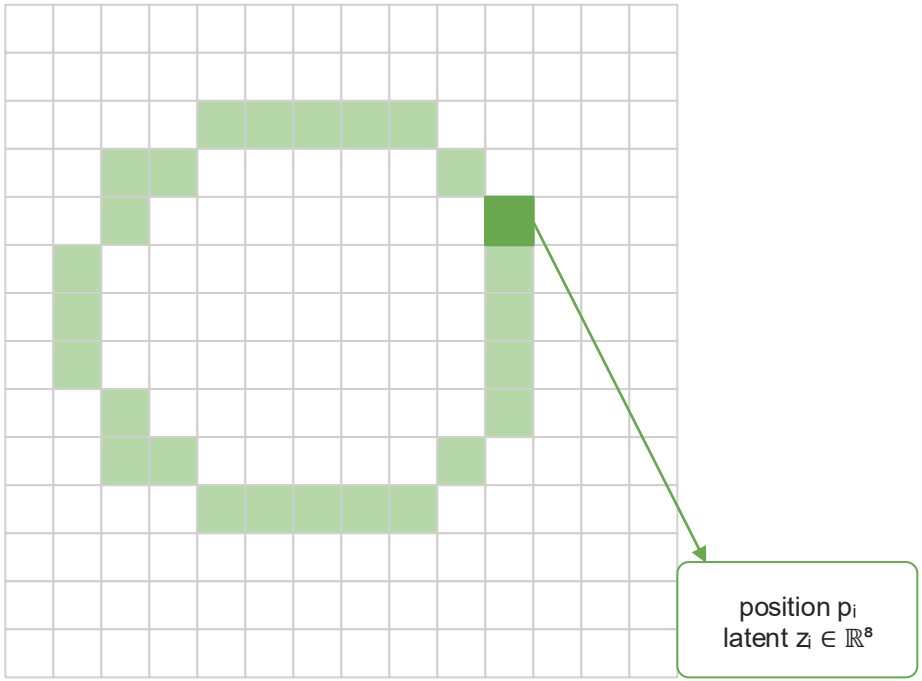
Geometry + appearance,
decodes to any format

Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

The Model

Key Idea: SLAT

- 64^3 grid, but only voxels touching the surface are “active”
- ~20K active voxels out of 262K
- Each active voxel gets an latent
- Structure = where, latent = what it looks like

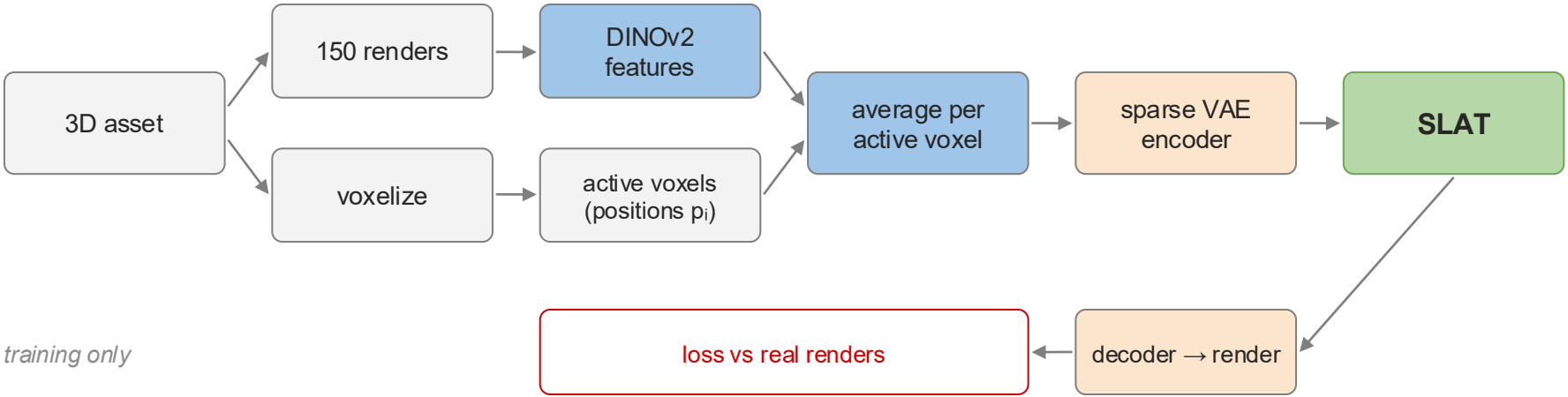


2D slice: shaded = active (surface) voxels, white = empty

Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

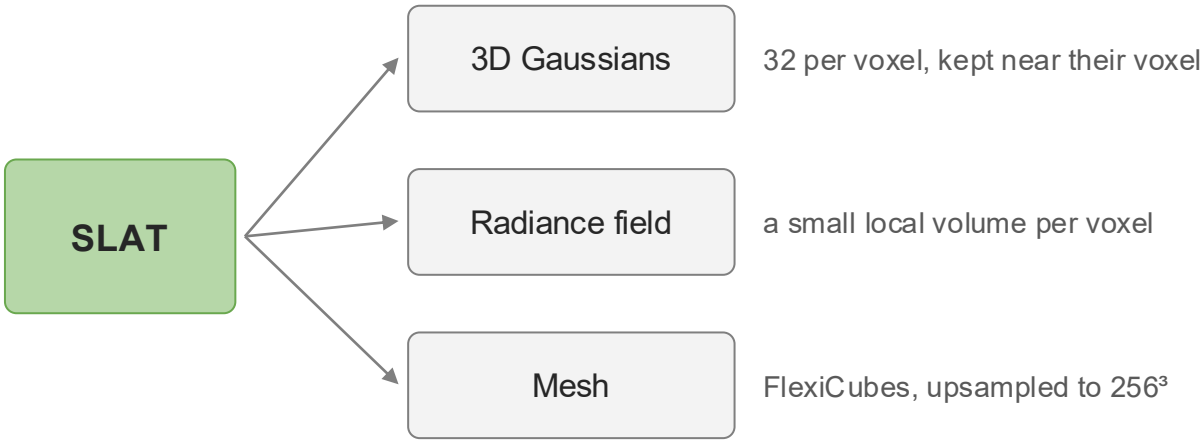
Encoding: Borrowing DINOv2

- Voxelize the asset → which voxels are active
- Render 150 views, run DINOv2, average features per active voxel
- Sparse VAE: compress to SLAT, decode, render, compare to the real renders



Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

Decoding to Any Format



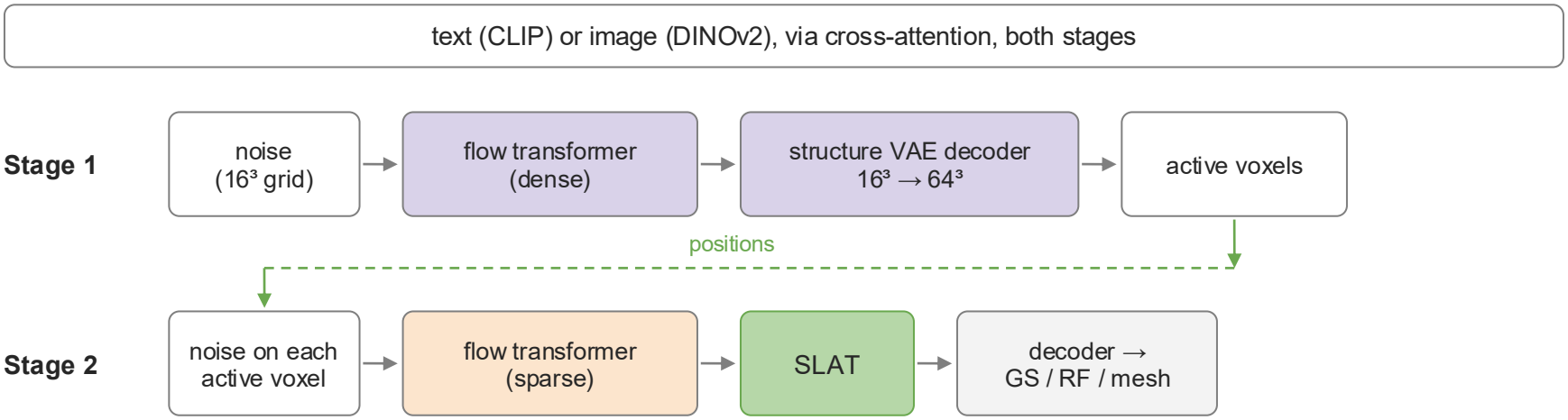
- Train encoder + Gaussian decoder together
- Then freeze the encoder, train the other decoders



same asset as Gaussians / radiance field / mesh (Fig. 12)

Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

Two-Stage Generation



- Stage 1: where the surface is (coarse shape)
- Stage 2: a latent for every active voxel (detail + appearance)
- Both are rectified flow models: predict (noise − data), MSE loss

Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

Editing

- Keep the structure, re-run stage 2 with a new prompt: variations
- Regenerate only a box: remove / add / replace parts
- No fine-tuning needed



Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

Results

How They Evaluate

What	Metrics	Data
Reconstruction	PSNR / LPIPS (appearance), Chamfer distance / F-score (shape)	500 Toys4k assets
Generation	CLIP score (prompt match), FD / KD with Inception, DINOv2, PointNet++	Toys4k (3,229 assets)
User study	pick the best asset (or “not sure”)	104 people, 135 prompts

- Trained on ~500K assets (Objaverse-XL, ABO, 3D-FUTURE, HSSD), GPT-4o captions
- Three sizes: 342M / 1.1B / 2B; image model only at 1.1B
- Baselines: 2D-lifting (InstantMesh, LGM) and native 3D (Shap-E, LN3Diff, 3DTopia-XL, GaussianCube)

Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

Is the Latent Any Good?

Encode → decode, compared to the original asset

Latent	PSNR ↑	LPIPS ↓	Chamfer ↓	F-score ↑
LN3Diff (triplane)	26.44	0.076	0.0299	0.9649
3DTopia-XL (primitives)	25.34*	0.074*	0.0128	0.9939
CLAY (vector set, shape only)	–	–	0.0124	0.9976
SLAT → Gaussians	**32.74**	**0.025**		
SLAT → radiance field	**32.19**	**0.029**		
SLAT → mesh			**0.0083**	**0.9999**

* evaluated on albedo color

– Radiance field + mesh decoders trained on a frozen encoder, still best

– Beats CLAY on geometry, even though CLAY only encodes shape

Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

Visual Comparison

One text prompt and one image prompt, every method



Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

Numbers + User Study

	TRELLIS	Best baseline
Text: CLIP ↑	**26.70**	25.56 InstantMesh
Text: FD Inception ↓	**20.48**	27.35 GaussianCube
Text: FD DINOv2 ↓	**237**	460 GaussianCube
Text: FD PointNet++ ↓	**5.21**	6.58 Shap-E
Image: CLIP ↑	**85.77**	84.43 InstantMesh
Image: FD Inception ↓	**9.35**	20.22 InstantMesh
Image: FD DINOv2 ↓	**67**	264 InstantMesh
Image: FD PointNet++ ↓	**2.03**	7.86 LN3Diff

text rows: 2B model, image rows: 1.1B model

Users picked TRELLIS

67%

text prompts

next best: 10%

95%

image prompts

next best: 2%

Best on every metric, biggest gaps for image-to-3D

Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

Ablations

Latent size (reconstruction)

Grid / channels	PSNR
32^3 / 16	31.64
32^3 / 32	31.80
32^3 / 64	31.85
64^3 / 8	**32.74**

Flow vs diffusion (FD ↓)

	Diff.	Flow
Stage 1	296	**270**
Stage 2	244	**240**

Model size (FD ↓)

Size	Train set	Toys4k
B	121	265
L	100	239
XL	94	237

- More channels at 32^3 plateaus, going to 64^3 helps a lot
- Rectified flow beats diffusion at both stages
- Bigger is better, but L → XL barely moves on Toys4k

Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

Strengths

- One latent, many output formats
- Geometry and appearance in the same latent
- Editing without retraining, thanks to locality
- Fully open: code, models, data
- Ablations on resolution, flow vs diffusion, and scale



Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

Weaknesses

- DINOv2 everywhere: features, image condition, and a metric
- No ablation of the feature extractor or number of views
- Baselines differ in training data and size
- Lighting baked into textures, no PBR materials (authors' own)
- Two stages: slower than end-to-end (authors' own), and can't revisit the shape
- Real photos only shown qualitatively

Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

Open Questions

- Why DINOv2? How much worse with CLIP or plain RGB?
- If DINOv2 matters that much, is FD-DINOv2 a fair metric?
- Trained only through Gaussians, so why do meshes come out well?
- Does a 64^3 grid hurt thin parts like wires or hair?
- How much does the aesthetic filtering matter?

Xiang, Jianfeng, et al. "Structured 3D Latents for Scalable and Versatile 3D Generation." CVPR (2025)

Follow-up: TRELLIS.2

	TRELLIS	**TRELLIS.2 (Dec 2025)**
Resolution	64^3 voxels	up to 1536^3
Latent built from	DINOv2 features of 150 renders	the mesh itself (“O-Voxel”)
Appearance	colors with baked-in lighting	PBR: color, metallic, roughness, opacity
Mesh output	FlexiCubes (signed distances)	also open and non-manifold surfaces
Generator	2 stages, up to 2B	3 stages (structure, shape, material), ~4B
Speed	~10 s (A100)	~3 s at 512^3, ~17 s at 1024^3 (H100)

- Addresses two weaknesses: baked-in lighting, and DINOv2 inside the latent
- User study (image-to-3D): 66.5% picked TRELLIS.2, 6.4% picked TRELLIS
- Still limited by voxel resolution, small holes, no part structure

Xiang, Jianfeng, et al. "TRELLIS.2: Native and Compact Structured Latents for 3D Generation." arXiv:2512.14692 (2025)

Questions?