Structured 3D Latents for Scalable and Versatile 3D Generation

Paper Presentation

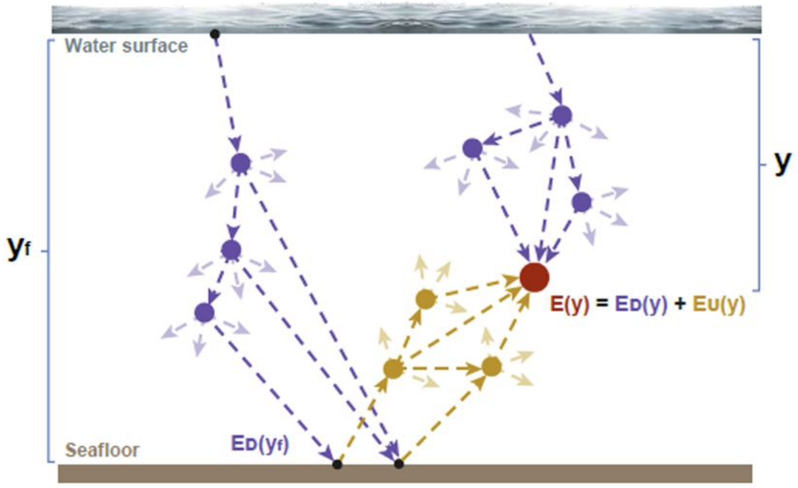
2025.11.24

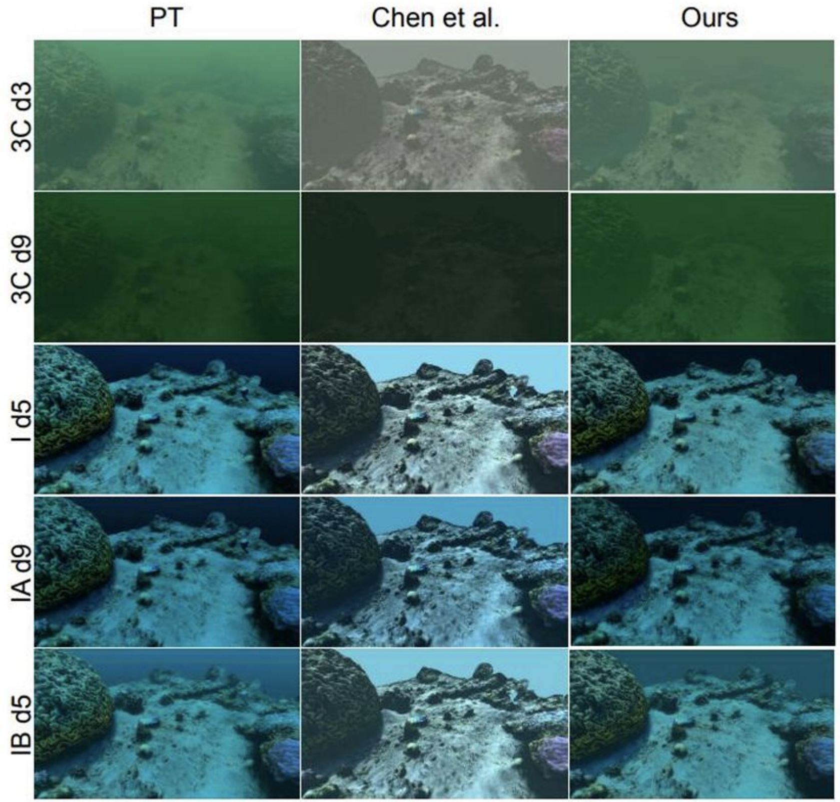
Minseo Park, Jewoo Shin, Sangmin Lee

Team 2

Review of previous lecture

Real-Time Underwater Spectral Rendering



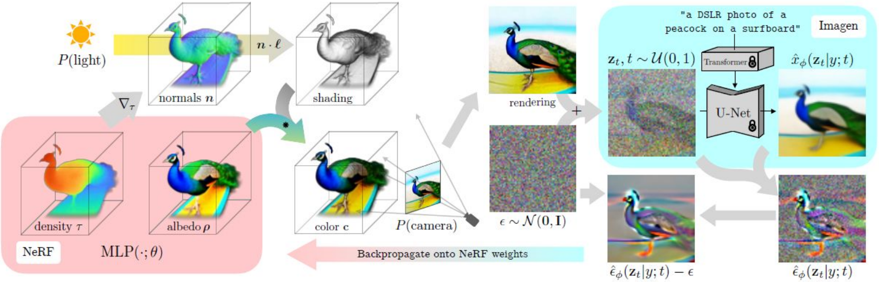


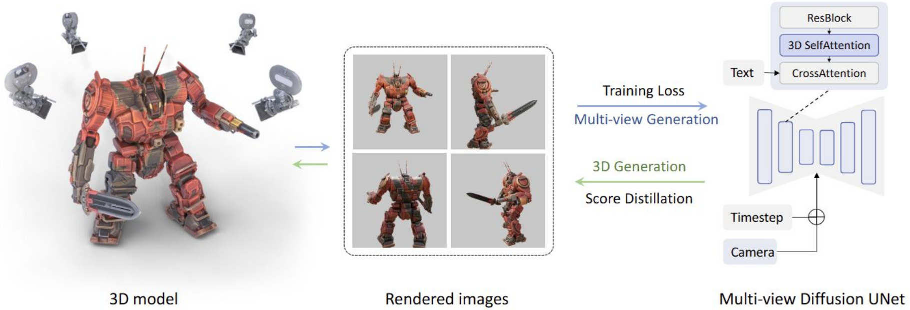
source: Monzon et al., Real-Time Underwater Spectral Rendering. 2024

Table of Contents

Introduction

3D generative models





Single-view Image Generation Model based Distillation (DreamFusion)

Multi-view Image Generation Model based Distillation (MVDream)

Source: B. Poole et al., "DreamFusion: Text-to-3D using 2D Diffusion", ICLR 2023.

3D generative models

Diffusion model + 3D representation:

- Pointcloud
- Voxel grid
- Triplane
- 3D gaussians

Challenges:

Efficiency for modeling in raw data space

Diffusion model + more compact latents:

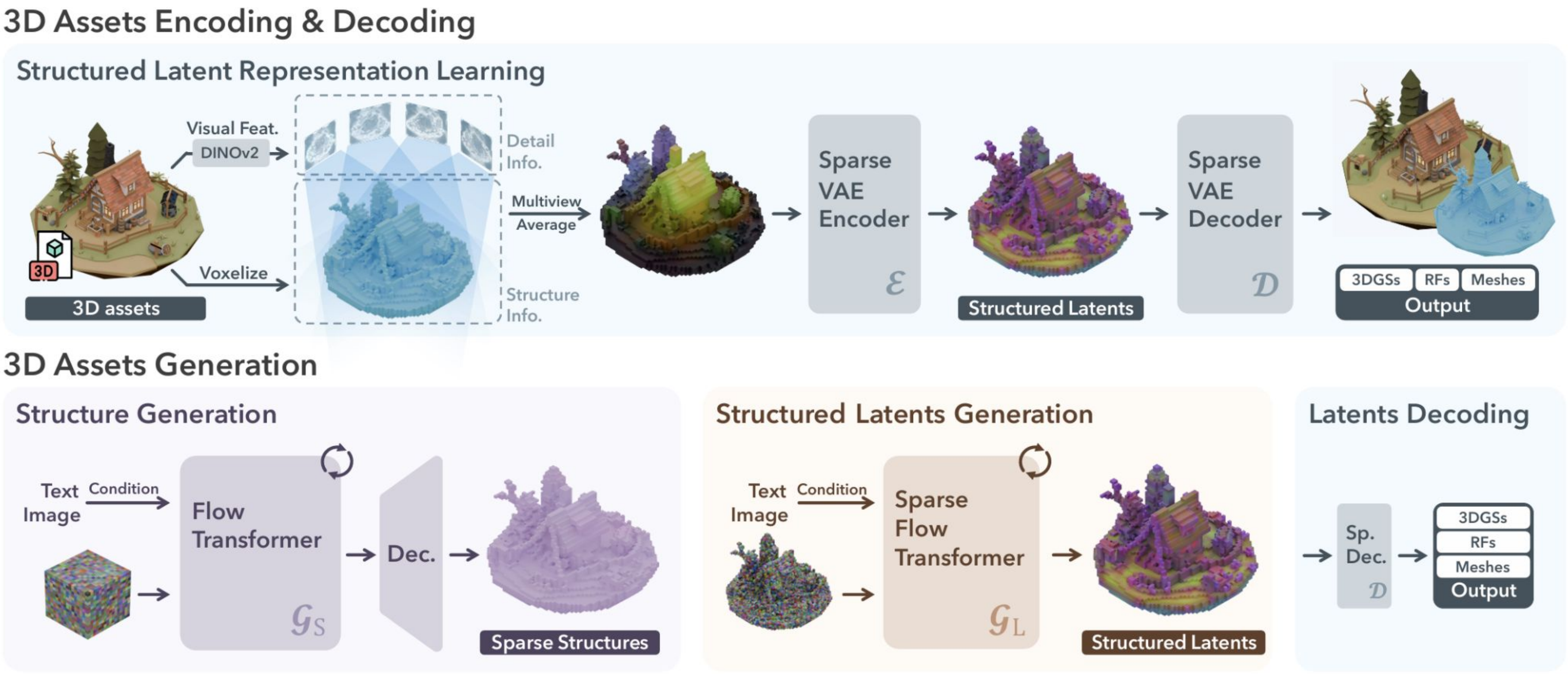
- Enhanced quality & efficiency

Challenges:

Accurate surface modeling

Methodology

Trellis: Structured 3D Latents for Scalable and Versatile 3D Generation



Source: Xiang et al., "Structured 3D Latents for Scalable and Versatile 3D Generation", CVPR 2025.

Structured Latent Representation

Structured Latent(SLat)

$$\boldsymbol{z} = \{(\boldsymbol{z}_i, \boldsymbol{p}_i)\}_{i=1}^{L}, \quad \boldsymbol{z}_i \in \mathbb{R}^C, \ \boldsymbol{p}_i \in \{0, 1, \ldots, N-1\}^3$$

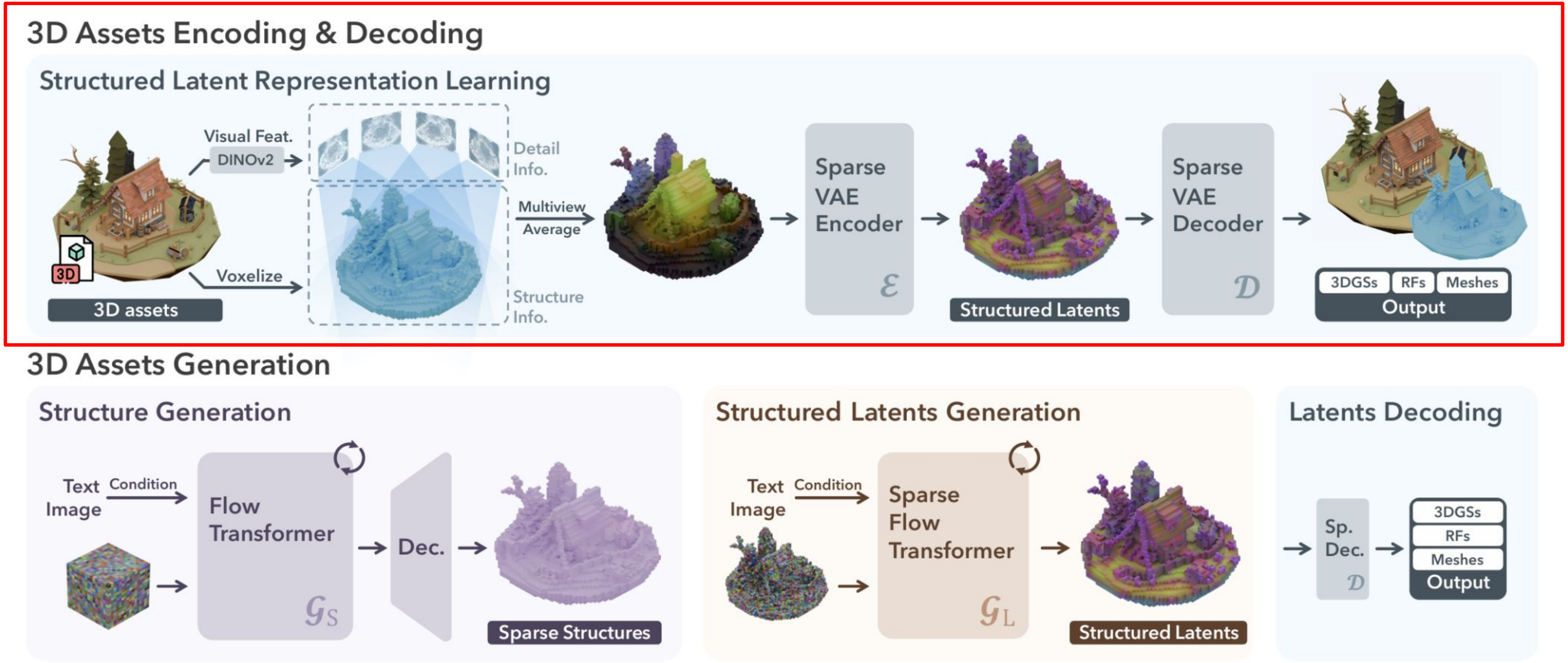
A set of local latents + positional index on 3D grid

- p_i: positional index of an active voxel on 3D grid intersecting with surface
- z_i: local latents attached to the corresponding voxel
- N: spatial length of the 3D grid
- L: total number of active voxels.
- By default, N=64, L=20k

Source: Xiang et al., “Structured 3D Latents for Scalable and Versatile 3D Generation”, CVPR 2025.

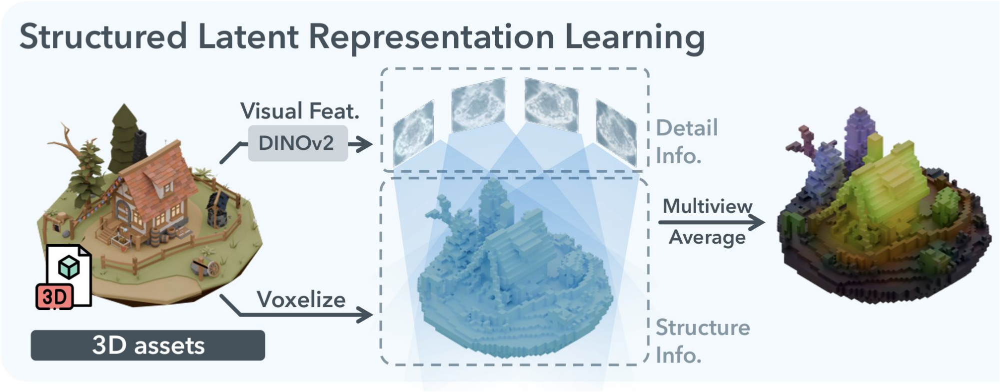
Structured Latents Encoding and Decoding

Structured Latents Encoding and Decoding



Source: Xiang et al., "Structured 3D Latents for Scalable and Versatile 3D Generation", CVPR 2025.

Visual Feature aggregation



- Initially, convert 3d assets into a **voxelized feature** *f*

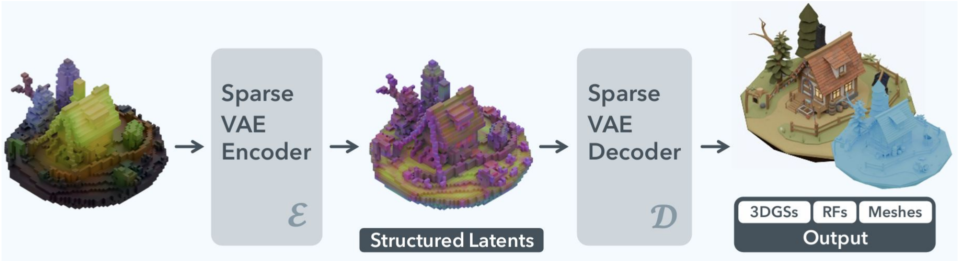
$$\boldsymbol{f} = \{(\boldsymbol{f}_i, \boldsymbol{p}_i)\}_{i=1}^{L}$$

- Feature maps are extracted from randomly sampled camera views with pre-trained **DINOv2** Encoder.

- It is sufficient to match resolution of voxelized feature with structured latents.

Source: Xiang et al., “Structured 3D Latents for Scalable and Versatile 3D Generation”, CVPR 2025.

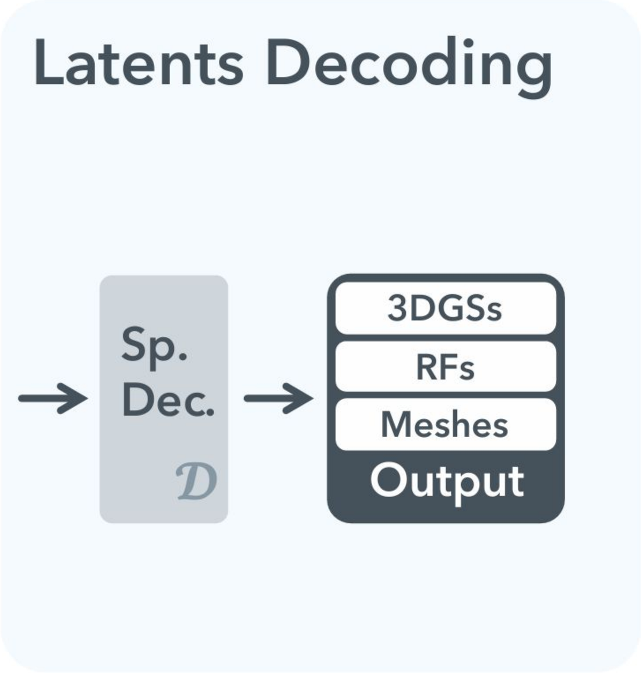
Sparse VAE for Structured Latents



- **Transformer-based VAE architecture** was introduced for 3d assets encoding.
- Training loss of KL-penalty is applied to train decoded 3D assets with ground truth.
- Sinusoidal **positional encodings** based on voxel position is processed through transformer blocks.
- **Shifted window attention** to enhance local information interaction.

Source: Xiang et al., “Structured 3D Latents for Scalable and Versatile 3D Generation”, CVPR 2025.

Structured latent Decoding

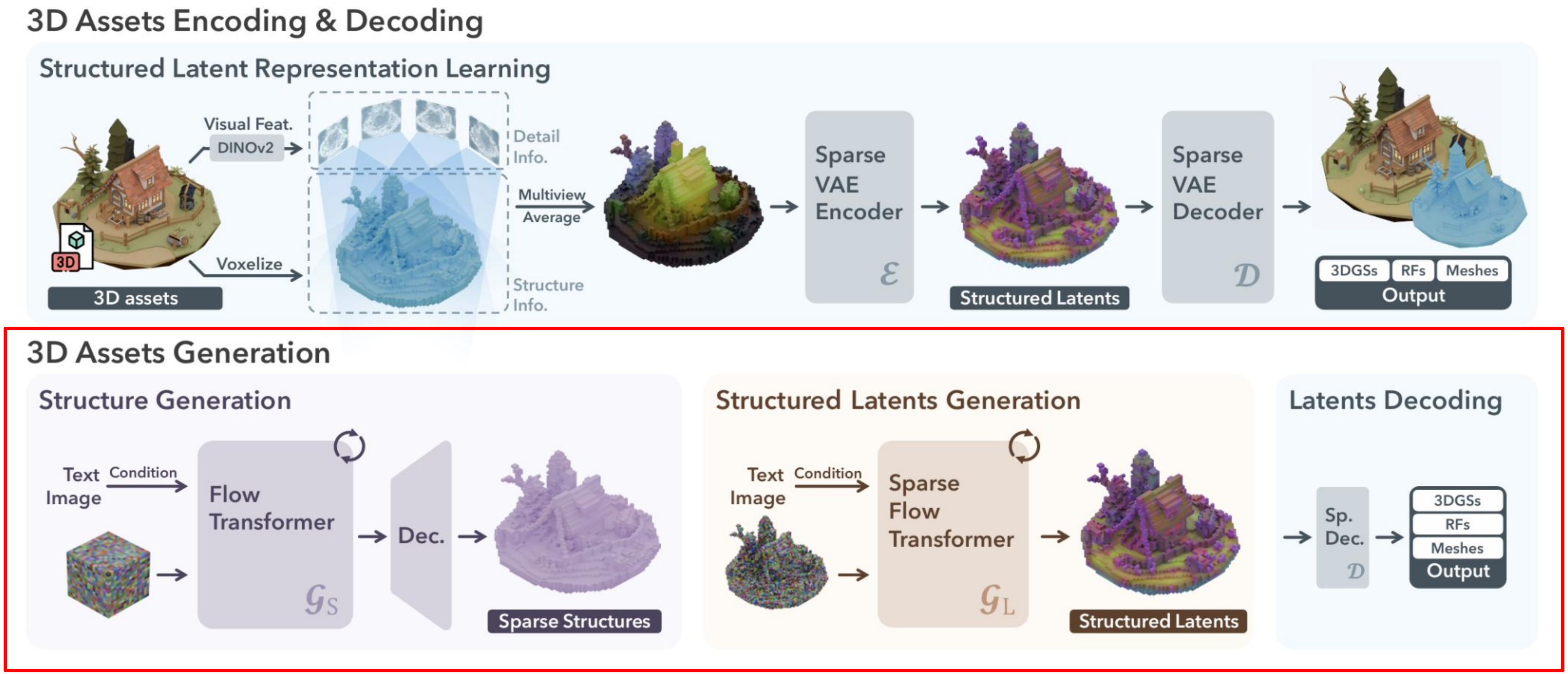


- Decoders for 3D gaussians, radiance fields, and meshes share the same architecture except output layers.
- **Reconstruction losses**: ex) L1, D-SSIM, LPIPS between rendered output & ground truth images.
- In practice, they adopt gaussians due to high fidelity & efficiency.

Source: Xiang et al., "Structured 3D Latents for Scalable and Versatile 3D Generation", CVPR 2025.

Structured Latents Generation

Structured Latents Generation



Source: Xiang et al., "Structured 3D Latents for Scalable and Versatile 3D Generation", CVPR 2025.

Rectified flow models

- **Linear Forward Process:** It uses a linear interpolation path to connect data samples x_0 and noise ϵ over a timestep t:

$$x(t) = (1 - t)x_0 + t\epsilon$$

- **Vector Field Definition:** The backward generation process is defined as a time-dependent vector field that moves noisy samples toward the data distribution:

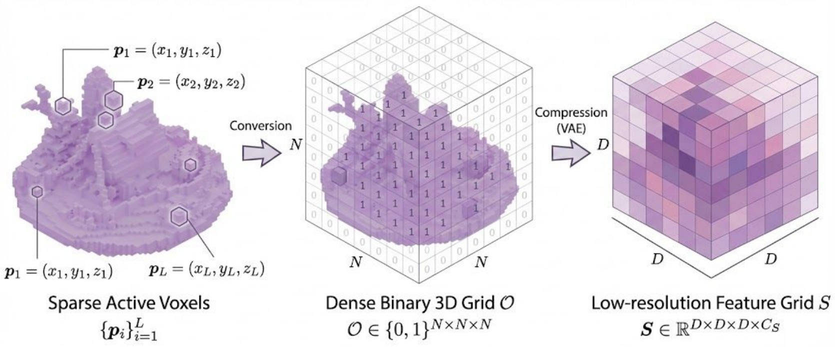
$$v(x, t) = \nabla_t x$$

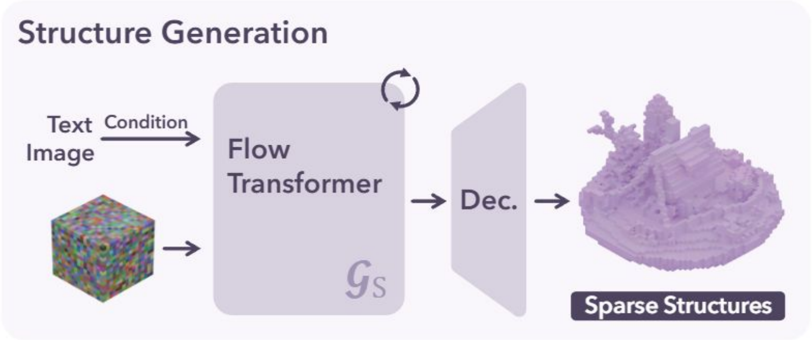
- **Flow Matching Objective:** A neural network v_θ is trained to approximate this vector field by minimizing the Conditional Flow Matching (CFM) loss:

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t,x_0,\epsilon}||v_\theta(x, t) - (\epsilon - x_0)||_2^2$$

Source: Xiang et al., “Structured 3D Latents for Scalable and Versatile 3D Generation”, CVPR 2025.

Sparse structure generation



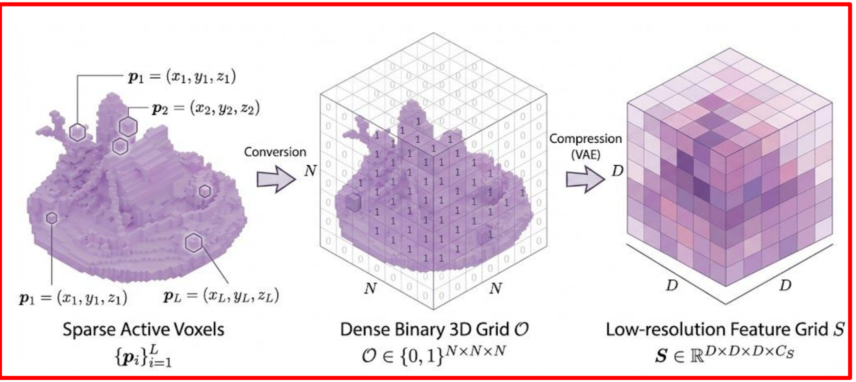


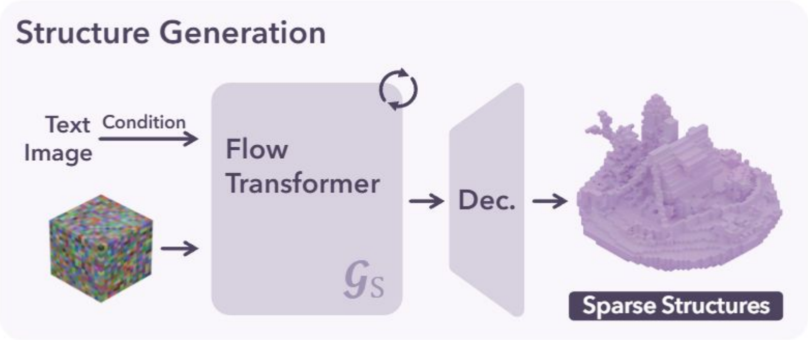
Noisy S → Clean S → O → P

- **Efficient Representation:** The method converts sparse active voxels $\{p_i\}_{i=1}^L$ into a dense binary grid $O \in \{0,1\}^{N\times N\times N}$, which is then compressed by a 3D VAE into a low-resolution, continuous feature grid $S \in \mathbb{R}^{D\times D\times D\times C_S}$ to facilitate computationally efficient rectified flow training.
- **Transformer Backbone:** A transformer model $\mathcal{G}_S$ is employed to generate S by processing serialized noisy grids combined with positional encodings, while timestep information is integrated using adaptive layer normalization (adaLN).
- **Conditioning Mechanism:** Conditions are injected via cross-attention layers, utilizing pre-trained CLIP features for text prompts and DINOv2 visual features for image prompts.
- **Final Output Decoding:** The generated denoised feature grid S is decoded back into the discrete grid O, which is subsequently converted into the final sparse structure $\{p_i\}_{i=1}^L$.

Source: Xiang et al., "Structured 3D Latents for Scalable and Versatile 3D Generation", CVPR 2025.

Sparse structure generation



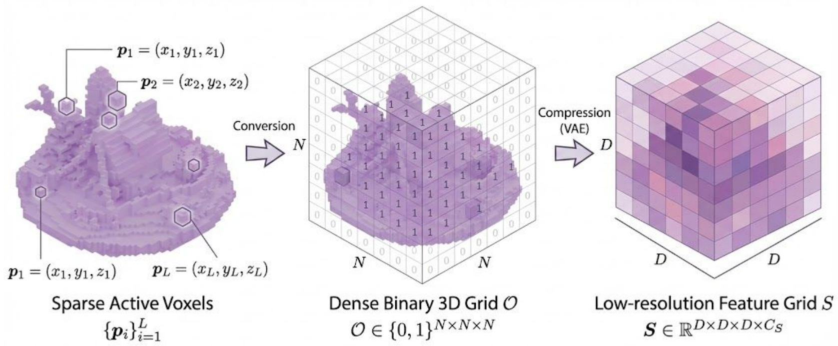


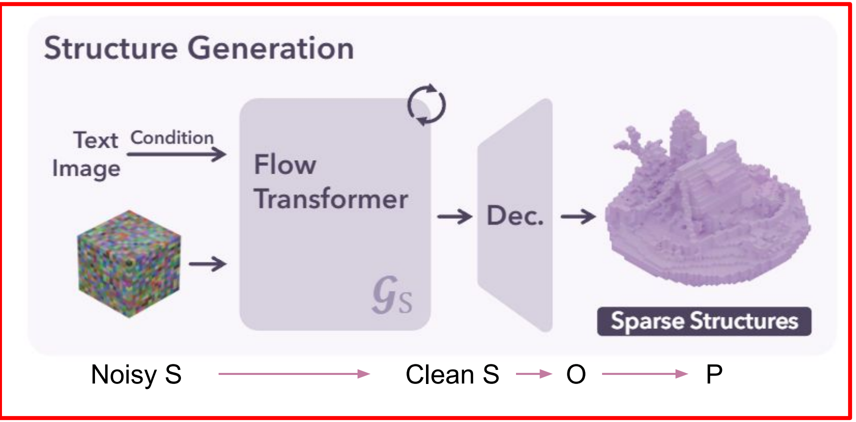
Noisy S → Clean S → O → P

- **Efficient Representation:** The method converts sparse active voxels $\{p_i\}_{i=1}^L$ into a dense binary grid $O \in \{0,1\}^{N\times N\times N}$, which is then compressed by a 3D VAE into a low-resolution, continuous feature grid $S \in \mathbb{R}^{D\times D\times D\times C_S}$ to facilitate computationally efficient rectified flow training.
- **Transformer Backbone:** A transformer model $\mathcal{G}_S$ is employed to generate S by processing serialized noisy grids combined with positional encodings, while timestep information is integrated using adaptive layer normalization (adaLN).
- **Conditioning Mechanism:** Conditions are injected via cross-attention layers, utilizing pre-trained CLIP features for text prompts and DINOv2 visual features for image prompts.
- **Final Output Decoding:** The generated denoised feature grid S is decoded back into the discrete grid O, which is subsequently converted into the final sparse structure $\{p_i\}_{i=1}^L$.

Source: Xiang et al., "Structured 3D Latents for Scalable and Versatile 3D Generation", CVPR 2025.

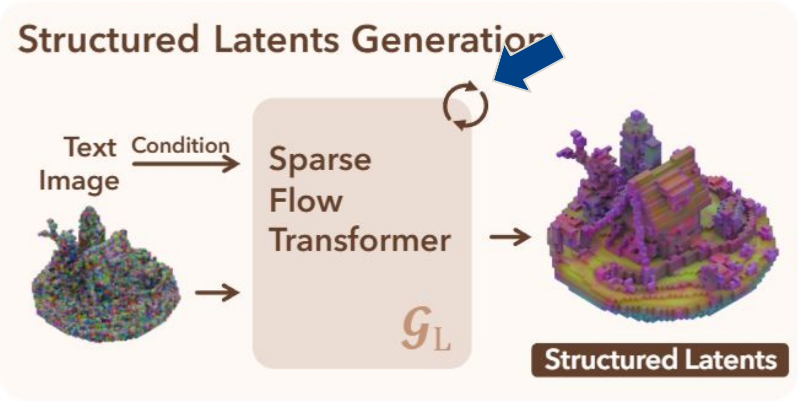
Sparse structure generation





- **Efficient Representation:** The method converts sparse active voxels $\{p_i\}_{i=1}^{L}$ into a dense binary grid $O \in \{0,1\}^{N\times N\times N}$, which is then compressed by a 3D VAE into a low-resolution, continuous feature grid $S \in \mathbb{R}^{D\times D\times D\times C_S}$ to facilitate computationally efficient rectified flow training.
- **Transformer Backbone:** A transformer model $\mathcal{G}_S$ is employed to generate S by processing serialized noisy grids combined with positional encodings, while timestep information is integrated using adaptive layer normalization (adaLN).
- **Conditioning Mechanism:** Conditions are injected via cross-attention layers, utilizing pre-trained CLIP features for text prompts and DINOv2 visual features for image prompts.
- **Final Output Decoding:** The generated denoised feature grid S is decoded back into the discrete grid O, which is subsequently converted into the final sparse structure $\{p_i\}_{i=1}^{L}$.

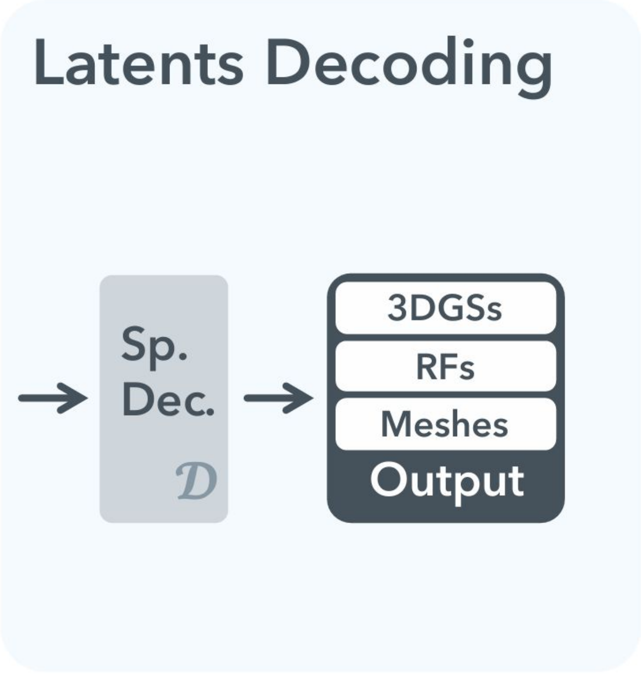
Structured latents generation



- **Conditional Generation:** The model generates local latents $\{z_i\}_{i=1}^{L}$ conditioned on the sparse structure $\{p_i\}_{i=1}^{L}$ generated in the previous stage, utilizing a specialized transformer $\mathcal{G}_L$.
- **Efficient Architecture:** To improve efficiency, input noisy latents are packed into shorter sequences using a downsampling block with **sparse convolutions** (grouping 2^3 local regions) before being processed by time-modulated transformer blocks.
- **Information Flow & Conditioning:** The network features a convolutional upsampling block with skip connections to facilitate spatial information flow, while integrating timesteps via adaLN and text/image conditions through cross-attention.
- **Independent Training:** The structure generator $\mathcal{G}_S$ and the latent generator $\mathcal{G}_L$ are trained separately utilizing the Conditional Flow Matching (CFM) objective.

Source: Xiang et al., “Structured 3D Latents for Scalable and Versatile 3D Generation”, CVPR 2025.

Structured latent Decoding

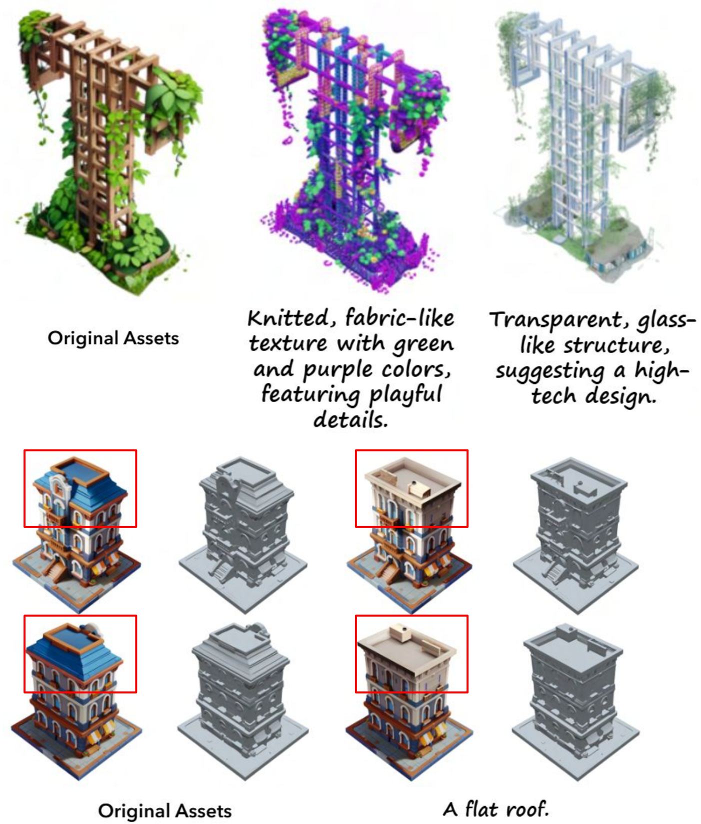


- SLat is decoded into diverse 3D formats (Gaussians, Radiance Fields, Meshes) using specific decoders $(\mathcal{D}_{GS}, \mathcal{D}_{RF}, \mathcal{D}_{M})$

Source: Xiang et al., “Structured 3D Latents for Scalable and Versatile 3D Generation”, CVPR 2025.

3D Editing with Structured Latents

Structured latent Decoding

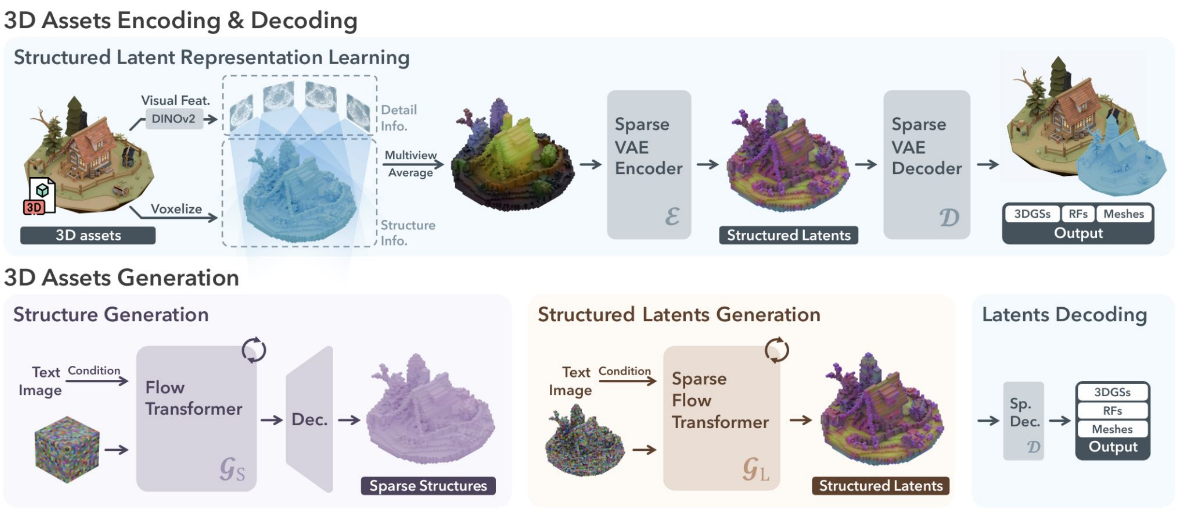


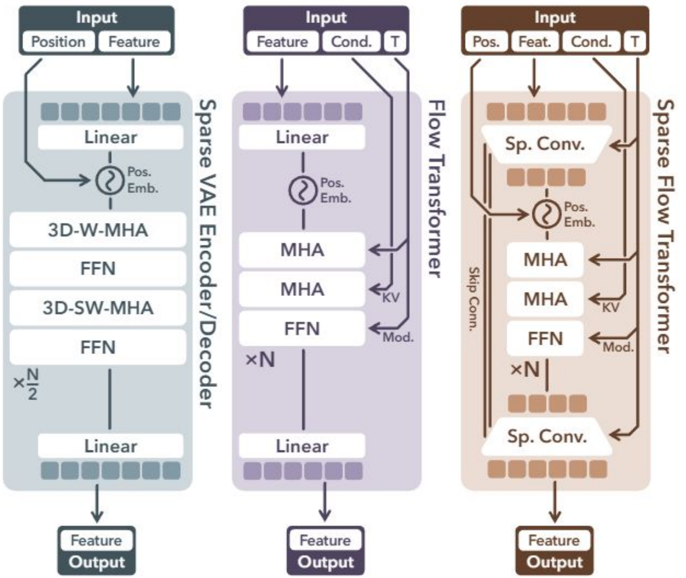
- **Detail Variation:** Modifies surface details while preserving coarse geometry by keeping the sparse structure and regenerating only the latents with new prompts.

- **Region-Specific Editing:** Regenerates both structure and details within a user-defined bounding box using inpainting techniques, conditioned on the unchanged surrounding context.

Source: Xiang et al., “Structured 3D Latents for Scalable and Versatile 3D Generation”, CVPR 2025.

The Network Structures for Encoding, Decoding, and Generation





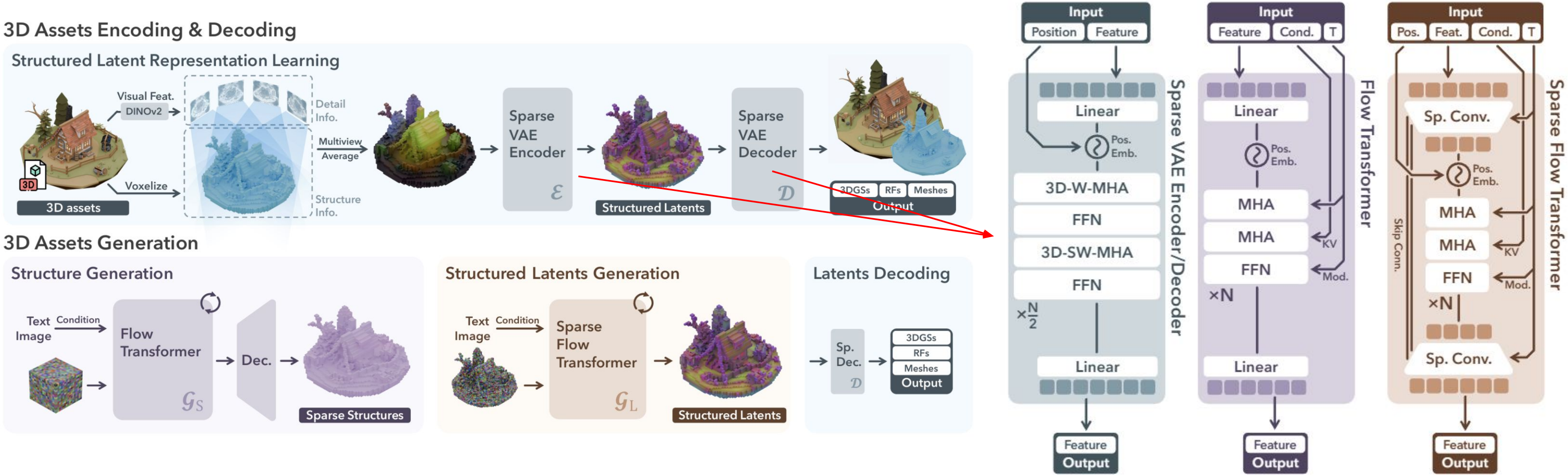
Encoding/Decoding

Sparse structure generation

Structured latents generation

Source: Xiang et al., “Structured 3D Latents for Scalable and Versatile 3D Generation”, CVPR 2025.

The Network Structures for Encoding, Decoding, and Generation



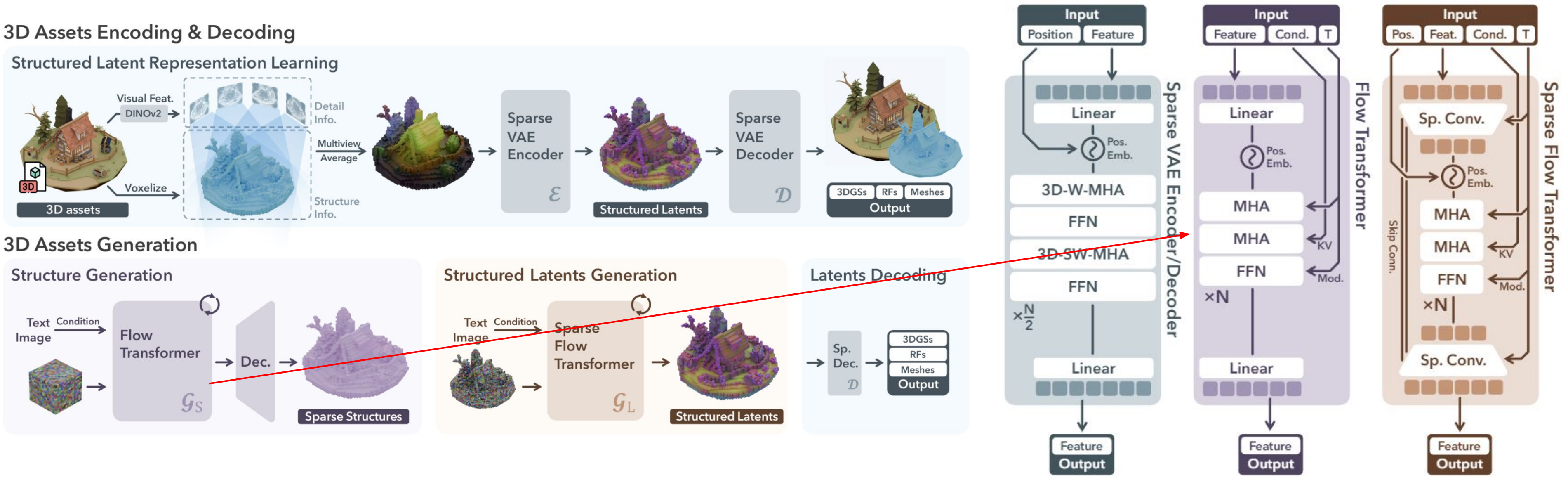
Encoding/Decoding

Sparse structure generation

Structured latents generation

Source: Xiang et al., "Structured 3D Latents for Scalable and Versatile 3D Generation", CVPR 2025.

The Network Structures for Encoding, Decoding, and Generation



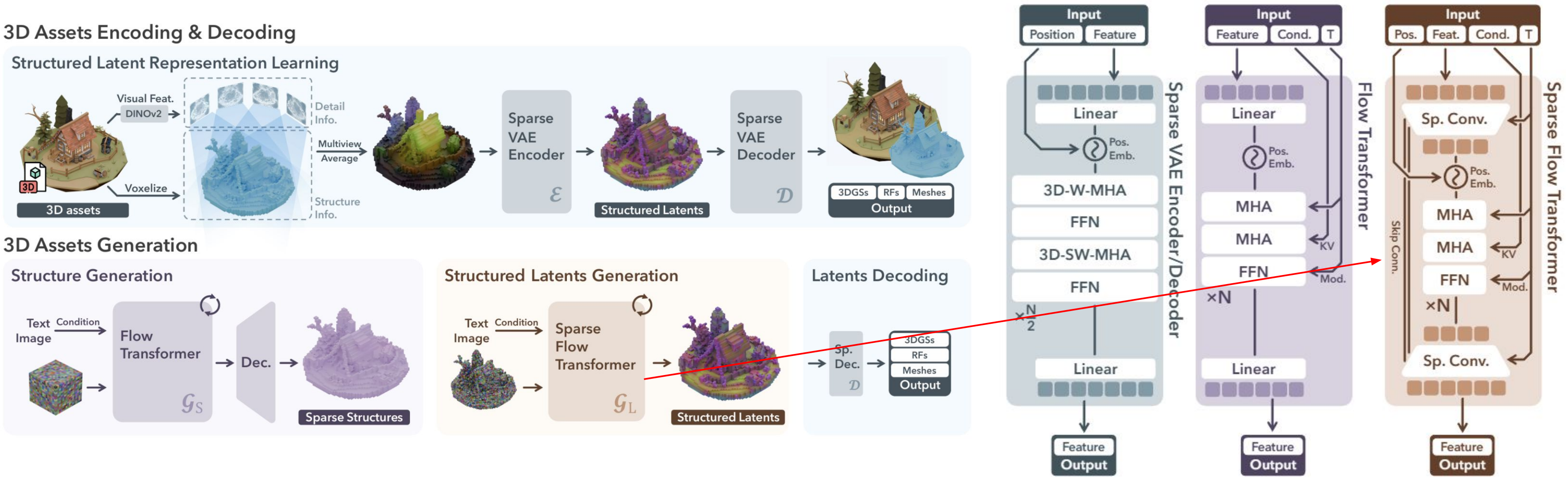
Encoding/Decoding

Sparse structure generation

Structured latents generation

Source: Xiang et al., "Structured 3D Latents for Scalable and Versatile 3D Generation", CVPR 2025.

The Network Structures for Encoding, Decoding, and Generation



Encoding/Decoding

Sparse structure generation

Structured latents generation

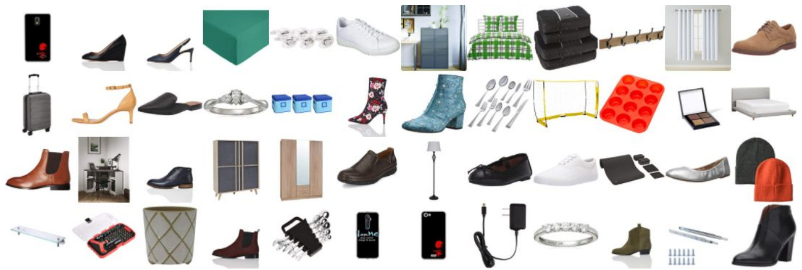
Source: Xiang et al., “Structured 3D Latents for Scalable and Versatile 3D Generation”, CVPR 2025.

Experiment

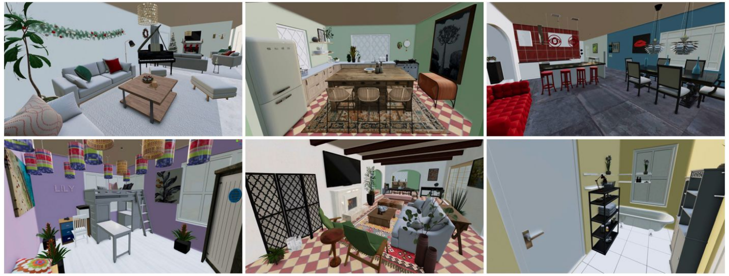
Training Set

Objaverse-XL

3D-Future



ABO



HSSD

31

Evaluation Set

Toys4K

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**

32

Results

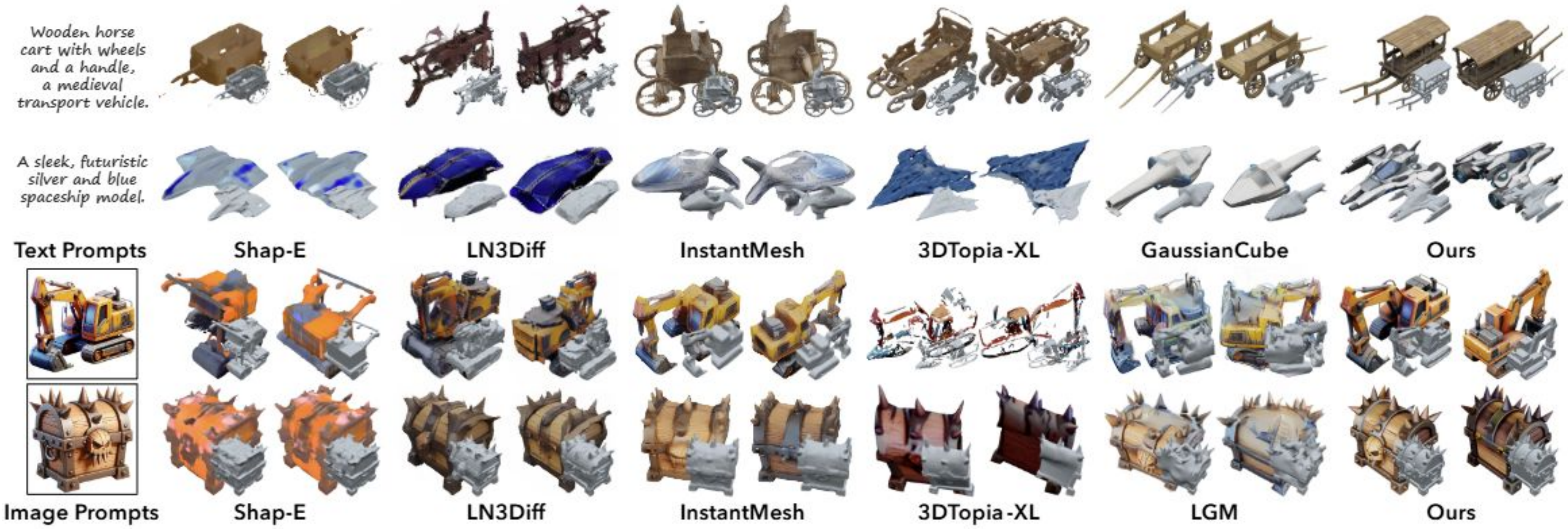


Figure 5. Visual comparisons of generated 3D assets between our method and previous approaches, given AI-generated prompts.

Results

Table 2. Quantitative comparisons using Toys4k [80]. (KD is reported ×100. †: evaluated using shaded images of PBR meshes.)

Method	Text-to-3D						Image-to-3D					
	CLIP ↑	FD_{incep} ↓	KD_{incep} ↓	$\text{FD}_{\text{dinov2}}$ ↓	$\text{KD}_{\text{dinov2}}$ ↓	FD_{point} ↓	CLIP ↑	FD_{incep} ↓	KD_{incep} ↓	$\text{FD}_{\text{dinov2}}$ ↓	$\text{KD}_{\text{dinov2}}$ ↓	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**	–	–	–	–	–	–

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 4. Ablation study on different generation paradigms.

	Method	Training set CLIP↑	Training set FD_{dinov2}↓	Toys4k CLIP↑	Toys4k 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**

Table 5. Ablation study on model size.

Method	Training set CLIP↑	Training set FD_{dinov2}↓	Toys4k CLIP↑	Toys4k 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**

- **Higher SLAT resolution (64^3) → major quality boost**; channel count matters less.
- **Rectified Flow > Diffusion** in both structure and latent generation (better quality + alignment).
- **Larger model size (B < L < XL) consistently improves results**.
- **Structured latents + flow models scale well**, enabling high-fidelity 3D generation.

Quiz

Thank you