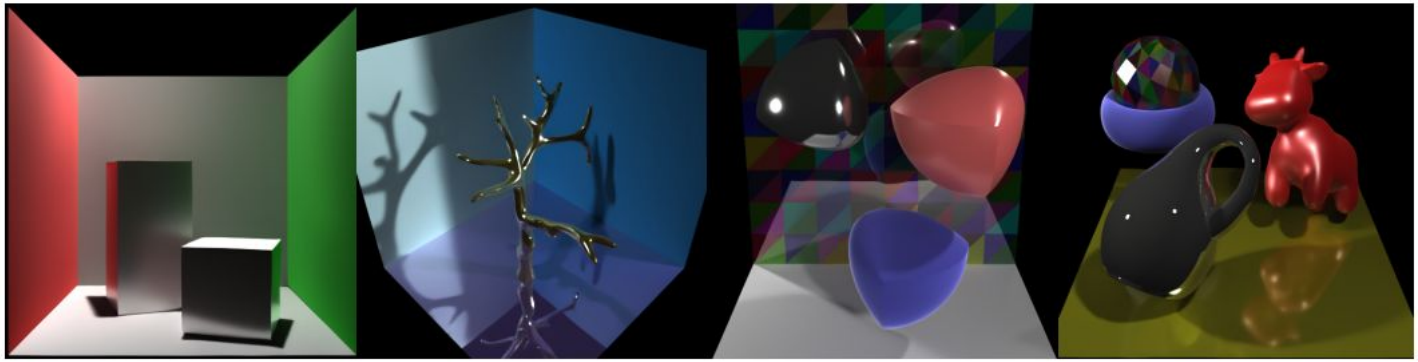
Quick recap: What is RenderFormer?

SIGGRAPH 2025

RenderFormer: Transformer-based Neural Rendering of Triangle Meshes with Global Illumination

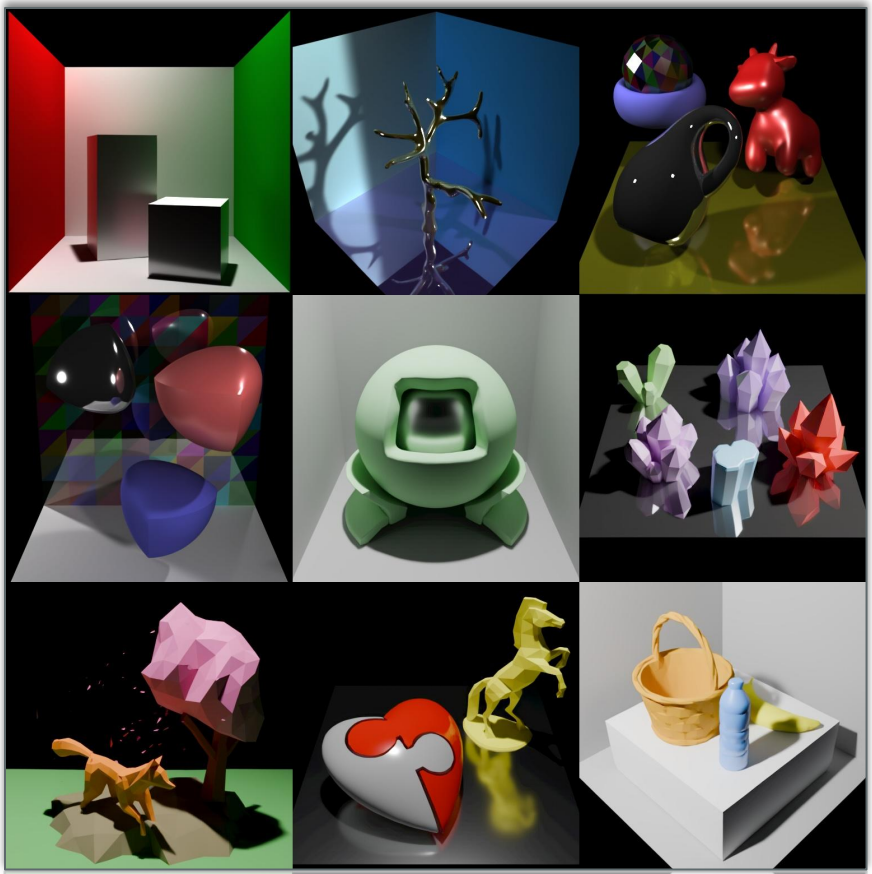
CHONG ZENG, State Key Lab of CAD & CG, Zhejiang University, China and Microsoft Research Asia, China
YUE DONG, Microsoft Research Asia, China
PIETER PEERS, College of William & Mary, USA
HONGZHI WU, State Key Lab of CAD & CG, Zhejiang University, China
XIN TONG, Microsoft Research Asia, China



https://microsoft.github.io/renderformer/

Summary

- PixelNeRF
- Instant-NGP
- Attention in Rendering
- Transformers' minimal 3D Inductive Bias



3D Generative Models

Student Presentation

2025.10.29

Minseo Park, Jewoo Shin, Sangmin Lee

Team 2

1. Single-view Image Generation Model based Distillation

Creating 3D object from 2D prior

“a highly detailed metal sculpture of a squirrel, wearing an elegant ballgown, riding a motorcycle”

Synthesizing a 3D object representation from a textual description

Creating 3D object from 2D prior

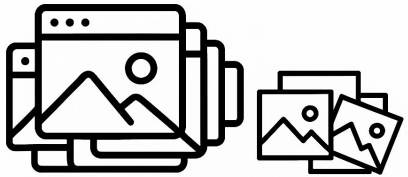
3D datasets

MVImgNet

Objaverse

Scarce!

2D datasets



Abundant

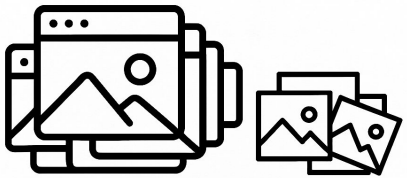
Internet-scale 2D images

Powerful diffusion model exists for generating 2D images

“Exhibit a severe imbalance, characterized by a scarcity of high-quality assets with complex geometric structures and rich surface details. [1]”

[1] Hi3DGen: High-fidelity 3D Geometry Generation from Images via Normal Bridging

Creating 3D object from 2D prior



Abundant

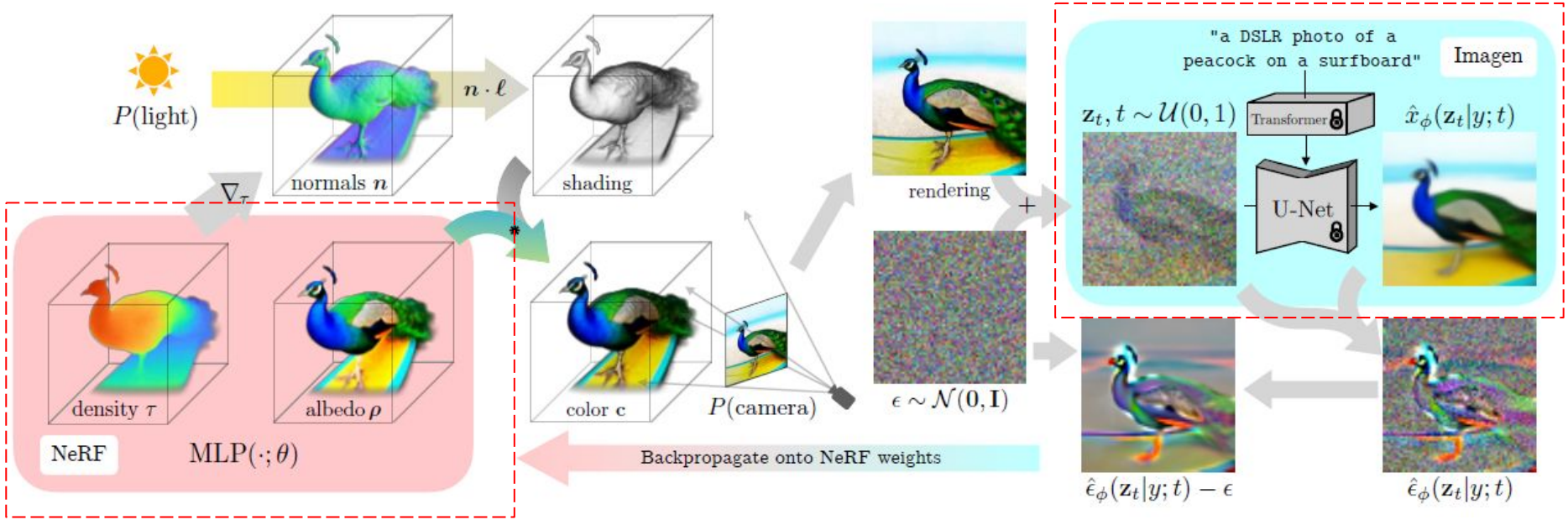
Internet-scale 2D images

Powerful diffusion model exists for generating 2D images

“Can 3D objects be generated using the prior knowledge of 2D diffusion models?”

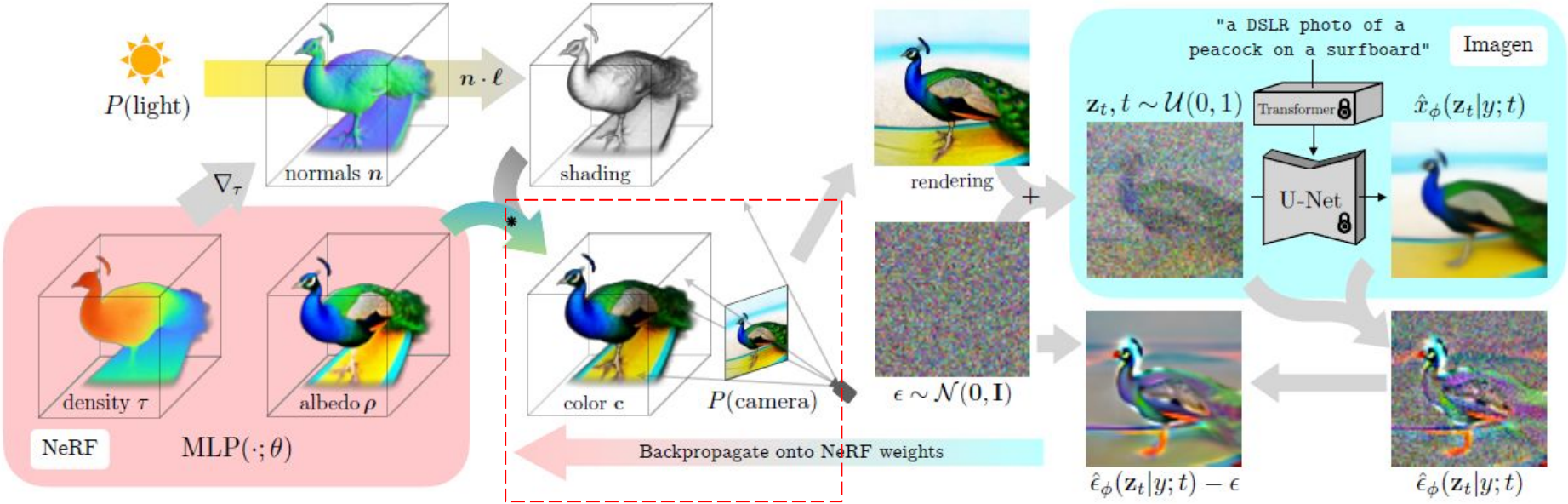
Inside the DreamFusion Pipeline

Well trained 2D diffusion model



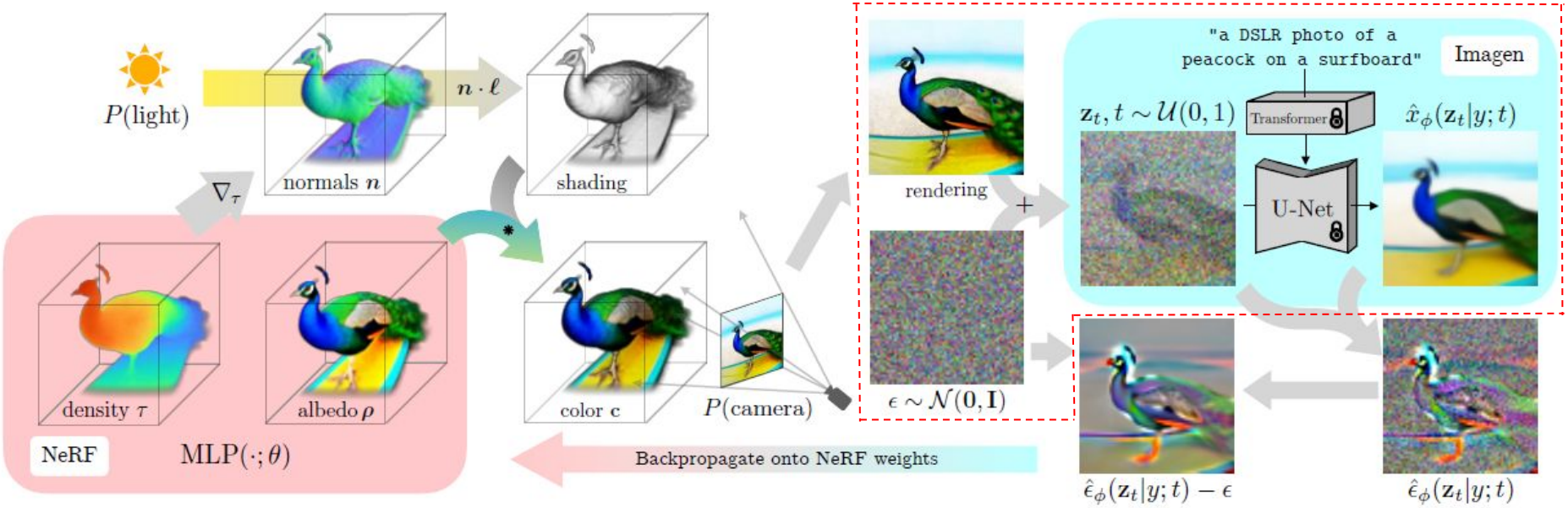
1. NeRF creates 3D mesh from current state

Inside the DreamFusion Pipeline



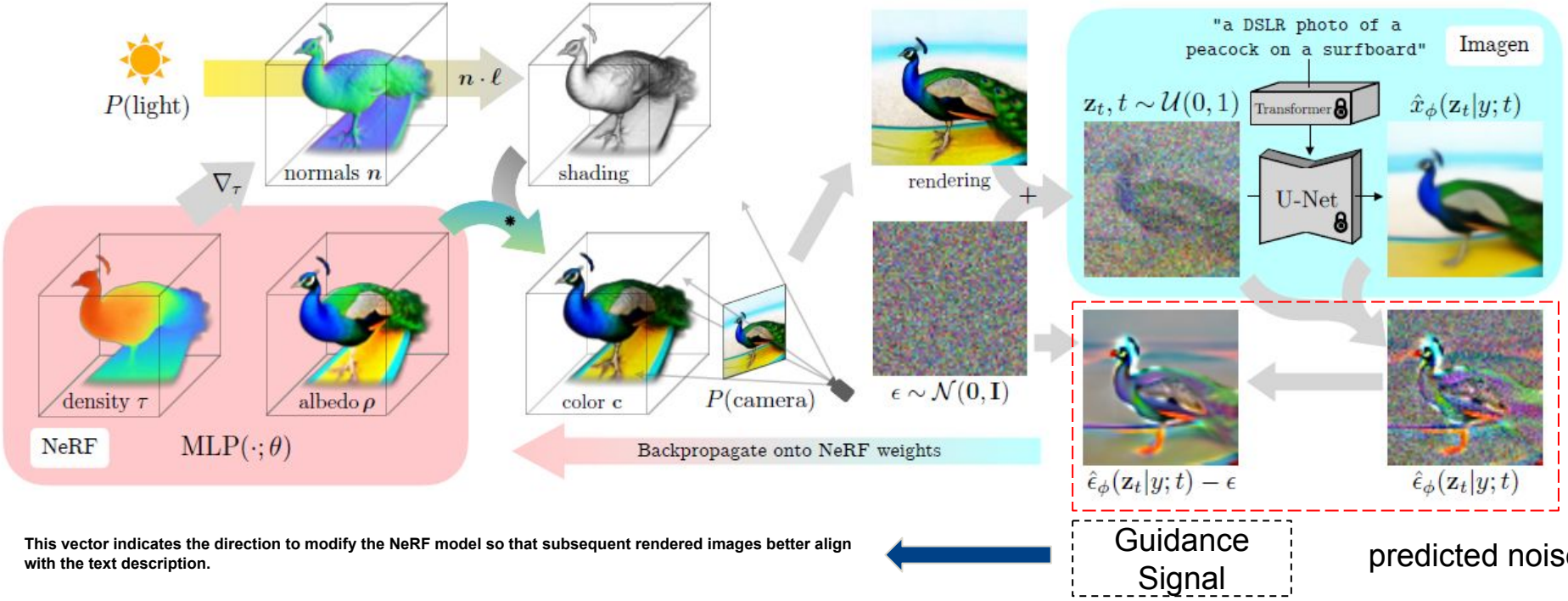
2. Render 2D images of the 3D object from multiple random viewpoints.

Inside the DreamFusion Pipeline



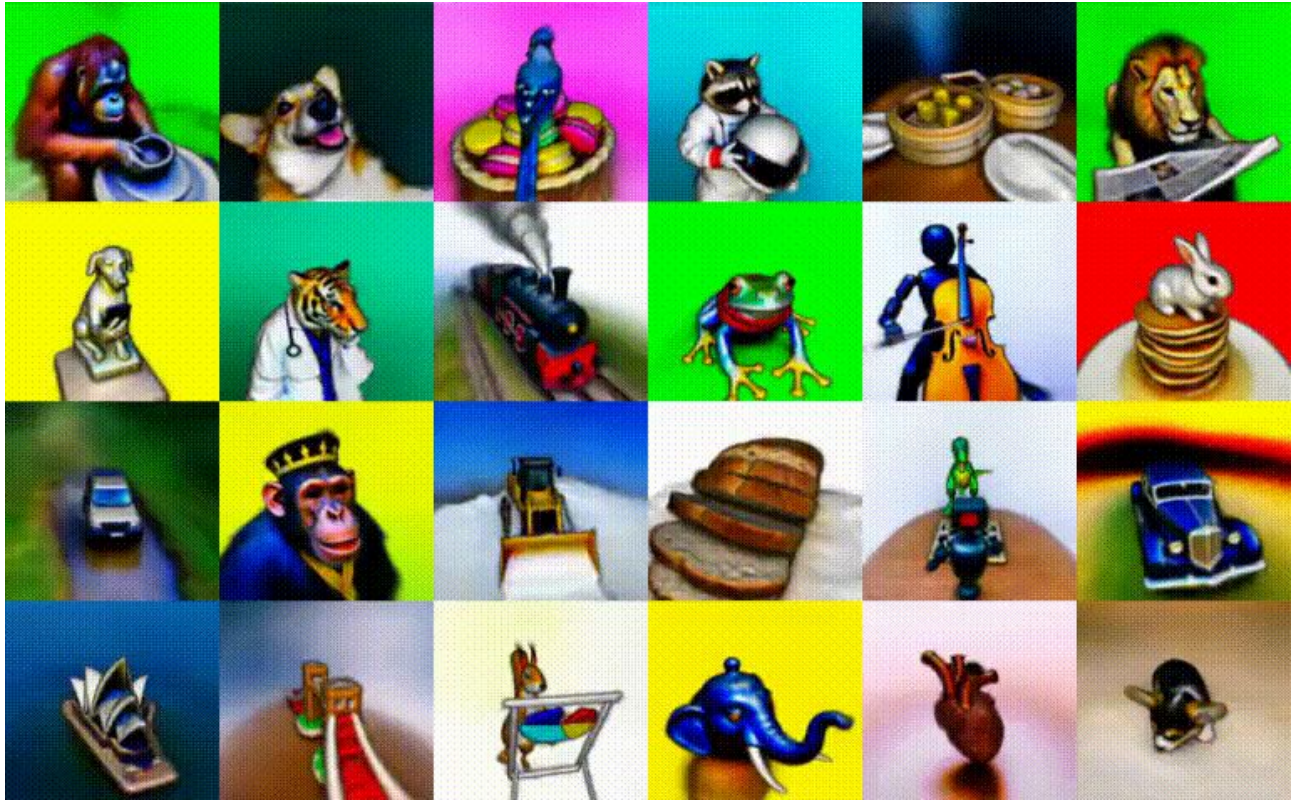
3. Rendered 2D view of the 3D scene is perturbed by adding Gaussian noise, forming a noisy latent representation z_t. The text-conditioned diffusion model (Imagen) predicts and removes the noise $\epsilon\hat{}\phi(z_t | y;t)$, progressively denoising the image to align with the text prompt and refine the 3D scene representation.

Inside the DreamFusion Pipeline



4. Refine the 3D model (NeRF) based on the 2D model's feedback

Results



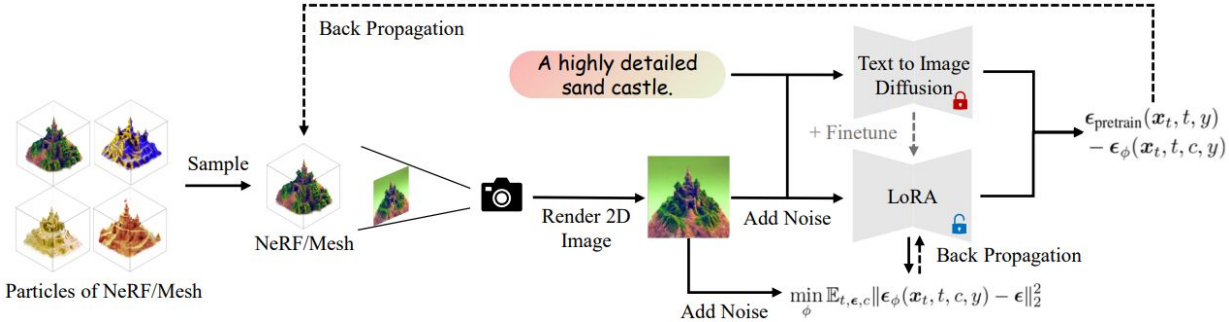
3D generated objectes from dreamfusion

2. Multi-view Image Generation Model based Distillation

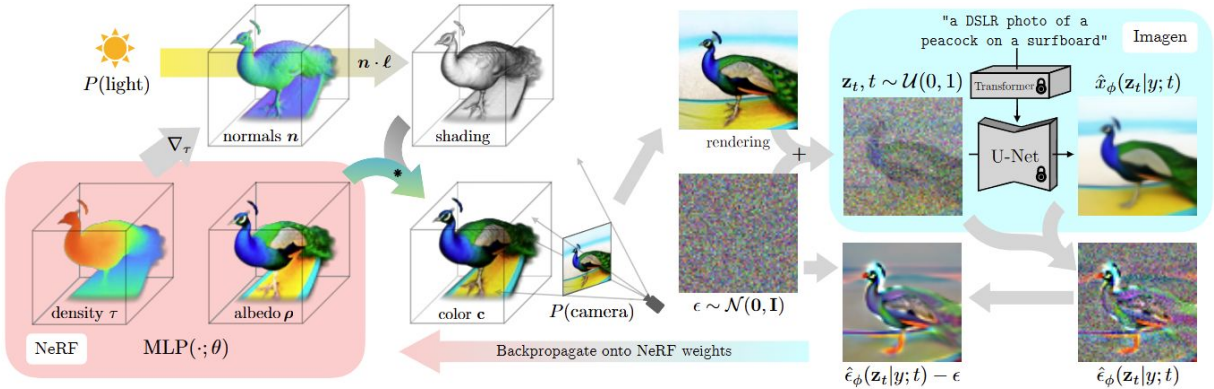
Lifting Single-view 2D-Diffusion Models

Previous Methods:

ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation



Single-view 2D-Diffusion Model (**ProlificDreamer**)

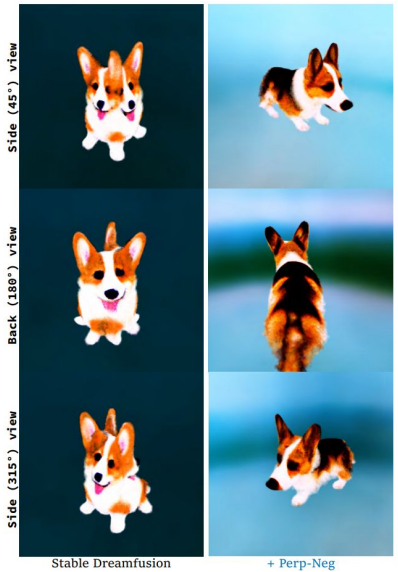


Single-view 2D-Diffusion Model (**DreamFusion**)

Previous Limitations

- **Problem**: Early Text-to-3D methods (e.g., DreamFusion) using single-view 2D priors suffer from critical flaws:
  - **Janus Problem & Content Drift**: Multi-view inconsistency (e.g., faces on both sides of a head).
  - **Bias**: A strong tendency to generate only the "front" or canonical view.
  - **Hollow face illusion**: Difficulty understanding 3D shapes (concave/convex errors).
- **Solution**: Replace the 2D "teacher" with a 3D-aware model. This is achieved by:
  - Training on 3D asset datasets (like Objaverse).
  - Or, more commonly, extending powerful 2D models into Multi-View Diffusion Models (e.g., MVDream, Zero1-to-3) that can generate consistent views.
- **Impact**: Using these multi-view models with distillation (like SDS) solves the Janus problem, improves geometric accuracy, and enables high-fidelity 3D generation.

Previous Limitations



Janus problem

Figure 2: **Viewpoint bias in text-to-image models.** We show samples from both Dall-E-2 and Stable Diffusion v2 from the prompt *"a chair"*. Most samples show a chair in a forward-facing canonical pose.

Bias problem

Re-imagine the Negative Prompt Algorithm: Transform 2D Diffusion into 3D, alleviate Janus problem and Beyond (Perp-Neg)

Zero-1-to-3: Zero-shot One Image to 3D Object

Previous Limitations

Hollow face illusion problem

"an austronaut"

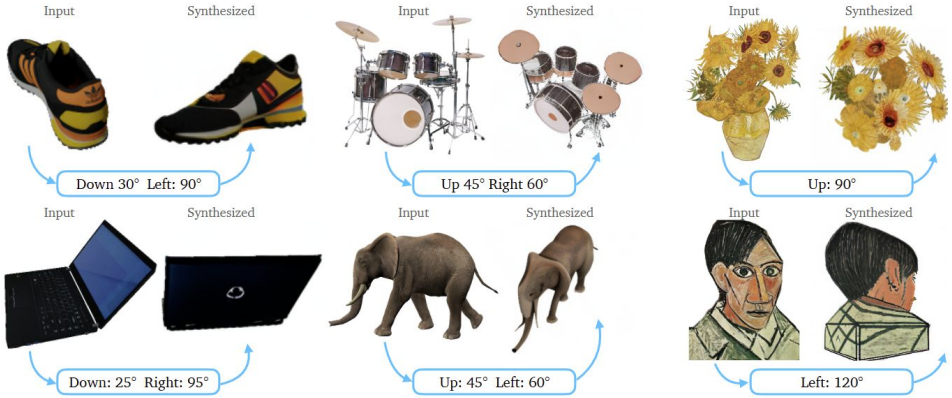
Figure 16: **Hollow Face Illusion:** some shapes might be generated with a concave hole in the geometry, while the rendering appears to be convex.

Content drift problem

"a DSLR photo of a plate of fried chicken and waffles with maple syrup on them"

Score Distillation via Reparametrized DDIM, MVDream: Multi-view Diffusion for 3D Generation

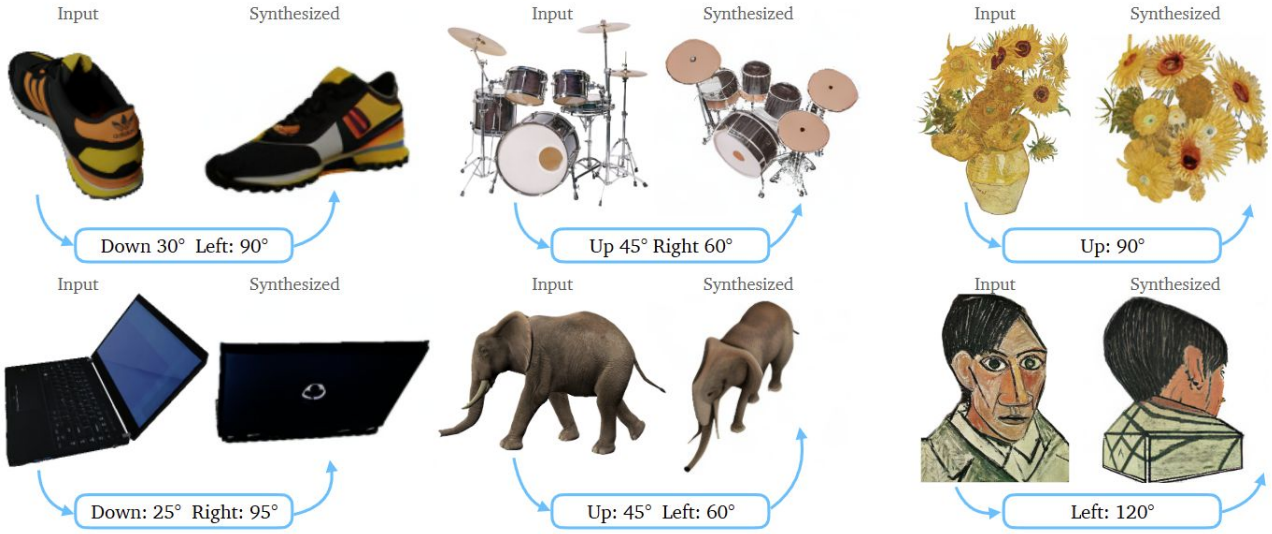
Zero-1-to-3: Zero-shot One Image to 3D Object



Novel View Synthesis

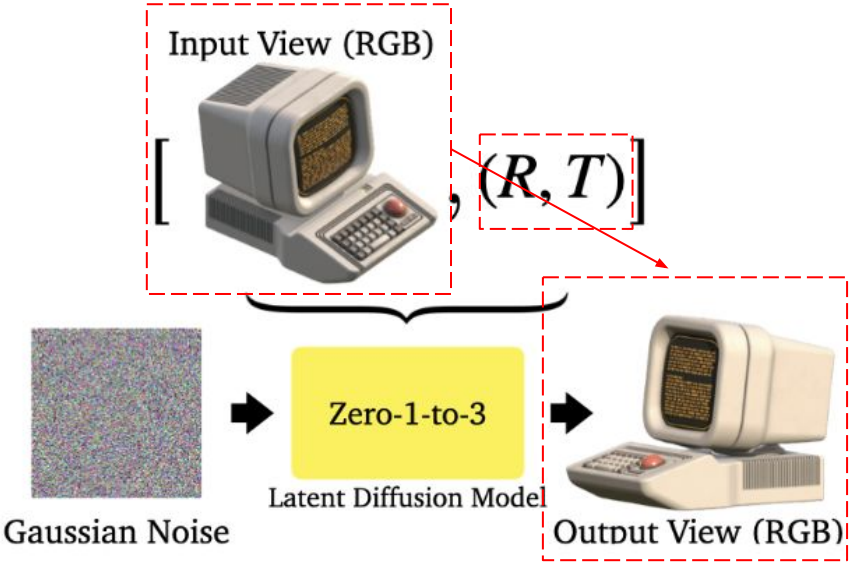
3D Reconstruction

Train view-conditioned diffusion model



1. Train view-conditioned diffusion model (or viewpoint-conditioned image translation model, since it controls camera extrinsic)

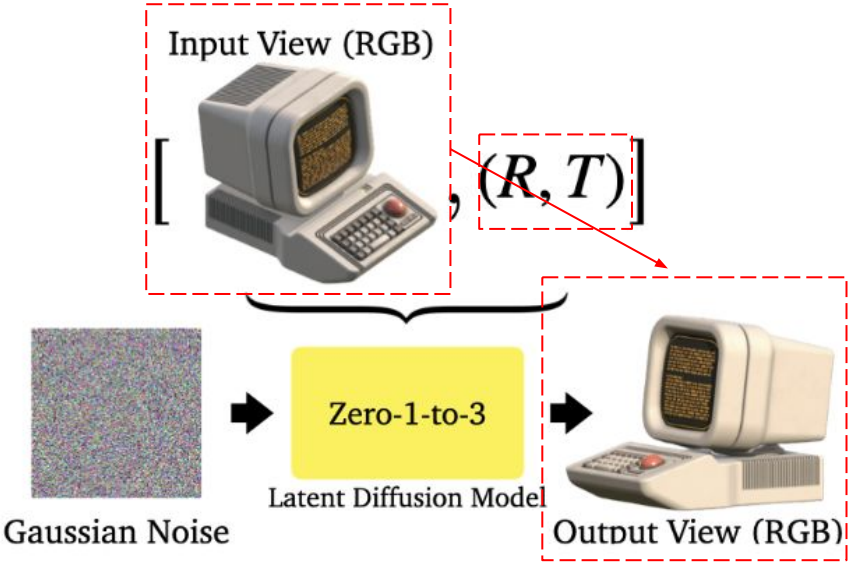
Train view-conditioned diffusion model



Novel View Synthesis

1. Train view-conditioned diffusion model (or viewpoint-conditioned image translation model, since it controls camera extrinsic)

Train view-conditioned diffusion model



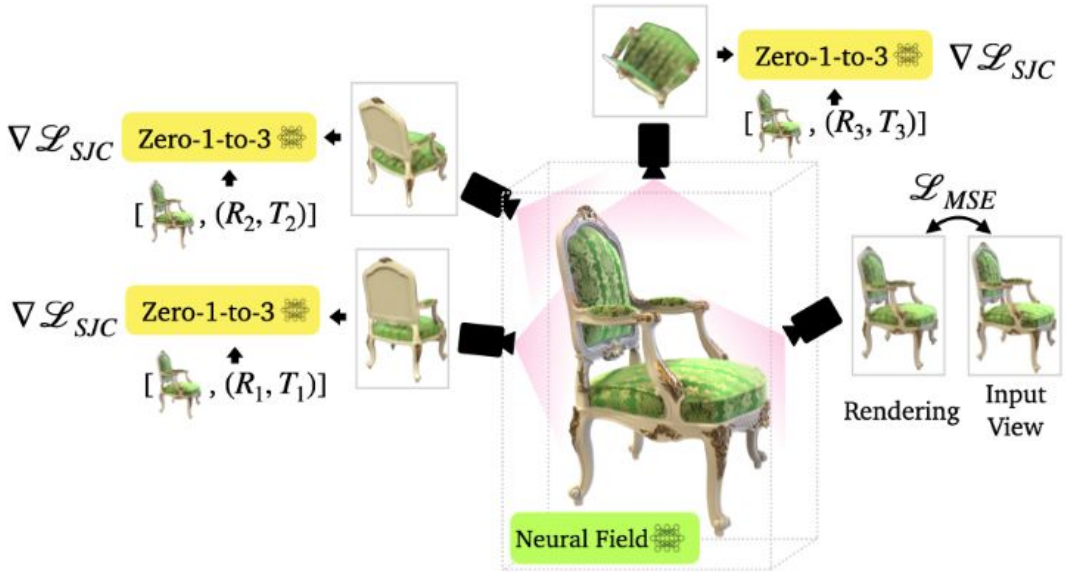
$c(x, R, T)$ → embedding of input view x and relative camera extrinsics (R, T)

$$\min_{\theta} \mathbb{E}_{z\sim\mathcal{E}(x),t,\epsilon\sim\mathcal{N}(0,1)}||\epsilon - \epsilon_\theta(z_t, t, c(x, R, T))||_2^2.$$

Novel View Synthesis

1. Training loss is same as original diffusion but text condition substituted with c(x, R, T) which is embedding of input view and relative camera extrinsics

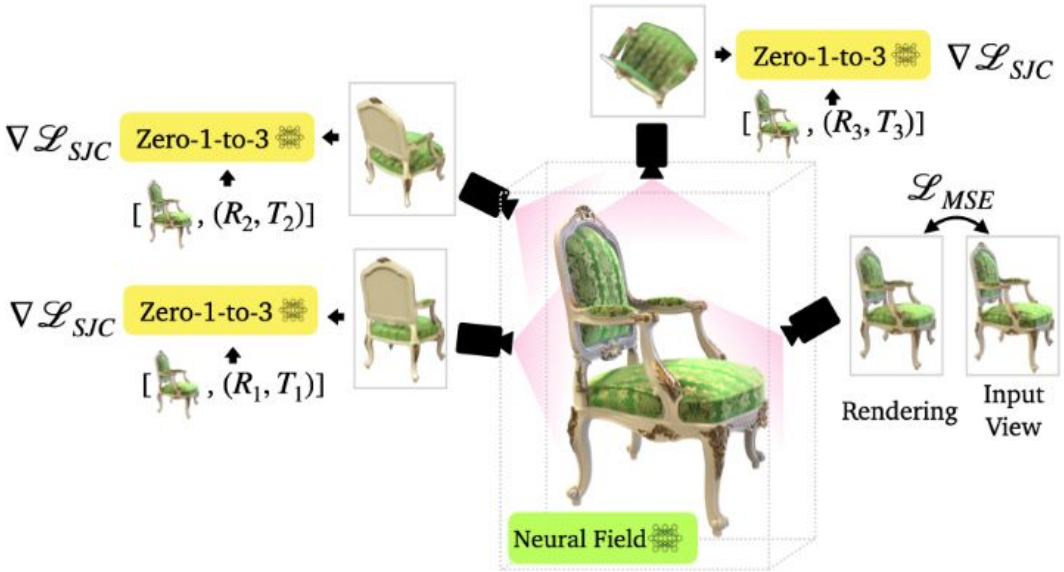
3D Reconstruction



3D Reconstruction

2. 3D Reconstruction, view-conditioned diffusion guides and refines a NeRF scene from multiple viewpoints to create a realistic 3D object.

SJC?



3D Reconstruction

2. 3D Reconstruction, view-conditioned diffusion guides and refines a NeRF scene from multiple viewpoints to create a realistic 3D object.

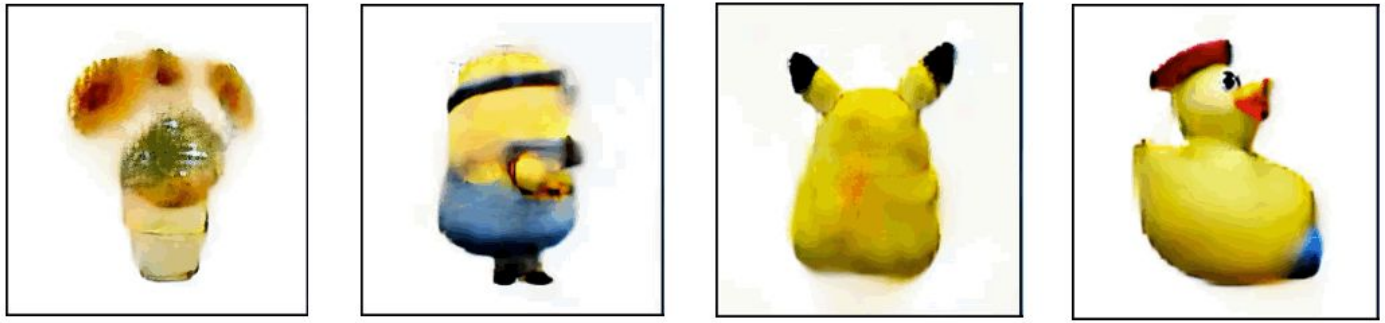
Results: Novel view synthesis on in-the-wild images

Stage 1
Results

Results: Single-View 3D Reconstruction (Texture)

Input View

3D Reconstruction (Stage 2 Results)



Results: Qualitative examples of 3D reconstruction

Comparison with other methods

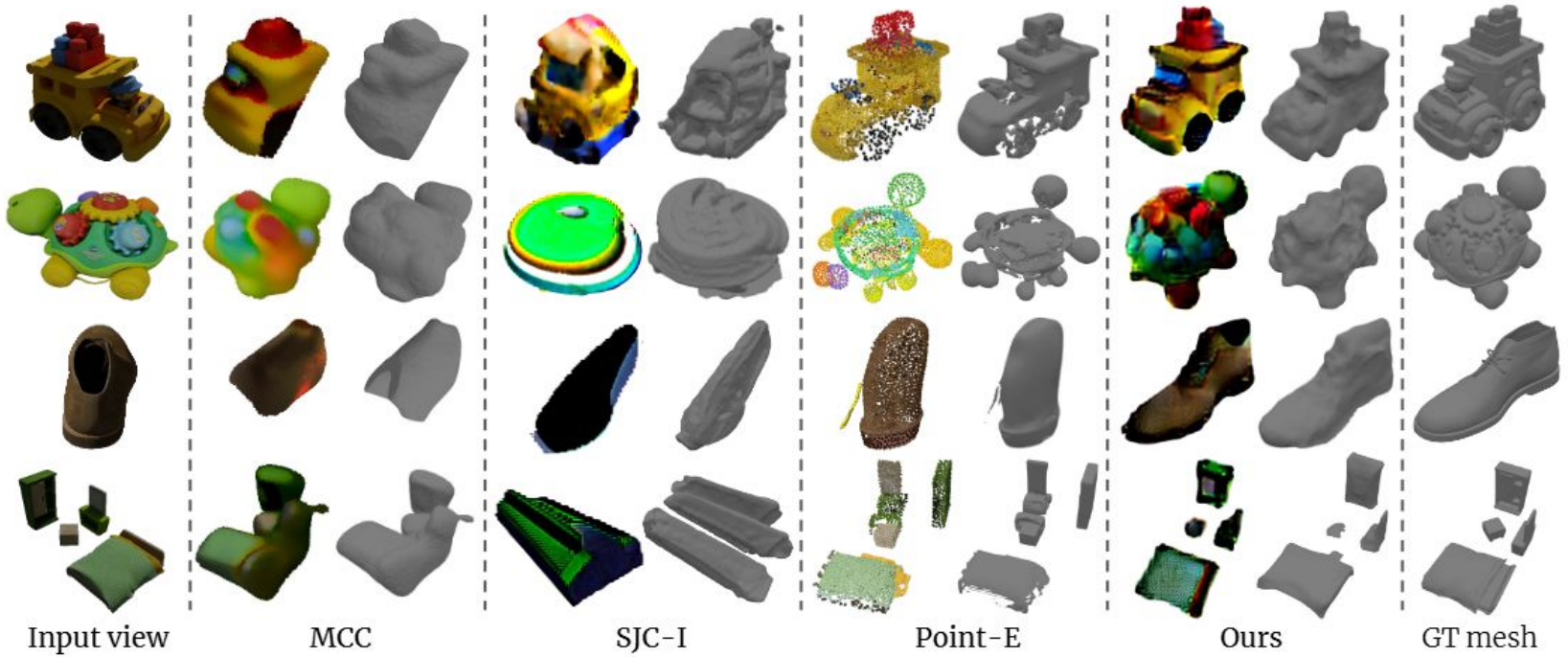
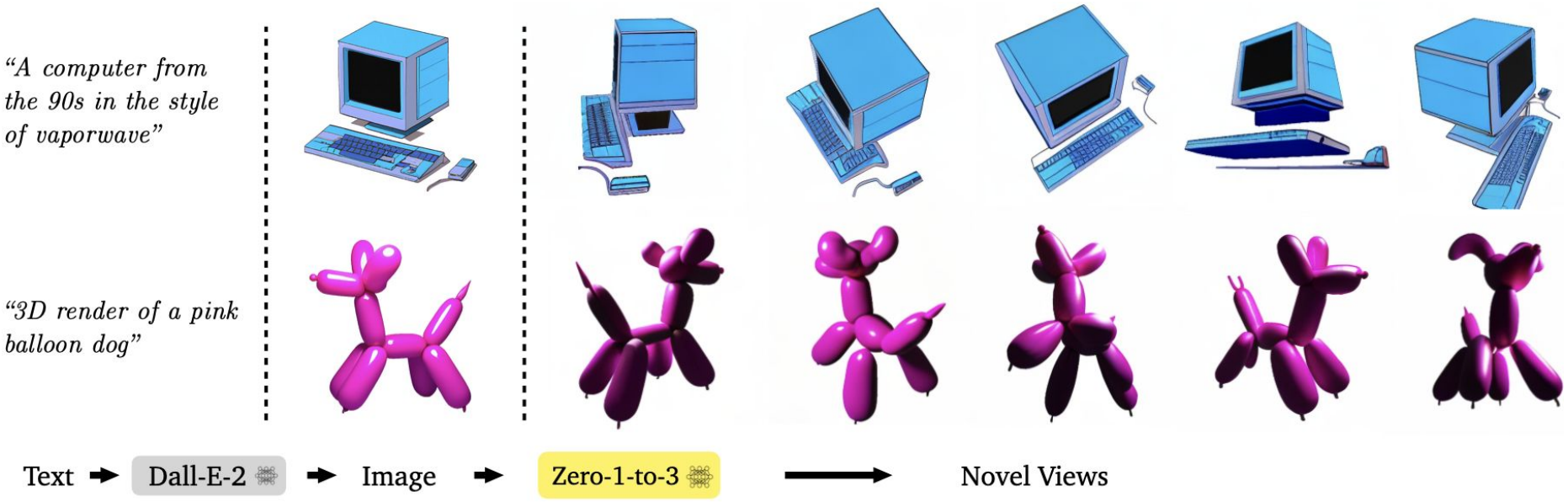
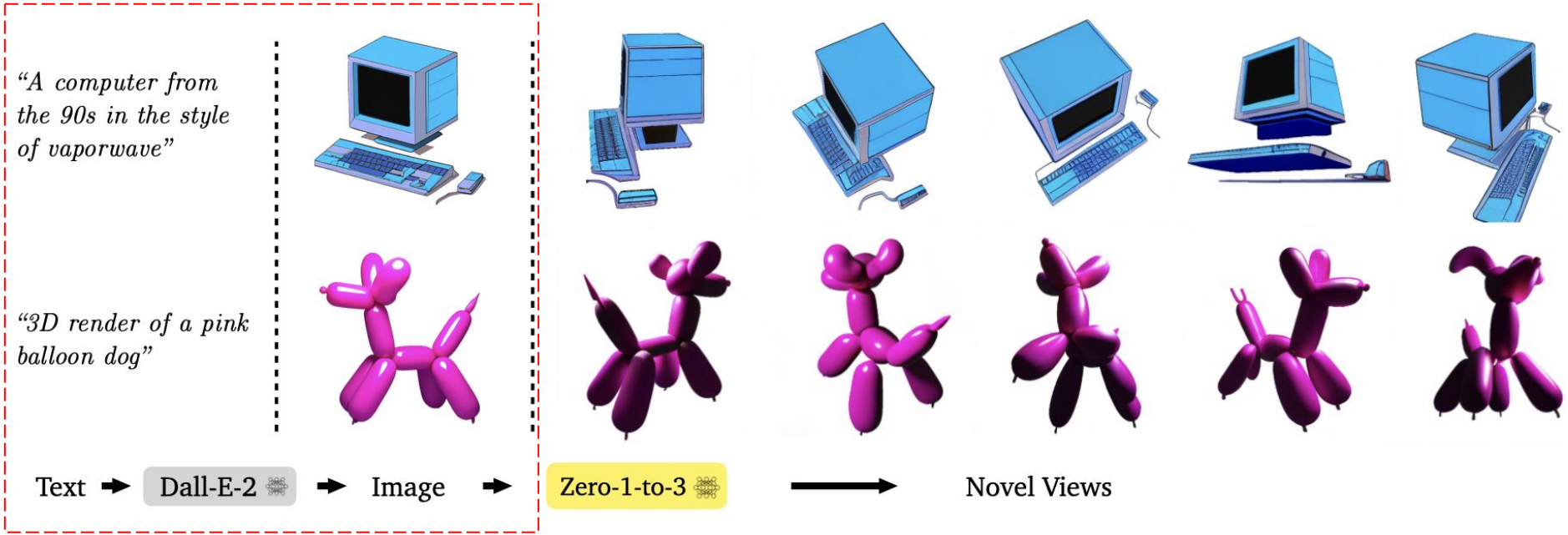


Figure 9: **Qualitative examples of 3D reconstruction.** The input view images are shown on the left. For each method, we show a rendered view from a different angle and the reconstructed 3D mesh. The ground truth meshes are shown on the right.

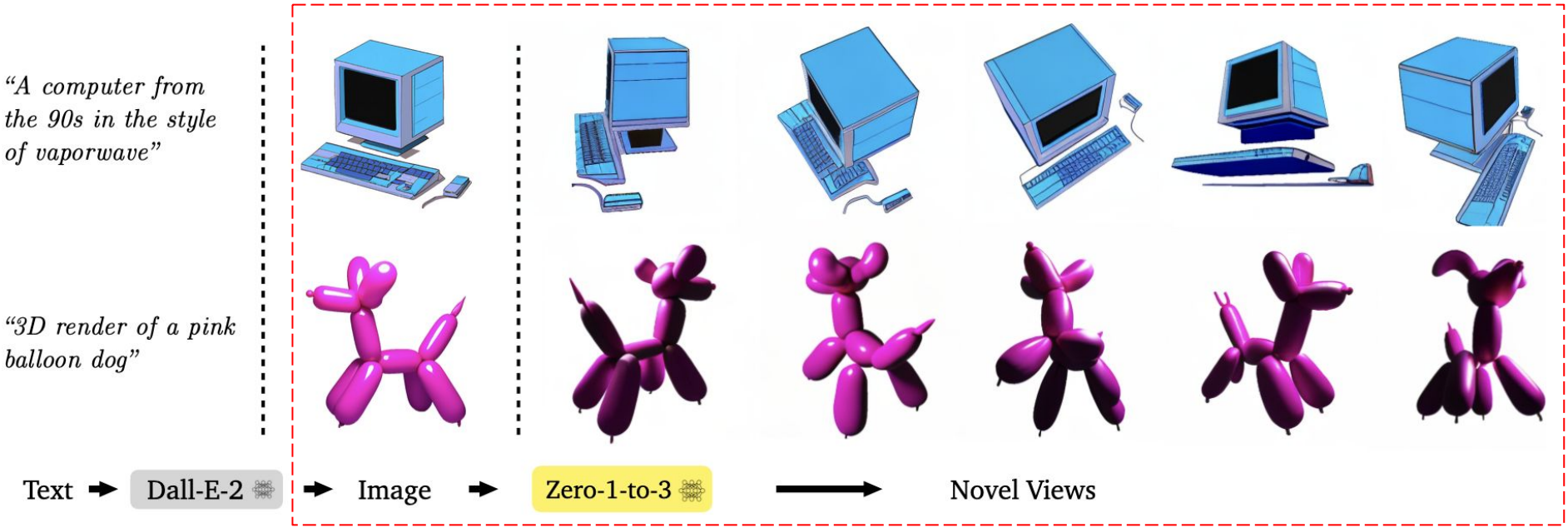
Results: Text to Image to Novel Views



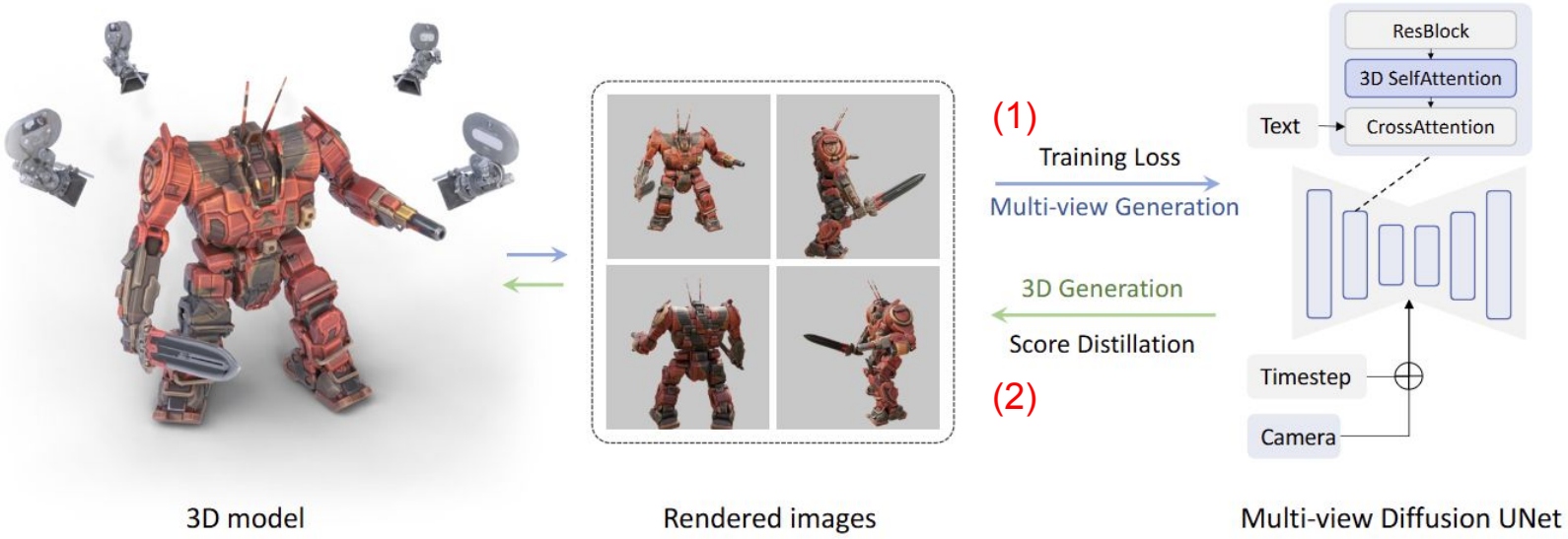
Results: Text to Image to Novel Views



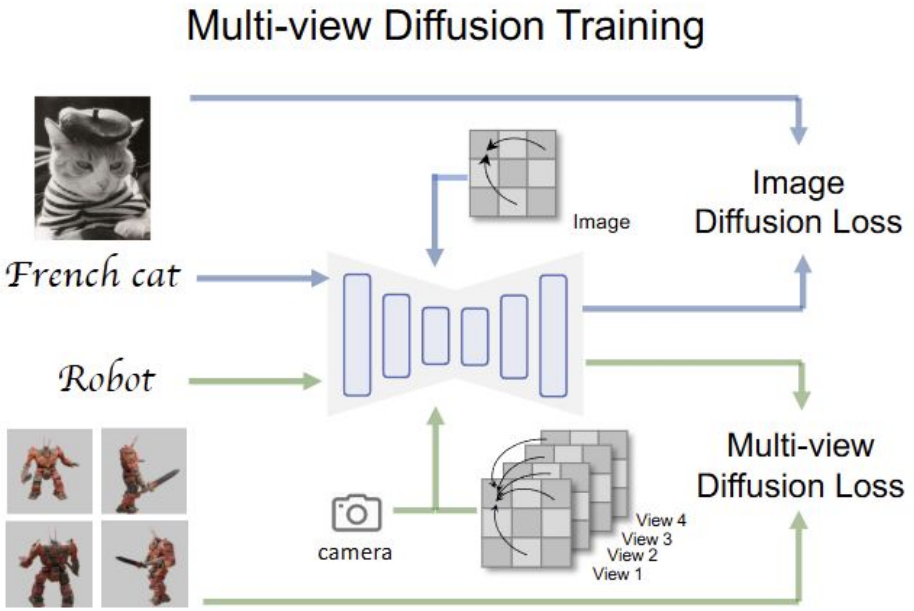
Results: Text to Image to Novel Views



MVDream: Multi-view Diffusion for 3D Generation

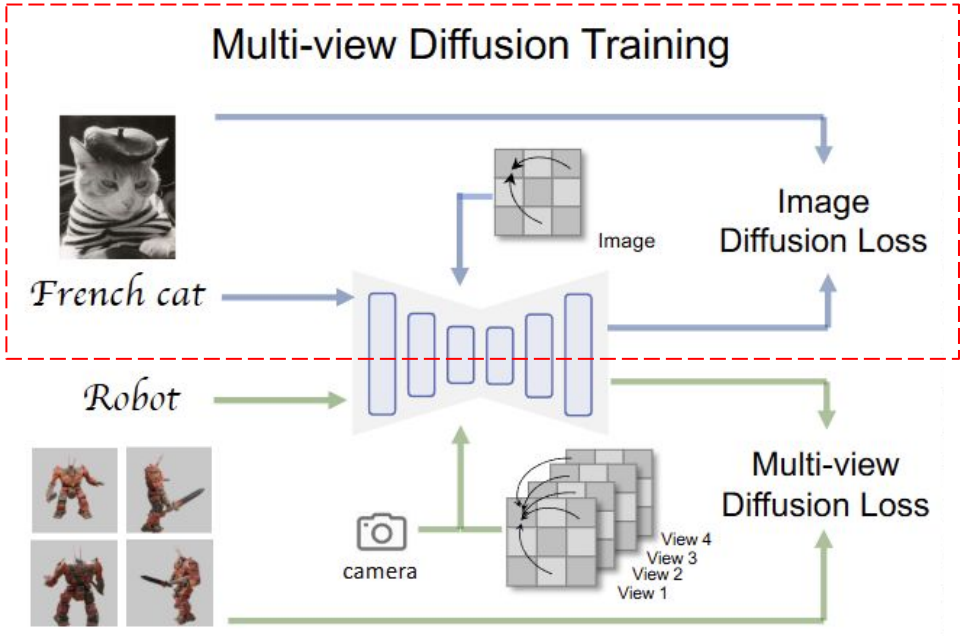


Train Multi-view Consistent Image Generation model



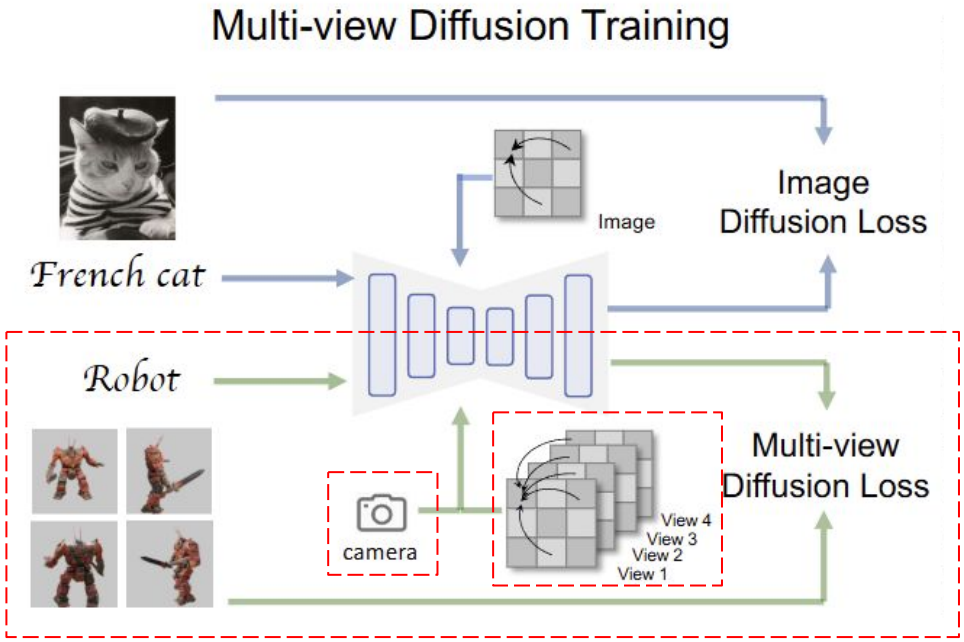
1. Train Multi-view Diffusion model with two modes: image mode with 2D attention (upper) and multi-view mode with 3D attention and camera embeddings (lower)

Train Multi-view Consistent Image Generation model



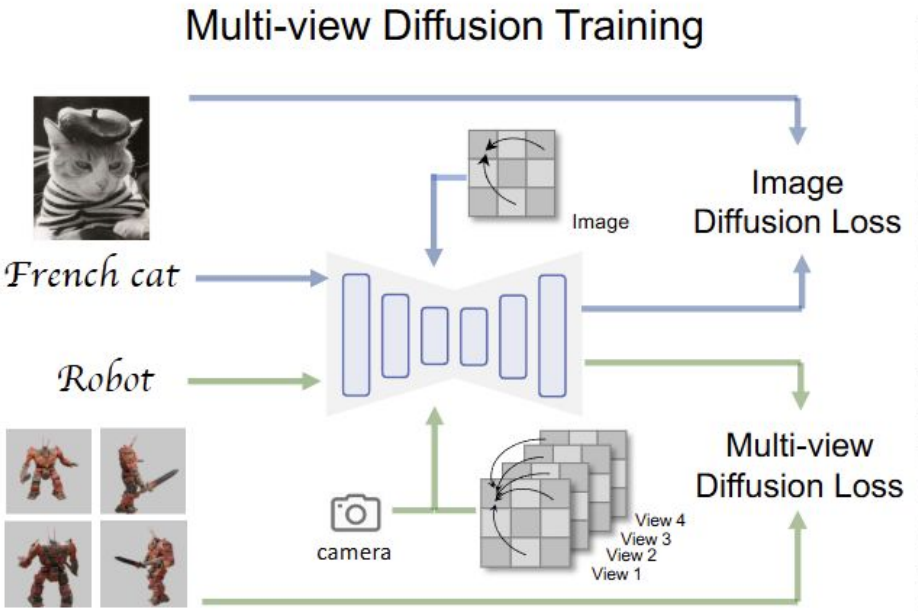
1. Train Multi-view Diffusion model with two modes: image mode with 2D attention (upper) and multi-view mode with 3D attention and camera embeddings (lower)

Train Multi-view Consistent Image Generation model



1. Train Multi-view Diffusion model with two modes: image mode with 2D attention (upper) and multi-view mode with 3D attention and camera embeddings (lower)

Train Multi-view Consistent Image Generation model

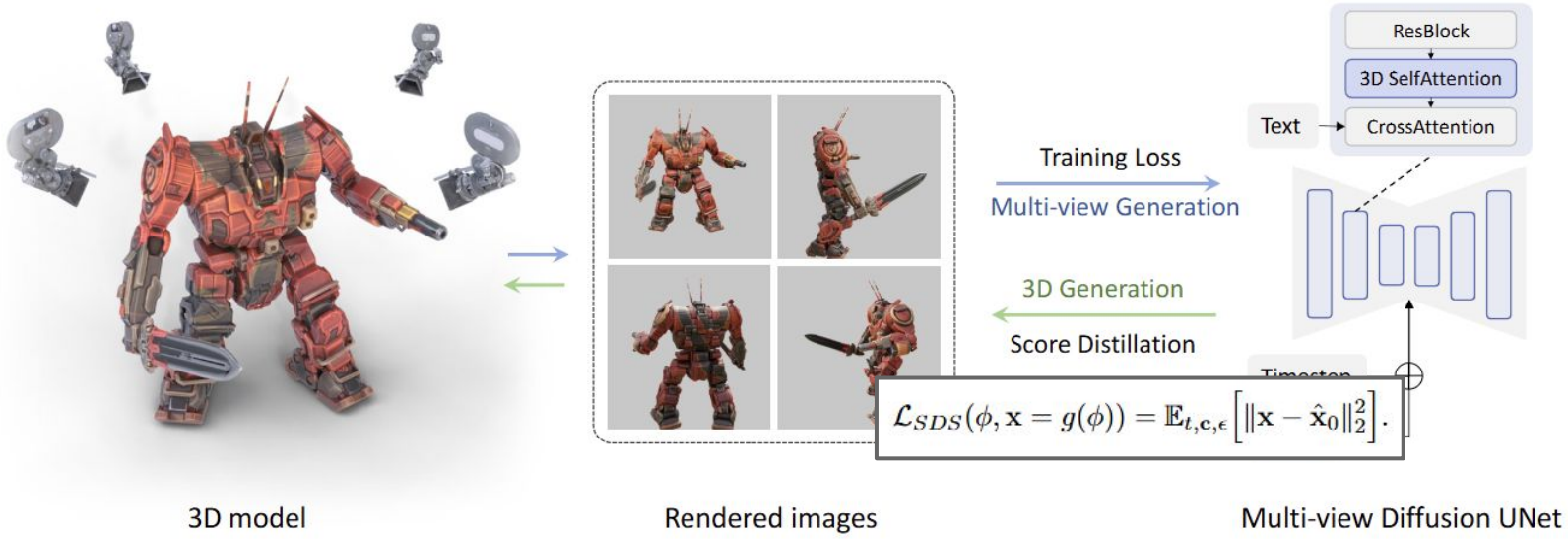


text-image dataset $\mathcal{X}$
multi-view dataset $\mathcal{X}_{mv}$
$\{\mathbf{x}, y, \mathbf{c}\} \in \mathcal{X} \cup \mathcal{X}_{mv}$

$$\mathcal{L}_{MV}(\theta, \mathcal{X}, \mathcal{X}_{mv}) = \mathbb{E}_{\mathbf{x},y,\mathbf{c},t,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t; y, \mathbf{c}, t)\|_2^2\right]$$

1. Training loss is same as original diffusion but with additional camera embeddings. At training we sample $\{\mathbf{x}, y, \mathbf{c}\} \in \mathcal{X} \cup \mathcal{X}_{mv}$ ($\mathbf{c}$ is empty for $\mathcal{X}$)

Text-to-3D Generation



2. Text-to-3D Generation, use multi-view diffusion model as prior of Score Distillation Sampling (SDS)

Results: Multi-view Generation

Stage 1 Results

Zombie bust, terror, 123dsculpt, bust, zombie

Battletech Zeus with a sword!, tabletop, miniature, battletech, miniatures, wargames, 3d asset

Medieval House, grass, medieval, vines, farm, middle-age, medieval-house, stone, house, home, wood, medieval-decor, 3d asset

Isometric Slowpoke Themed Bedroom, fanart, pokemon, bedroom, assignment, isometric, pokemon3d, isometric-room, room-low-poly, 3d asset

An astronaut riding a horse

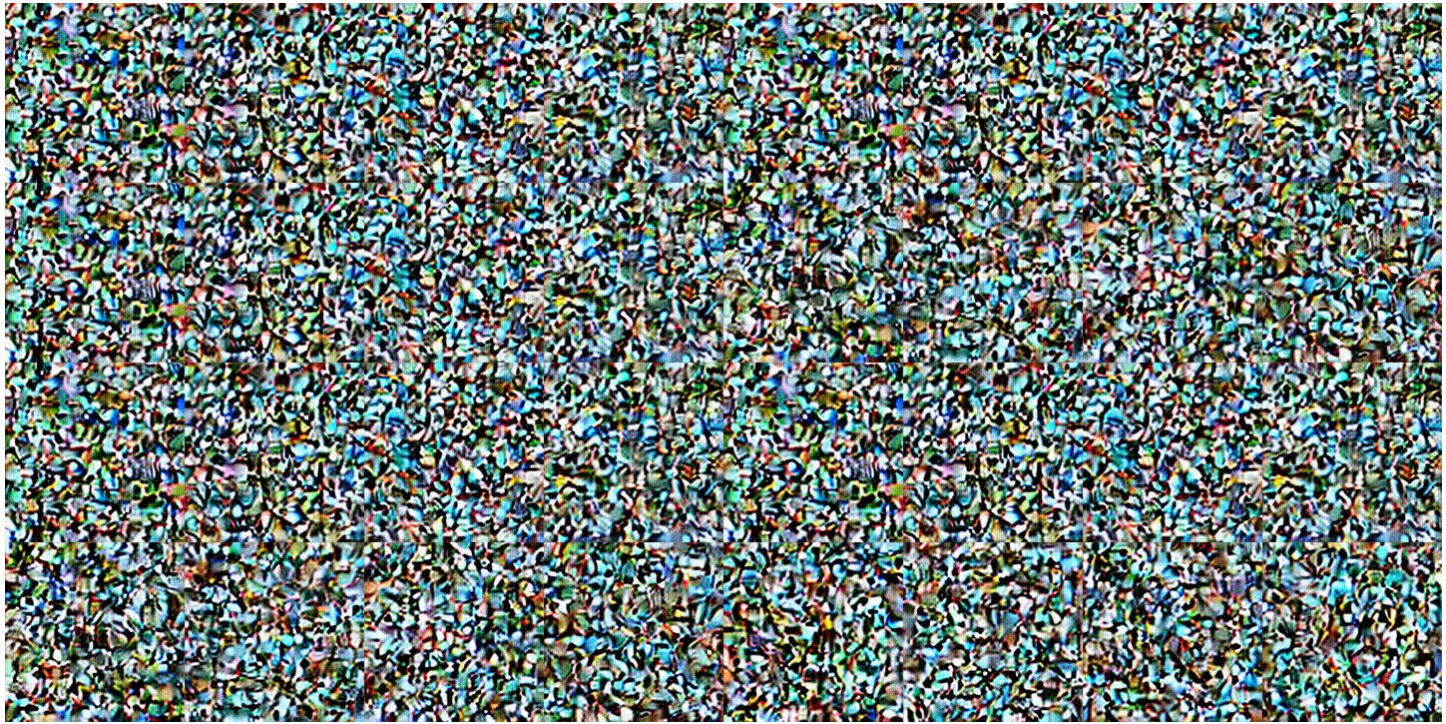
A bald eagle carved out of wood

A bull dog wearing a black pirate hat

a DSLR photo of a ghost eating a hamburger

Results: Multi-view Generation

Stage 1
Results



Results: Multi-view Score Distillation (Text-to-3D)

Stage 2
Results

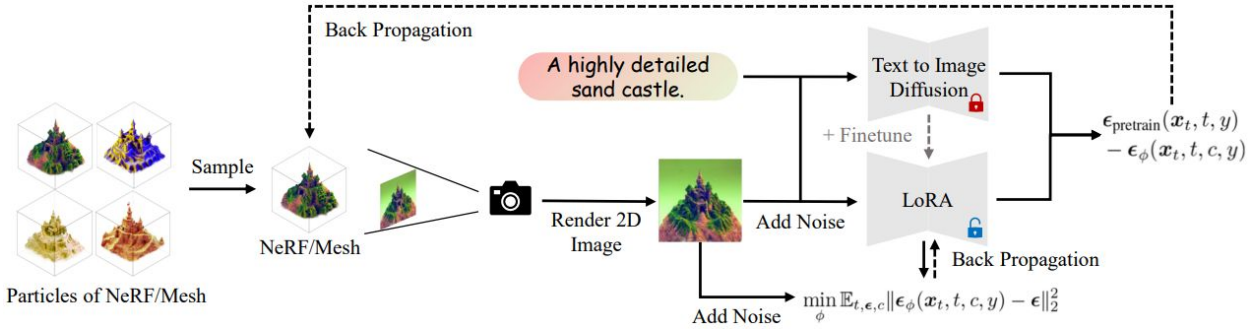
3. Feed-Forward Models

Previous Limitations

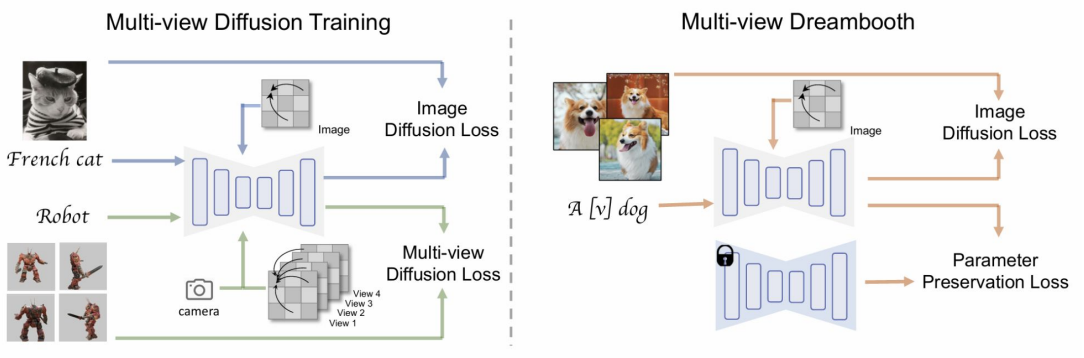
- **Problem**
  - Too slow, rendering time takes from **1 hours ~ 8 hours** for single scene.
  - 3D Gaussian Splatting reduced a lot, but since there is quality-time tradeoff we need more amount of time for better quality.
- **Solution**
  - Nowadays, many large 3D datasets exists -> it’s time for **training**!
  - We have learned a lot of recipes for consistent 3D generation. And feedforward method will reduce the time significantly if the model is well trained.

Lifting 2D-Diffusion Models

Previous Methods:



Single-view 2D-Diffusion Model (**ProlificDreamer**)



Multi-view 2D-Diffusion Model (**MVDream**)

Recent 3D Generative Models

MVDream

Hunyuan 3D

- From a Single Image
- High-Quality Geometry
- 3D Consistency
- Fast & Versatile
- High-Resolution Texture
  (Not covered in this lecture)

How?

Key Ideas

Neural Representation of 3D Shapes

- Primary representation methods: mesh, pointcloud, voxels, implicit functions, …
- Vector Sets
- Structured Latents (triplane, sparse volume, ..)

Shape Generative models

- Transformer-based architecture
- VAE

3D Datasets

- Growth of open-source datasets with high-resolution 3D assets.

Neural Representation of 3D Shapes

- Primary representation methods: mesh, pointcloud, voxels, implicit functions, …

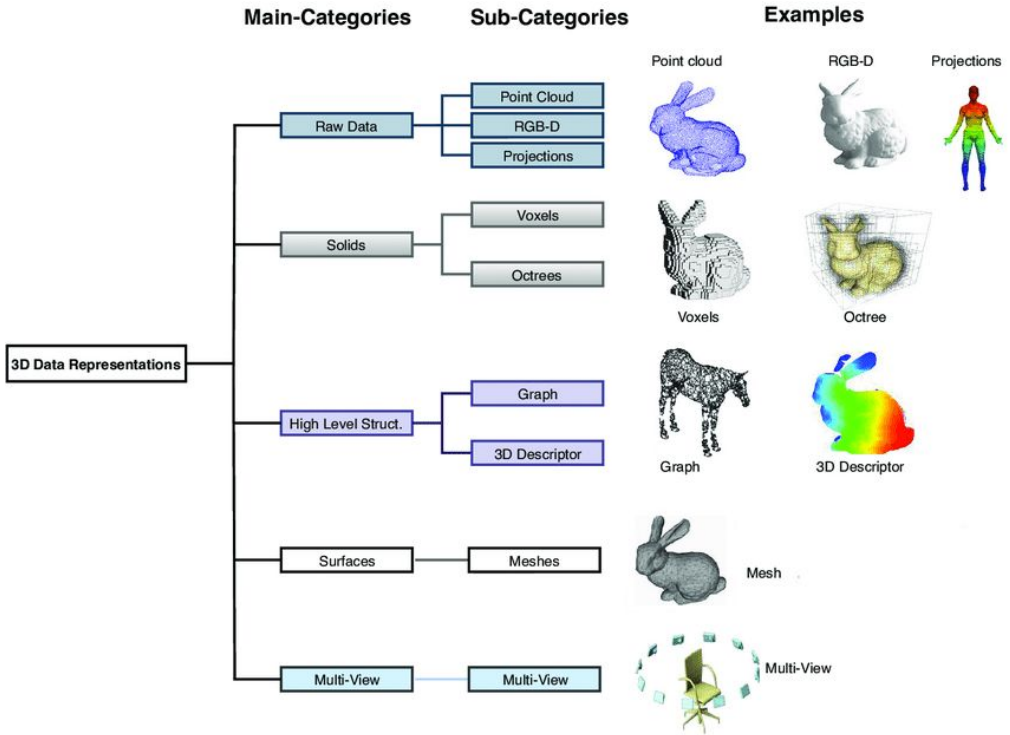
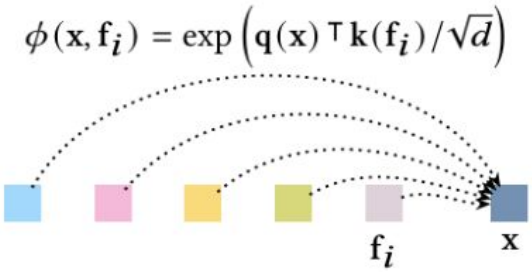


Image source: **A Review on Deep Learning Approaches for 3D Data Representations in Retrieval and Classifications.** Abubakar Gezawa,Yan Zhang, Qicong Wang, Lei Yunqi

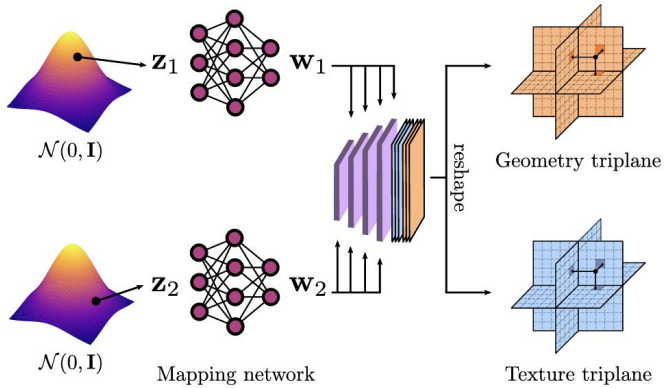
Neural Representation of 3D Shapes

- Vector Sets
- Structured Latents (triplane, sparse volume, ..)



Latent vector set

3DShape2VecSet (SIGGRAPH 2023)
Zhang et al.

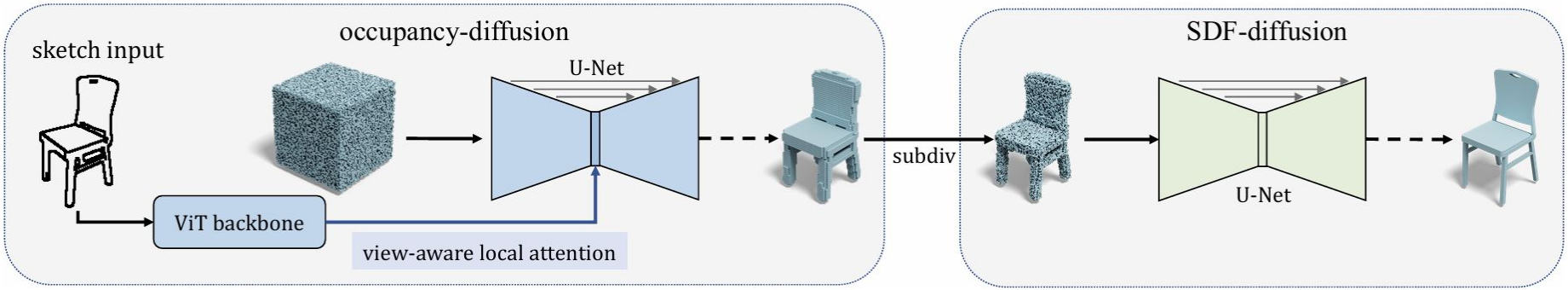


Triplane

GET3D (NeurIPS 2022)
Gao et al.

Neural Representation of 3D Shapes

- Vector Sets
- Structured Latents (triplane, sparse volume, ..)

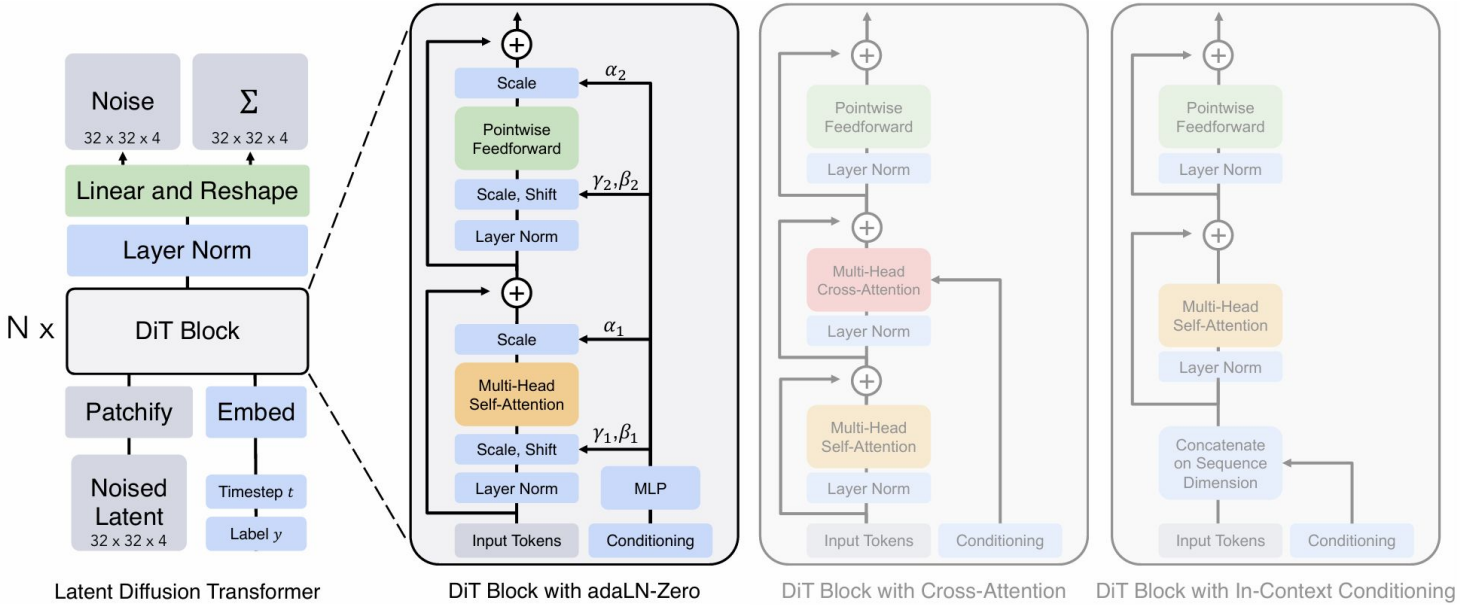


Sparse Voxel Grid

Locally Attentional SDF Diffusion for Controllable 3D Shape Generation (SIGGRAPH 2023) Zheng et al.

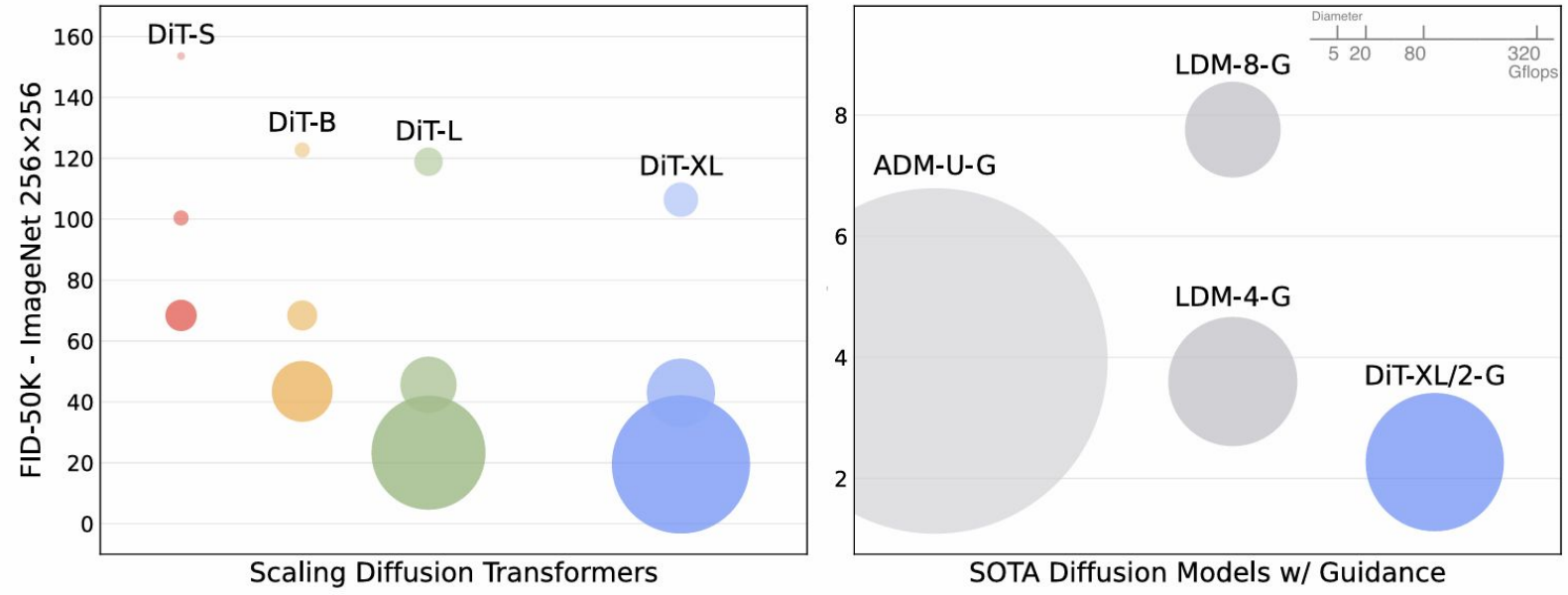
Shape Generative models

Scalable Diffusion Models with Transformers (DiT) (ICCV 2023)

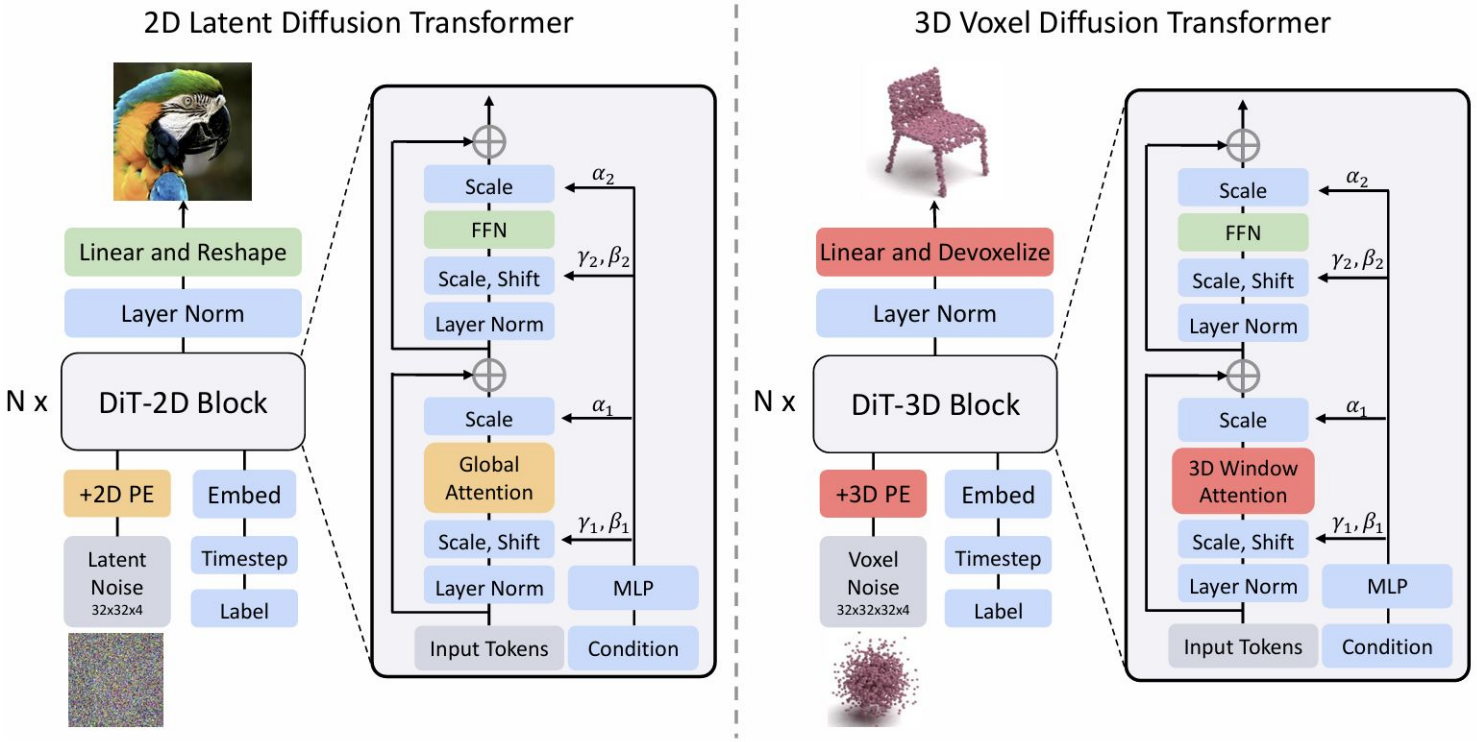


Shape Generative models

Scalable Diffusion Models with Transformers (DiT) (ICCV 2023)



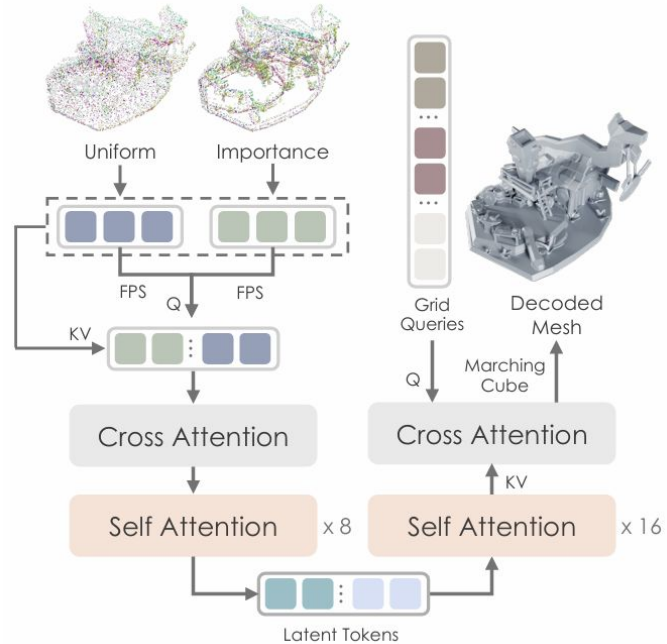
Shape Generative models



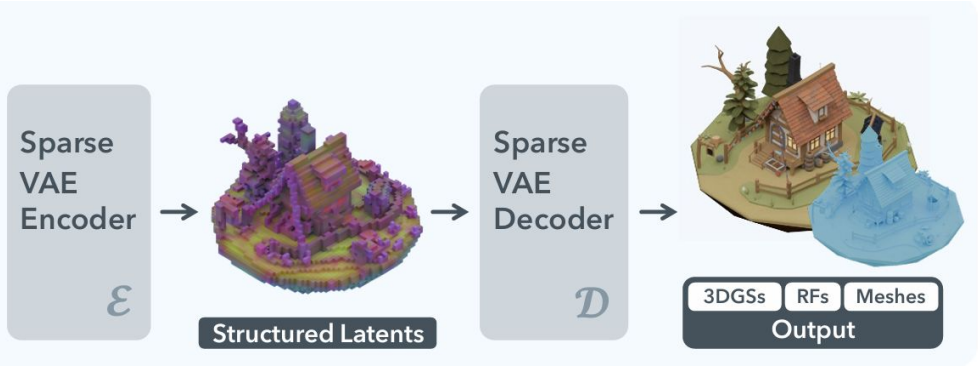
DiT-3D: Exploring Plain Diffusion Transformers for 3D Shape Generation, Mo et al.

Shape Generative models

VAE (Variational AutoEncoder) + Neural representation



ShapeVAE from Hunyuan3D



SparseVAE from Trellis3D

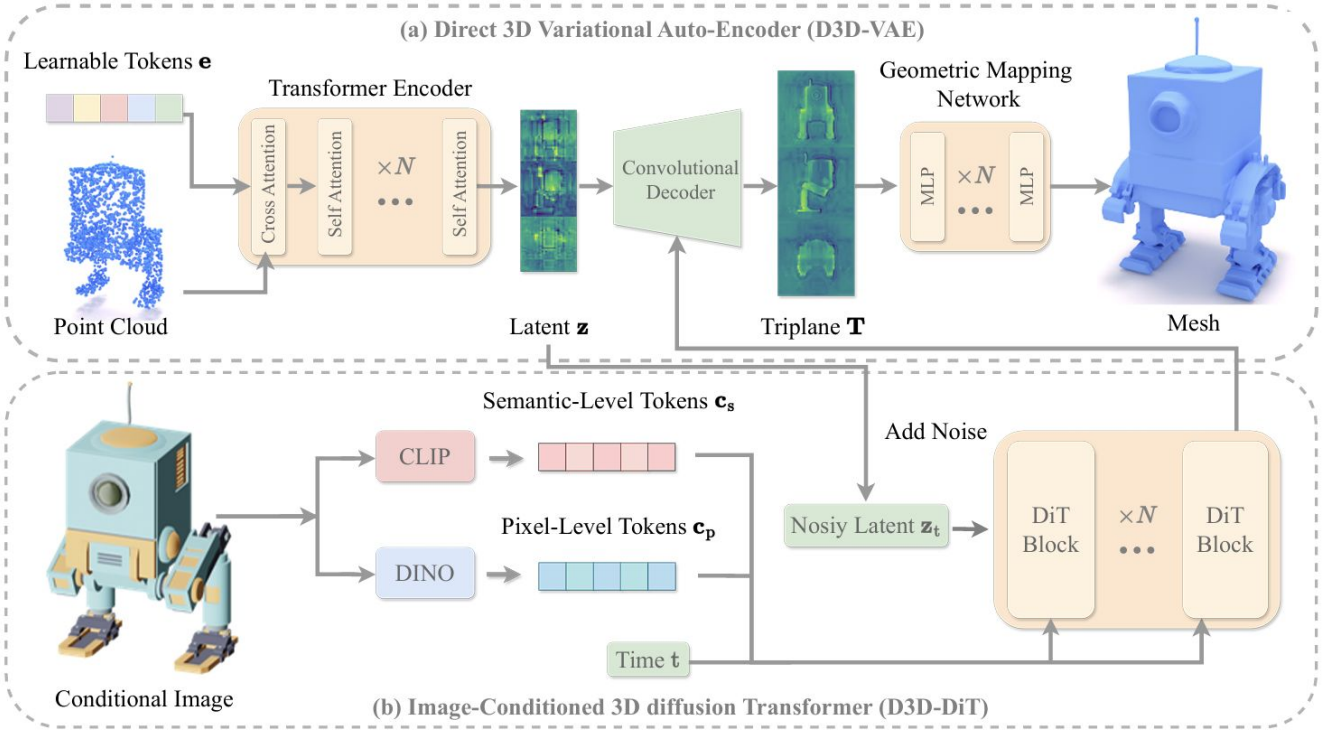
3D Datasets

- Growth of open-source datasets with high-resolution 3D assets.

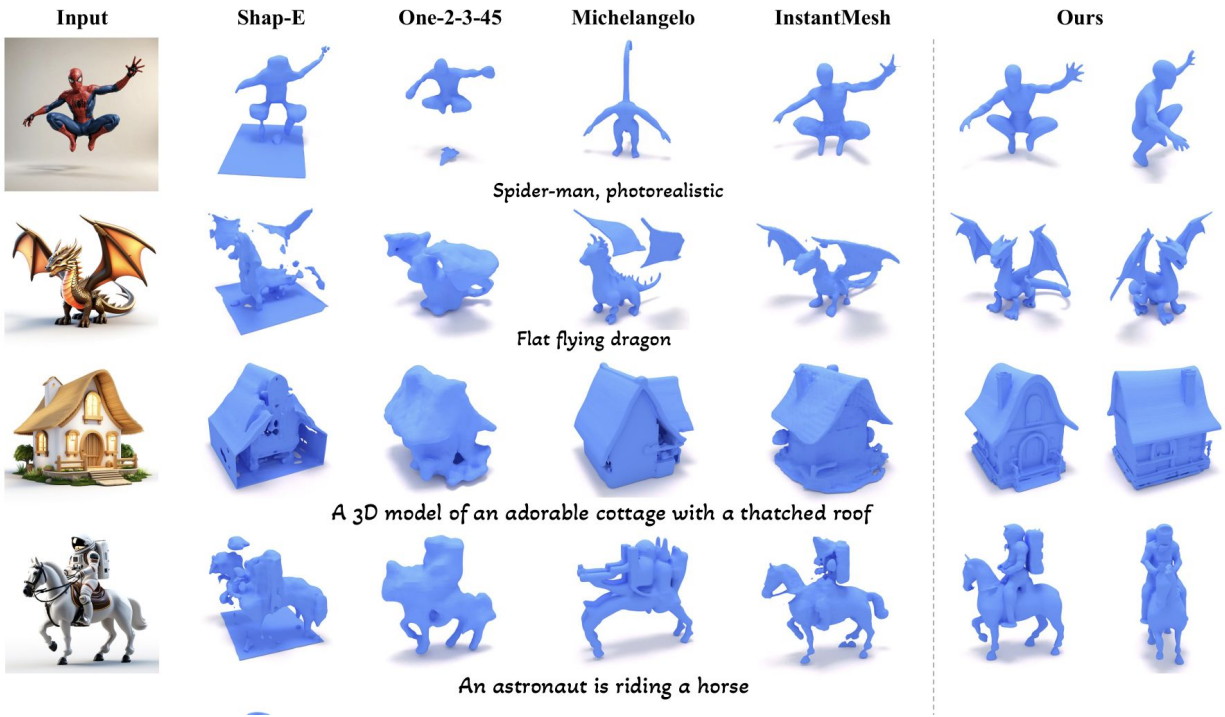
Source	# Objects
IKEA [32]	219
GSO [17]	1K
EGAD [41]	2K
OmniObject3D [63]	6K
PhotoShape [46]	5K
ABO [13]	8K
Thingi10K [67]	10K
3d-Future [19]	10K
ShapeNet [9]	51K
Objaverse 1.0 [14]	800K
Objaverse-XL	**10.2M**

Objaverse-XL https://objaverse.allenai.org/

Direct3D: Scalable Image-to-3D Generation via 3D Latent Diffusion Transformer (NeurIPS 2024)



Direct3D: Scalable Image-to-3D Generation via 3D Latent Diffusion Transformer (NeurIPS 2024)



Remaining Limitations by **ChatGPT**

- **Data scarcity and inconsistency:** High-quality 3D datasets are limited and often lack consistent topology or texture information, constraining generalization and realism.
- **Computational and scalability challenges:** Generating high-resolution 3D outputs (meshes, NeRFs, or implicit fields) requires significant GPU memory and long inference times, hindering real-time or large-scale applications.
- **Limited physical and structural accuracy:** Many models produce geometrically plausible but physically unrealistic or topologically inconsistent shapes that need post-processing.
- **Weak multi-view and texture consistency:** Cross-view alignment and material fidelity still lag behind 2D diffusion models, especially for complex or occluded regions.
- **Evaluation and standardization issues:** There is no unified benchmark for perceptual quality, mesh usability, or text-3D alignment, making fair comparison and progress tracking difficult.

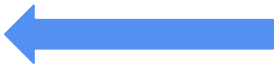
Conclusion

- **Single-view Image Generation Model based Distillation**

  3D Consistency ❌, Generation time(too long) ❌

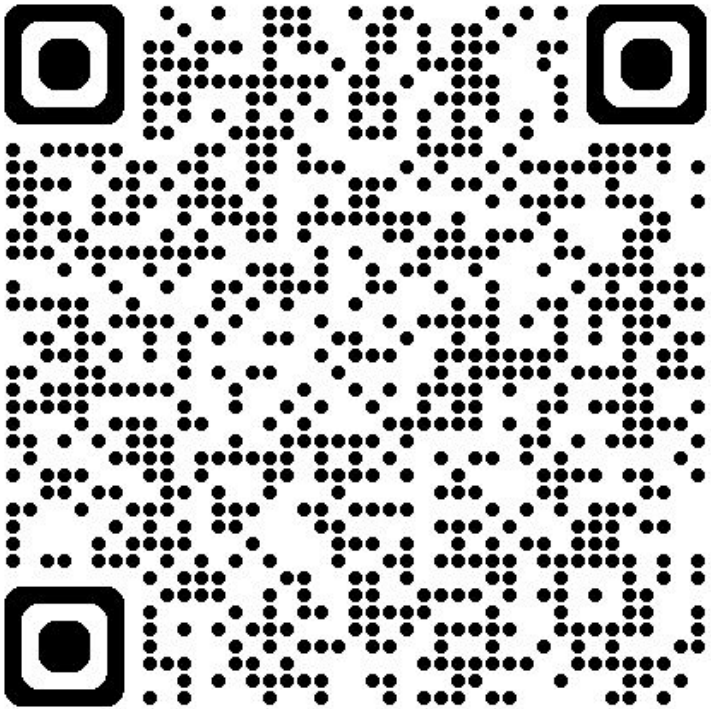
- **Multi-view Image Generation Model based Distillation**

  3D Consistency ✅, Generation time(too long) ❌

- **Feed-forward Models**  **Current Trend!**

  3D Consistency ✅, Generation time(reasonable) ✅

Quiz



Thank you