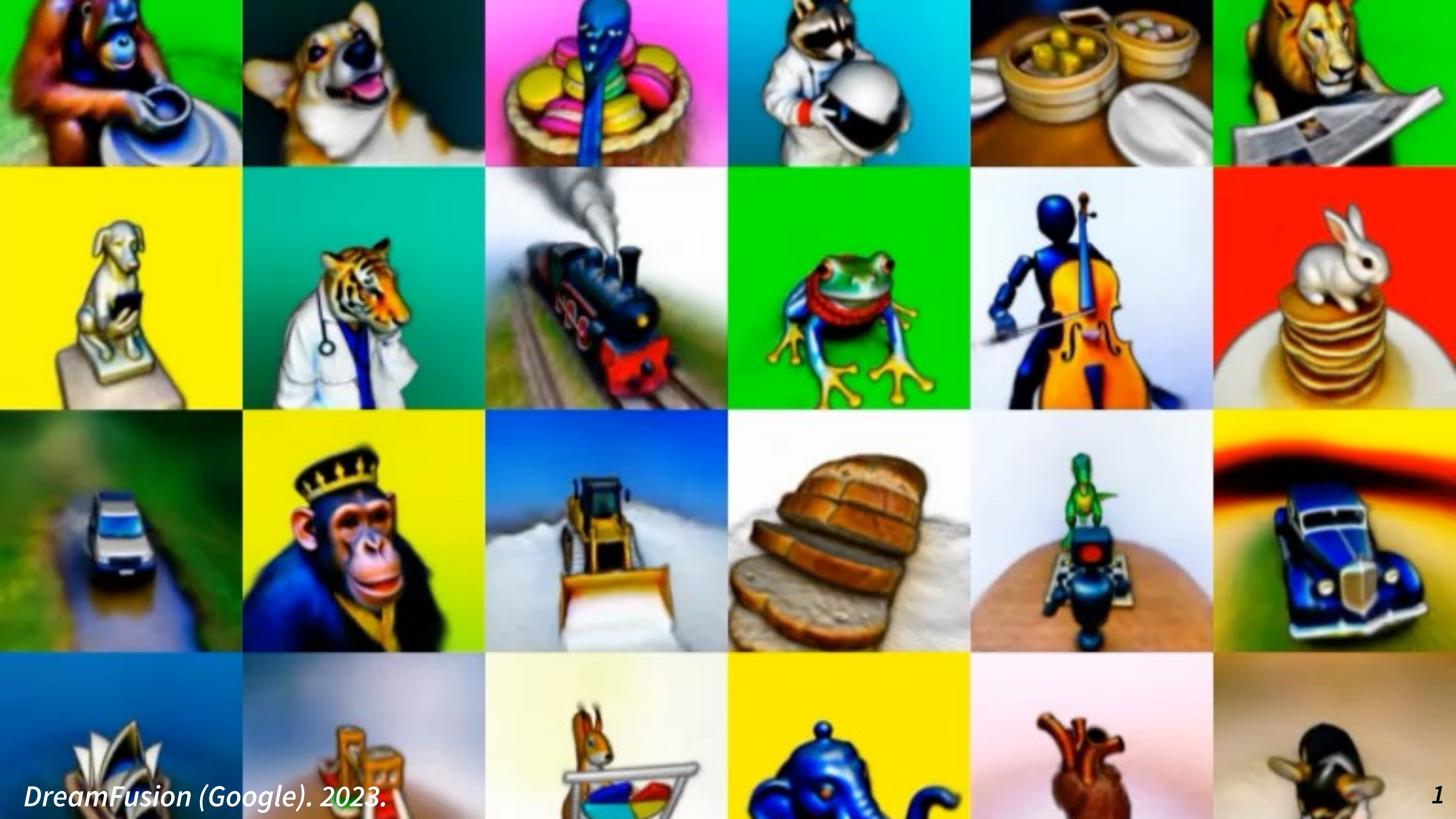
Diffusion Models for Visual Content Generation

3D Generation

Presenter: Minhyuk Sung

Eurographics 2024 Tutorial



DreamFusion (Google). 2023.

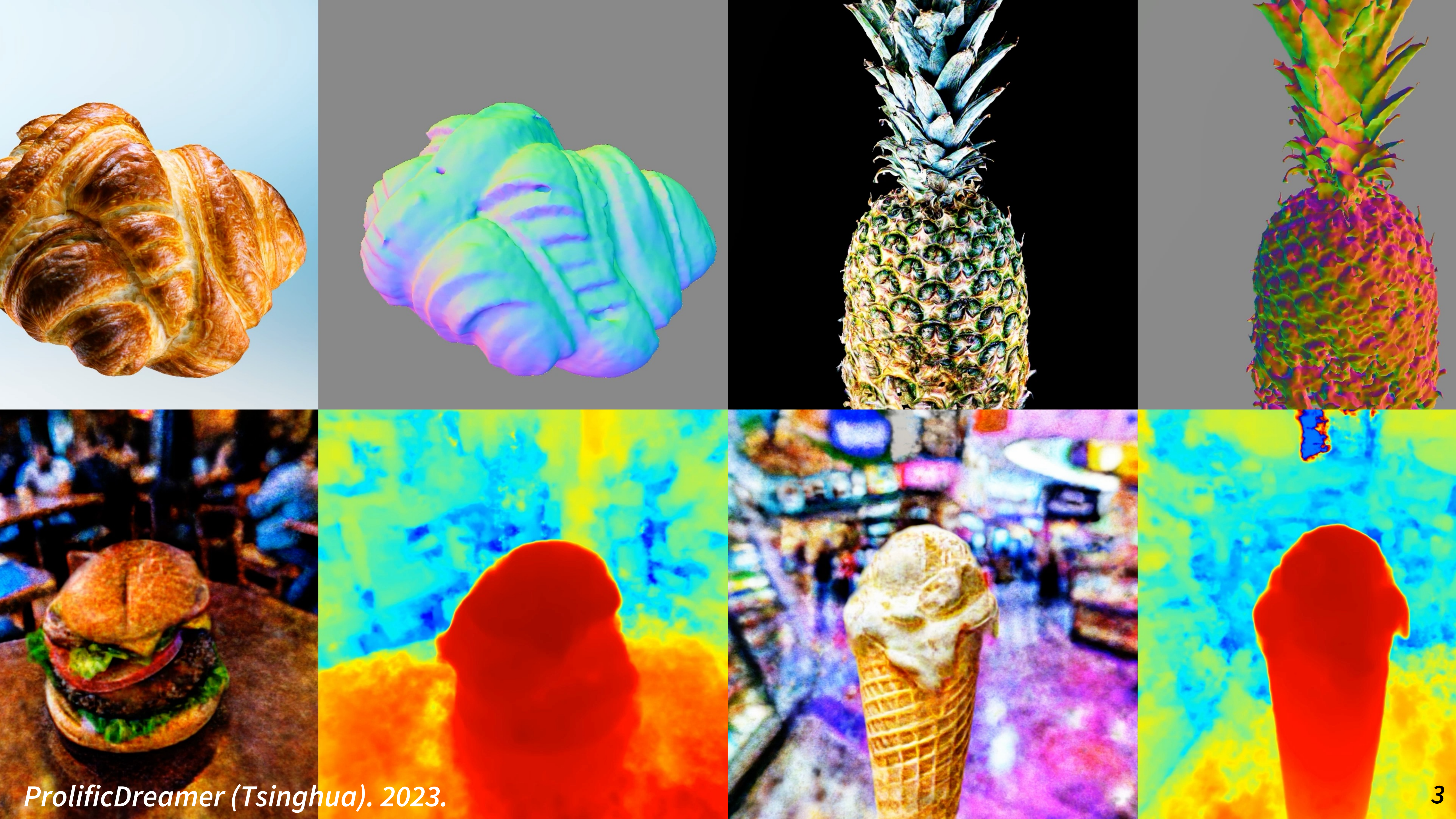
Magic3D:
High-Resolution Text-to-3D Content Creation

Chen-Hsuan Lin* Jun Gao* Luming Tang* Towaki Takikawa* Xiaohui Zeng*
Xun Huang Karsten Kreis Sanja Fidler# Ming-Yu Liu# Tsung-Yi Lin

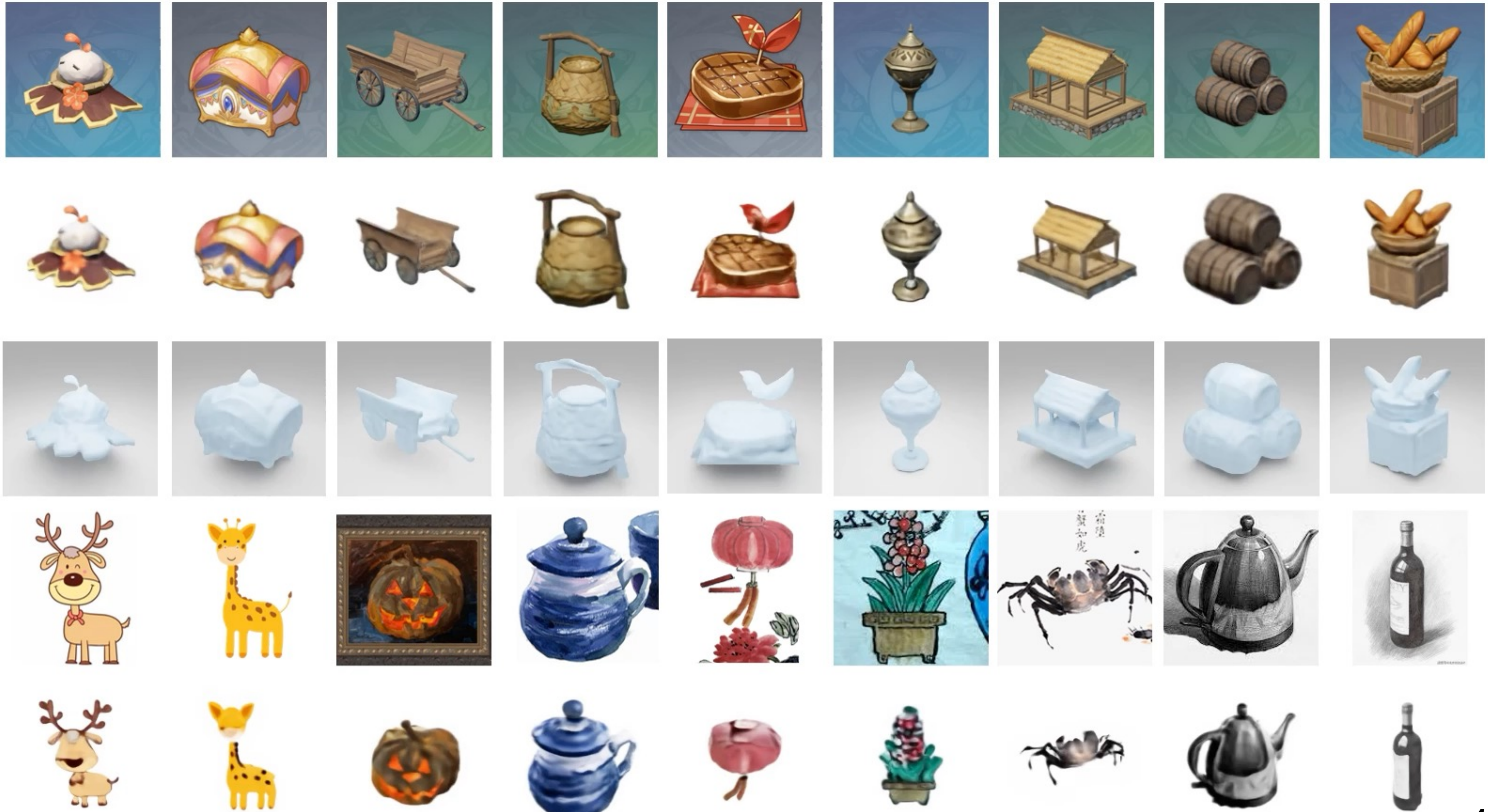
* # : equal contributions

NVIDIA Corporation

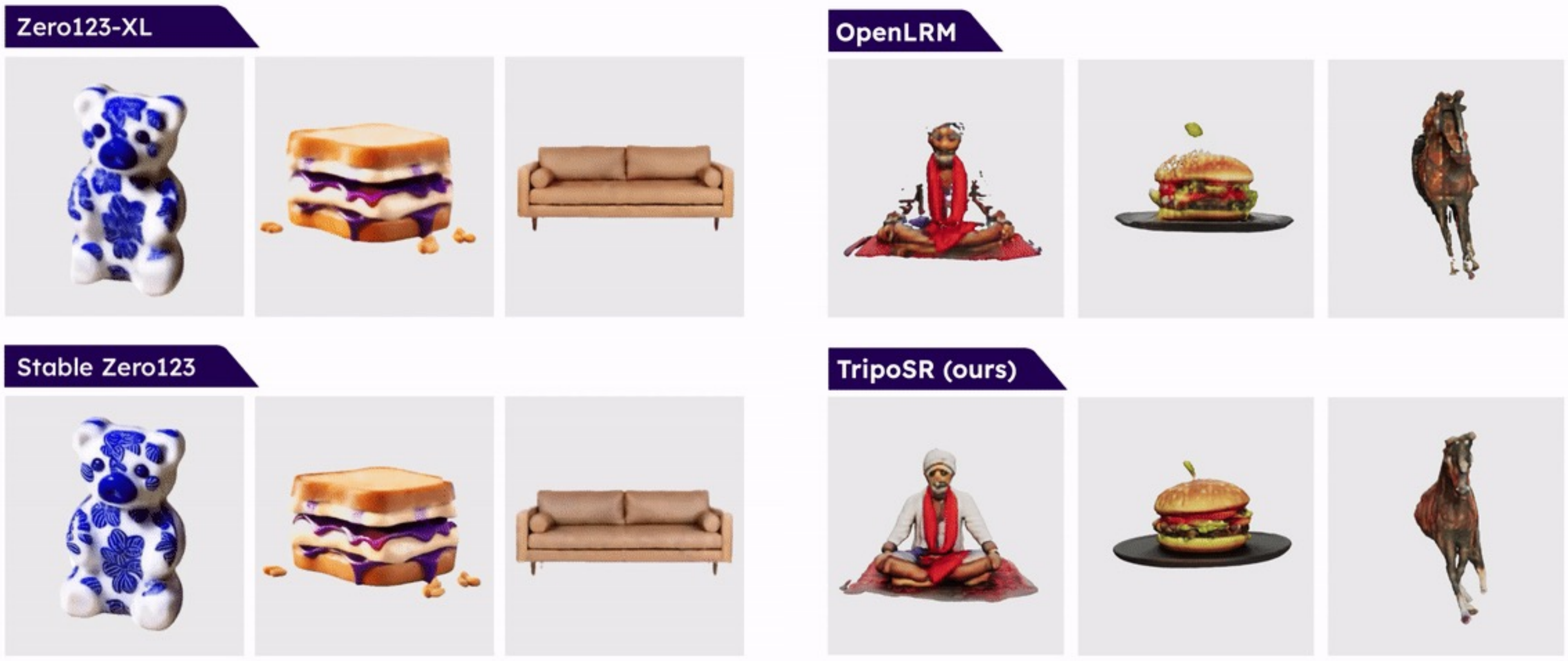
Magic3D (NVIDIA). 2023.



ProlificDreamer (Tsinghua). 2023.



Liu et al., SyncDreamer: Generating Multiview-consistent Images from a Single-view Image, arXiv 2023.

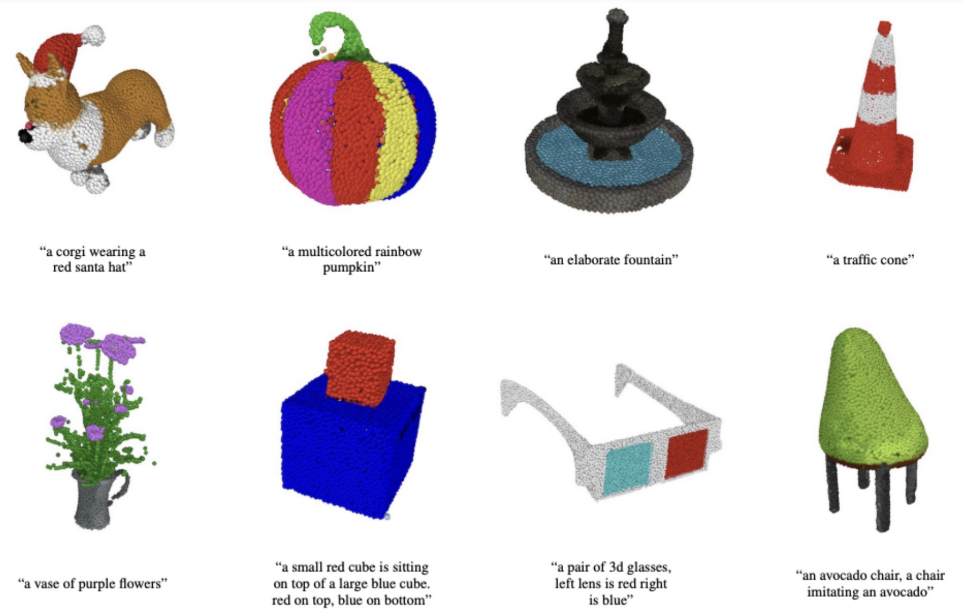

Zero123-XL
OpenLRM
Stable Zero123
TripoSR (ours)

3D Diffusion Models

3D Diffusion Models

Point clouds

- *ShapeGF (Cai et al., 2020)*
- *DPM (Luo and Hu, 2021)*
- *LION (Nichol et al., 2022)*

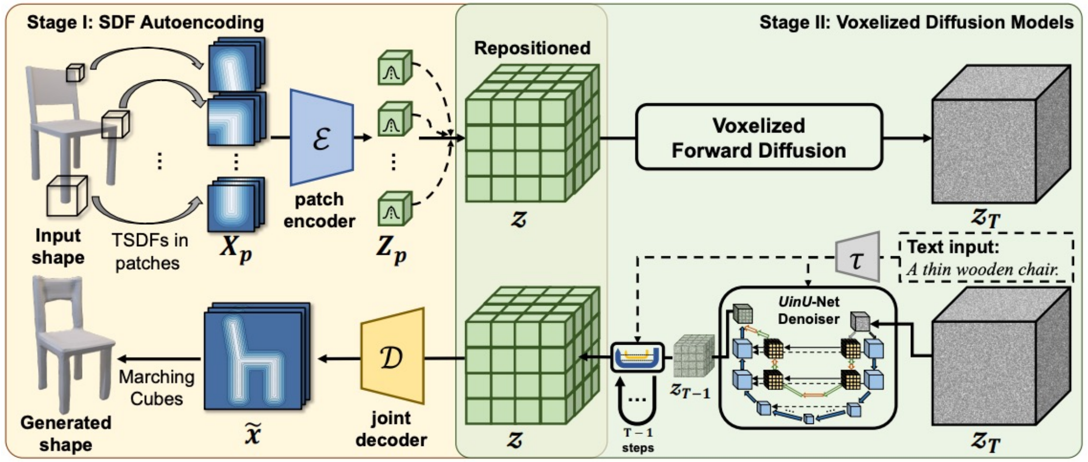


LION, Nichol et al. 2022.

3D Diffusion Models

Voxel representation

- *PVD (Zhou et al., 2022)*
- *Diffusion-SDF (Li et al, 2022)*

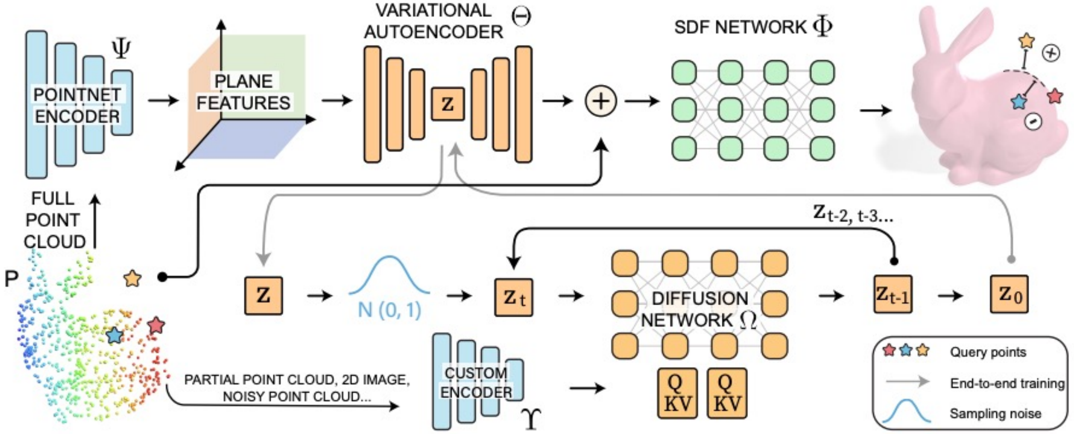


Diffusion-SDF, Li et al., 2022.

3D Diffusion Models

Latent representation

- *SDFusion (Cheng et al., 2022)*
- *DiffusionSDF (w/o hyphen, Chou et al., 2023)*



DiffusionSDF, *Chou et al., 2023*

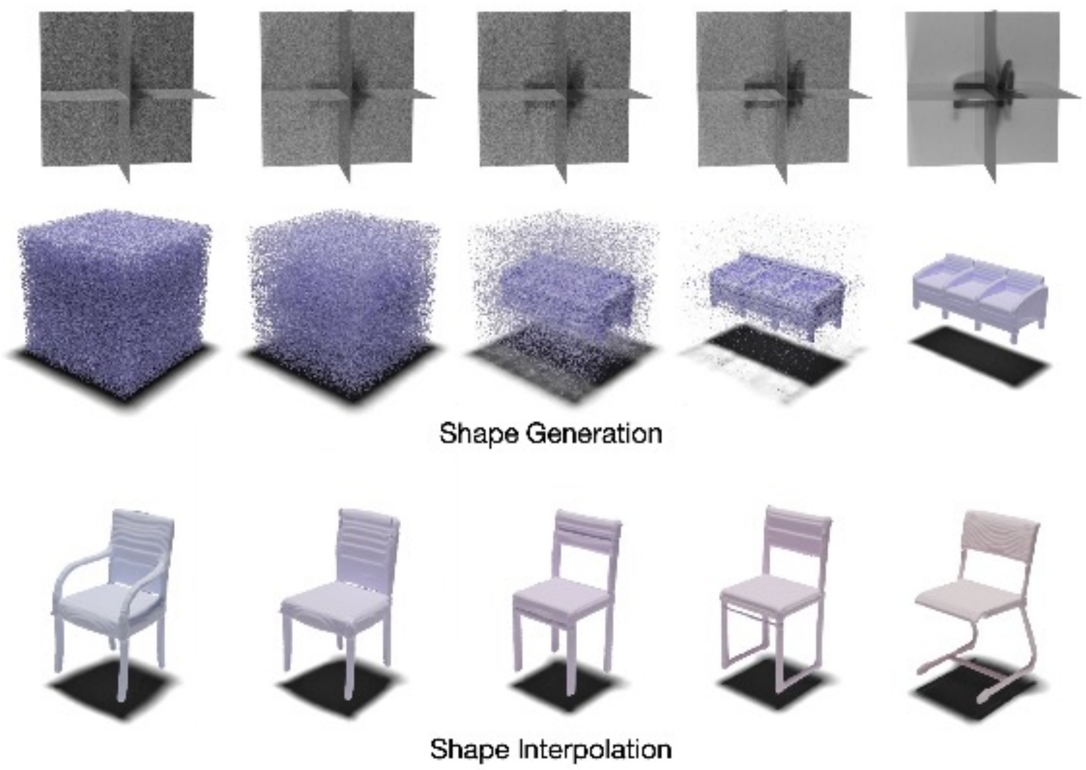
3D Diffusion Models

Triplane representation

- *NFD (Shue et al., 2022)*

Diffusion in the spectral domain:

- *NeuralWavelet (Hui et al., 2022)*



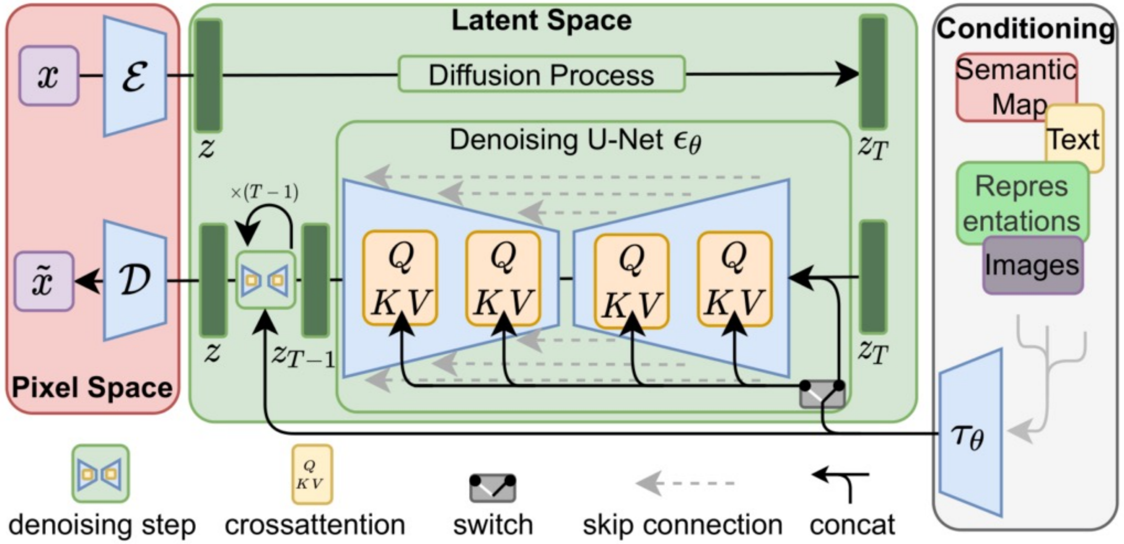
NFD, Shue et al., 2022.

Diffusion w/ Different Representations

- *Implicit representation*
- *Explicit representation*

Diffusion w/ Different Representations

- *Implicit representation (I.e., latent features)*
  - *E.g., Latent diffusion (Rombach et al., 2022)*
  - *(+) Best quality of the generated data.*
  - *(−) Requires retraining for each conditional generation setup.*



Latent Diffusion, Rombach et al., 2022.

Diffusion w/ Different Representations

- *Explicit representation (E.g., pixels in images)*
  - *E.g.: The original DDPM model (Ho et al., 2020) for images.*
  - *(−) Suboptimal performance due to the high dimensionality.*
  - *(−) Cannot change the resolution of the data.*
  - *(+) Can be directly leveraged in conditional generation setups in a zero-shot manner.*

SDEdit *[Meng et al., 2022]*

Image editing using a pretrained pixel-space diffusion model.

$$\mathbf{x}^{(0)} = m \odot \mathbf{x}_a^{(0)} + (1-m) \odot \mathbf{x}_b^{(0)}$$

$$\mathbf{x}^{(t)} = \text{denoise}\left(\mathbf{x}^{(0)}, t\right)$$

$$\mathbf{x}'^{(0)} = \text{add\_noise}\left(\mathbf{x}^{(t)}, t\right)$$

RePaint *[Lugmayr et al., 2022]*

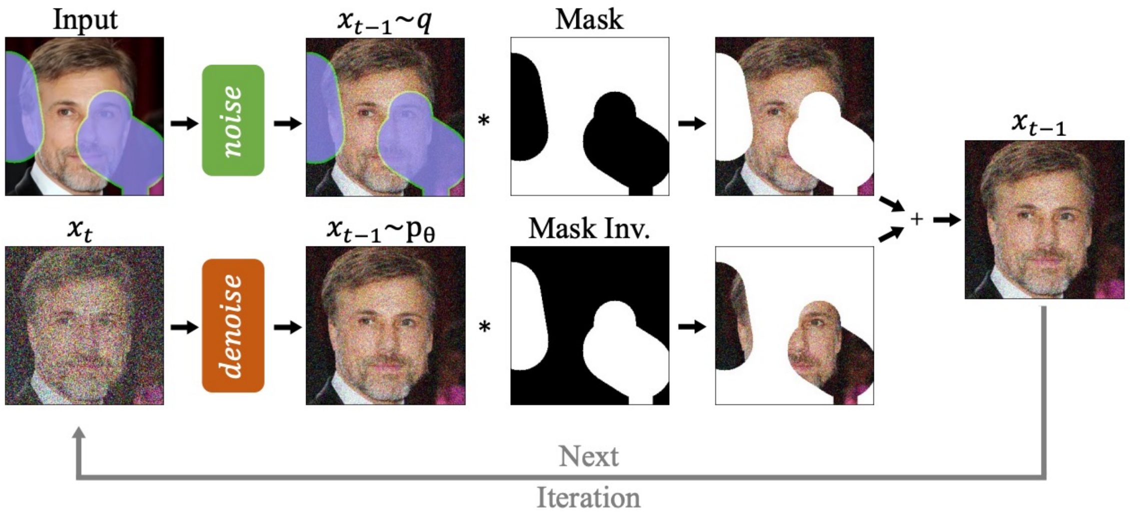
Image inpainting using a pretrained pixel-space diffusion model.

$$\mathbf{x}_f^{(t-1)} = \text{denoise}\left(\mathbf{x}^{(t)}, 1\right)$$

$$\mathbf{x}_b^{(t-1)} = \text{add\_noise}\left(\mathbf{x}^{(0)}, t\right)$$

$$\mathbf{x}^{(t-1)} = m \odot \mathbf{x}_f^{(t-1)} + (1-m) \odot \mathbf{x}_b^{(t-1)}$$

Repeat for $t = T, \cdots, 1.$



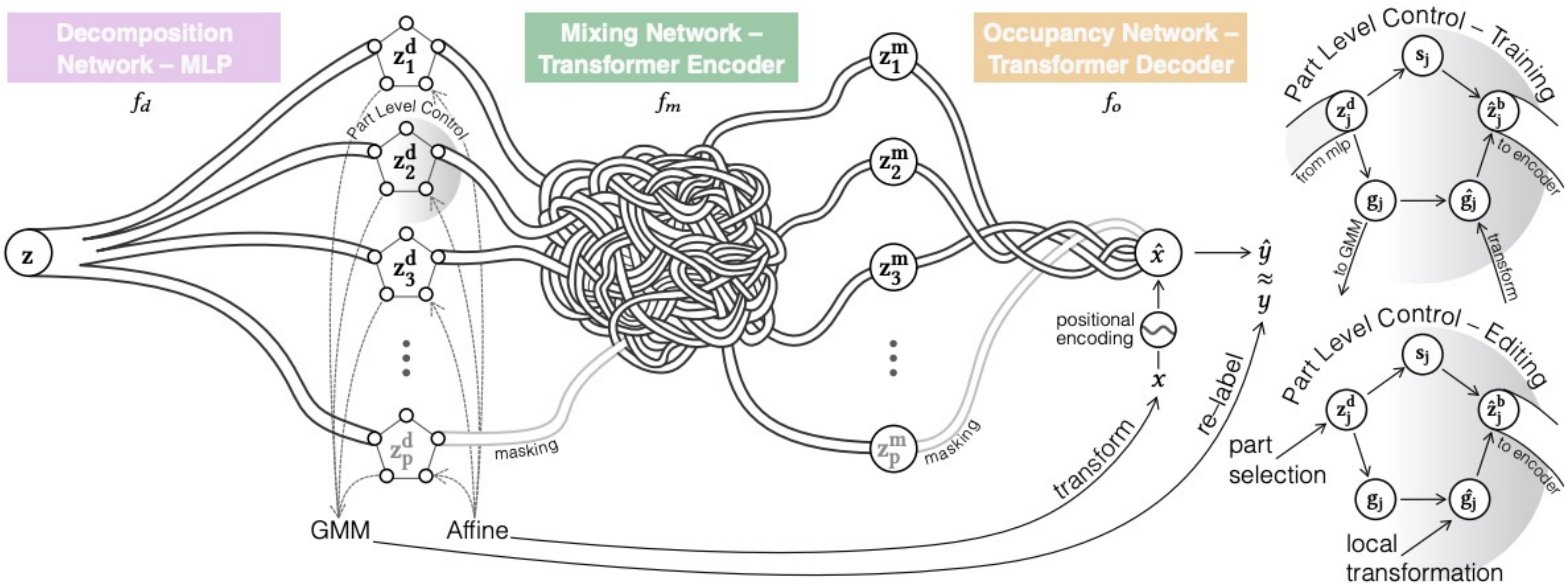
Hybrid Representation

Leverages a novel hybrid representation describing

- *global part-level structure explicitly, and*
- *local geometry implicitly.*

SPAGHETTI [Hertz et al., 2022]

The part decomposition and the explicit/implicit representations are learned in an unsupervised way.



Part-Level Representation

For each part of an object learned in an unsupervised way,

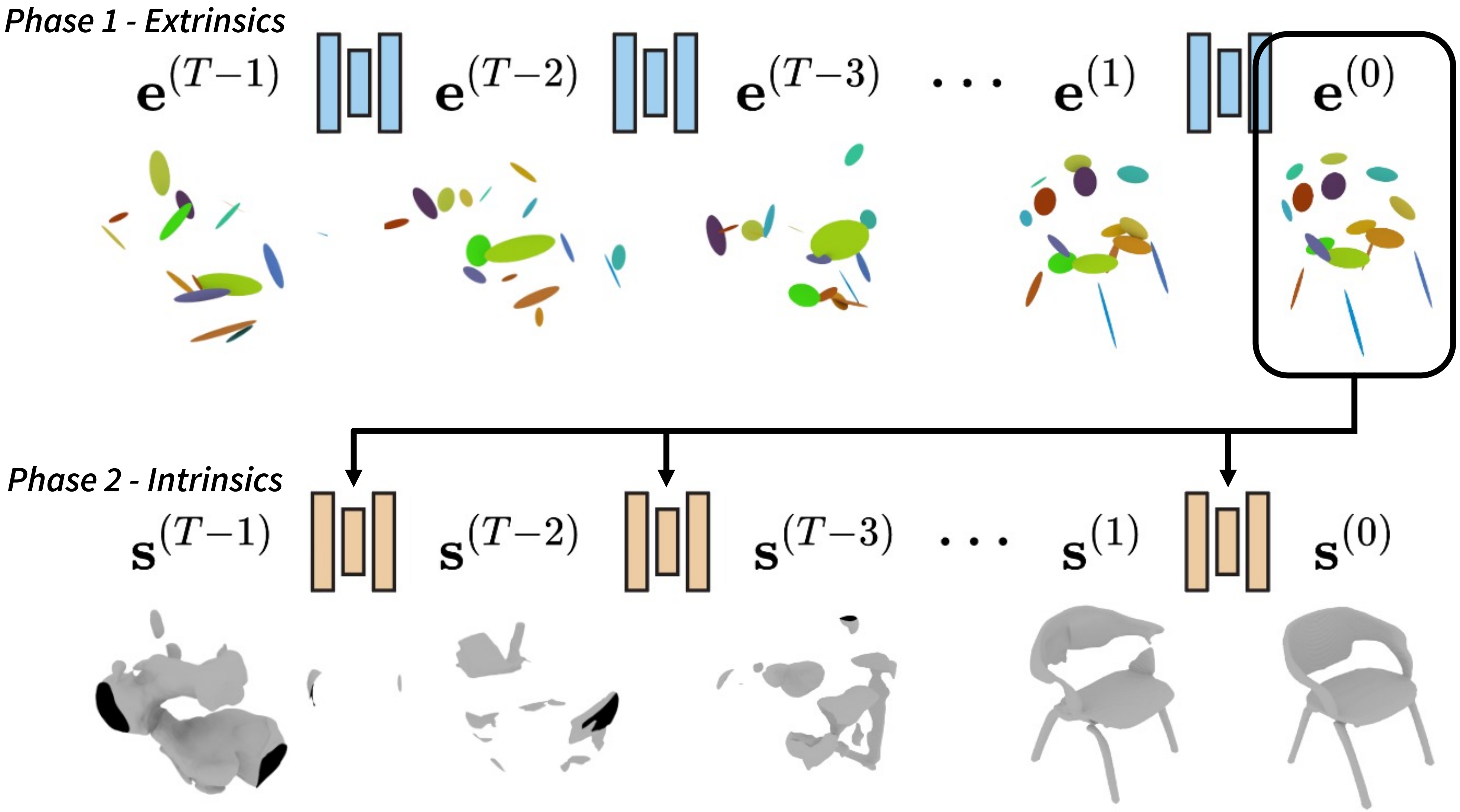
- *Explicit parameters of Gaussian blubs indicate position, scale, and rotation.*

Part-Level Representation

For each part of an object learned in an unsupervised way,

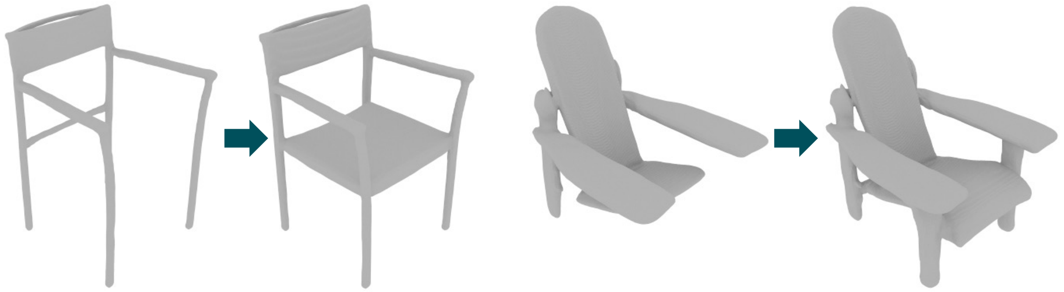
- *Implicit latent feature is decoded into an occupancy function.*

Two-Phase Cascaded Diffusion

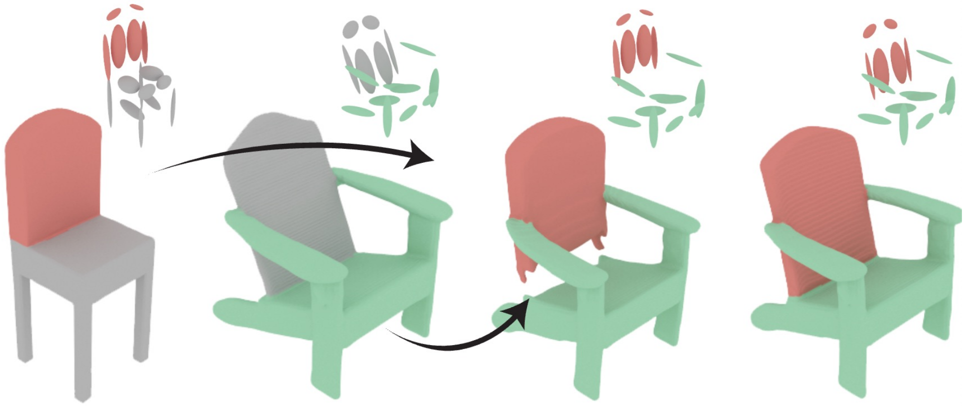


Diversity of 3D Shapes

Applications



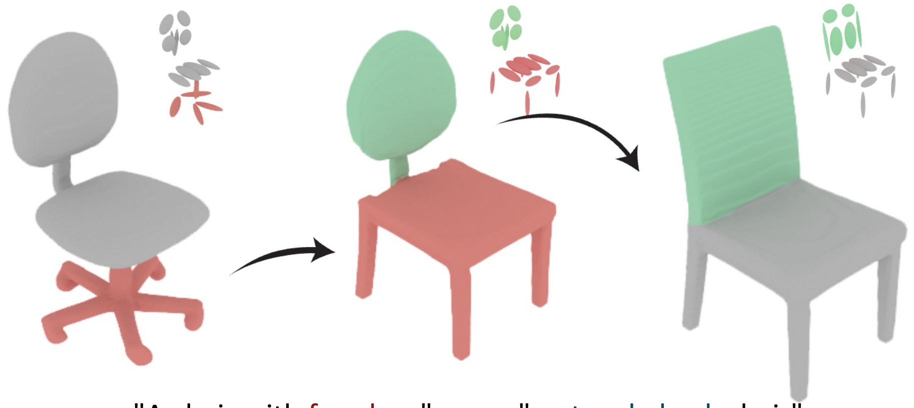
Part Completion



Part Mixing

"Chair has round arms and wheels." "Its the one with gaps in the back."

Text-to-3D Generation

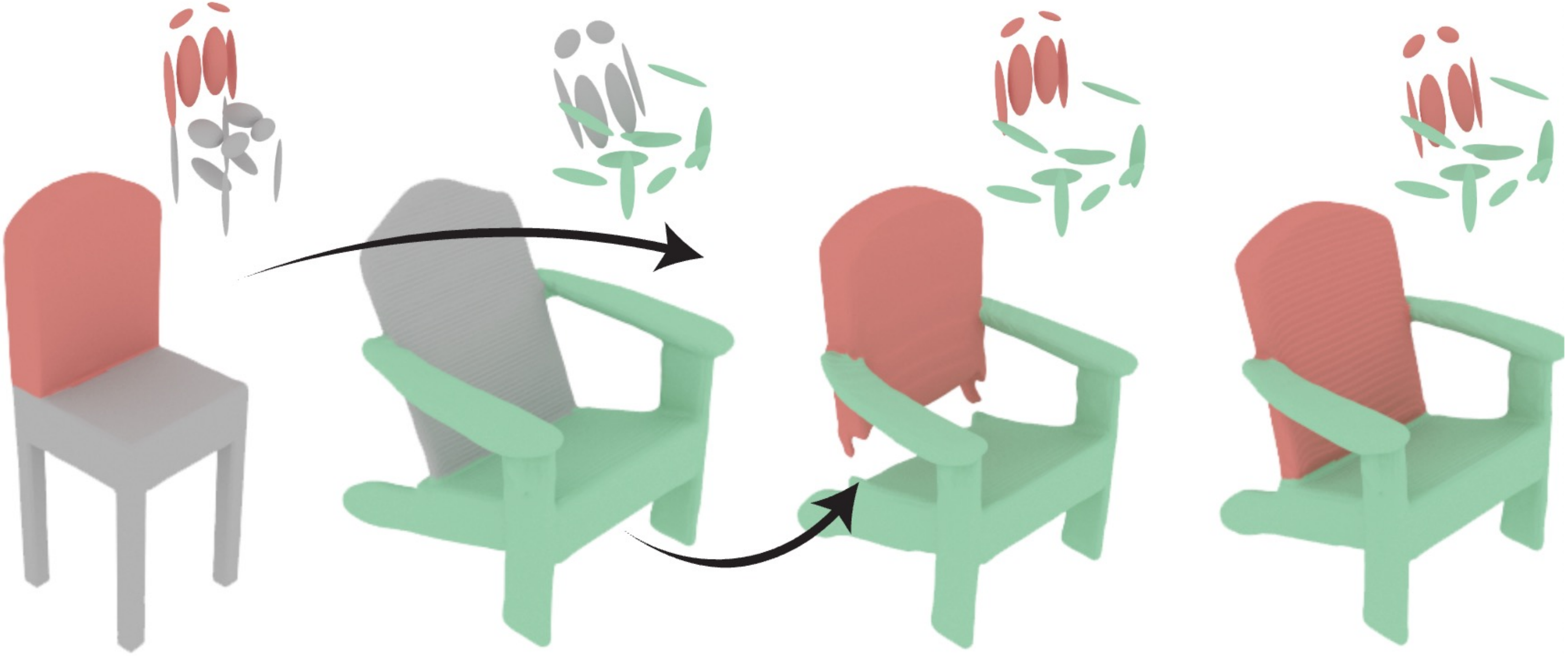


"A chair with four legs" "rectangle back chair"

Text-Guided Part Editing

Part Mixing

$$\mathbf{x}^{(0)} = m \odot \mathbf{x}_a^{(0)} + (1-m) \odot \mathbf{x}_b^{(0)}$$
$$\mathbf{x}^{(t)} = \mathrm{denoise}\left(\mathbf{x}^{(0)}, t\right)$$
$$\mathbf{x}'^{(0)} = \mathrm{add\_noise}\left(\mathbf{x}^{(t)}, t\right)$$



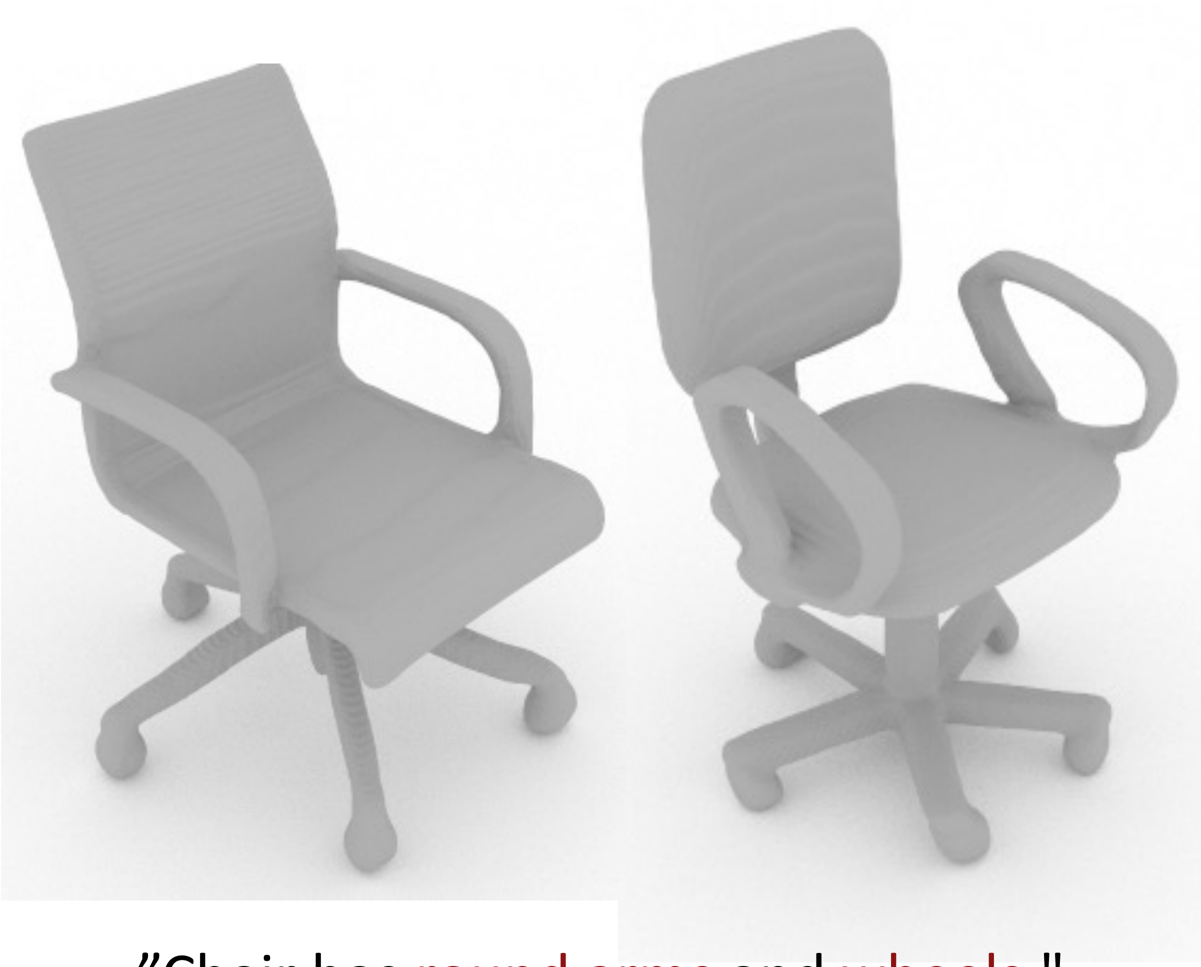
Part Completion

$$\mathbf{x}_f^{(t-1)} = \mathrm{denoise}\left(\mathbf{x}^{(t)}, 1\right)$$
$$\mathbf{x}_b^{(t-1)} = \mathrm{add\_noise}\left(\mathbf{x}^{(0)}, t\right)$$
$$\mathbf{x}^{(t-1)} = m \odot \mathbf{x}_f^{(t-1)} + (1-m) \odot \mathbf{x}_b^{(t-1)}$$

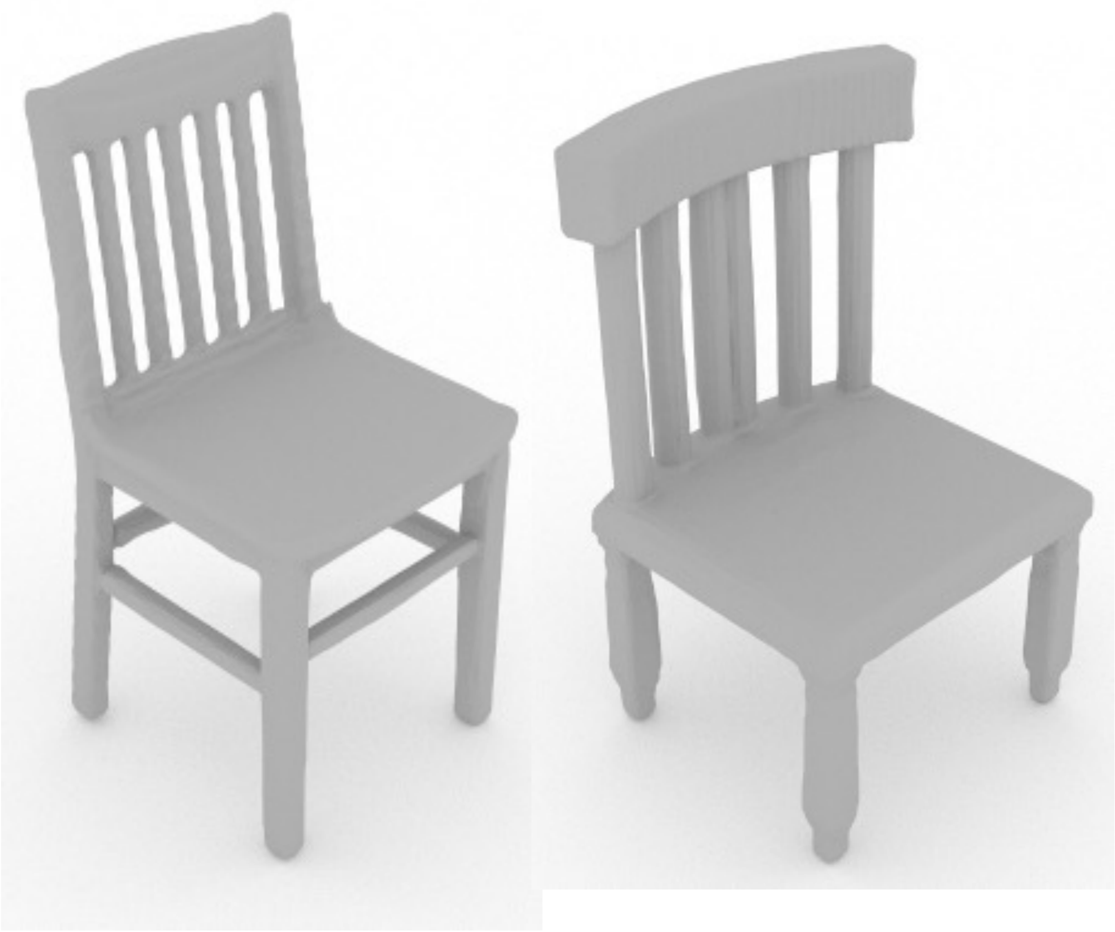
Repeat for
$t = T, \cdots, 1.$

Koo et al., SALAD: Part-Level Latent Diffusion for 3D Shape Generation and Manipulation, ICCV 2023.

Text-to-3D Generation

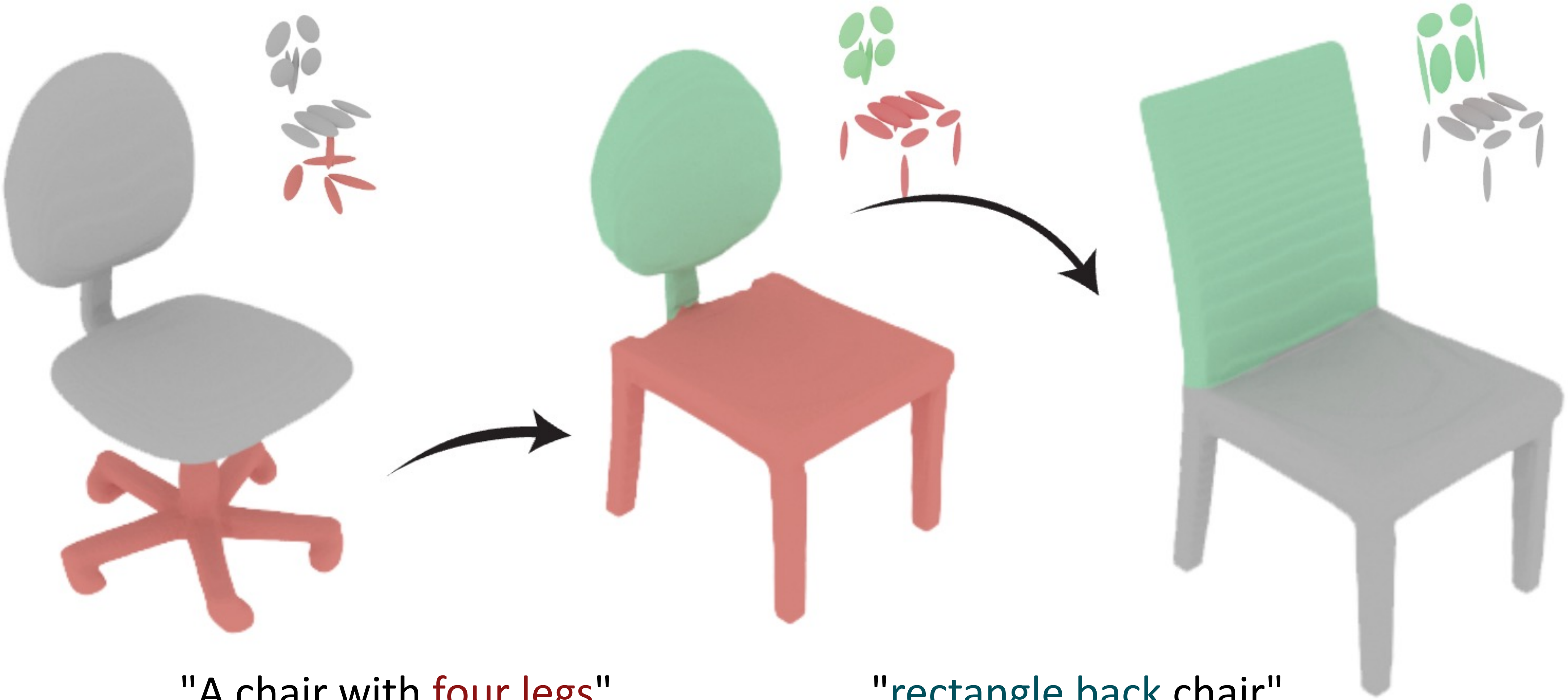


”Chair has round arms and wheels."



”Its the one with gaps in the back."

Text-Guided Part Editing



"A chair with four legs"

"rectangle back chair"

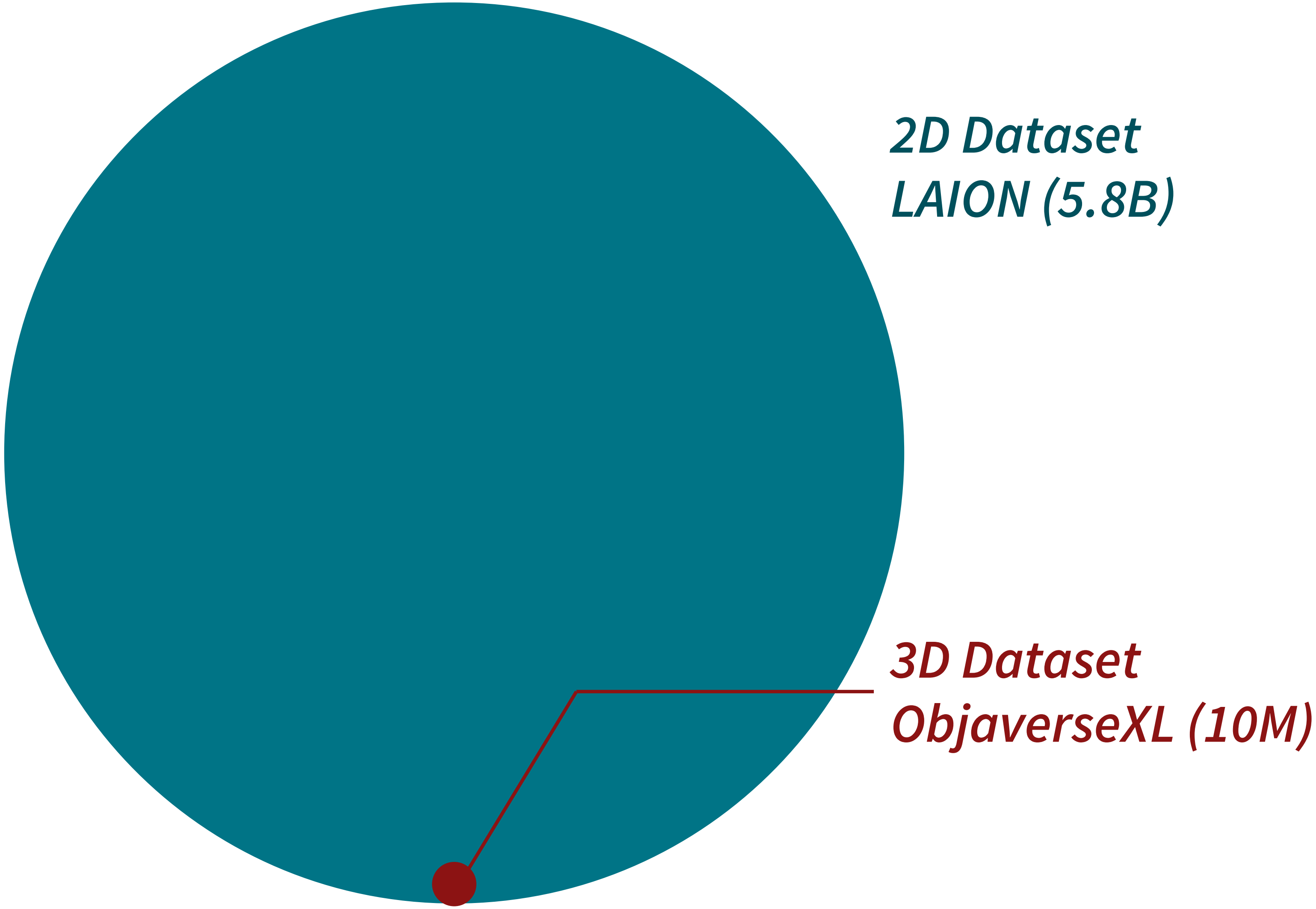
Limitation?

Koo et al., SALAD: Part-Level Latent Diffusion for 3D Shape Generation and Manipulation, ICCV 2023.

Challenge: Lack of Large-Scale 3D Dataset

Objaverse-XL
Allen Institute

Data Deficiency



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

Diversity of Imaginable 3D Shapes

"frog wearing a sweater"

"eggshell broken in two with an adorable chick standing next to it"

"ghost eating a hamburger"

"a pig wearing a backpack"

We have a small-scale 3D dataset

but images on an internet scale.

Images are projections of 3D from specific angles.

https://d2jv9003bew7ag.cloudfront.net/uploads/Tim-Noble-and-Sue-Webster-Wild-Mood-Swings-2009.jpg

3D Generation

How to generate a 3D object

from a collection of 2D images?

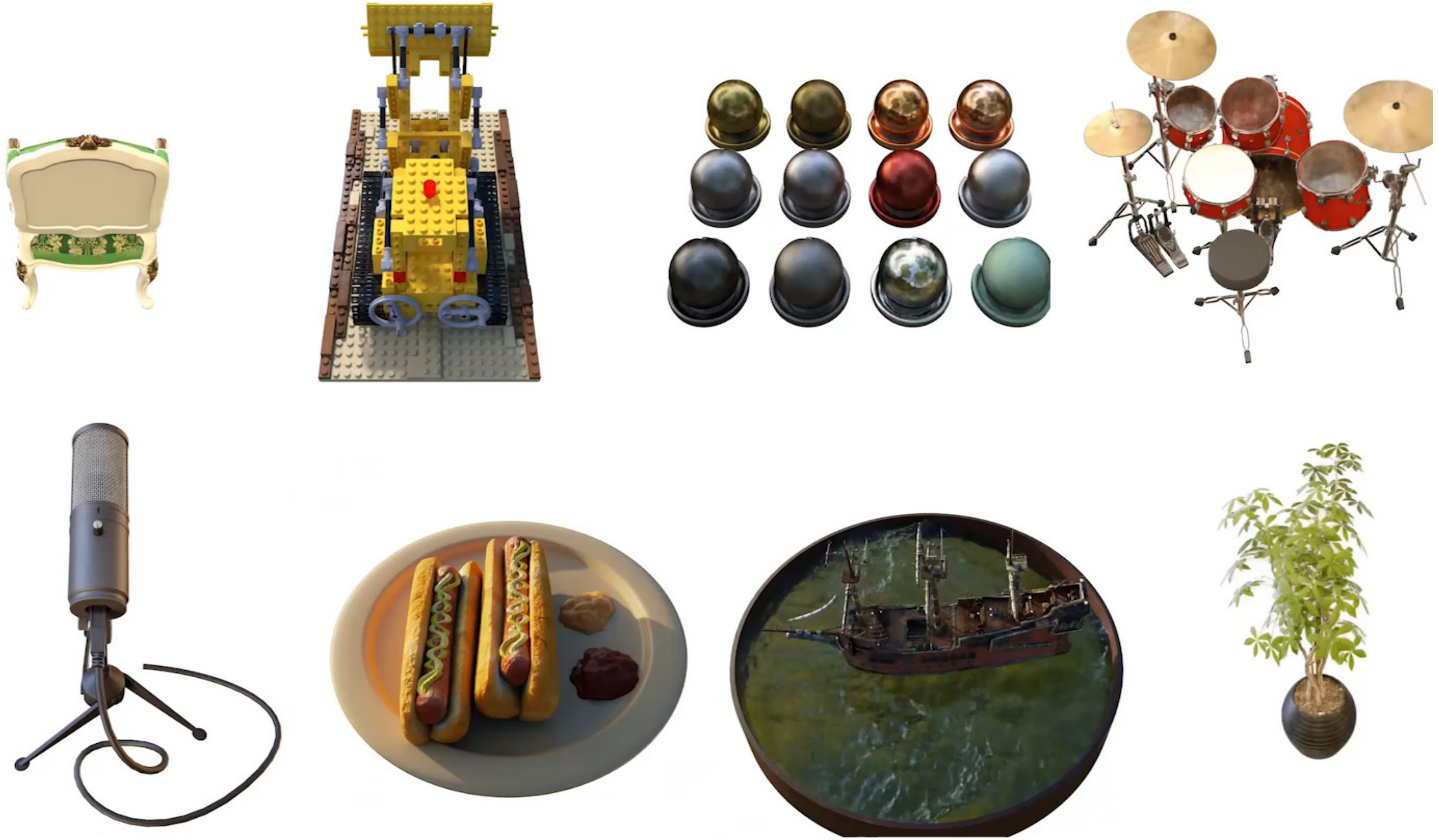
3D Reconstruction

How to reconstruct a 3D object

from a collection of 2D images

of a specific object?

NeRF



3D Reconstruction

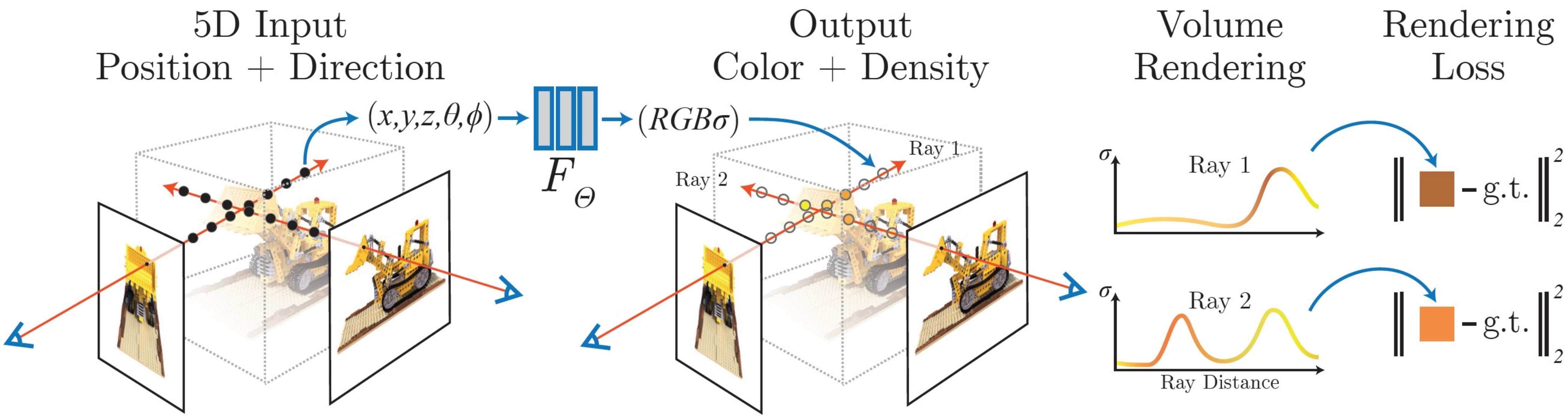
- *Input: A set of images with camera poses.*
- *Output: A representation of the 3D object.*

Kerbl et al., 3D Gaussian Splatting for Real-Time Radiance Field Rendering, SIGGRAPH 2023.

Gaussian Splatting

NeRF Optimization

1. Render a NeRF representation into a specific view.
2. Compute the difference with the given image.
3. Update the NeRF using gradient descent.



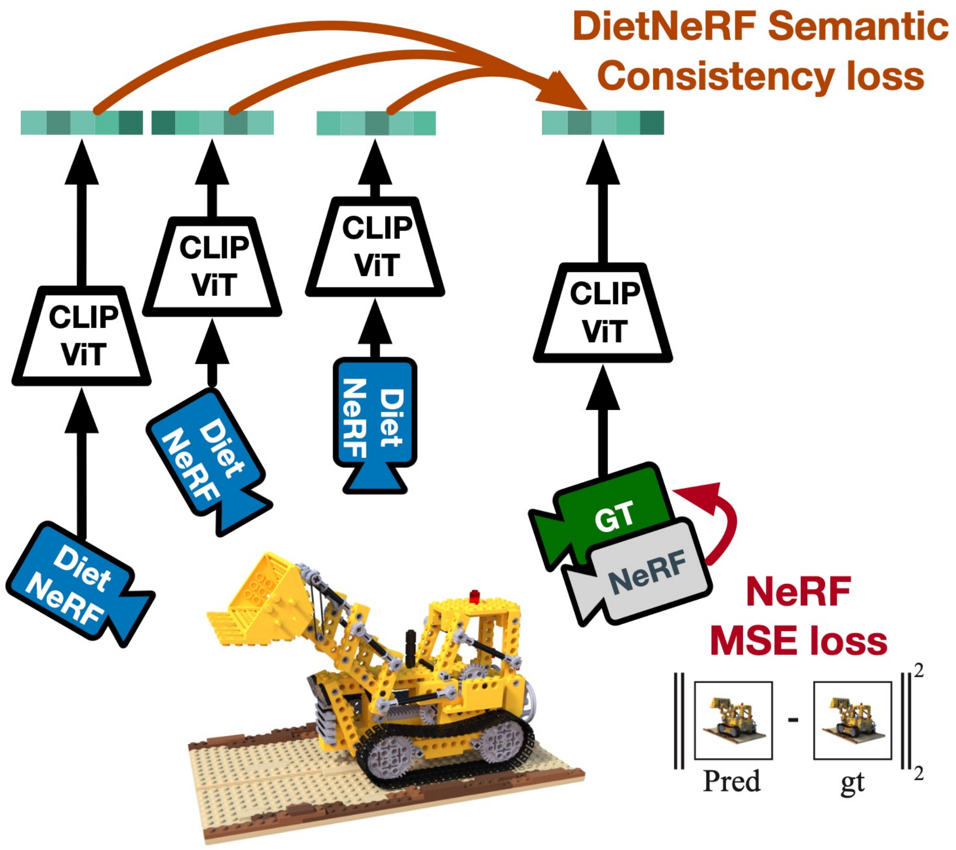
Can we perform NeRF reconstruction with a few images?

Then, we need additional information!

Few-Shot NeRF

DietNeRF [Jain et al., ICCV 2021]

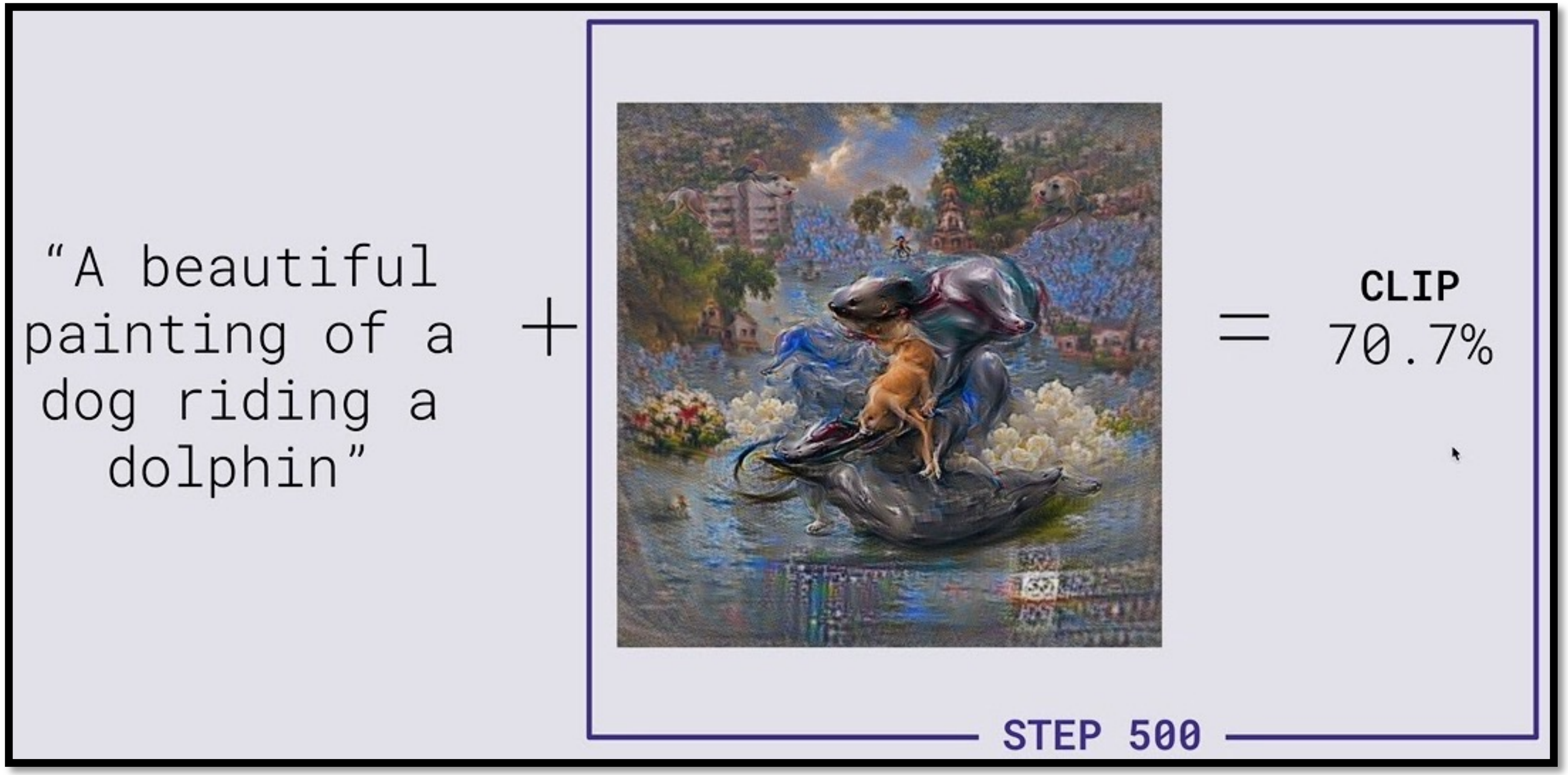
Use priors learned by CLIP [Radford et al., 2021], a text-image model.



"a bulldozer is a bulldozer
from any perspective"

CLIP *[Radford et al., 2021]*

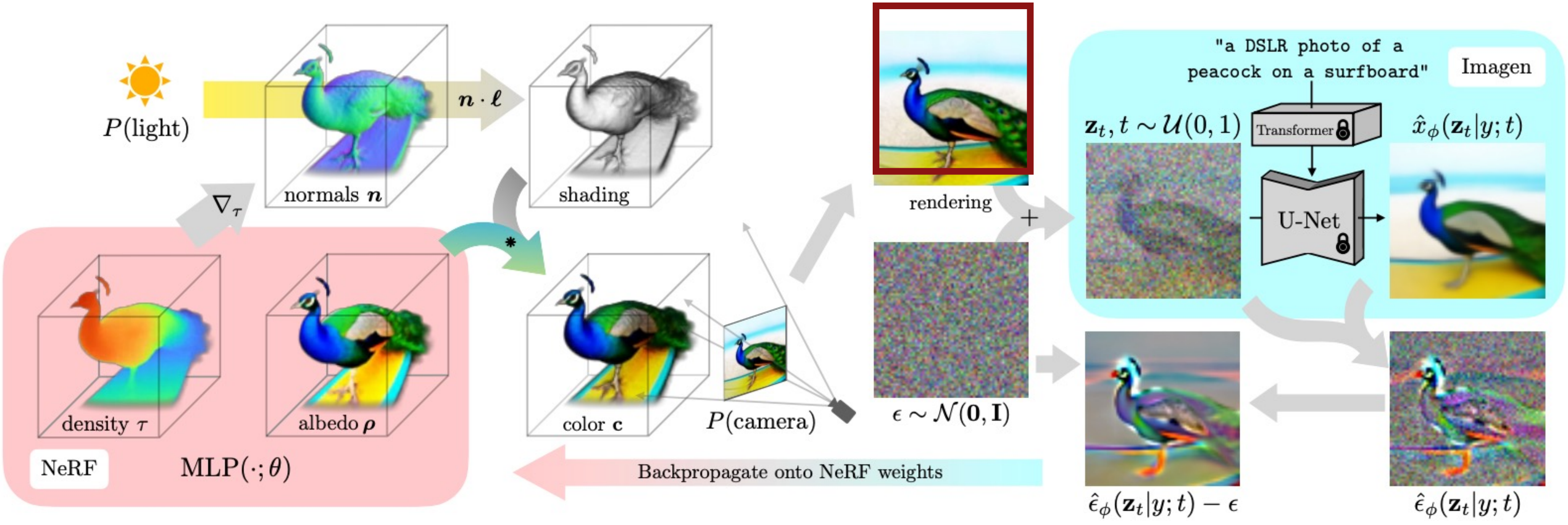
CLIP takes a text-image pair as input and assesses the alignment between the text and the image.



https://www.youtube.com/watch?app=desktop&v=-b7xKWeADHQ

Knowledge Distillation in 3D Generation

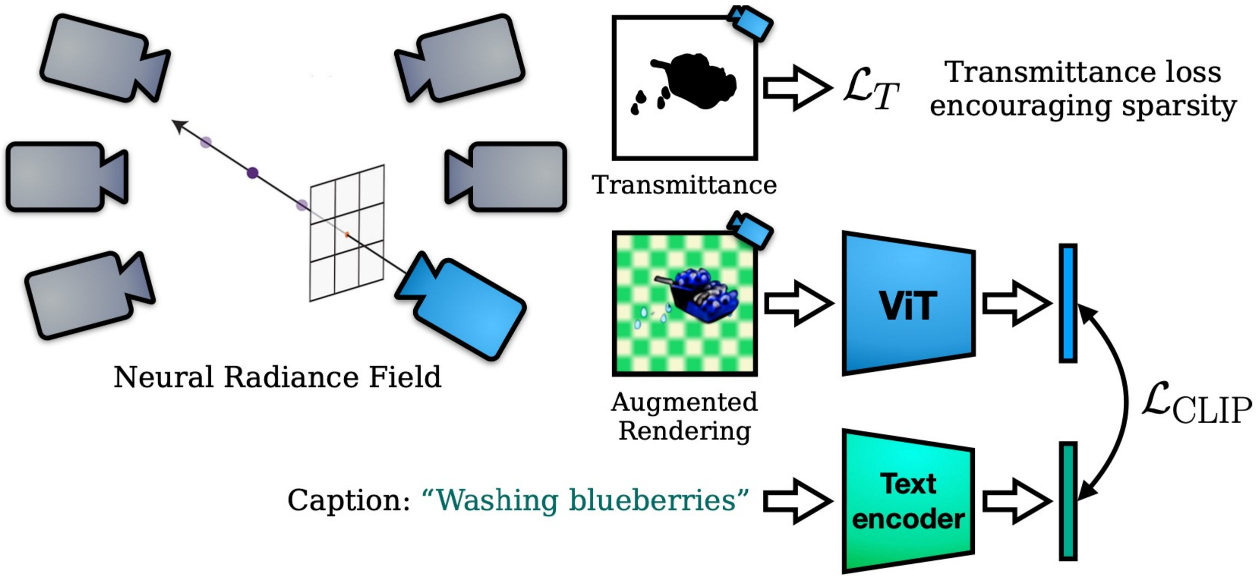
1. ***Render** the NeRF representation into a specific view*
2. ***Compute the alignment to the given text.***
3. ***Update the NeRF using the gradient descent.***

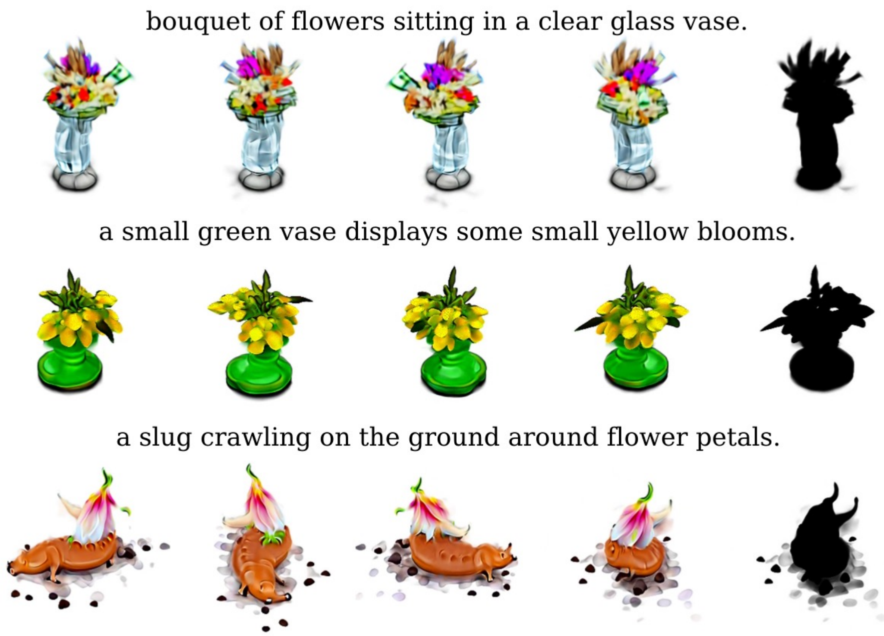


Extrem Case: Zero-Shot NeRF

DreamFields *[Jain et al., CVPR 2022]*

Given a text prompt but no images, generate a 3D shape by maximizing similarity between a rendered image and the input prompt in the CLIP embedding space.





Jain et al., Zero-Shot Text-Guided Object Generation with Dream Fields, CVPR 2022.

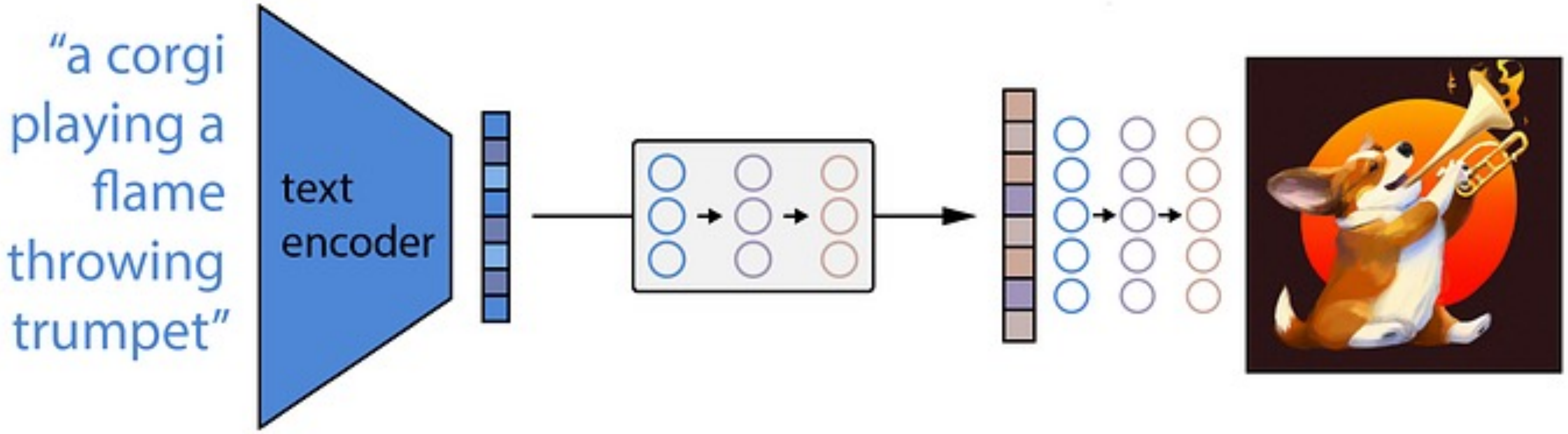
DreamFields [Jain et al., CVPR 2022]

The first 3D generative model using 2D priors!

an archair in the shape of a ____.
an archair imitating a ____.

a teapot in the shape of a ____.
a teapot imitating a ____.

Can we use an image diffusion model instead of CLIP?

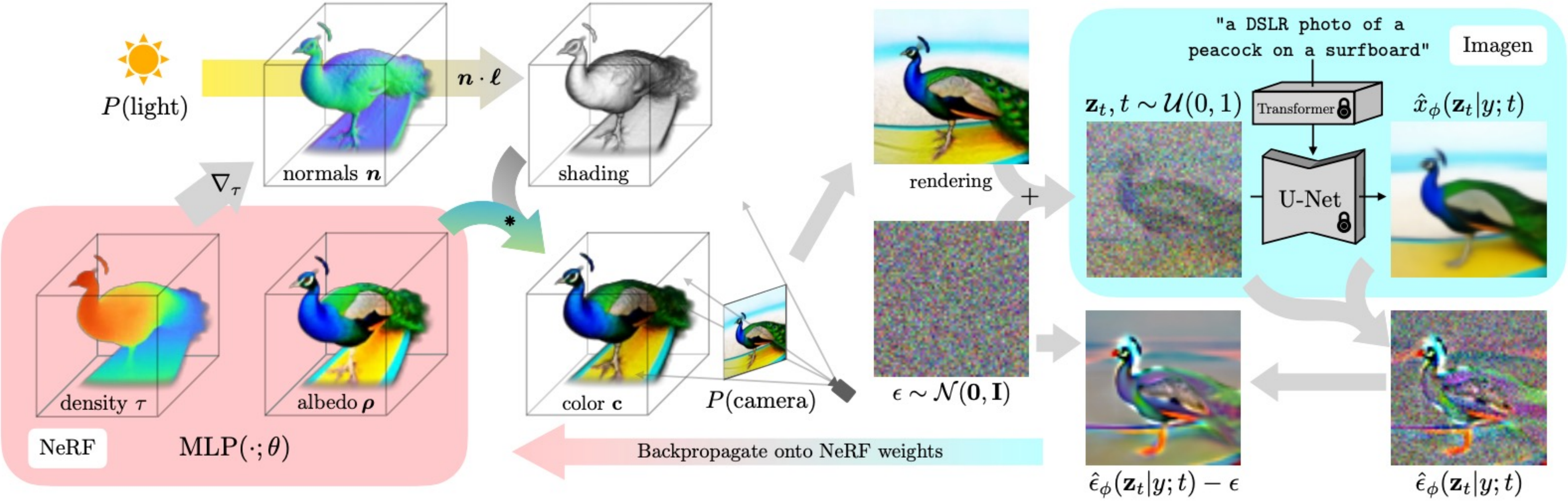


https://medium.com/latinxinai/text-to-image-with-stable-diffusion-4df16da2cfd5

Score Distillation Sampling (SDS)

DreamFusion [Poole et al., ICLR 2023]

Proposed the idea of Score Distillation Sampling (SDS), leveraging a pretrained diffusion model to measure the plausibility of rendered images.



Score Distillation Sampling (SDS)

- *How can we utilize a pretrained diffusion model to measure the plausibility of rendered images?*

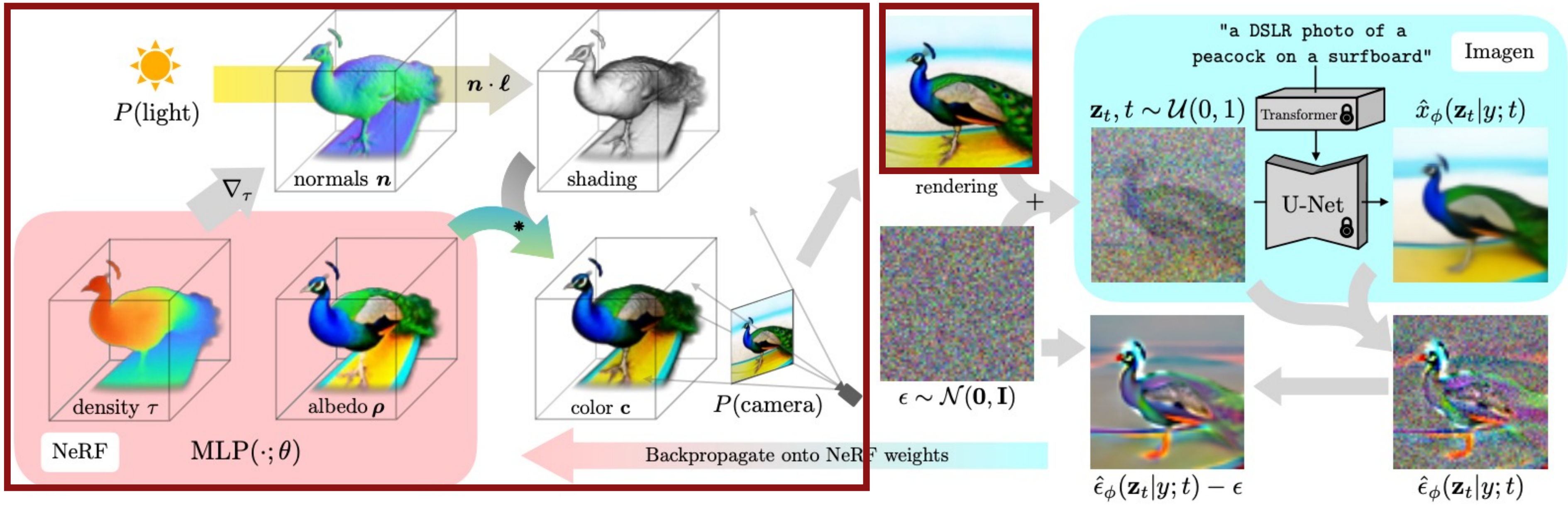
- *Review the loss function:*

$$\mathcal{L} = \mathbb{E}_{t\sim[1,T],\mathbf{x}_0,\boldsymbol{\varepsilon}_t}\left[\left\|\boldsymbol{\varepsilon}_t - \boldsymbol{\varepsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_t, t\right)\right\|^2\right]$$

If the training of the diffusion model has converged, the loss for real data $\mathbf{x}_0$ will be close to zero.

Score Distillation Sampling (SDS)

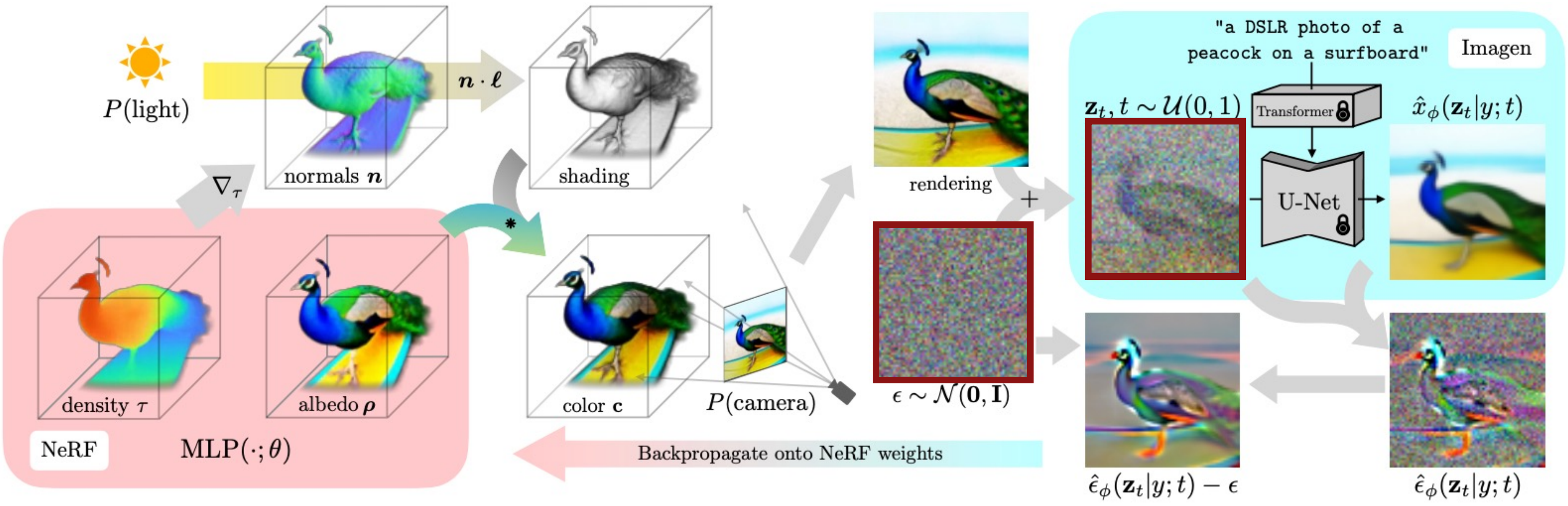
1. Render the NeRF representation into a specific view.
   Let ϕ denote the NeRF parameter, and
   $\mathbf{x}_0 = g(\phi)$ denote the rendered image.



Score Distillation Sampling (SDS)

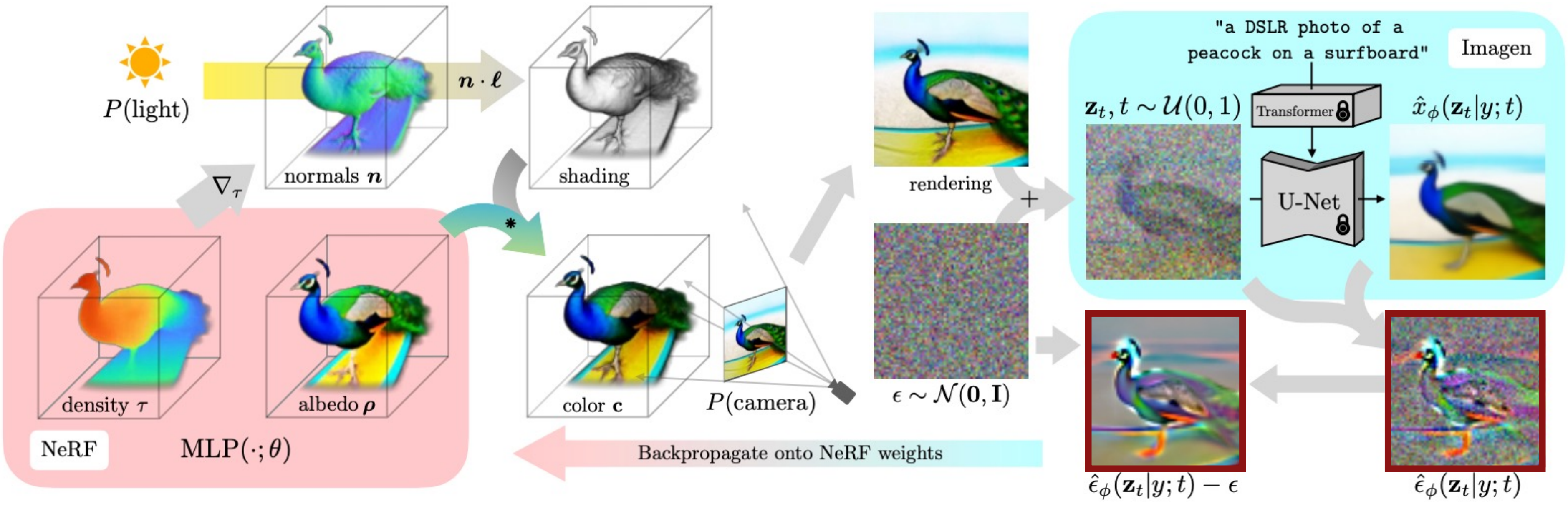
2. *Add noise to the rendered image* $\mathbf{x}_0$*:*

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_t.$$



Score Distillation Sampling (SDS)

3. ***Perform gradient descent on $\mathcal{L}$ with respect to the NeRF parameters ϕ.***



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

DreamFusion Results

“frog wearing a sweater”

“eggshell broken in two with an adorable chick standing next to it”

“ghost eating a hamburger”

“a pig wearing a backpack”

Why SDS Instead of Reverse Diffusion?

This is a scenario where the images are parameterized differently from how they were represented during the training of the diffusion model.

- *Training: Per-pixel colors.*
- *Inference: NeRF rendering.*

Example: Vector Images / Sketches

The same idea but with a different parameterization of images.

Example: Mesh Editing

The same idea but with a different parameterization of images.

Source

ARAP
[Sorkine and Alexa, 2007]

APAP
[Yoo et al., 2024]

Stable-DreamFusion

Stable-Dreamfusion

A pytorch implementation of the text-to-3D model **Dreamfusion**, powered by the Stable Diffusion text-to-2D model.

ADVERTISEMENT: Please check out threestudio for recent improvements and better implementation in 3D content generation!

NEWS (2023.6.12):

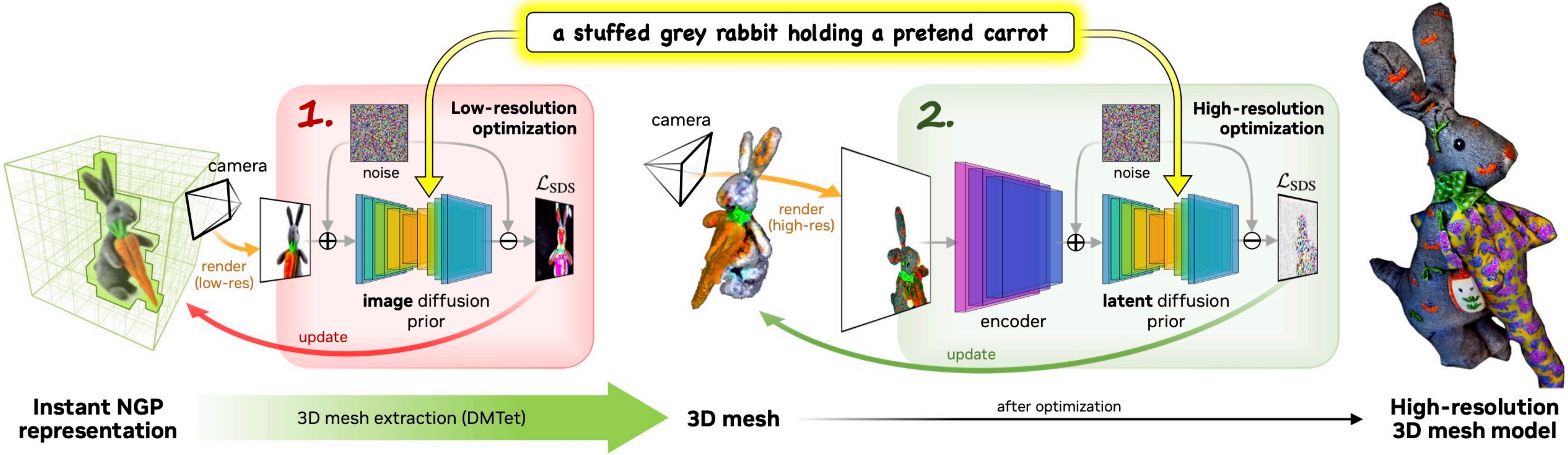
- Support of Perp-Neg to alleviate multi-head problem in Text-to-3D.
- Support of Perp-Neg for both Stable Diffusion and DeepFloyd-IF.

Text-to-3D

a rabbit, animated movie character, high detail 3d model | a DSLR photo of a delicious hamburger | a highly detailed stone bust of Theodoros Kolokotronis | a small saguaro cactus planted in a clay pot

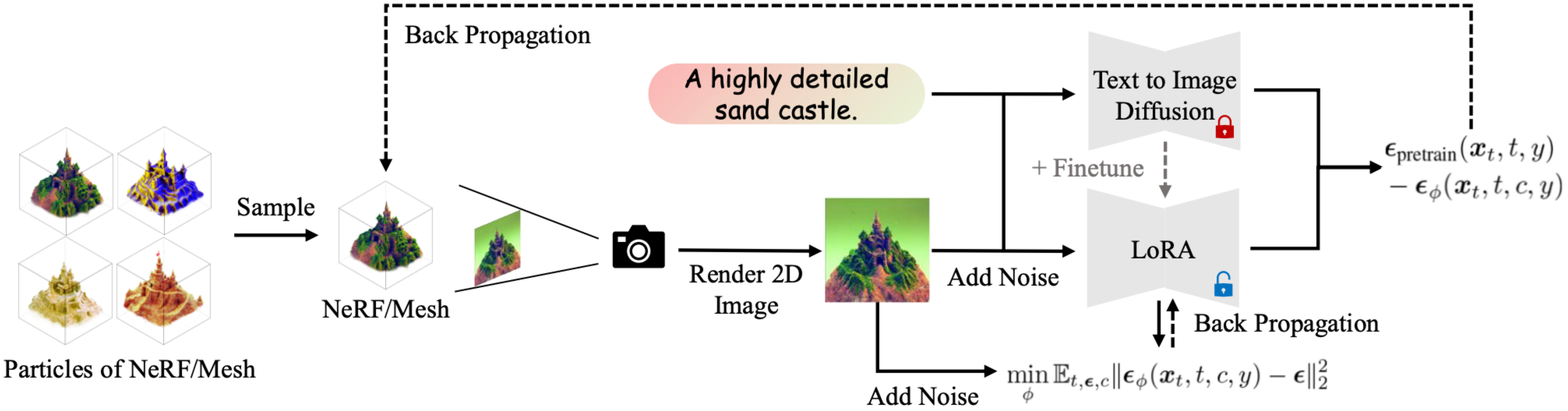
Magic3D [Lin et al., CVPR 2023]

- *Two-stage approach.*
- *Extract a coarse mesh in the first stage, and then texture the mesh in the second stage.*



ProlificDreamer [Wang et al., arXiv 2023]

- *Minimize the SDS loss for the multiple samples of the NeRF parameters ϕ.*
- *Finetune the diffusion model with the Low Rank Adaptation (LoRA) technique.*



Limitation of SDS

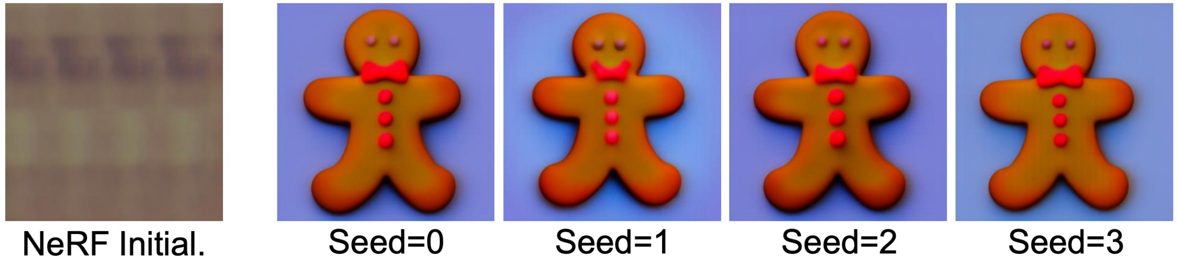
It does not converge well without a high CFG weight (e.g., $w = 400$) and thus suffers from model collapse.

“a delicious hamburger”

$w = 400$

$w = 7.5$

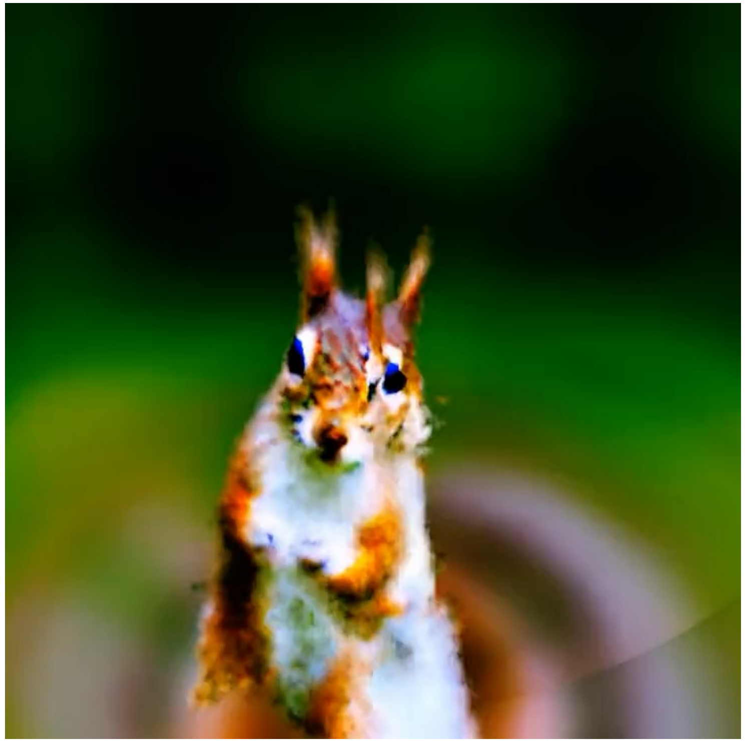
Credit: Jaihoon Kim

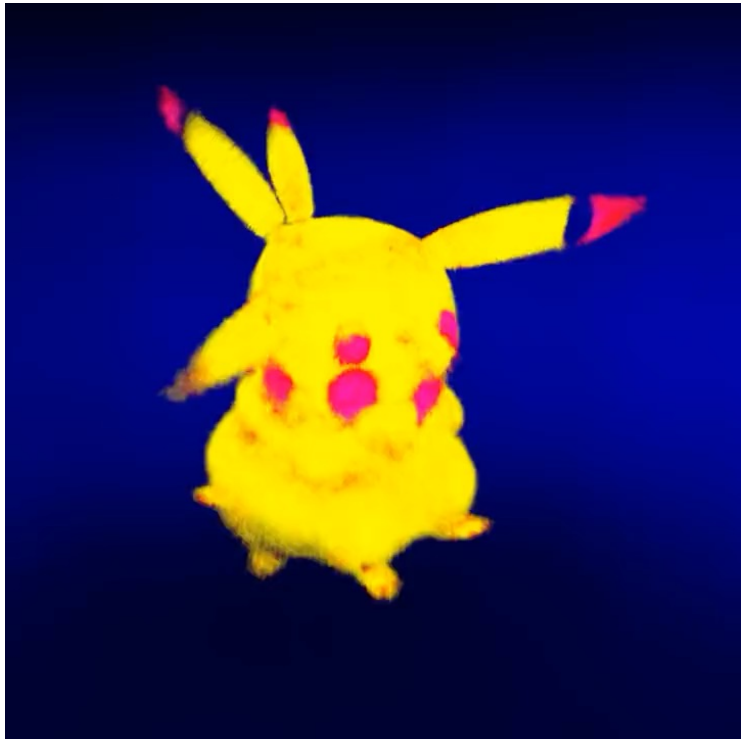


“gingerbread man”

Huang et al., DreamTime: An Improved Optimization Strategy for Text-to-3D Content Creation, arXiv 2023.

Limitation of 3D Generation from 2D Priors





Limitation of 3D Generation from 2D Priors

Supervision for geometry is still needed!

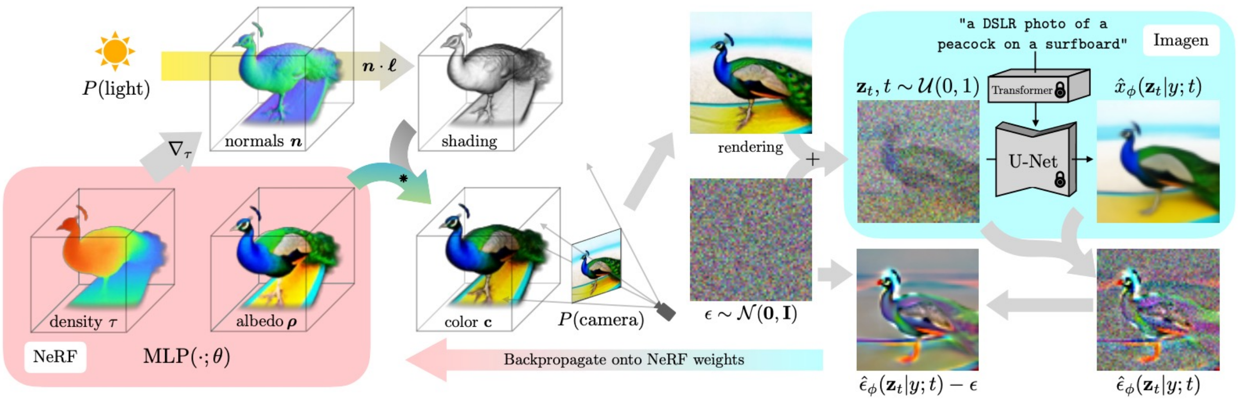
ProlificDreamer, Wang et al., 2023.

StableDreamFusion

What's Next for 3D Generative Models?

1. Combining 3D Supervision

While pretrained image generative models will continue to be valuable for 3D generation, the key to producing realistic and solid 3D shapes would lie in utilizing small-scale 3D priors.



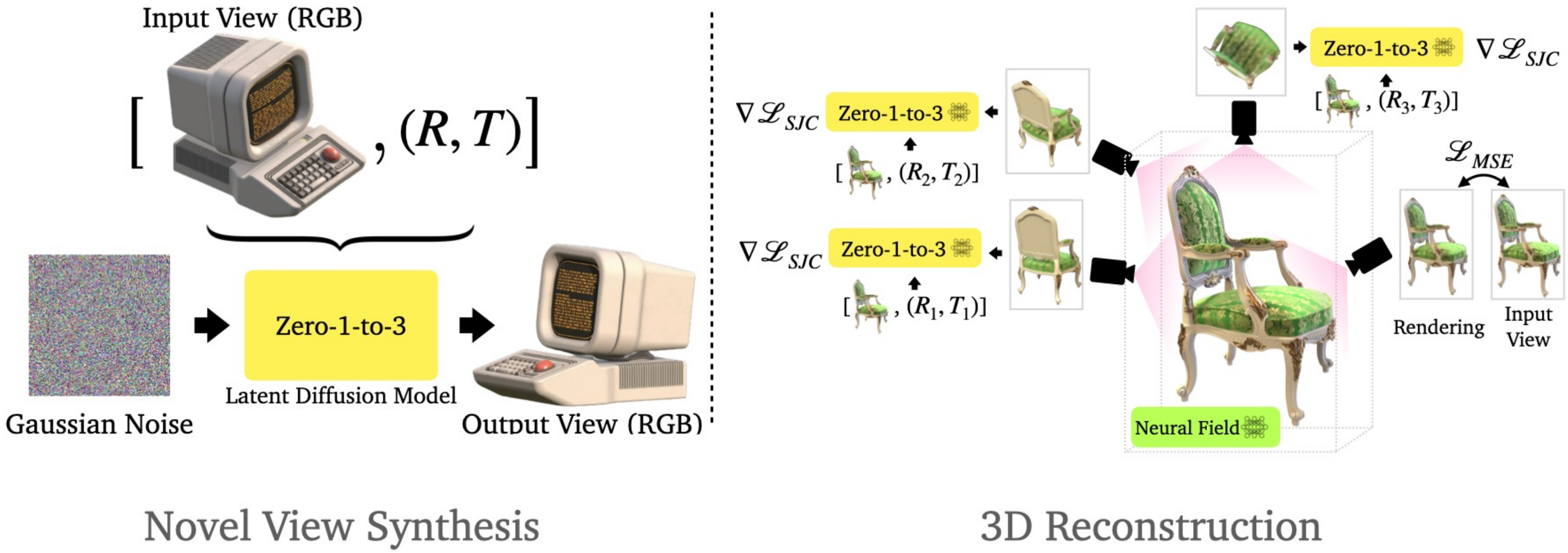
DreamFusion
Google

Objaverse
Allen Institute

Zero-1-to-3 [Liu et al., 2023]

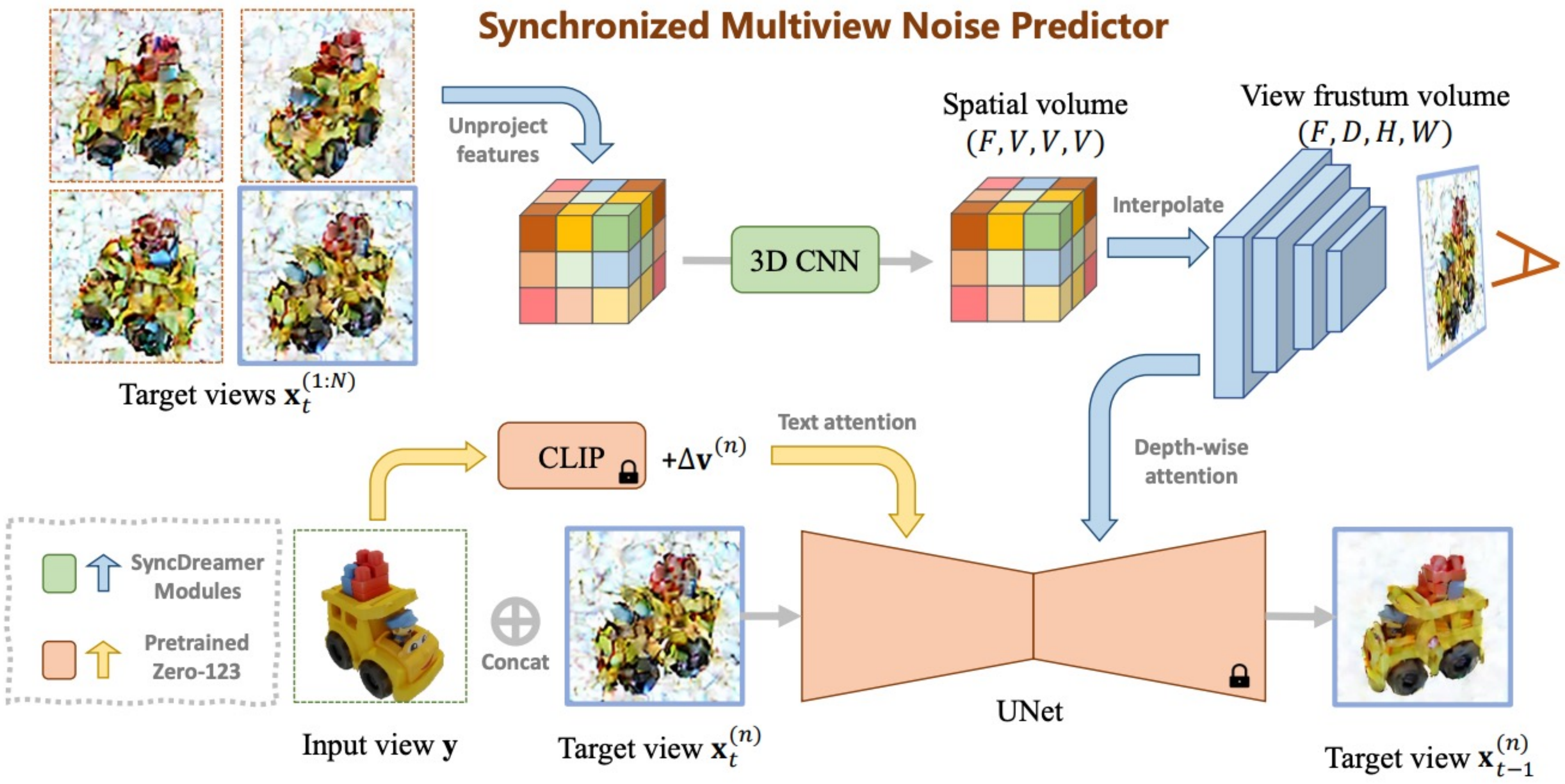
Novel view generation

An image diffusion model generating a novel view image conditioned by another view image and camera pose.



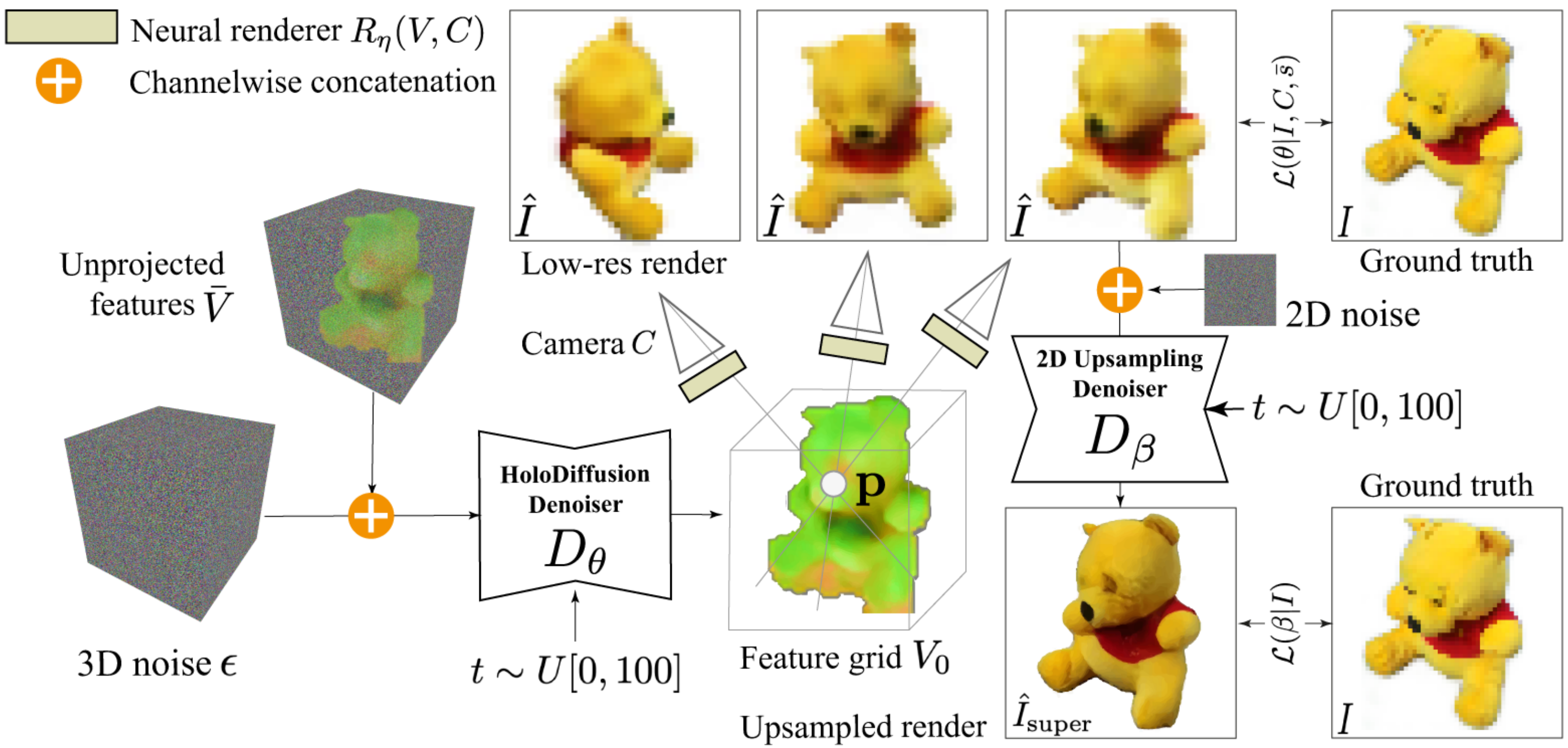
SyncDreamer *[Liu et al., 2023]*

Utilize Zero 1-to-3 to learn the joint probability distribution of multi-view images.



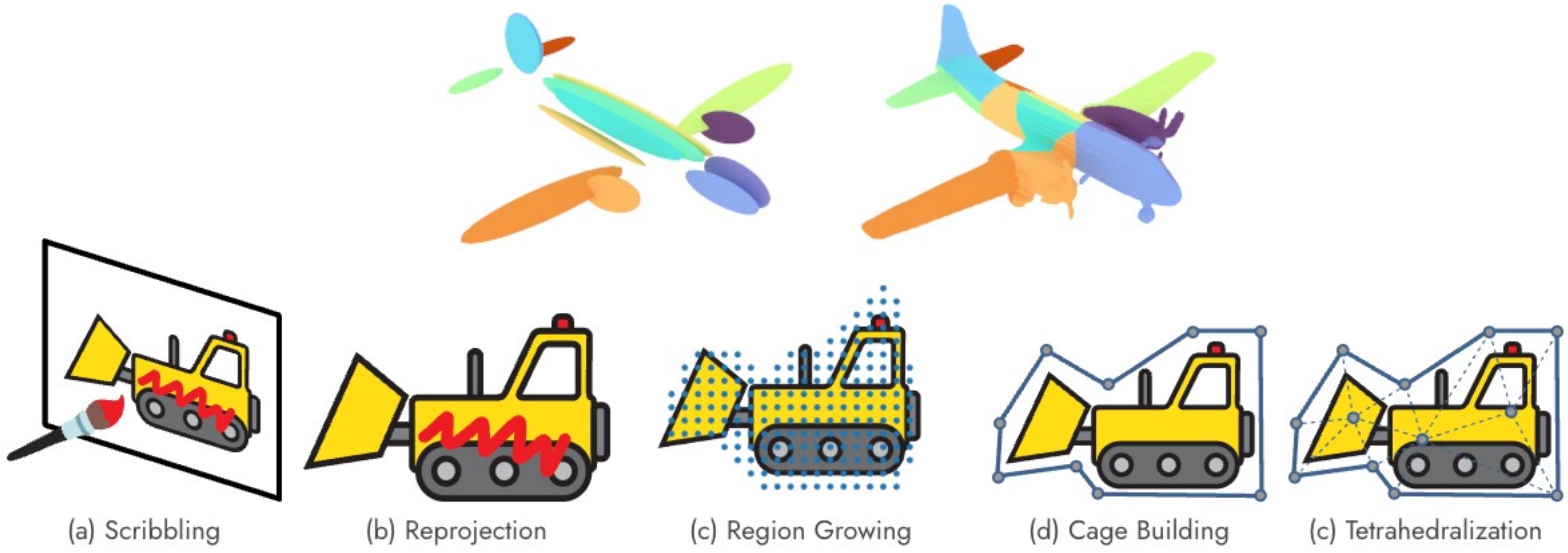
HoloFusion [Karnewar et al., 2023]

- Train a 3D diffusion model using multi-view images only.
- Can be extended to integrate 2D priors.



2. Generation → Editing

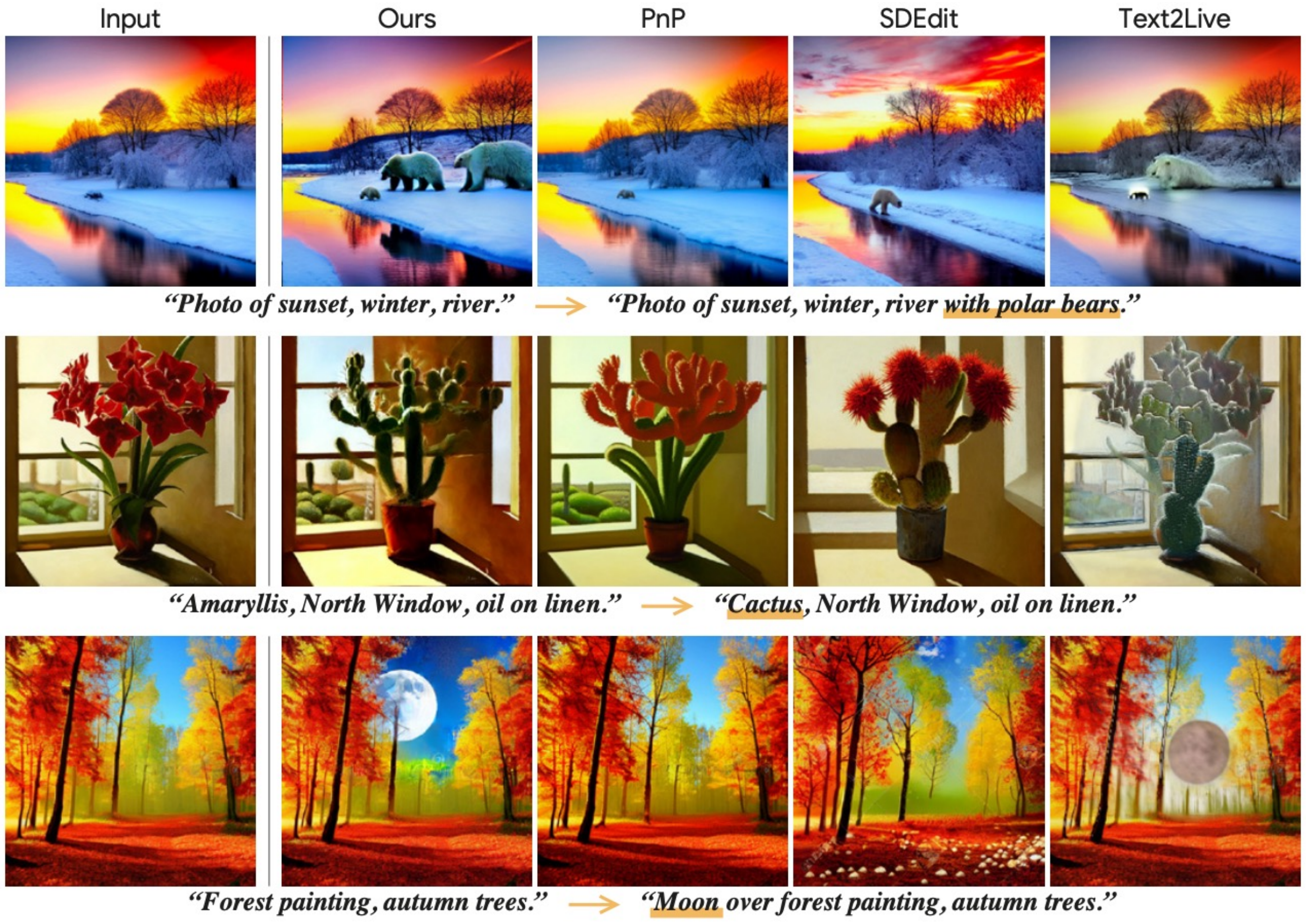
The focus of 3D generative models will shift towards creating versatile models capable of not only generating but editing and manipulating 3D shapes.



NeRFshop, Jambon et al., I3D 2023.

Delta Denoising Score [Hertz et al., 2023]

A new loss function for zero-shot image editing.



Koo *et al.*, *Posterior Distillation Sampling, CVPR 2024.*

Posterior Distillation Sampling [Koo et al., 2024]

A new loss function for zero-shot NeRF editing.

"Leonardo DiCaprio"

Posterior Distillation Sampling [Koo et al., 2024]

A new loss function for zero-shot NeRF editing.

"Raising his arms"

3. Texture Generation

How can we generate photorealistic texture given a 3D mesh or Gaussian splats?