भारतीय प्रौद्योगिकी संस्थान दिल्ली
Indian Institute of Technology Delhi

COV877
Special Module on Visual Computing

Generative AI for Visual Content Creation: Image, Video, and 3D

3D Generation

Instructor:

Dr. Lokender Tiwari

Research Scientist

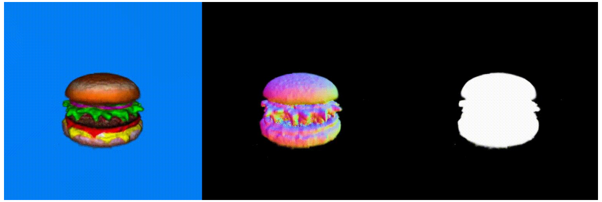
Text to 2D Generation

... many more

[1] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." CVPR 2022.
[2] Saharia, Chitwan, et al. "Photorealistic text-to-image diffusion models with deep language understanding." NeurIPS (2022)
[3] https://www.midjourney.com/

Some existing methods ...

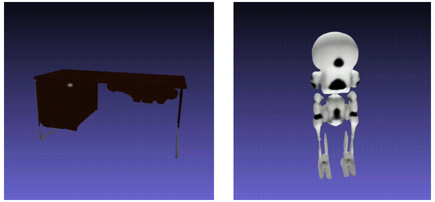
Magic3D / ProlificDreamer / DreamFusion



Input : Text / Image prompt
Output: Mesh with texture

- Time consuming (2-3 hours for 1 object)
- Poor quality mesh
- No editing, articulation
- Janus problem

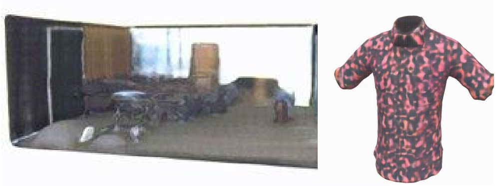
SHAP-E



Input : Text / Image prompt
Output: Mesh with texture

- Very poor quality mesh
- No editing, articulation

GENIE by LUMA AI



Input : Text prompt
Output: textured mesh of scene /object

- Poor quality mesh
- No editing, articulation

Text2Room

Input : Text prompt and initial camera coordinates
Output : mesh of a room

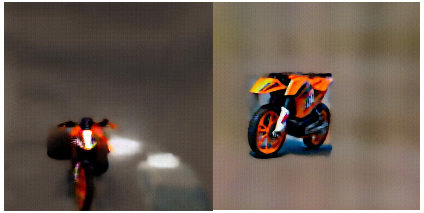
- Non intuitive, require multiple negative prompts
- Poor quality mesh
- No editing, articulation

Neural Articulation Prior

Input : Object part latents
Output: Articulation between object latents

- Dataset specific
- Non-intuitive, every object needs to be converted to its part latents
- Poor quality missing parts

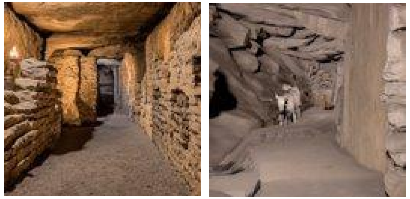
Set the Scene



Input : Text prompt and object meshes
Output: textured mesh of the room resembling the text prompt and given mesh

- Poor quality mesh
- Time consuming (3-4 hours for 1 object)
- No editing, articulation

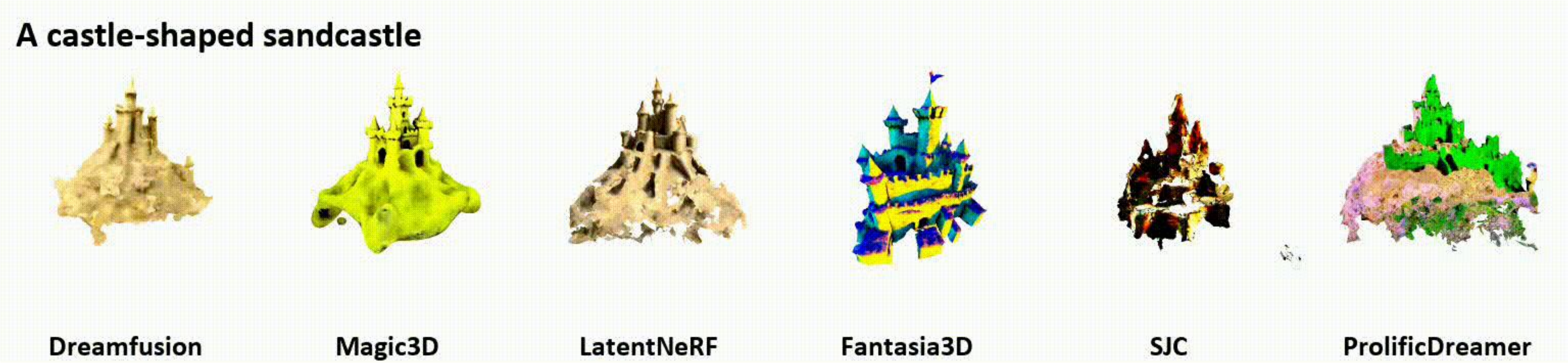
Scene Scape



Input : Text prompt
Output: textured mesh of scene

- Poor quality mesh
- No consistency
- Time consuming and computationally expensive
- No editing, articulation

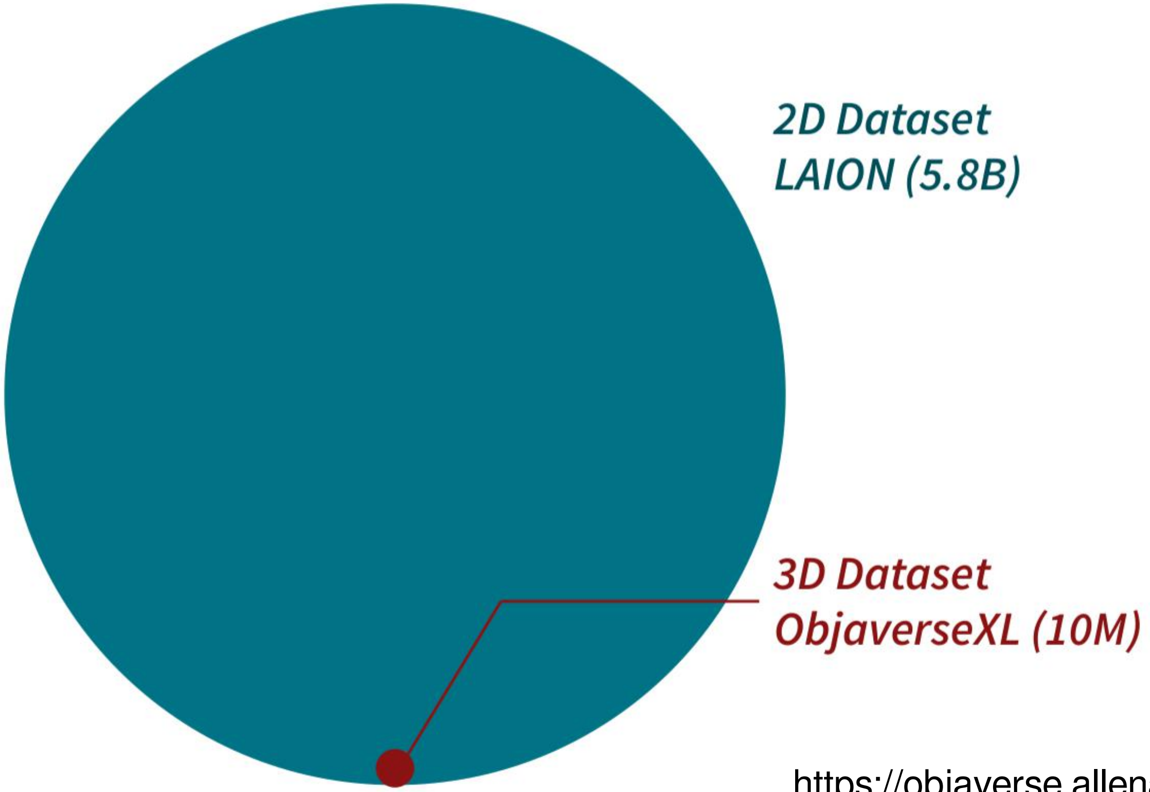
Some existing methods ...



[1] Poole, Ben, et al. "Dreamfusion: Text-to-3d using 2d diffusion." arXiv preprint arXiv:2209.14988 (2022).
[2] Lin, Chen-Hsuan, et al. "Magic3d: High-resolution text-to-3d content creation." CVPR 2023.
[3] Metzer, Gal, et al. "Latent-nerf for shape-guided generation of 3d shapes and textures." CVPR , 2023.
[4] Chen, Rui, et al. "Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation." ICCV, 2023.
[5] Wang, Haochen, et al. "Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation." CVPR 2023.
[6] Wang, Zhengyi, et al. "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation." NeurIPS (2023).

Why Text to 3D not progressed ....

Data scarcity

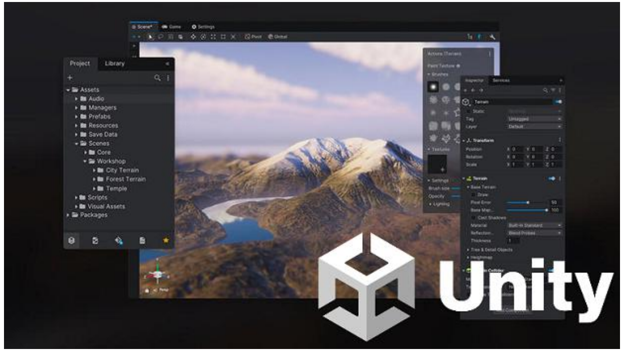


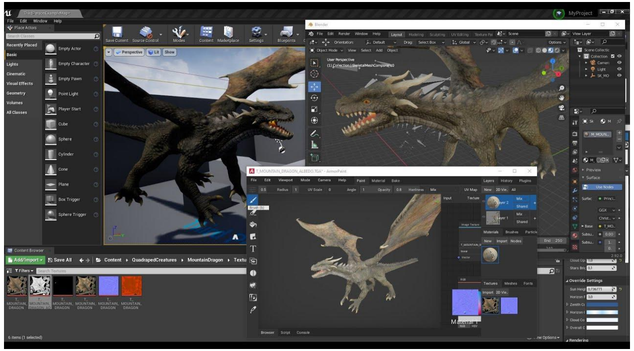
https://objaverse.allenai.org/

Existing tools for 3D creation

blender

UNREAL ENGINE





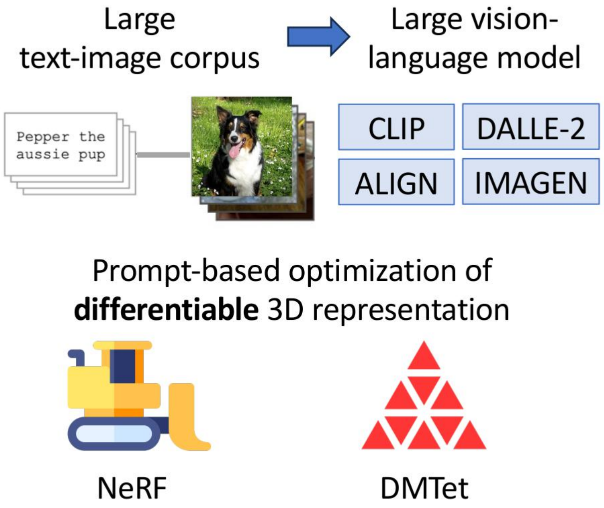
- High quality 3D assets
- Expert 3D artist required
- Time consuming and expensive

Assistive tools are coming up

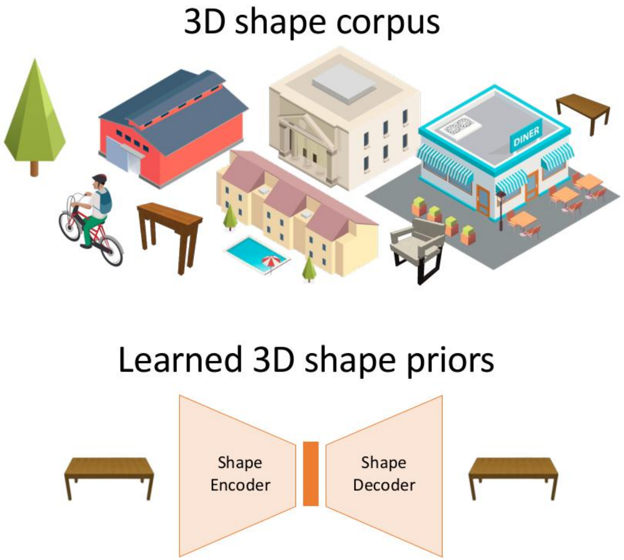
Different Categories

Hybrid3D

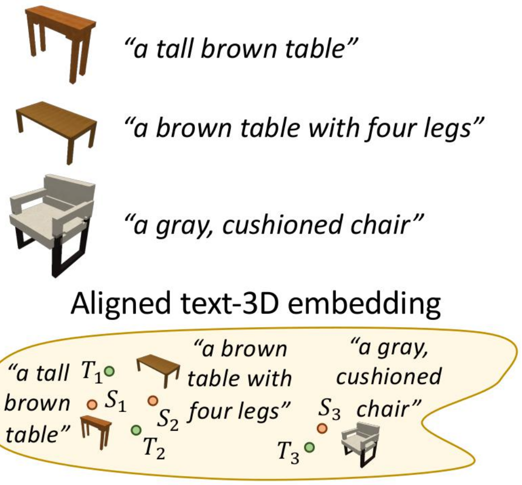
No 3D Data



Unpaired Text-3D Data

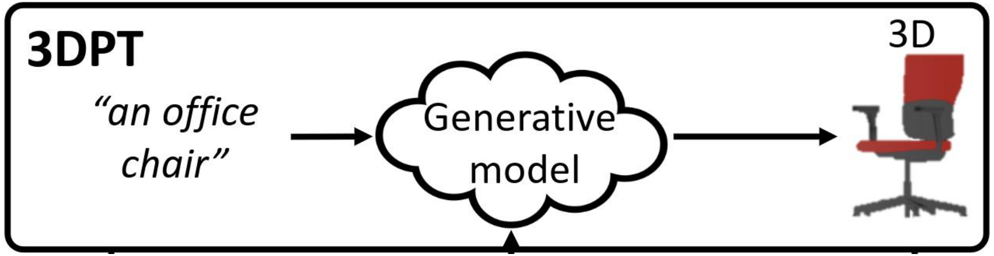


Paired Text-3D Data



[1] Lee, Hanhung, Manolis Savva, and Angel X. Chang. "Text-to-3D Shape Generation." Computer Graphics Forum. Vol. 43. No. 2. 2024.

Different Categories - Training Paradigm

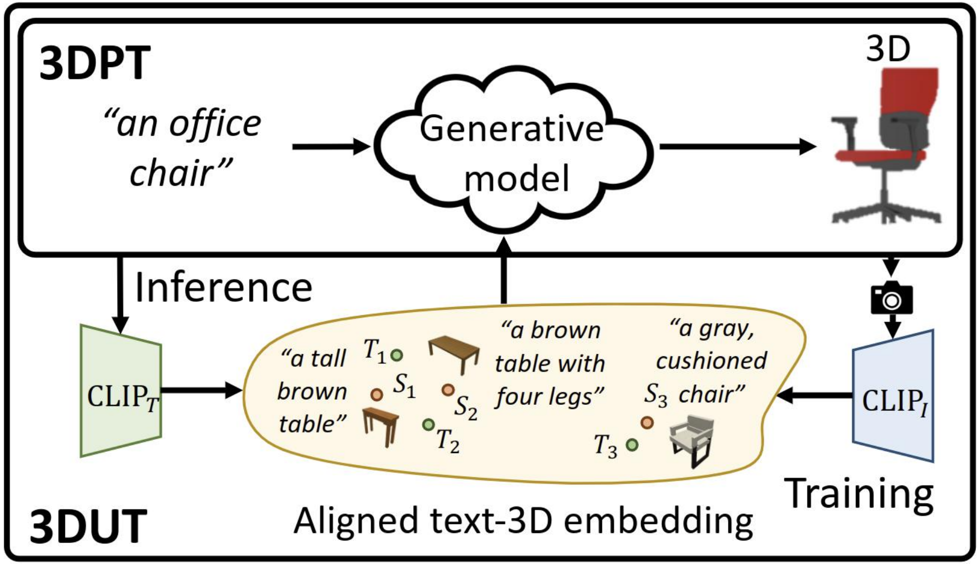


3D Paired Text (3DPT)

- Requires paired text-3D data which is limited.
- Generation limited to observed data.

[1] Lee, Hanhung, Manolis Savva, and Angel X. Chang. "Text-to-3D Shape Generation." Computer Graphics Forum. Vol. 43. No. 2. 2024.

Different Categories - Training Paradigm

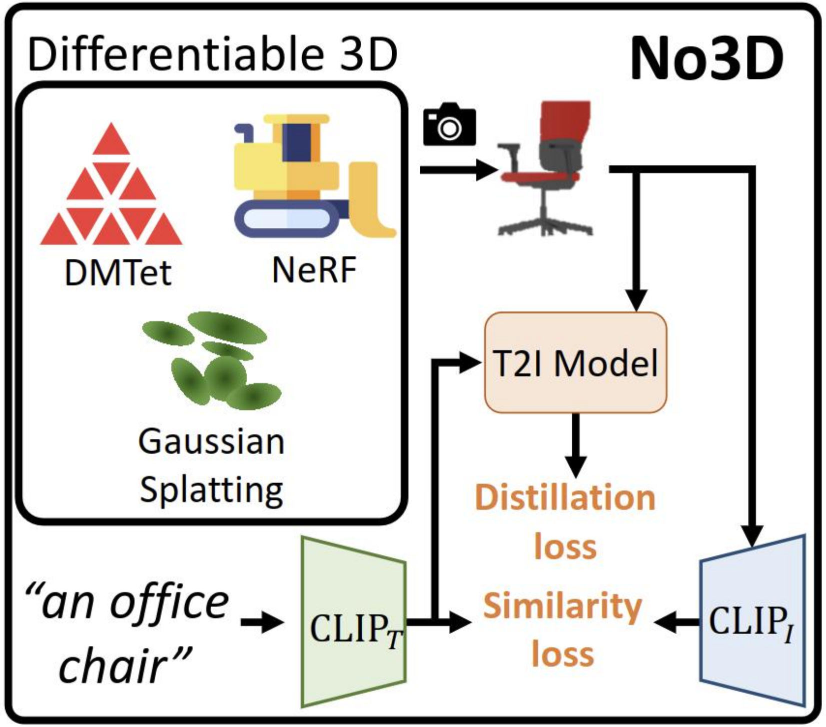


3D Unpaired Text (3DUT)

- Leverages 3D data to train 3D generative model.
- Bridges text and 3D using images.
- Can use vision-language models to generate captions for 3D data, reducing to "Paired" scenario.

[1] Lee, Hanhung, Manolis Savva, and Angel X. Chang. "Text-to-3D Shape Generation." Computer Graphics Forum. Vol. 43. No. 2. 2024.

Different Categories - Training Paradigm

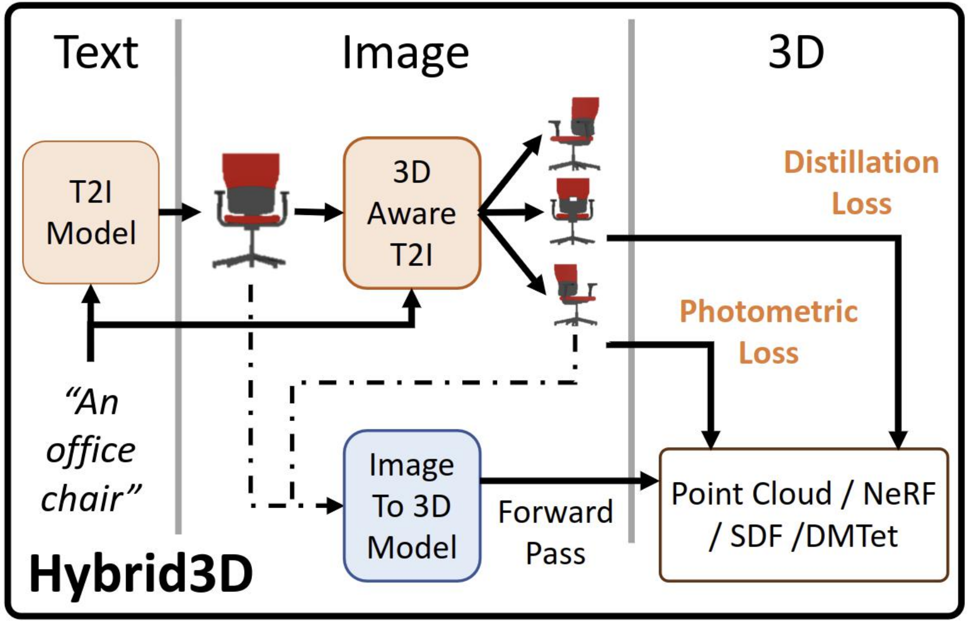


No 3D Data (NO3D)

- No 3D data for training.
- Multi-view and structure consistency is an issue.
- Uses images as bridge, typically with differentiable rendering.
- Conceptually can generate arbitrary 3D content.
- Per-prompt optimization, slow.

[1] Lee, Hanhung, Manolis Savva, and Angel X. Chang. "Text-to-3D Shape Generation." Computer Graphics Forum. Vol. 43. No. 2. 2024.

Different Categories - Training Paradigm



Hybrid3D

- Combine text-to-image and image-to-3D methods.
- Enforce 3D consistency using 3D–aware text-to-image models or multi-view images.

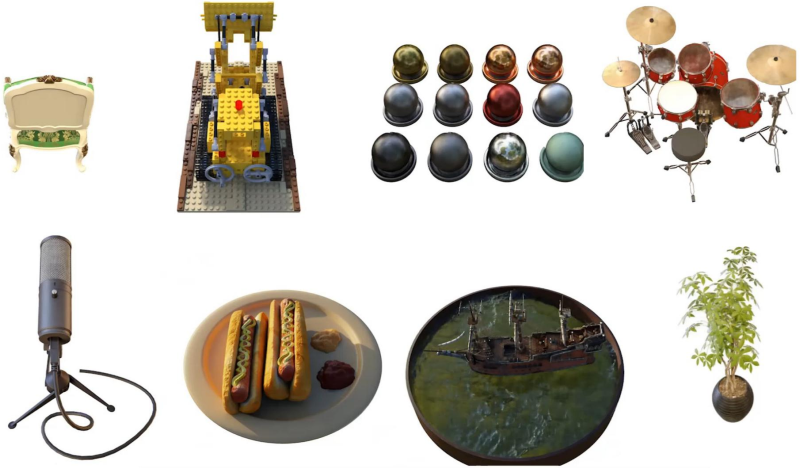
[1] Lee, Hanhung, Manolis Savva, and Angel X. Chang. "Text-to-3D Shape Generation." Computer Graphics Forum. Vol. 43. No. 2. 2024.

Text-to-3D Diffusion Models - No 3D Data

Can we generate a 3D object from its 2D images ?

NeRF

Gaussian Splatting



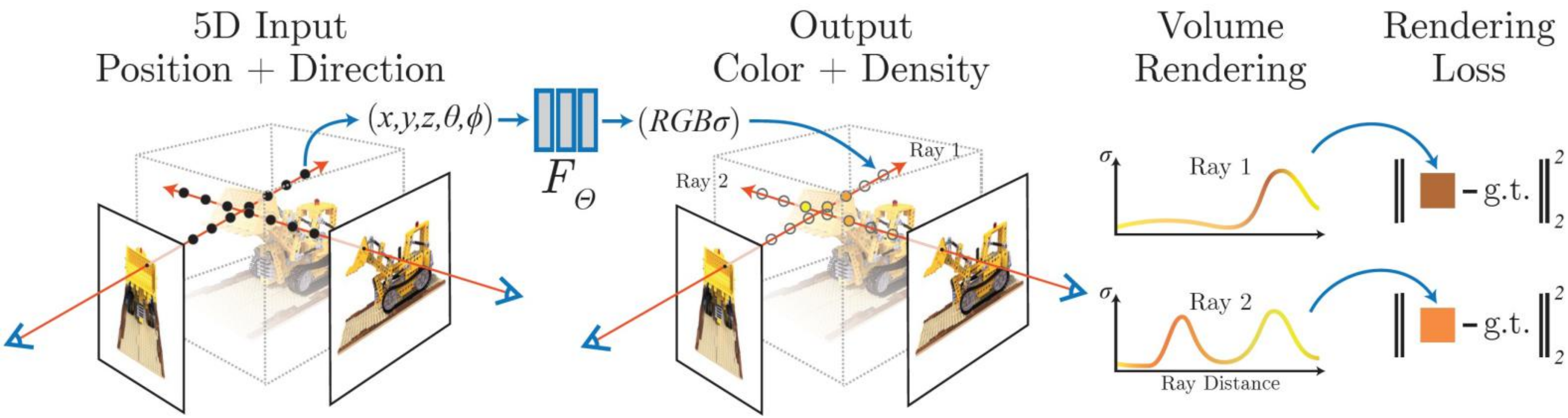
Text-to-3D Diffusion Models - No 3D Data

NeRF

Input : Set of images with camera poses
Output : An implicit representation of 3D object

The optimization loop

1. Render an image from a specific view using NeRF
2. Compute the loss between rendered and ground truth image
3. Compute the gradient and update the NeRF using gradient descent



Text-to-3D Diffusion Models - No 3D Data

Few-shot view synthesis

DietNeRF

Do we need many images ? **No, but...** **additional information would be required**

Leverage **CLIP's** prior knowledge.
CLIP (ICML 2021) : A text-to-image model

It takes a text-image pair as input and compute the alignment between them.

$$\mathcal{L}_{SC}(I, \hat{I}) = \lambda \phi(I)^T \phi(\hat{I})$$

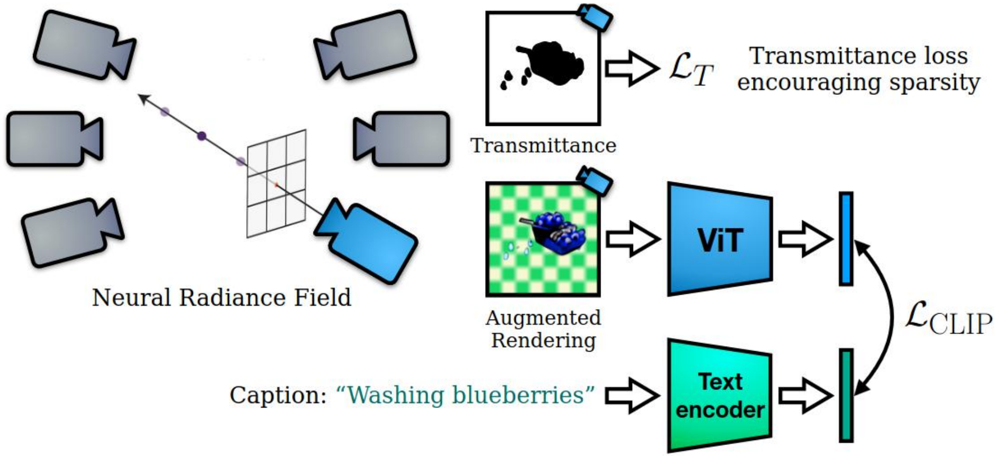
"a bulldozer is a bulldozer from any perspective"

Text-to-3D Diffusion Models - No 3D Data

Zero-shot view synthesis

Dream Fields (CVPR 2022)

Input : A text prompt but no images,
Output : A 3D shape
How ? : Maximize the **similarity** between a **rendered image** and the input prompt in the **CLIP** embedding space.



sample outputs

a robotic dog. a robot in the shape of a dog.

A boat on the water tied down to a stake.

matte painting of a castle made of cheesecake surrounded by a moat made of ice cream

matte painting of a bonsai tree

Text-to-3D Diffusion Models - No 3D Data

Can we do better than CLIP , in terms of **evaluating** the **similarity/plausibility** of the **rendered image** ?

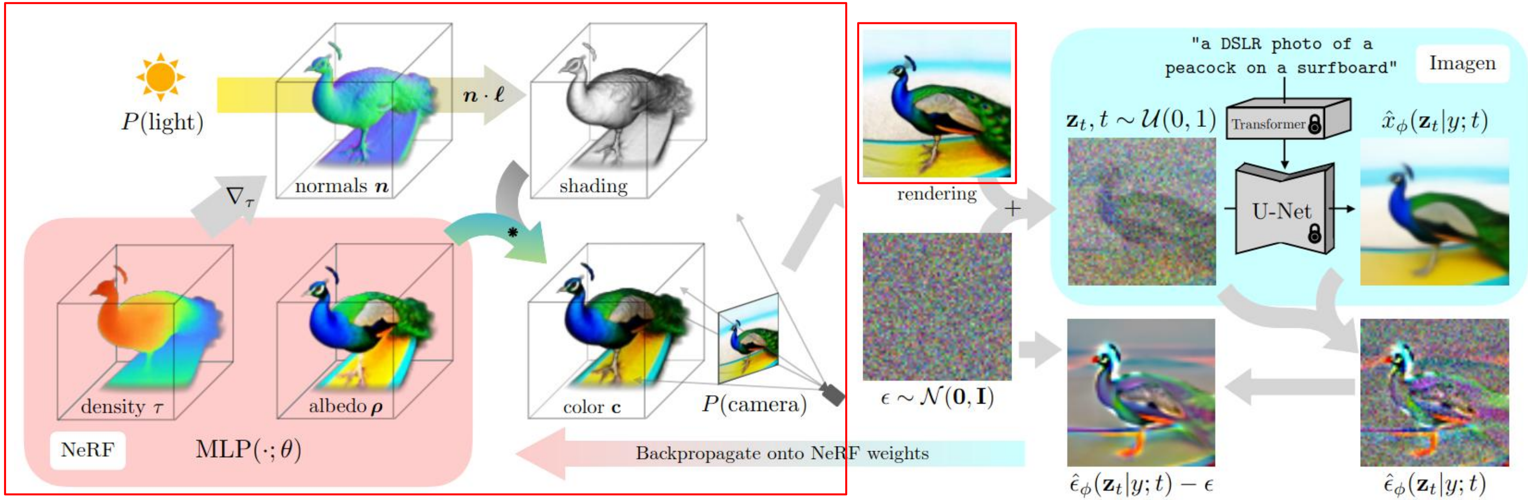
Image Diffusion Models?   DreamFusion (ICLR 2023) : Text-to-3D using 2D Diffusion

Score Distillation Sampling (SDS) : A concept proposed to measure the plausibility of the rendered image, leveraging the 2D diffusion.

How?

...want to create 3D models that look like good images when rendered from random angles...

1. Render an image from a specific view using NeRF

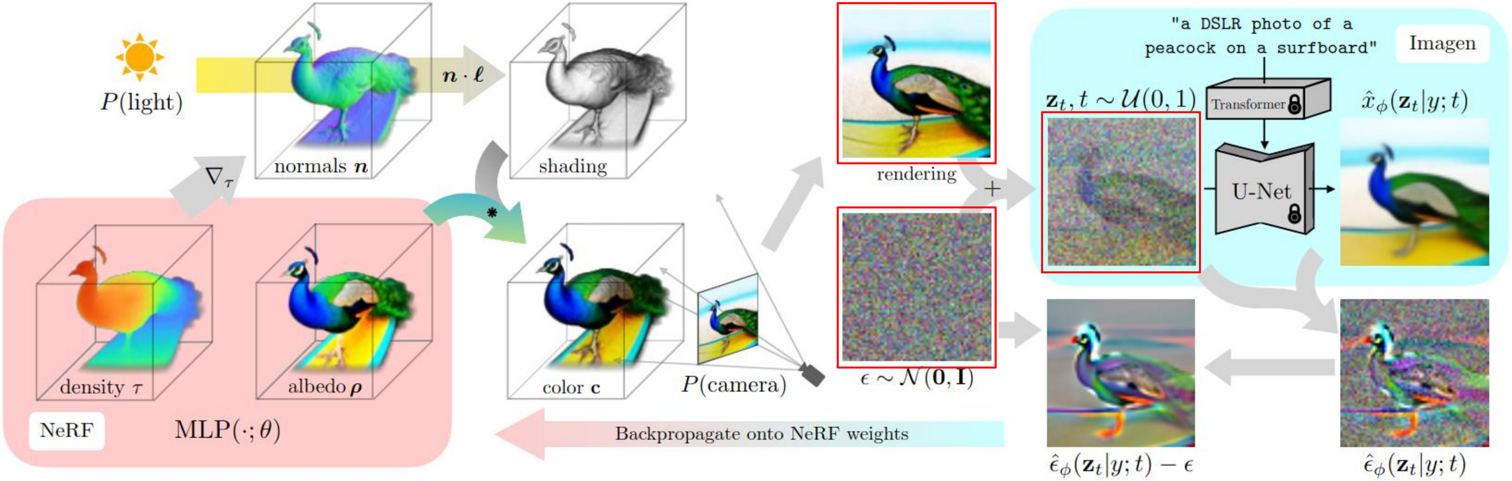


Text-to-3D Diffusion Models - No 3D Data

How?

2. Add noise to the rendered image

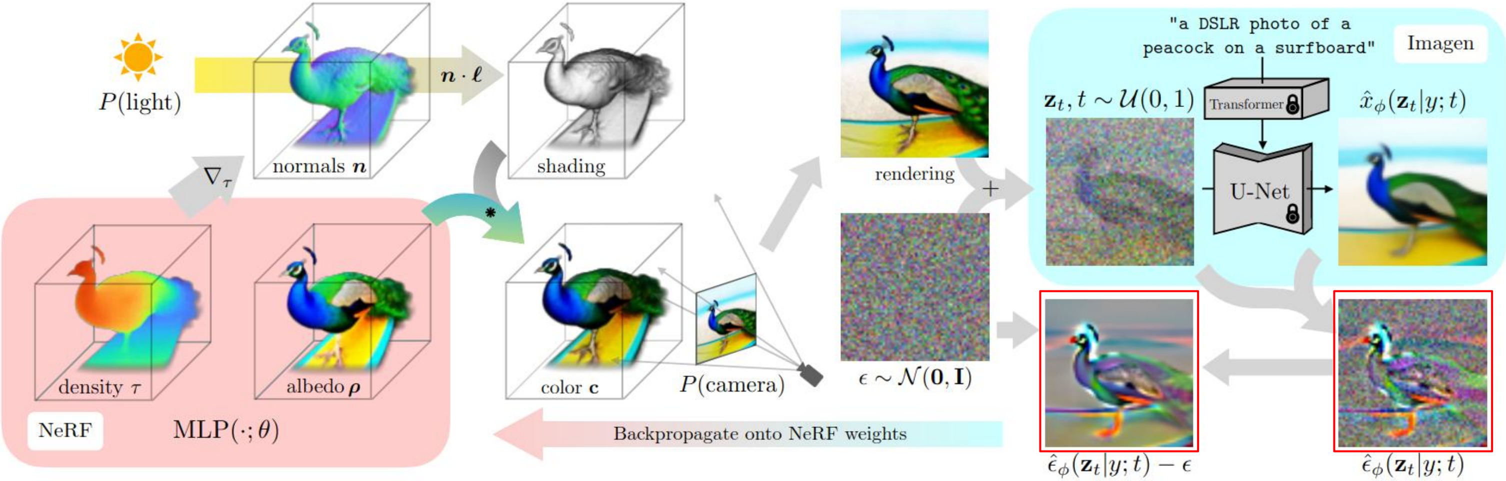
$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_t$$



Text-to-3D Diffusion Models - No 3D Data

How?

3. Perform gradient descent on loss L with respect to the NeRF parameters



Text-to-3D Diffusion Models - No 3D Data

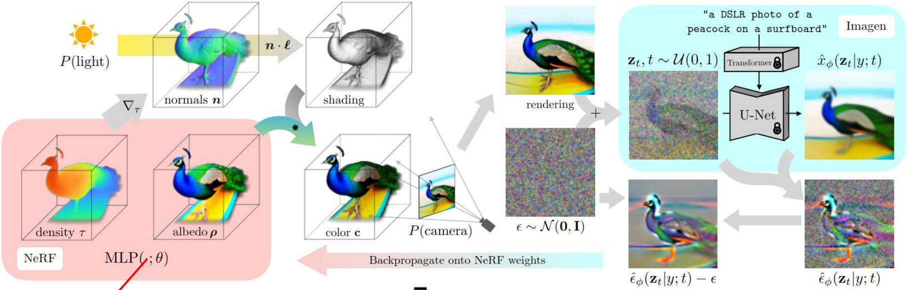
How? - The maths

loss perturbs x with a random amount of noise, and estimates an update direction that follows the score function of the diffusion model to move to a higher density region

$$\mathcal{L}_{\text{Diff}}(\phi, \mathbf{x}) = \mathbb{E}_{t\sim\mathcal{U}(0,1),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[w(t)\|\epsilon_\phi(\mathbf{z}_t; t, y) - \epsilon\|_2^2\right]$$

classifier-free guidance

$$\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) = (1+\omega)\epsilon_\phi(\mathbf{z}_t; y, t) - \omega\epsilon_\phi(\mathbf{z}_t; t)$$



$$\theta^* = \arg\min_\theta \mathcal{L}_{\text{Diff}}(\phi, \mathbf{x} = g(\theta))$$

g differentiable generator

$$\nabla_\theta \mathcal{L}_{\text{Diff}}(\phi, \mathbf{x} = g(\theta)) = \mathbb{E}_{t,\epsilon}\Bigg[w(t)\underbrace{(\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) - \epsilon)}_{\text{Noise Residual}}\underbrace{\cancel{\frac{\partial \hat{\epsilon}_\phi(\mathbf{z}_t; y, t)}{\mathbf{z}_t}}}_{\cancel{\text{U-Net Jacobian}}}\underbrace{\frac{\partial \mathbf{x}}{\partial \theta}}_{\text{Generator Jacobian}}\Bigg]$$

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x} = g(\theta)) \triangleq \mathbb{E}_{t,\epsilon}\left[w(t)(\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) - \epsilon)\frac{\partial \mathbf{x}}{\partial \theta}\right]$$

gradient of a weighted probability density distillation

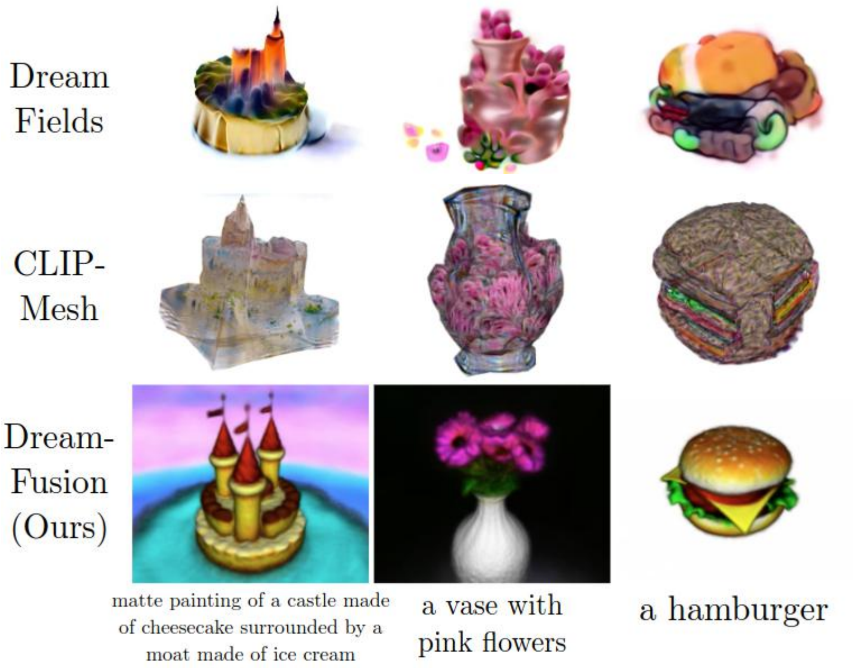
expensive to compute (requires backpropagating through the diffusion model U-Net), and poorly conditioned for small noise levels as it is trained to approximate the scaled Hessian of the marginal density. We found that omitting the U-Net Jacobian term leads to an effective gradient for optimizing

Text-to-3D Diffusion Models - No 3D Data

DreamFusion : Sample Results

More results

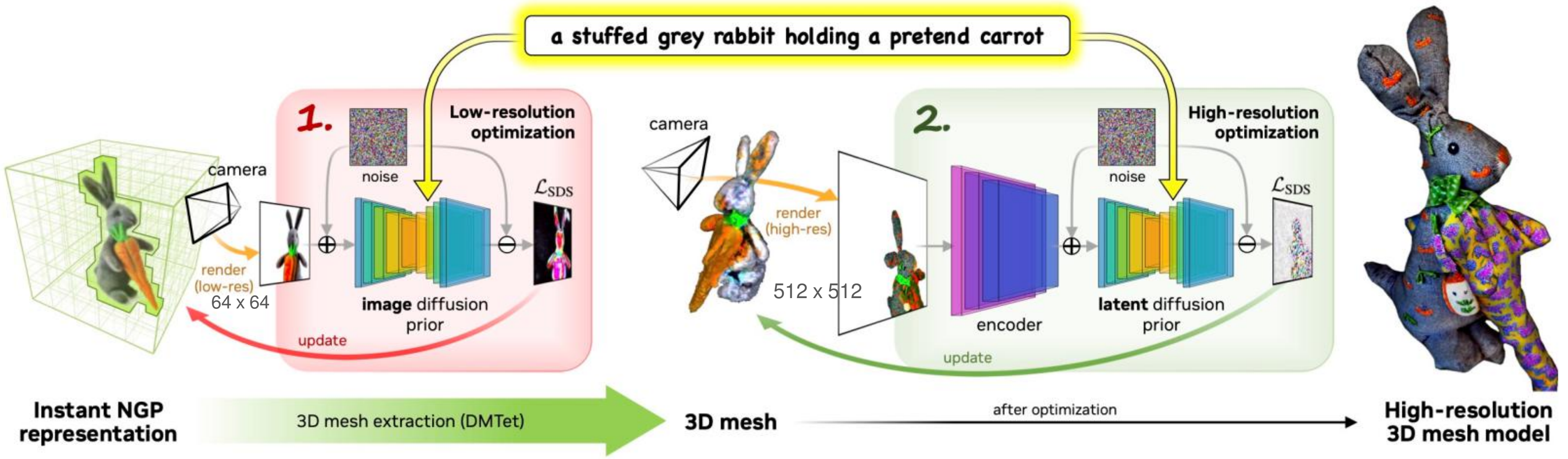
https://dreamfusion3d.github.io/



Text-to-3D Diffusion Models - No 3D Data

Magic3D (CVPR 2023)

- Two stage approach
- Stage 1: Coarse level
- Stage 2: Fine level textured mesh



Text-to-3D Diffusion Models - No 3D Data

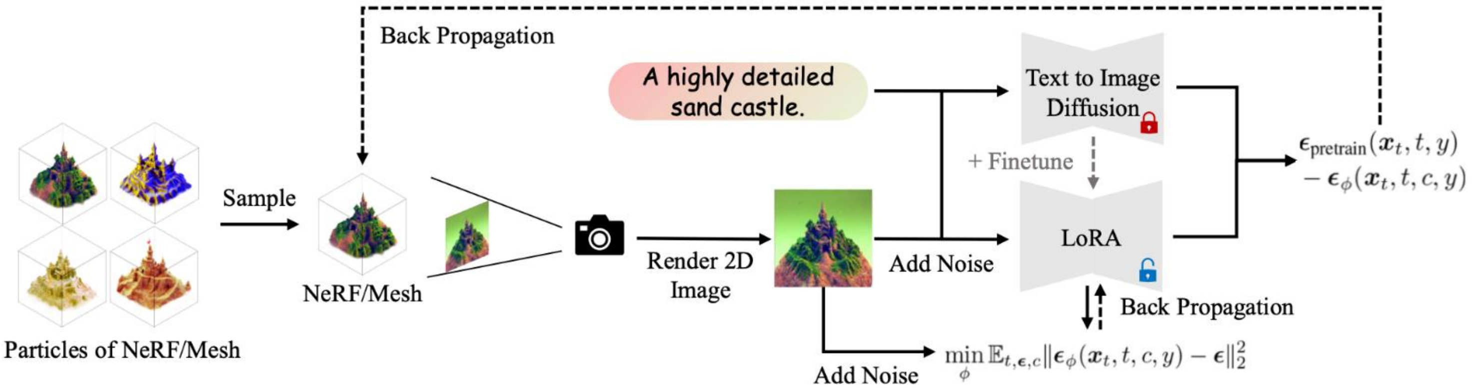
(ProlificDreamer NeurIPS 2023)

- SDS suffers from over-saturation, over-smoothing, and low-diversity problems
- Minimize the SDS loss for multiple sample of NeRF - 3D scene given a textual prompt as a **random variable** instead of a single point as in SDS
- VSD optimizes a distribution of 3D scenes such that the distribution induced on images rendered from all views aligns as closely as possible
- Finetune the diffusion model using low rank adaptation

Variational Score Distillation

$$\nabla_\theta \mathcal{L}_{\mathrm{VSD}}(\theta) \triangleq \mathbb{E}_{t,\boldsymbol{\epsilon},c}\left[\omega(t)\left(\boldsymbol{\epsilon}_{\mathrm{pretrain}}(\boldsymbol{x}_t, t, y^c) - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t, c, y)\right)\frac{\partial \boldsymbol{g}(\theta, c)}{\partial \theta}\right]$$

SDS is a special case of VSD



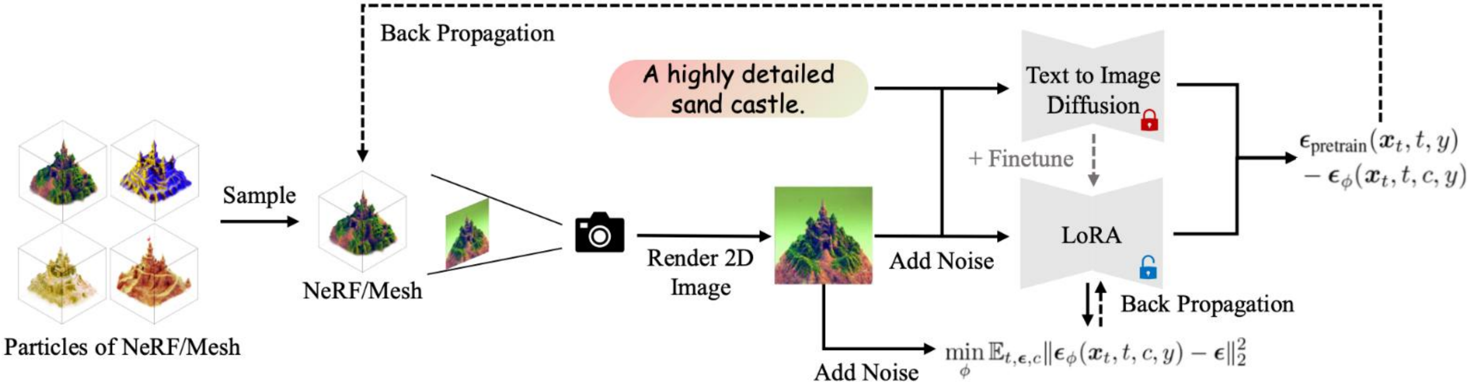
Text-to-3D Diffusion Models - No 3D Data

Algorithm 1 Variational Score Distillation

Input: Number of particles n (≥ 1). Large text-to-image diffusion model $\epsilon_{\text{pretrain}}$. Learning rate η_1 and η_2 for 3D structures and diffusion model parameters, respectively. A prompt y.

initialize n 3D structures $\{\theta^{(i)}\}_{i=1}^{n}$, a noise prediction model ϵ_ϕ parameterized by ϕ.
while not converged **do**
  Randomly sample $\theta \sim \{\theta^{(i)}\}_{i=1}^{n}$ and a camera pose c.
  Render the 3D structure θ at pose c to get a 2D image $\boldsymbol{x}_0 = \boldsymbol{g}(\theta, c)$.
  $\theta \leftarrow \theta - \eta_1 \mathbb{E}_{t,\boldsymbol{\epsilon},c}\left[\omega(t)\left(\boldsymbol{\epsilon}_{\text{pretrain}}(\boldsymbol{x}_t, t, y^c) - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t, c, y)\right)\frac{\partial \boldsymbol{g}(\theta,c)}{\partial \theta}\right]$
  $\phi \leftarrow \phi - \eta_2 \nabla_\phi \mathbb{E}_{t,\boldsymbol{\epsilon}}||\boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t, c, y) - \boldsymbol{\epsilon}||_2^2$.
end while
return

To model the score of the variational distribution, we train an additional diffusion model parameterized by LoRA



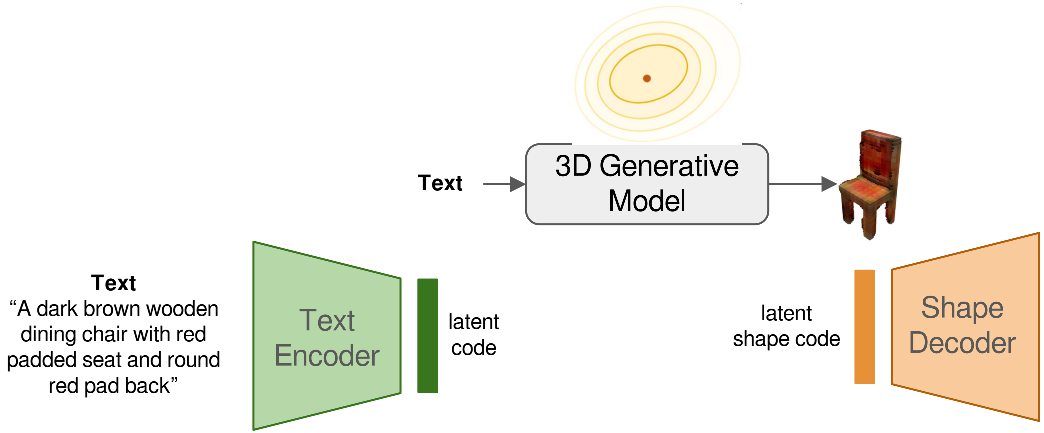
Text-to-3D Diffusion Models - No 3D Data

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

Other issues : “**Janus problem**”

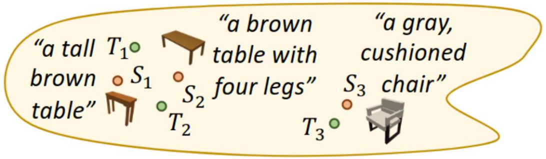
Mitigation ? : Add 3D consistency e.g. MVDream, SweetDreamer

Text-to-3D Diffusion Models - Paired Text-3D Data



When paired text-3D data is available ?

- Train a joint text-shape embedding by specifying the **alignment**



When paired text-3D data is not available ?

- Use **image** as a bridge between text and shape
- Pre-trained vision-language models : aligned text-image embedding space.
- Train shape encoder to align embedding into the same space, and use them to train shape decoder

Text-to-3D Diffusion Models - Paired Text-3D Data

SDFusion (CVPR 2023)

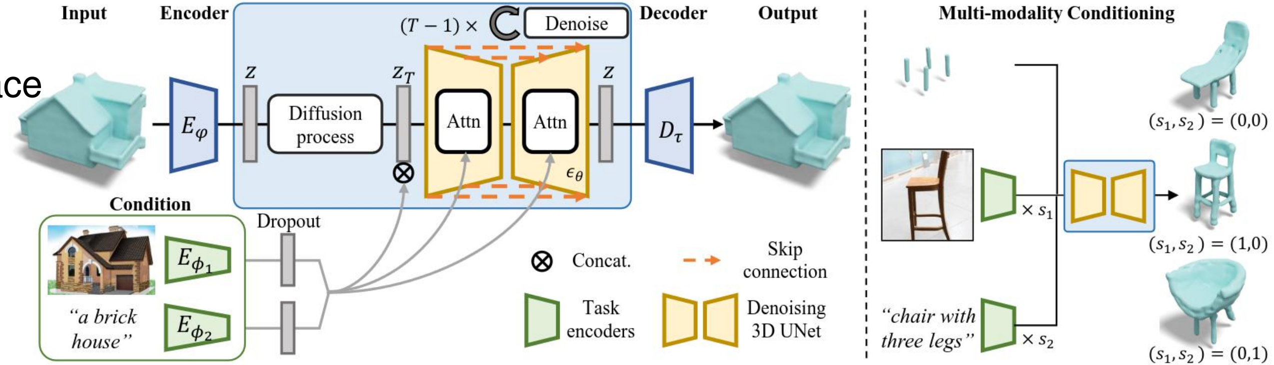
inference time, we can control the importance of each conditioning modality.

Three steps
- compress the 3D shape into a discretized and compact latent space
- Latent diffusion model
- Include user conditions

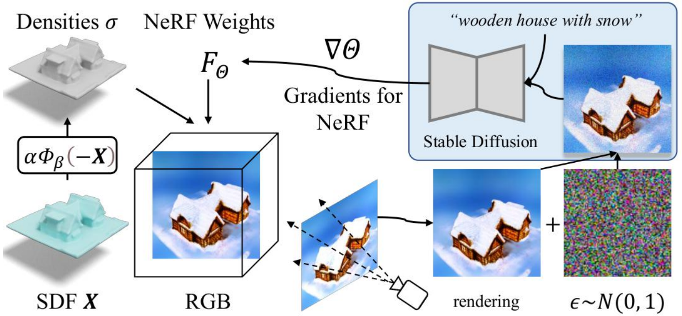
$$L_{\text{simple}}(\theta) := \mathbb{E}_{\mathbf{z},\epsilon\sim N(0,1),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t)\|^2\right]$$

$$L(\theta, \{\phi_i\}) := \mathop{\mathbb{E}}_{\mathbf{z},\mathbf{c},\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, F\{D \circ E_{\phi_i}(\mathbf{c}_i)\})\|^2\right]$$

$$\epsilon_\theta(\mathbf{z}_t, t, F\{E_{\phi_i}\ \forall i\}) = \epsilon_\theta(\mathbf{z}_t, t, \emptyset) + \sum_i s_i \left(\epsilon_\theta(\mathbf{z}_t, t, F\{E_{\phi_i}(\mathbf{c}_i), E_{\phi_j}(\mathbf{c}_j) : \mathbf{c}_j = \emptyset\ \forall j \neq i\}) - \epsilon_\theta(\mathbf{z}_t, t, \emptyset)\right),$$

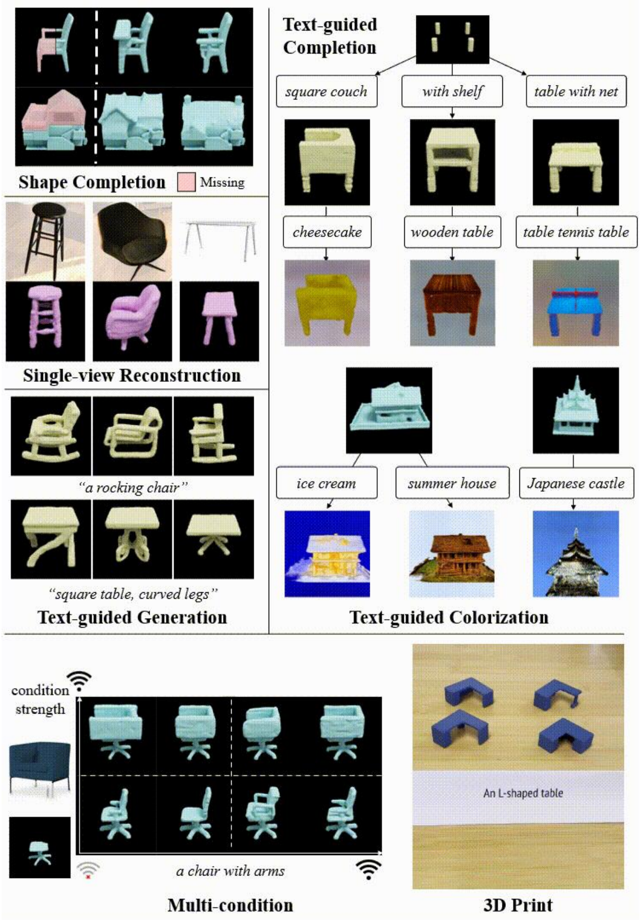


class-specific encoders along with classifier-free guidance to enable multi-modality conditioning



Text-to-3D Diffusion Models - Paired Text-3D Data

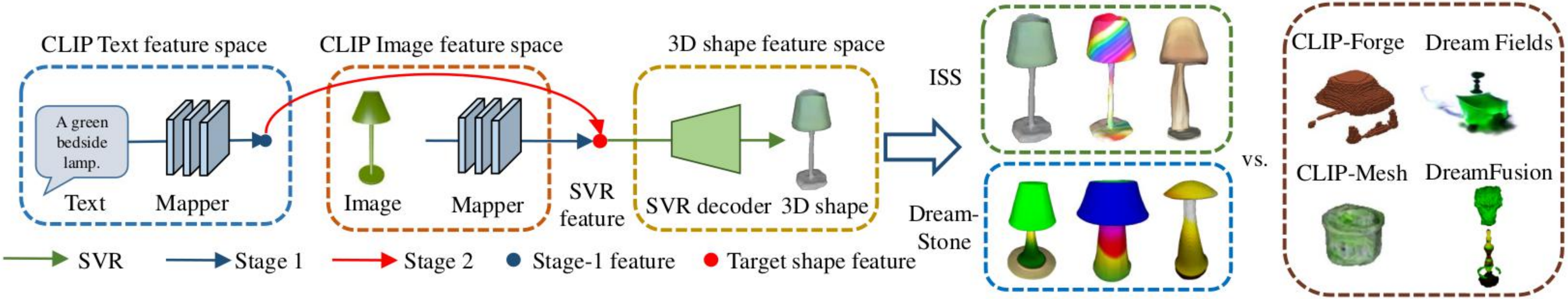
SDFusion - Sample outputs



Text-to-3D Diffusion Models - Un-paired Text-3D Data

DreamStone (TPAMI 2023)

- Two stage feature space alignment approach
- leverages a pre-trained single-view reconstruction (SVR) model to map CLIP features to shapes
- A text-guided shape stylization module that can enhance the output shapes with novel structures and textures



(a) Our DreamStone Framework (b) Text: A {green, rainbow, mushroom} bedside lamp.

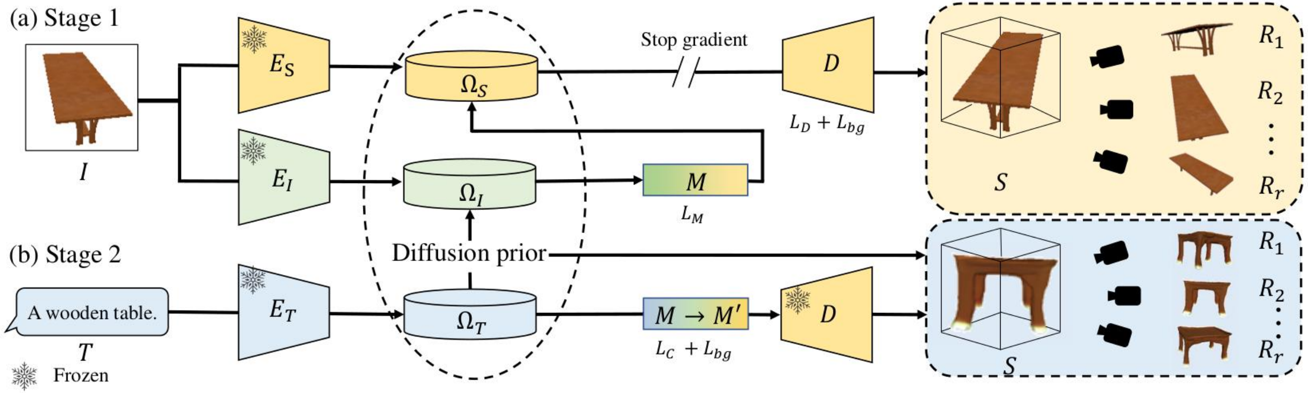
Map the CLIP image feature to the detail-rich shape space in the SVR model

map the CLIP text feature to the shape space by encouraging consistency between the input text and rendered images of the generated shape

Text-to-3D Diffusion Models - Un-paired Text-3D Data

$$\mathcal{L}_M = \sum_{i=1}^{N} ||E_S(I_i) - M(f_{I,i})||_2^2 \qquad \mathcal{L}_{bg} = \sum_p ||D_c(p) - 1||_2^2 \mathbb{1}(F \cap \text{ray}(o,p) = \emptyset)$$

Reduce the semantic gap



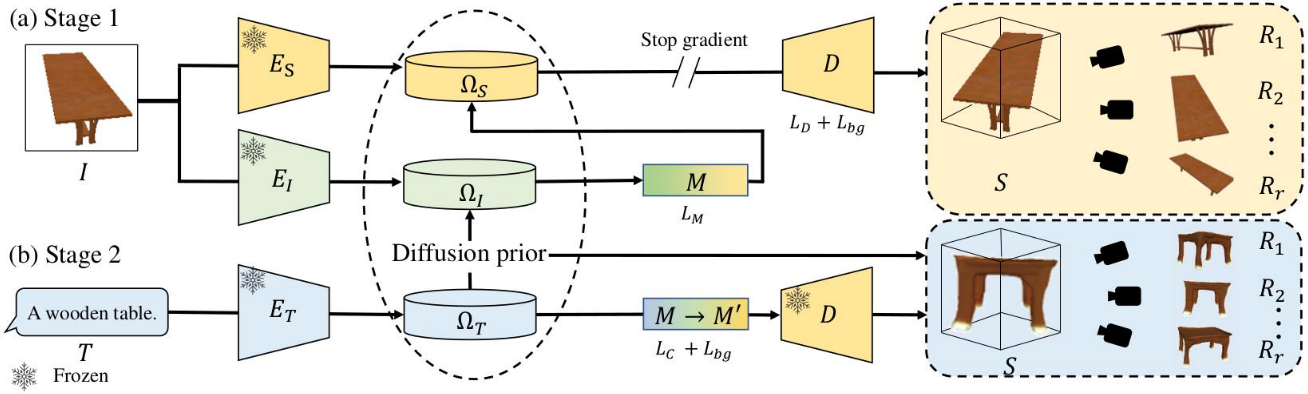
Stage -1

- Leverage a pre-trained single-view reconstruction (SVR) model to align the feature spaces of the CLIP image feature space and the shape space of the SVR model.
- Train CLIP2Shape mapper to map images to shapes while keeping encoder frozen, and
- Fine-tune the decoder using an additional background loss
- During training, we stop the gradients from the SVR loss and the background loss propagating to mapper

Text-to-3D Diffusion Models - Un-paired Text-3D Data

$$\mathcal{L}_M = \sum_{i=1}^{N} ||E_S(I_i) - M(f_{I,i})||_2^2 \qquad \mathcal{L}_{bg} = \sum_p ||D_c(p) - 1||_2^2 \mathbb{1}(F \cap \mathrm{ray}(o,p) = \emptyset)$$

Reduce the semantic gap



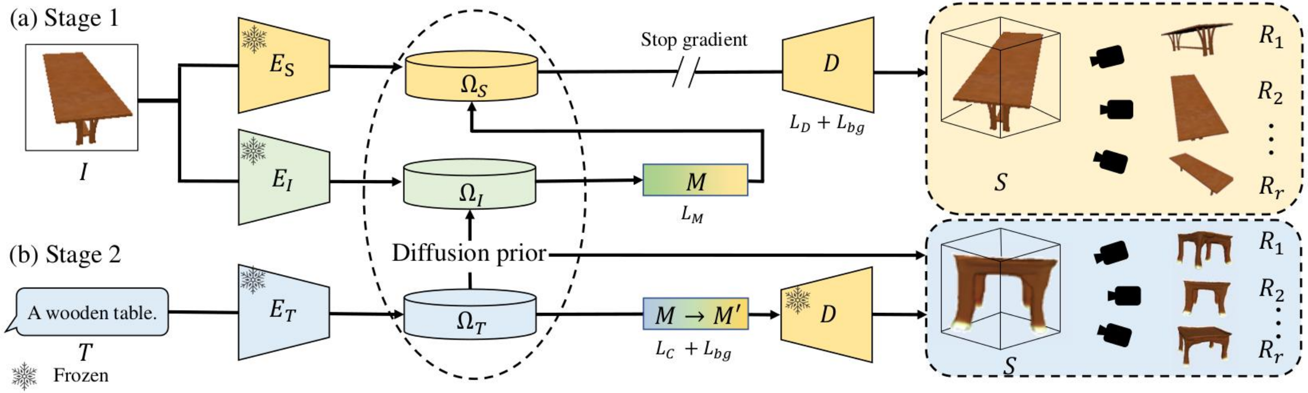
Stage -2

- Fix the decoder D and fine-tuning the mapper M to M ′,
- Encourage the CLIP consistency between the rendered images of the generated shape and the input text T .

Text-to-3D Diffusion Models - Un-paired Text-3D Data

$$\mathcal{L}_M = \sum_{i=1}^{N} ||E_\text{S}(I_i) - M(f_{\text{I},i})||_2^2 \qquad \mathcal{L}_\text{bg} = \sum_p ||D_c(p) - 1||_2^2 \mathbb{1}(F \cap \text{ray}(o,p) = \emptyset)$$

Reduce the semantic gap



fine-tune the mapper M using a **CLIP consistency loss** to reduce the gap between the input text T and **m rendered images** captured from **random camera viewpoints** of the output shape S

$$\mathcal{L}_\text{C} = \sum_{i=1}^{m} \langle f_\text{T} \cdot \frac{E_\text{I}(R_i)}{||E_\text{I}(R_i)||} \rangle$$

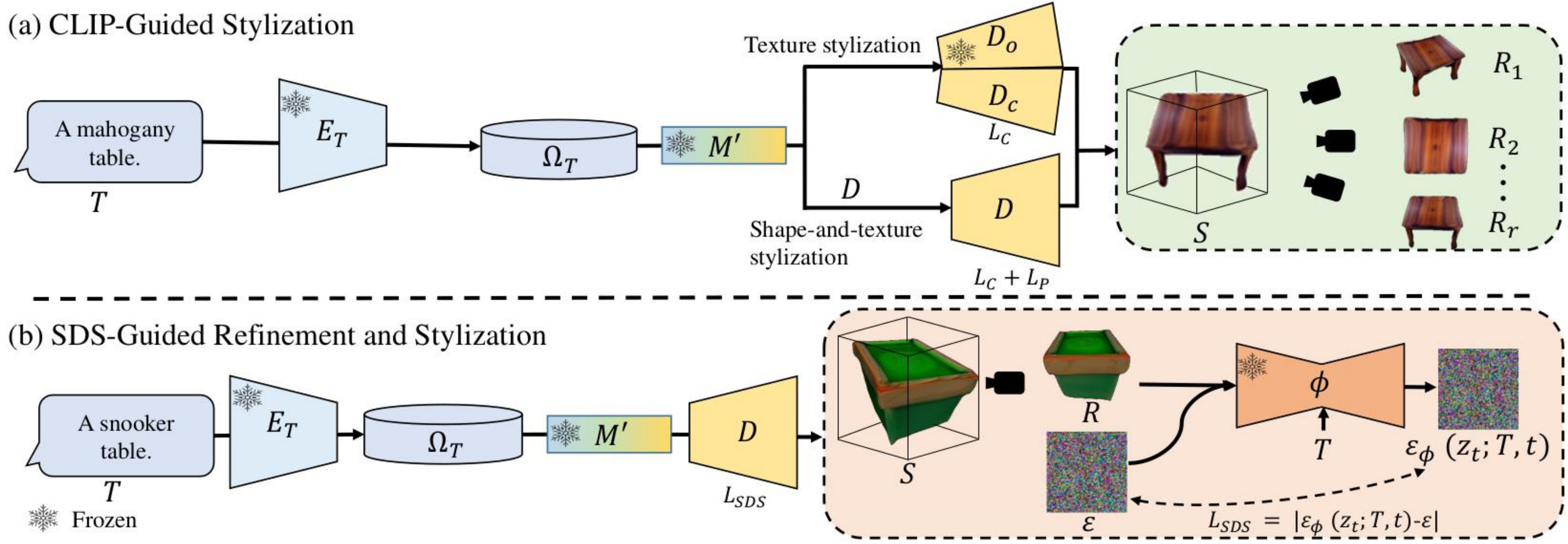
Diverse 3D shape generation? - **use diffusion prior**

$$\mathcal{L}_\text{C} = \sum_{i=1}^{m} \langle (\tau f_{\text{T}\rightarrow\text{I}} + (1-\tau) f_\text{T}) \cdot \frac{E_\text{I}(R_i)}{||E_\text{I}(R_i)||} \rangle$$

text-to-image feature by sampling a random noise

Text-to-3D Diffusion Models - Un-paired Text-3D Data

- The generative space and quality are still limited by the pre-trained SVR model in use
- Further refinement can be done using CLIP or SDS guided stylization/refinement



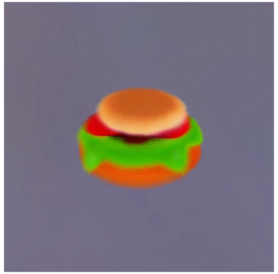
Text-to-3D Diffusion Models - Un-paired Text-3D Data

DreamStone sample outputs

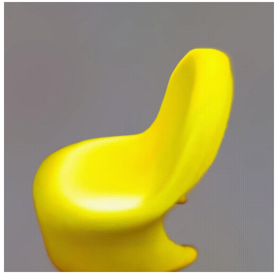
"A chair imitating avocado"

"A 3D model of an adorable cottage with a thatched roof"

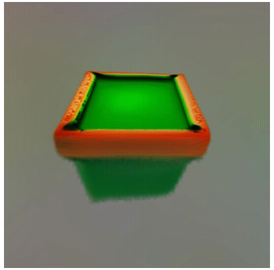
"A_hamburger"



"A chair imitating banana"



"Snooker table"



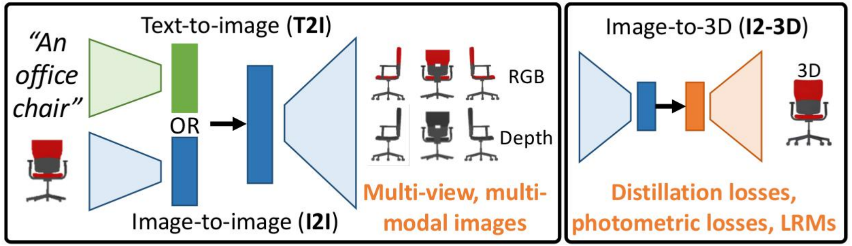
"A swivel chair with wheels"

"This is a bar stool with metal arches as a design feature"

More results

https://liuzhengzhe.github.io/DreamStone.github.io/

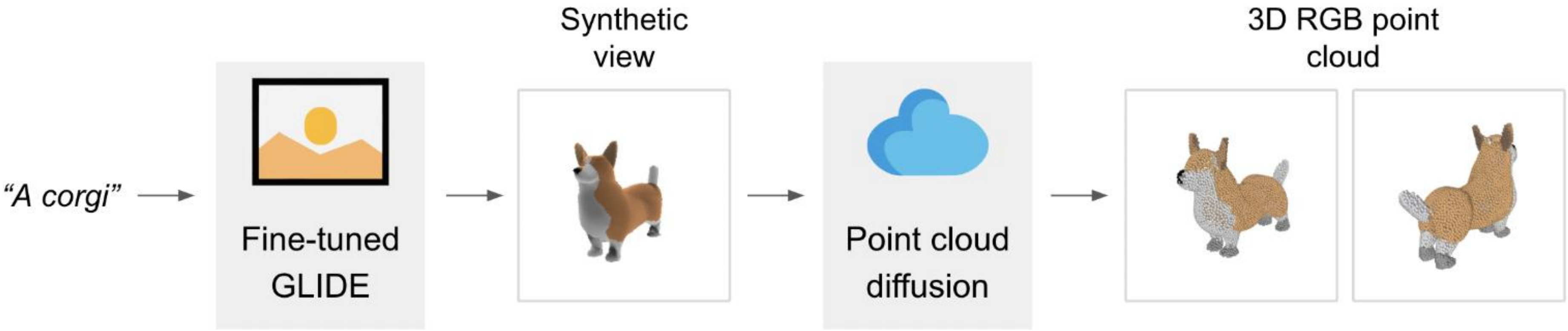
Text-to-3D Diffusion Models - Hybrid 3D



Point-E by OpenAI

- a hybrid text-image-3D model
- Fast 3D generation (1-2 mins)

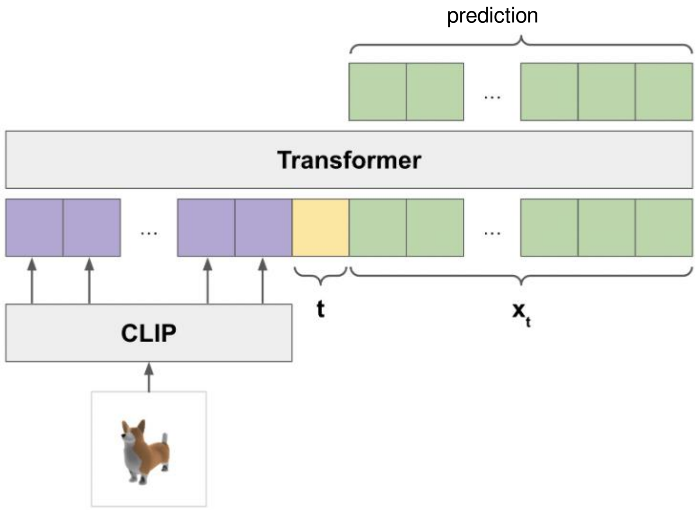
Step 1: generate a synthetic view conditioned on a text caption.
Step 2: generate a coarse point cloud (1,024 points) conditioned on the synthetic view.
Step 3: generate a fine point cloud (4,096 points) conditioned on the low-resolution point cloud and the synthetic view.



Text to image synthesis models

Text-to-3D Diffusion Models - Hybrid 3D

- Point Cloud Diffusion : Represent point cloud as a tensor of shape K × 6
  - (coordinates + colors)
- run each point in point cloud through a linear layer and obtain a K × D input tensor.
- run the timestep t through a small MLP, obtaining another D-dimensional vector to prepend to the context
- Run the image to CLIP (ViT-L/14 CLIP model)

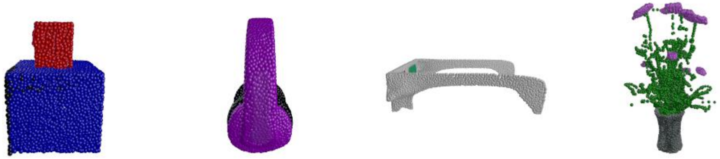


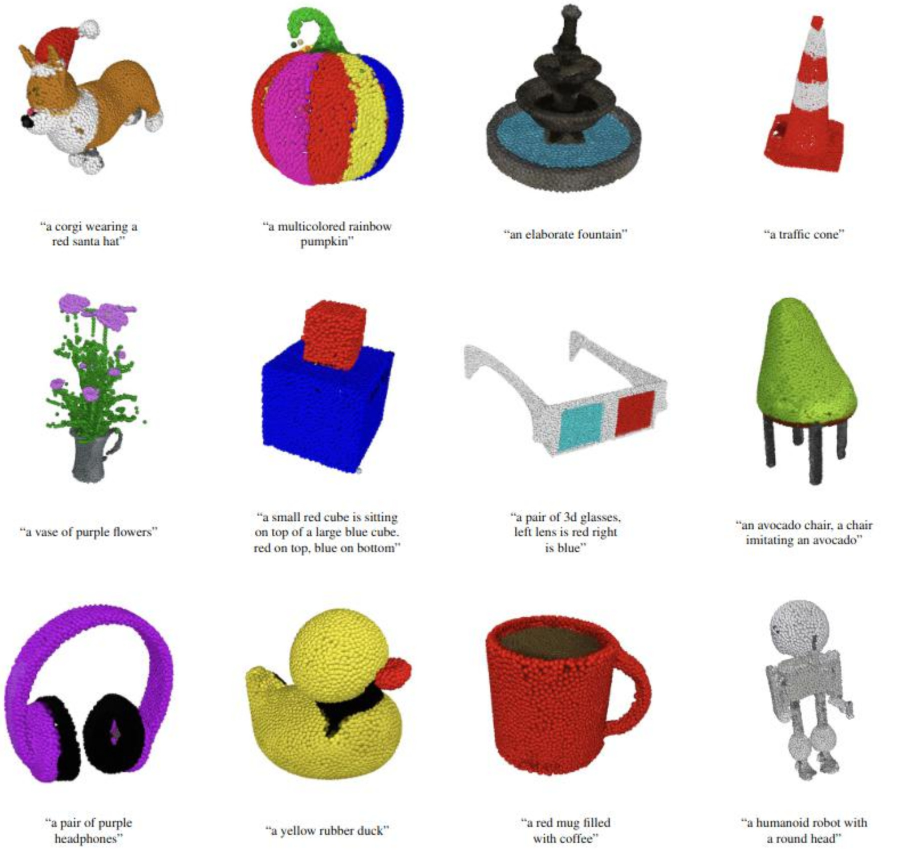
Pointcloud upsampler

- same architecture as the base model, but with extra conditioning tokens for the low-resolution point cloud.

Text-to-3D Diffusion Models - Hybrid 3D

Sample outputs

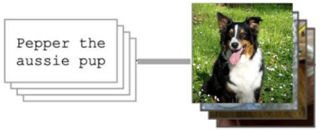




Conclusion

No 3D Data

Large text-image corpus → Large vision-language model



CLIP	DALLE-2
ALIGN	IMAGEN

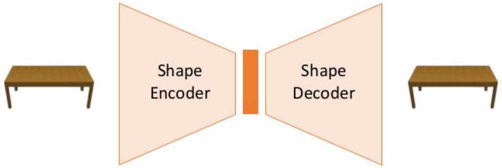
Prompt-based optimization of **differentiable** 3D representation

NeRF DMTet

Unpaired Text-3D Data

3D shape corpus

Learned 3D shape priors



Paired Text-3D Data

"a tall brown table"

"a brown table with four legs"

"a gray, cushioned chair"

Aligned text-3D embedding