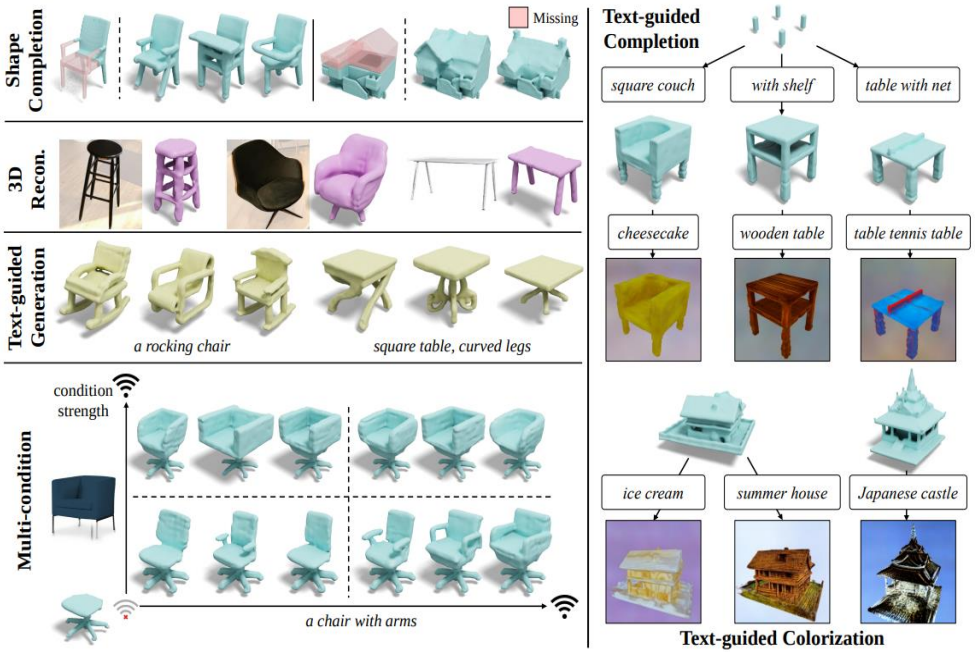
SDFusion: Multimodal 3D Shape Completion, Reconstruction, and Generation

Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulykov, Alexander Schwing*, Liangyan Gui*

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Overview

- Task: learn a diffusion model over 3D shapes. Given some condition **c**, generate the corresponding 3D shape **X**
  - 3D shape: voxelized SDF
- Input: partial shape, image, text, or any subset of their union
- Output: 3D shape



Motivation

- Why SDF?
  - Compared to point cloud: do not have surfaces information
  - Compared to mesh: cannot easily adopt many existing techniques from 2D images
  - Compared to volume: memory consumption for volume rendering is high

- Why Diffusion Model?
  - Compare to autoregressive model: more difficult to scale (memory usage v.s. sequence length)
  - Compare to GANs: training is more stable and easier to converge, GANs are known to have mode-collapsing
  - Great success in 2D image synthesis

  - Classifier-free guidance: provides flexible controls for the conditional generation (non-trivial to do for others such as GANs)

Method

- Problem formulation: learn a diffusion model over 3D shape – SDF, $X \in \mathbb{R}^{D \times D \times D}$ (D=64 or 128)
- Challenge: 3D data is complex and high-dimensional – computationally intractable to train on raw data

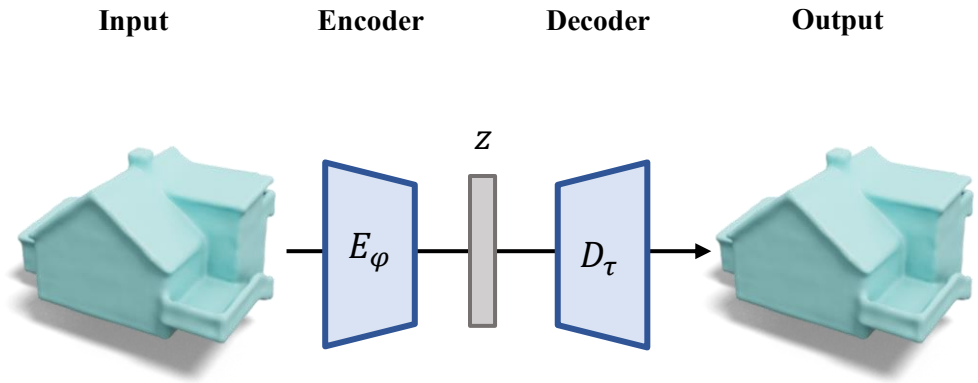
Table 4. **Comparisons of computation (MACs) and memory.**

Method	MACs (G)	Memory (MB)
Raw Voxel	15745	OOM (> 48685)
Ours	725	4845

- Solution: compressed the 3D shape with VQ-VAE, then train a diffusion model in the latent space

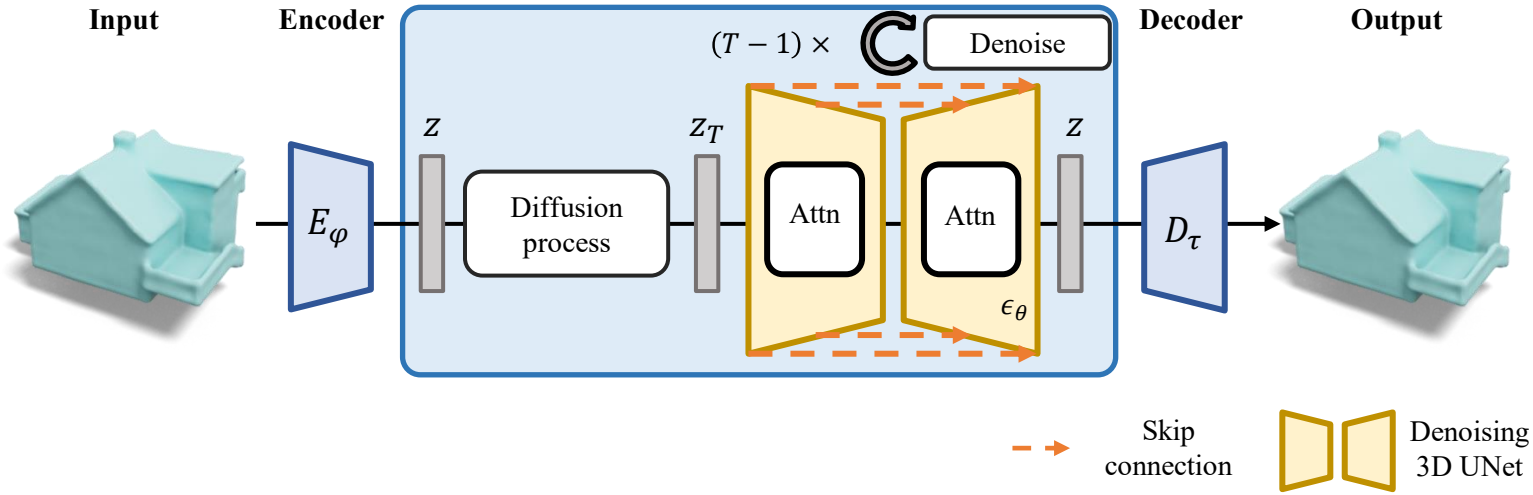
Method - Shape Compression via VQ-VAE

- **First stage**: learn the 3D shape compression with VQ-VAE



Method - Latent Diffusion Model for SDF

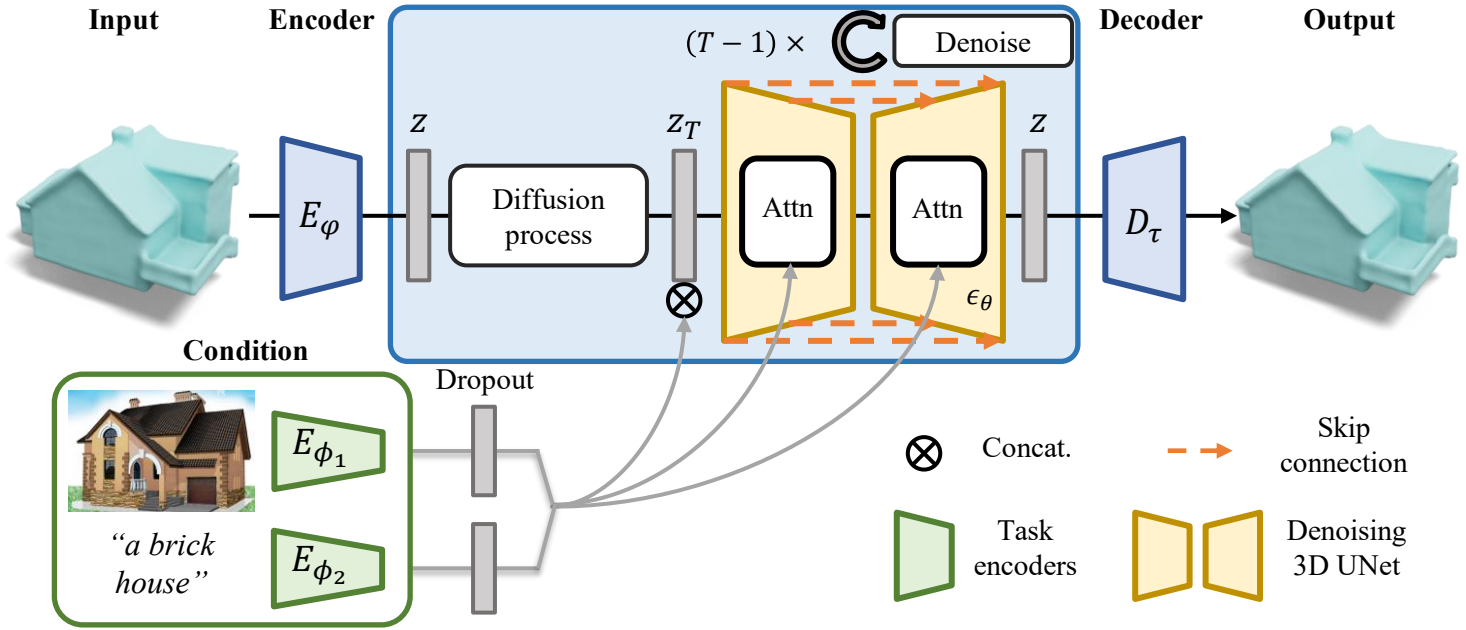
- **Second stage**: train latent diffusion model for SDF



Method - Conditional Generation

- Adopt task specific encoders for different modalities
- Cross-attention and classifier-free guidance [1]

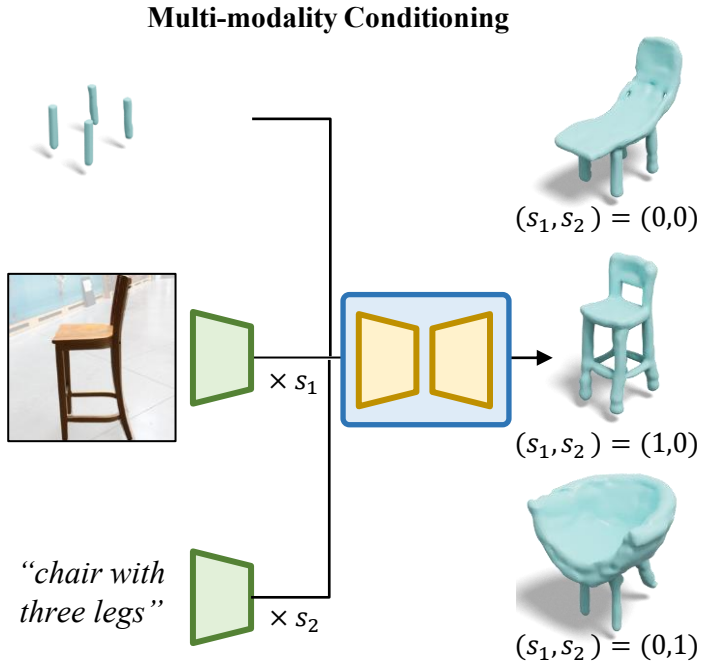
[1] Ho et al. Classifier-Free Diffusion Guidance, 2022.



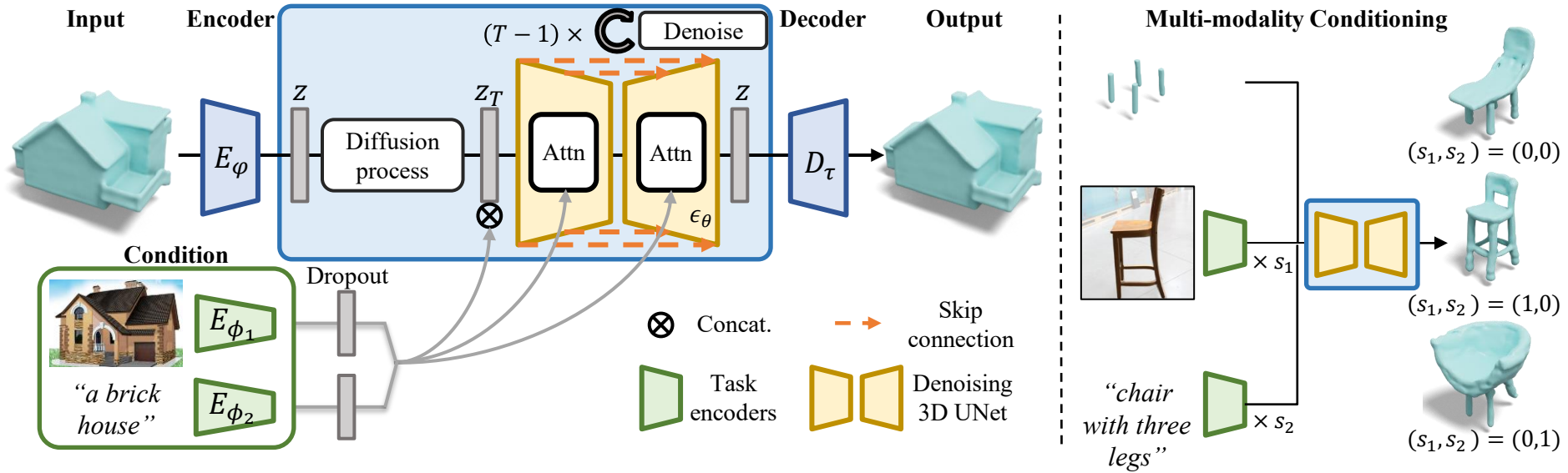
Method - Multi-modality Conditioning

- Adopt task specific encoders for different modalities
- Cross-attention and classifier-free guidance [1]

[1] Ho et al. Classifier-Free Diffusion Guidance, 2022.

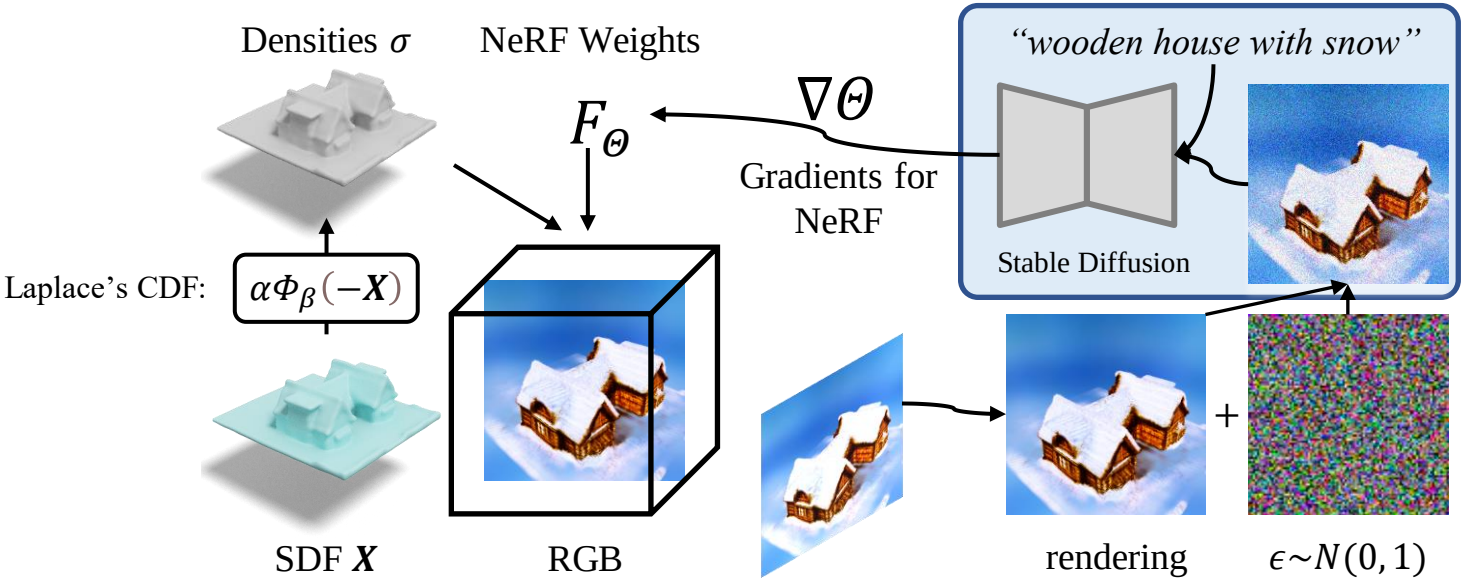


Method – Overview



Method - Text-guided Texturing

- Inspired by DreamFusion [1], given a text, we can add textures for the generated SDF with NeRF and a 2D diffusion model



[1] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. arXiv:2209.14988, 2022
[2] Lior Yariv, Jiatao Gu, Yoni Kasten, Yaron Lipman. Volume Rendering of Neural Implicit Surfaces. In NeurIPS 2021.

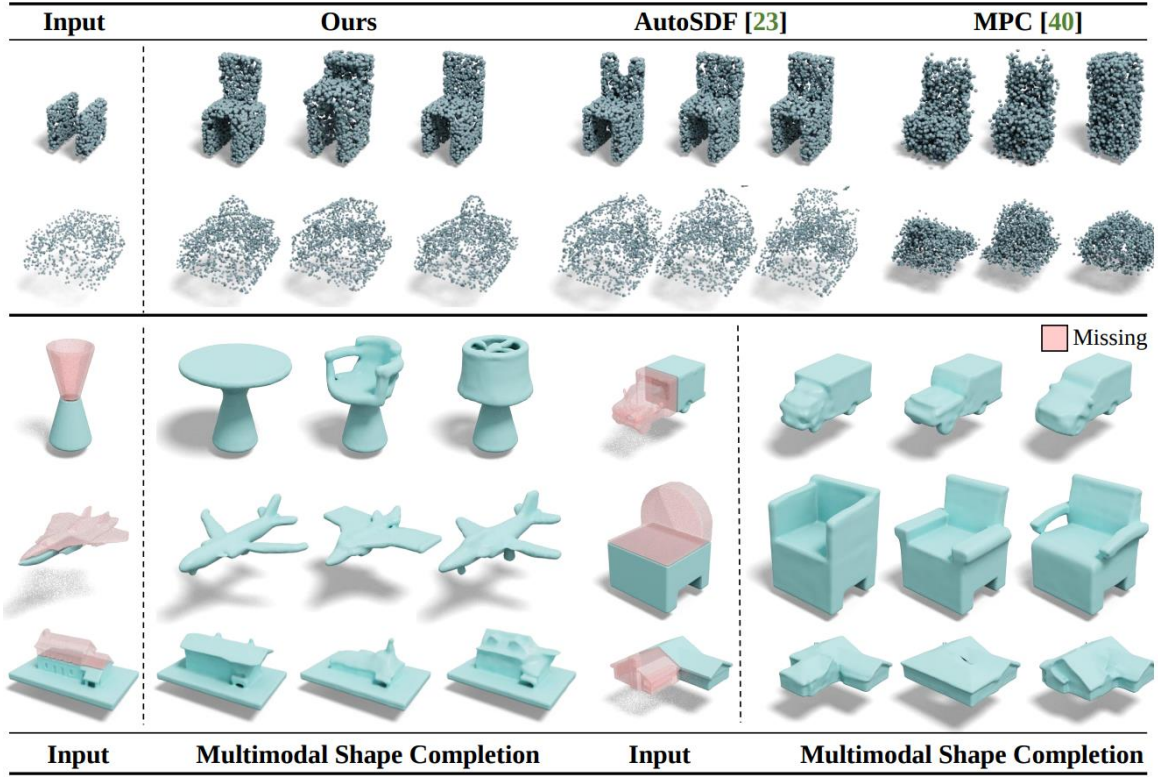
Experiments

- Evaluate on five tasks:
  - Shape completion
  - Single-view reconstruction
  - Text-guided generation
  - Multi-modality condition generation
  - Text-guided texturing

- Datasets
  - Shape: ShapeNet, BuildingNet
  - Image-Shape: ShapeNet Rendering, Pix3D
  - Text-Shape: text2shape

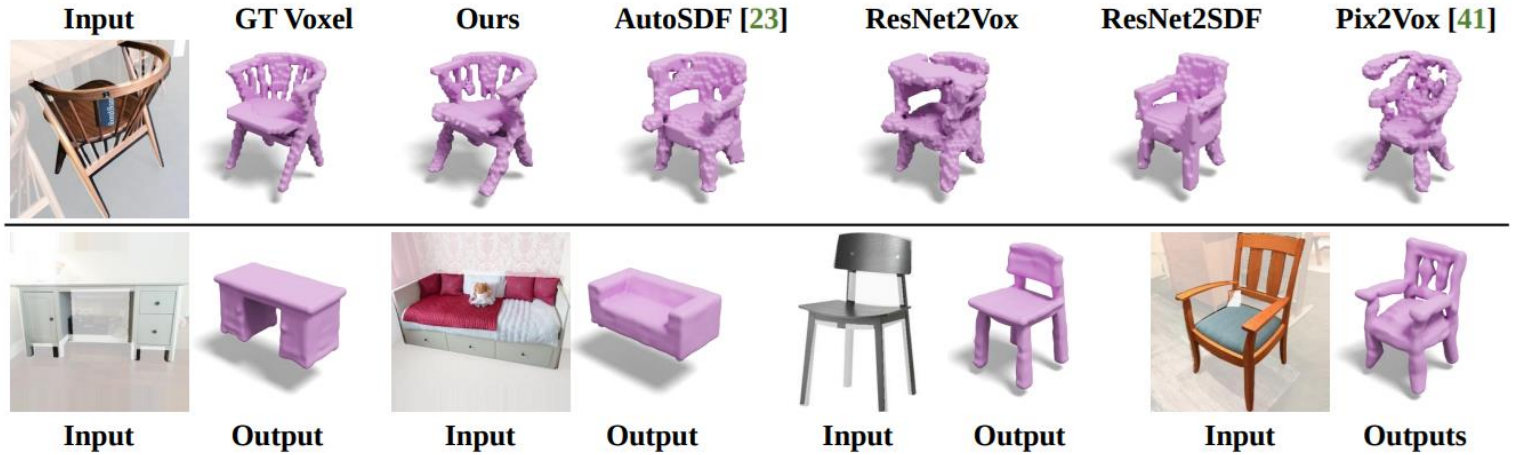
Results: Shape Completion

- Input: partial shape
- Output: complete shape
- Dataset: ShapeNet & BuildingNet



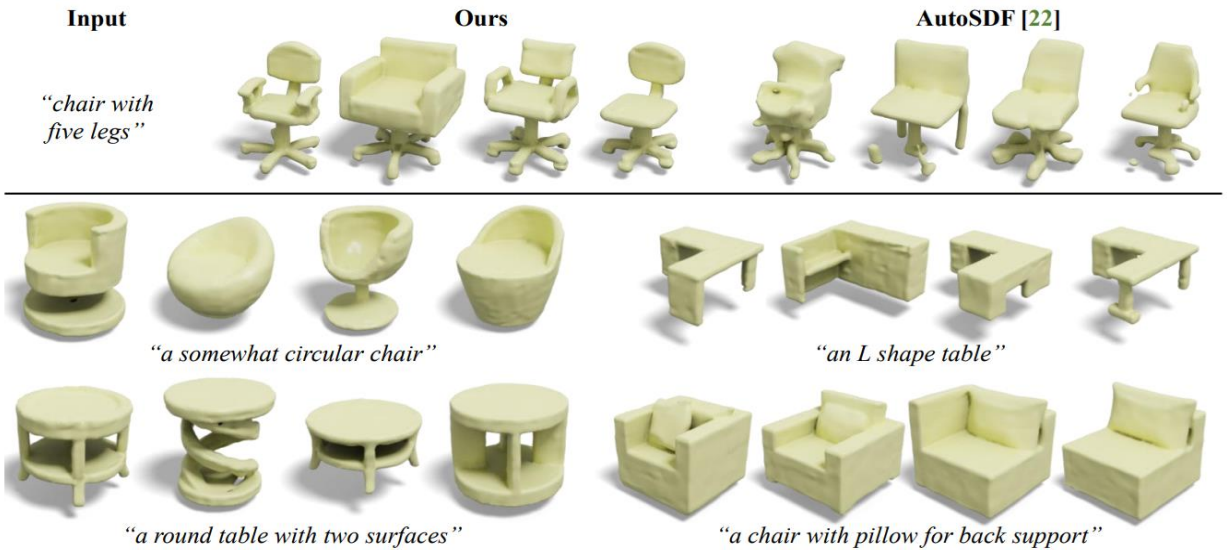
Results: Single-view Reconstruction

- Input: image
- Output: 3D shape
- Dataset: Pix3D



Results: Single-view Reconstruction

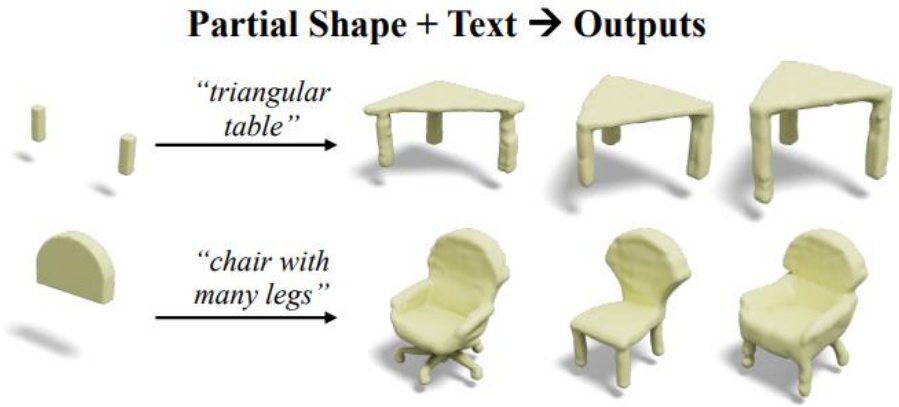
- Input: text
- Dataset: text2shape (ShapeNet Chair & Table)
- Output: 3D shape

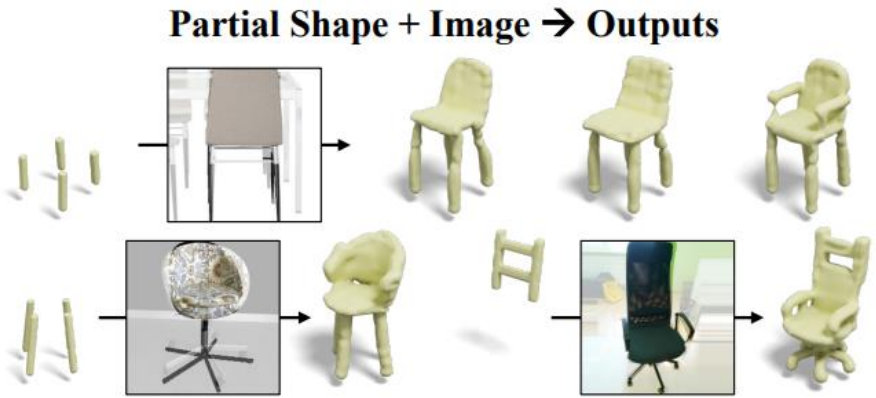


[3] Chen et al. Text2Shape: Generating Shapes from Natural Language by Learning Joint Embeddings, 2018

Results: Condition-guided Shape Completion

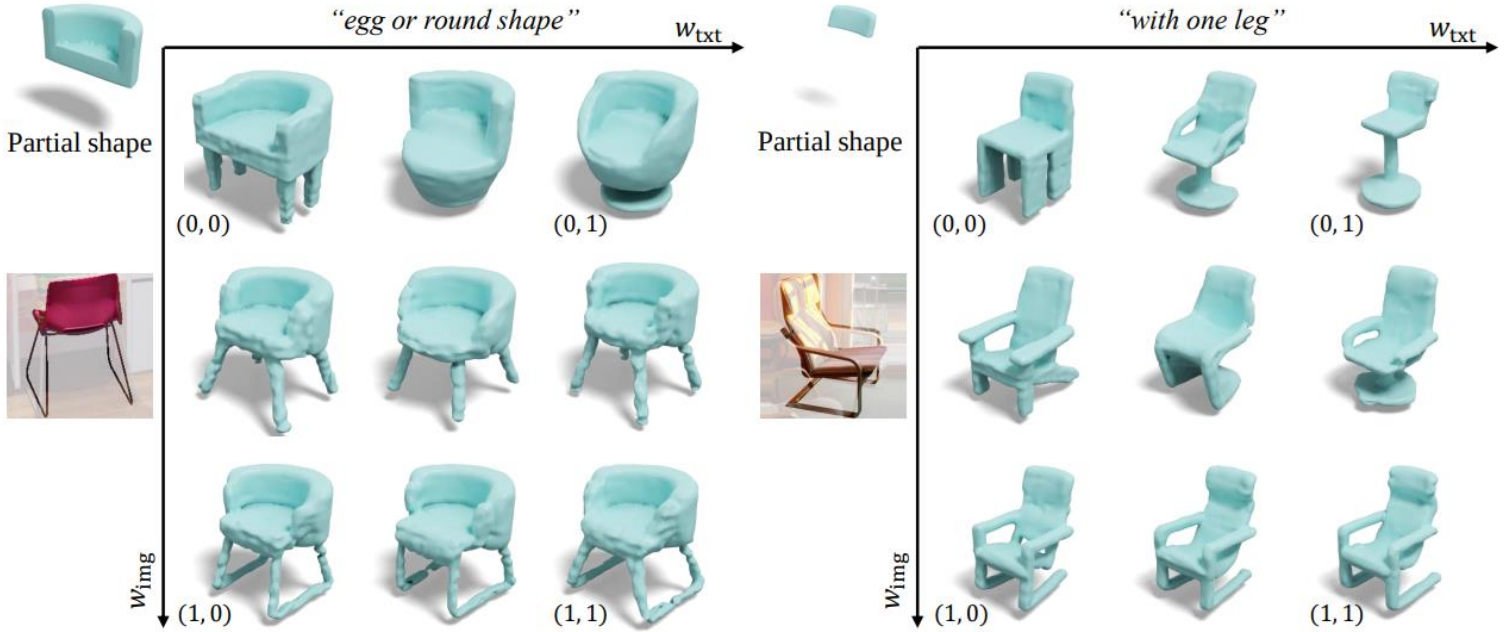
- Input: partial shape + [image or text]
- Output: 3D shape





Results: Multi-modality Conditional Generation

- Input: partial shape + image + text
- Output: 3D shape



Results: Text-guided Texturing

- Input: 3D shape + text
- Output: 3D shape with textures

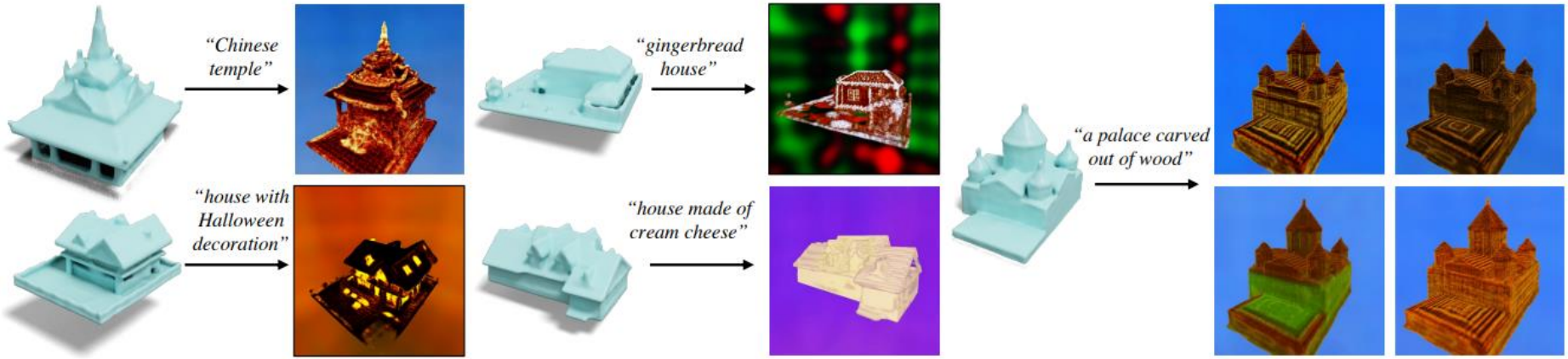


Figure 9. **3D Shape Texturing.** We texture the generated 3D shapes with a 2D diffusion model trained on large-scale data. This permits to generate textures from diverse textual inputs, including style and material descriptions. The pipeline can also generate diverse results given the same input description.

Summary

- We propose SDFusion – a diffusion based 3D shape generative model
- We adopt cross-attention for modulating the conditional signal
- By leveraging classifier-free guidance, SDFusion enables controllability for multi-modality conditional generation
- Using NeRF and an off-the-shelf 2D diffusion model, we can add textures onto the generated 3D shape (inspired from DreamFusion)