SALAD: Part-Level Latent Diffusion for 3D Shape Generation and Manipulation

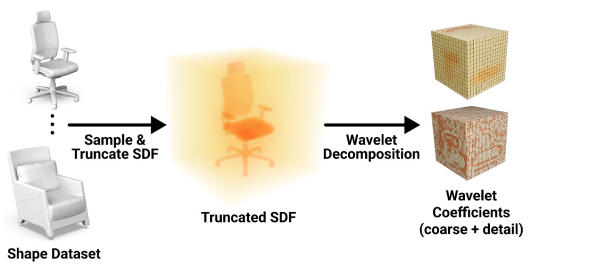
Juil Koo* Seungwoo Yoo* Minh Hieu Nguyen* Minhyuk Sung

KAIST

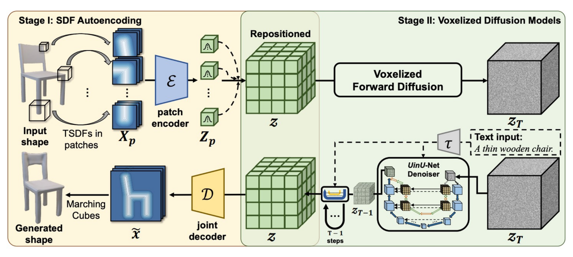
* equal contribution.

Motivation

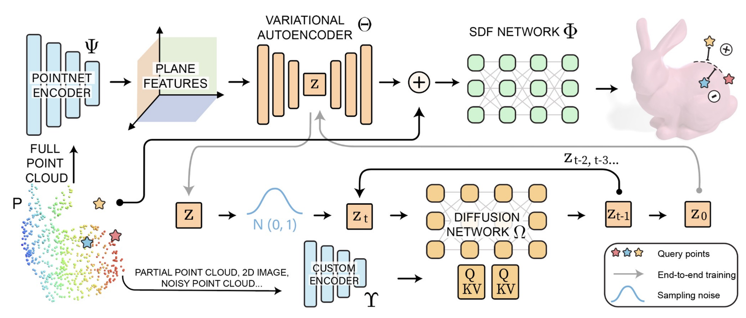
The existing 3D diffusion models are based on either voxel grid features or global latent features.



Neural Wavelet [Hui et al., 2022]



Diffusion-SDF [Li et al., CVPR 2023]



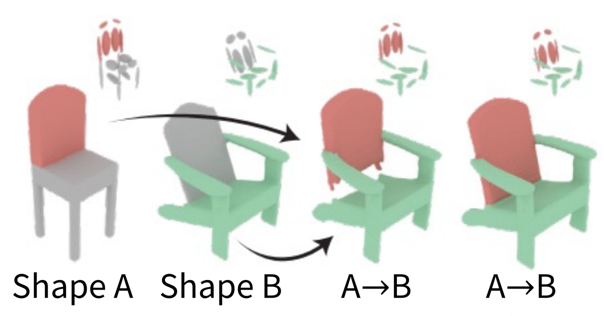
Diffusion-SDF [Chou et al., CVPR 2023]

Due to the data representation, they do **not fully realize the zero-shot editing capability of diffusion models** in shape manipulation. For this, it is essential to leverage a part-level representation.

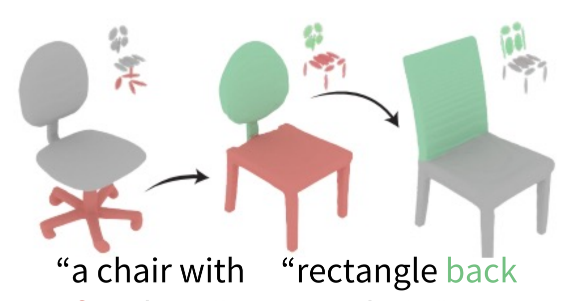
Goal

Achieve state-of-the-art shape generation quality and zero-shot part-level shape manipulation.

Shape Generation



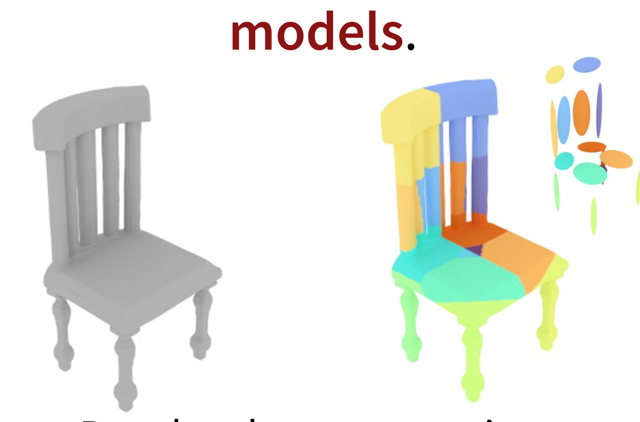
Part Mixing and Refinement



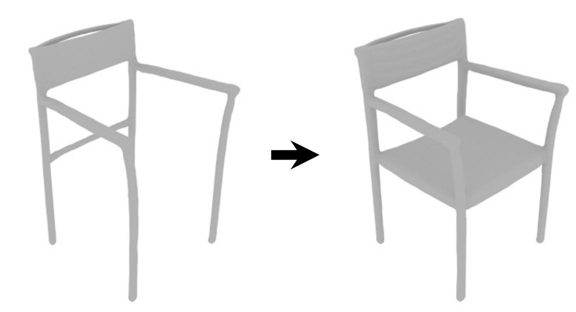
Text-Guided Part Completion

Key Idea

Combine the **expressivity of part-level implicit representation** with **flexible manipulation capabilities of diffusion models**.



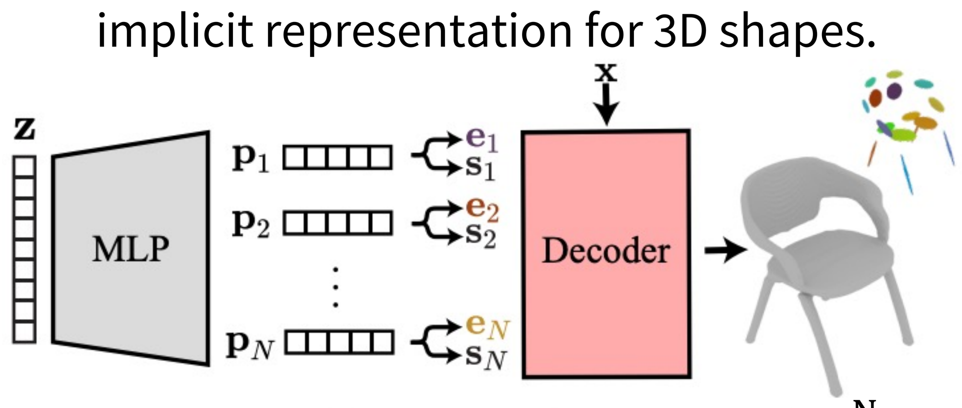
Part-level representation



Zero-shot manipulation capability of diffusion models

Part-Level Implicit Representation

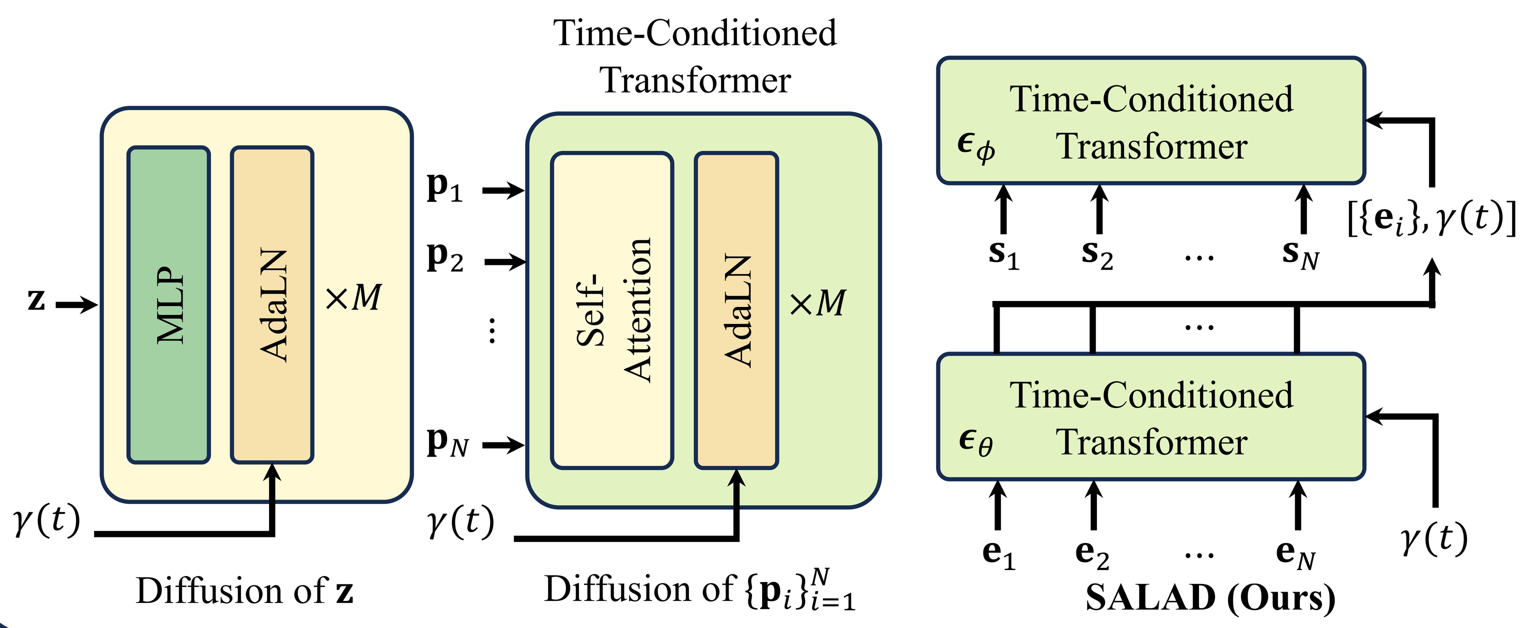
Hertz et al. [SPAGHETTI, SIGGRAPH 2022] proposes a part-level implicit representation for 3D shapes.



A 3D shape is represented by a set of extrinsics $\{e_i\}_{i=1}^N$ and a set of intrinsics $\{s_i\}_{i=1}^N$. $e_i \in \mathbb{R}^{16}$ represents a 3D gaussian primitive and s_i encodes fine details of shapes in a 512-dimensional space. The decoder finally predicts an occupancy value at point x given $\{e_i, s_i\}_{i=1}^N$.

SALAD Architecture Diagrams

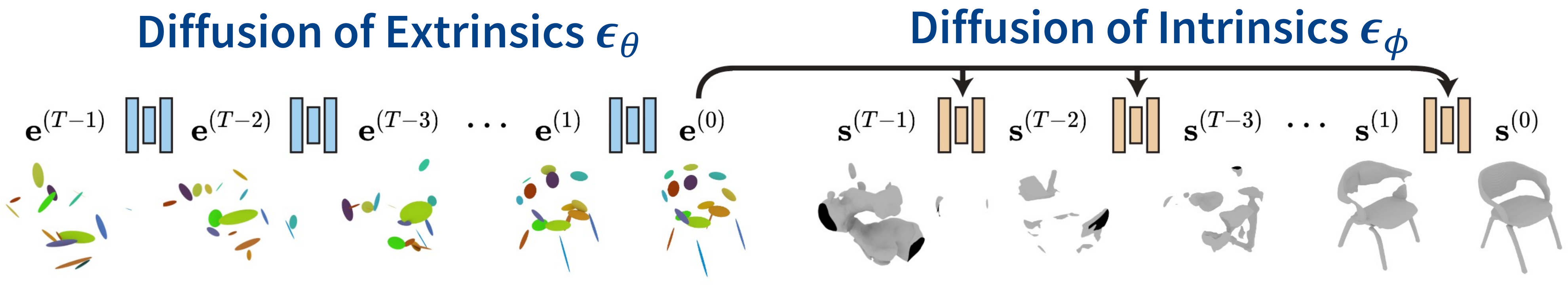
We employ Transformers to process embeddings represented as set. SALAD consists of two cascaded latent diffusion models, $\boldsymbol{\epsilon}_\theta$ for the diffusion of extrinsics and $\boldsymbol{\epsilon}_\phi$ for the diffusion of intrinsics.



Cascaded Diffusion Models

We propose a two-phase cascaded diffusion training framework, learning diffusion first in a low-dimensional subspace and subsequently in the other high-dimensional subspace.

Diffusion of Extrinsics ϵ_θ / Diffusion of Intrinsics ϵ_ϕ



In the first phase, SALAD learns the **rough structures of shapes** encoded with extrinsics. In the second phase, it captures the **fine details of shape surfaces** encoded with intrinsics. To ease the training of the diffusion of intrinsics in a high-dimensional space, we feed extrinsics as condition.

SALAD Shape Manipulation Method

Part Completion

$$\mathbf{x}^{(t-1)} = \text{add\_noise}(\mathbf{x}^{(0)}, t-1)$$
$$\tilde{\mathbf{x}}^{(t-1)} = \text{denoise}(\tilde{\mathbf{x}}^{(t)}, t)$$
$$\tilde{\mathbf{x}}^{(t-1)} = \text{combine}(\mathbf{x}^{(t-1)}, \tilde{\mathbf{x}}^{(t-1)}, m)$$

Given a data $\mathbf{x}^{(0)}$ and a part mask m, the completed shape $\tilde{\mathbf{x}}^{(t-1)}$ at $t-1$ is obtained by combining the original and generated data.

Part Mixing and Refinement

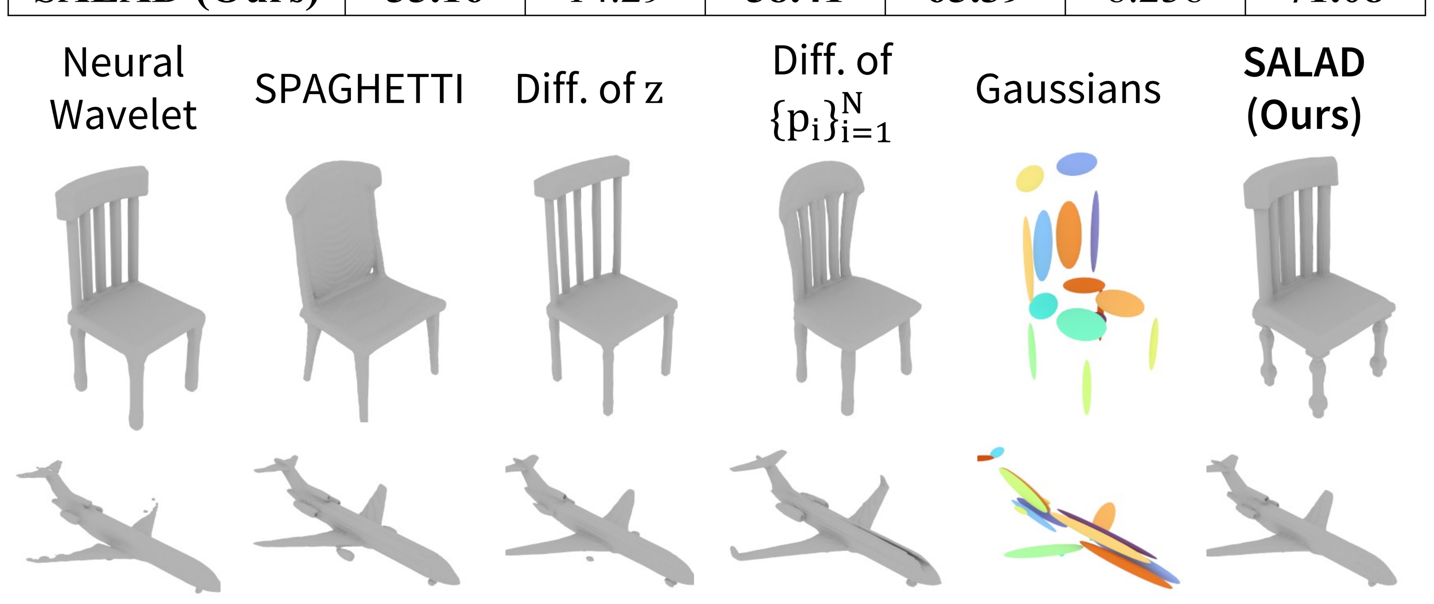
$$\tilde{\mathbf{x}}^{(0)} = \text{combine}(\mathbf{x}_a^{(0)}, \mathbf{x}_b^{(0)}, m)$$
$$\tilde{\mathbf{x}}^{(S)} = \text{add\_noise}(\tilde{\mathbf{x}}^{(0)}, S)$$
for $t = S, \dots, 1$ **do**
$$\tilde{\mathbf{x}}^{(t-1)} = \text{denoise}(\tilde{\mathbf{x}}^{(t)}, t)$$
end for

Given the mixed data $\tilde{\mathbf{x}}^{(0)}$ from $\mathbf{x}_a^{(0)}$ and $\mathbf{x}_b^{(0)}$, we add noise sampled at $t = S \le T$ to $\tilde{\mathbf{x}}^{(0)}$ and run the reverse process to refine $\tilde{\mathbf{x}}^{(0)}$.

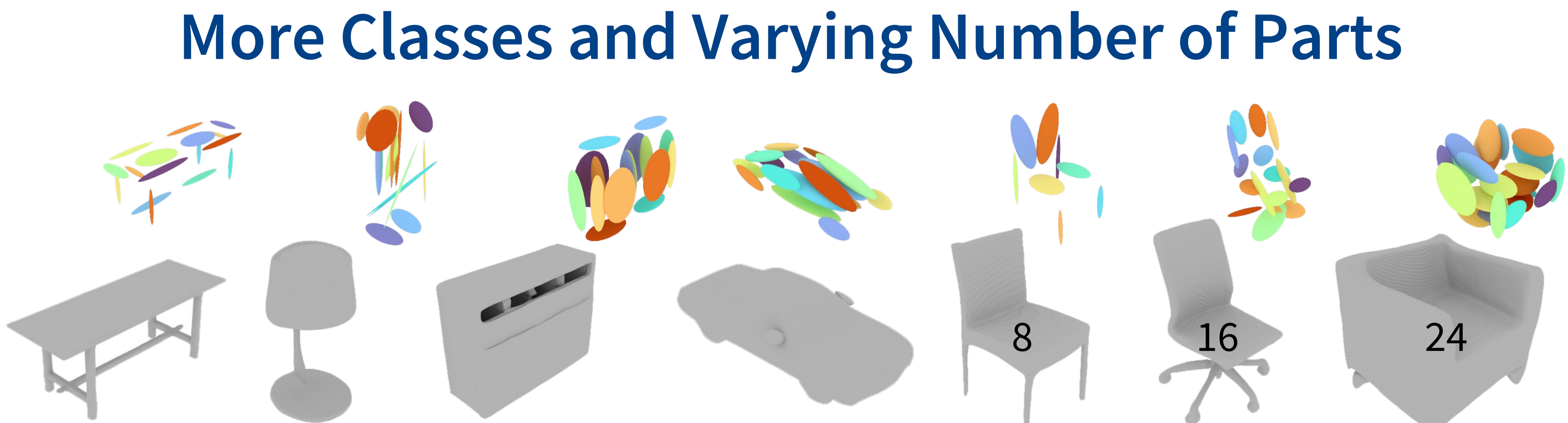
Shape Generation

SALAD achieves **state-of-the-art shape generation** across widely used metrics while capturing fine details.

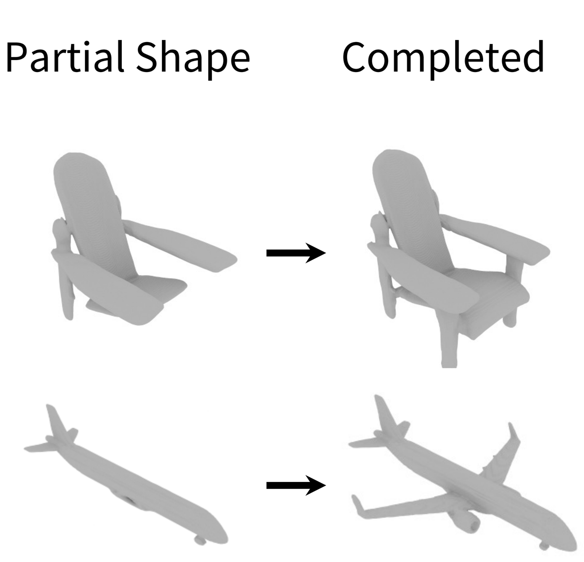
Method	Chair			Airplane		
	COV ↑	MMD ↓	1-NNA ↓	COV ↑	MMD ↓	1-NNA ↓
Neural Wavelet	50.15	**14.25**	62.87	59.09	**7.964**	72.93
SALAD (Ours)	**55.16**	14.29	**58.41**	**65.39**	8.238	**71.08**



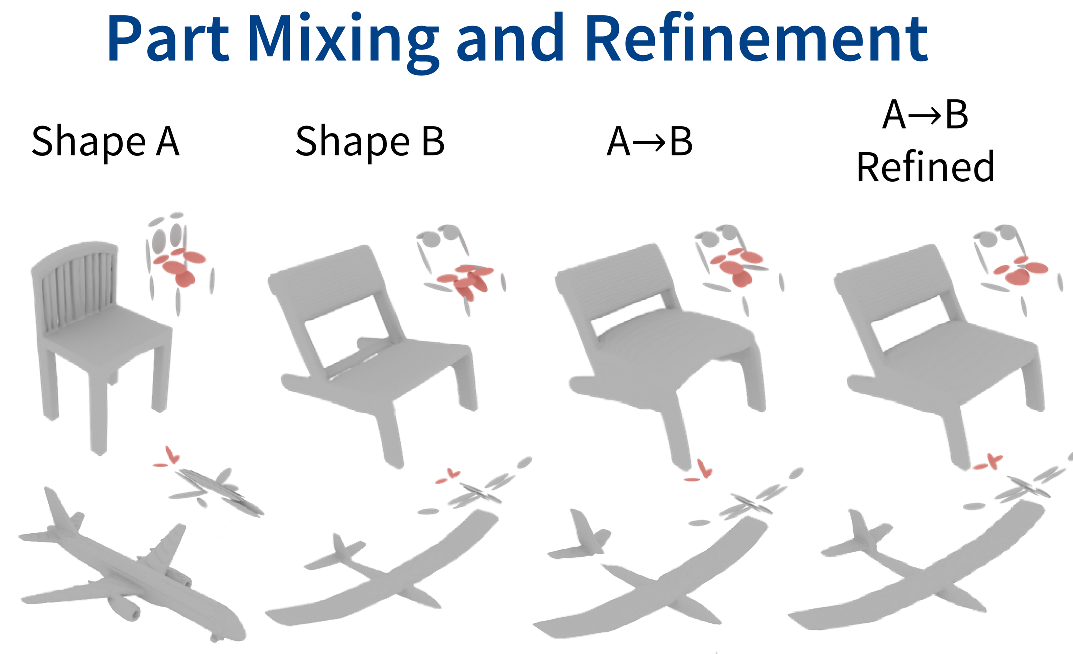
More Classes and Varying Number of Parts



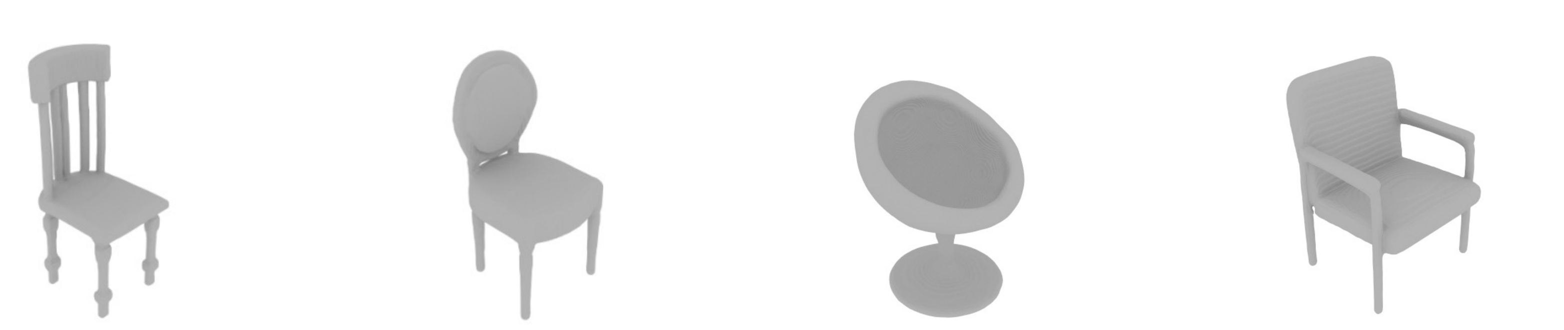
Part Completion



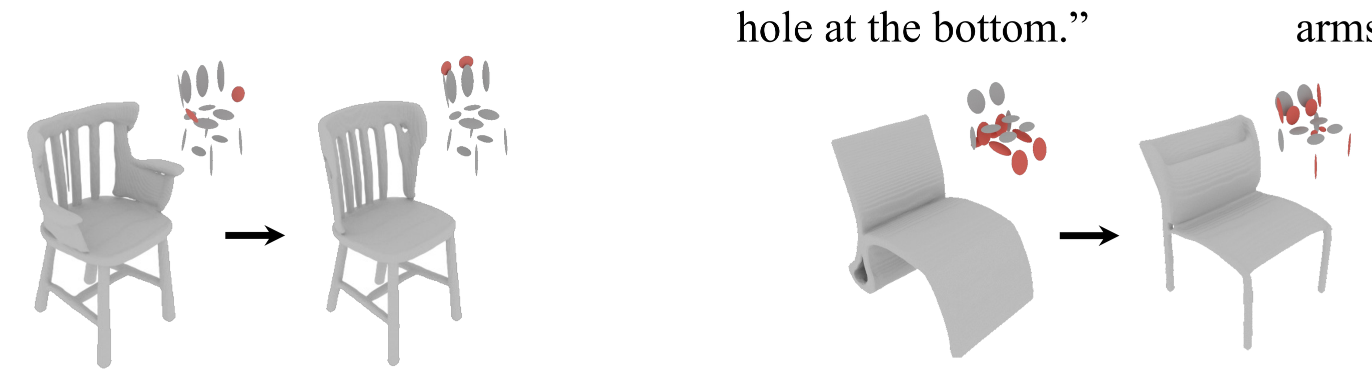
Part Mixing and Refinement



Text-Guided Shape Generation and Part Editing



"the oval shaped back." "5 lines, with curve." "a wide circle with a hole at the bottom." "a chair has very tiny arms."



"chair with no arms." " four thin legs."