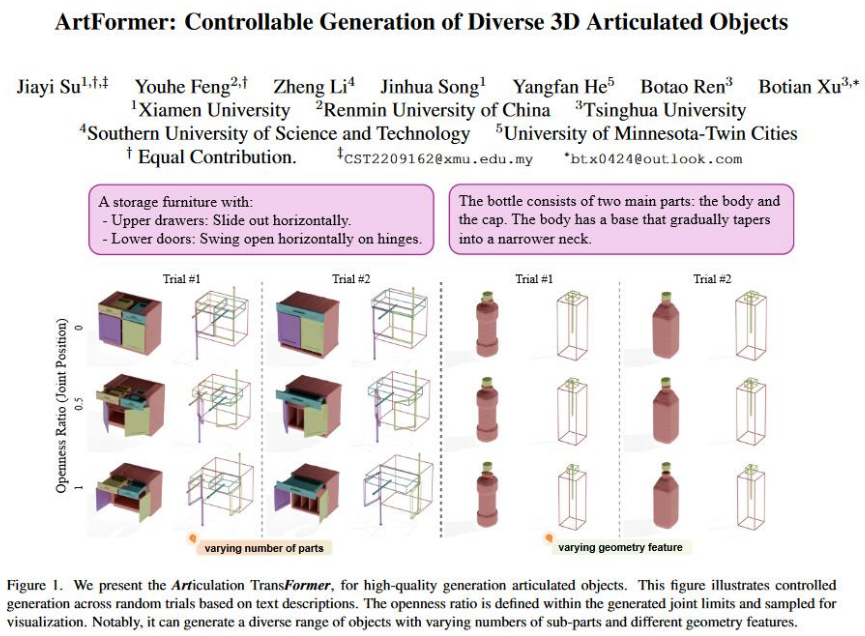
Appendix

ArtFormer: Controllable Generation of Diverse 3D Articulated Objects



ArtFormer: Controllable Generation of Diverse 3D Articulated Objects

Jiayi Su[1,†,‡] Youhe Feng[2,†] Zheng Li[4] Jinhua Song[1] Yangfan He[5] Botao Ren[3] Botian Xu[3,*]
[1]Xiamen University [2]Renmin University of China [3]Tsinghua University
[4]Southern University of Science and Technology [5]University of Minnesota-Twin Cities
† Equal Contribution. ‡CST2209162@xmu.edu.my *btx0424@outlook.com

A storage furniture with:
- Upper drawers: Slide out horizontally.
- Lower doors: Swing open horizontally on hinges.

The bottle consists of two main parts: the body and the cap. The body has a base that gradually tapers into a narrower neck.

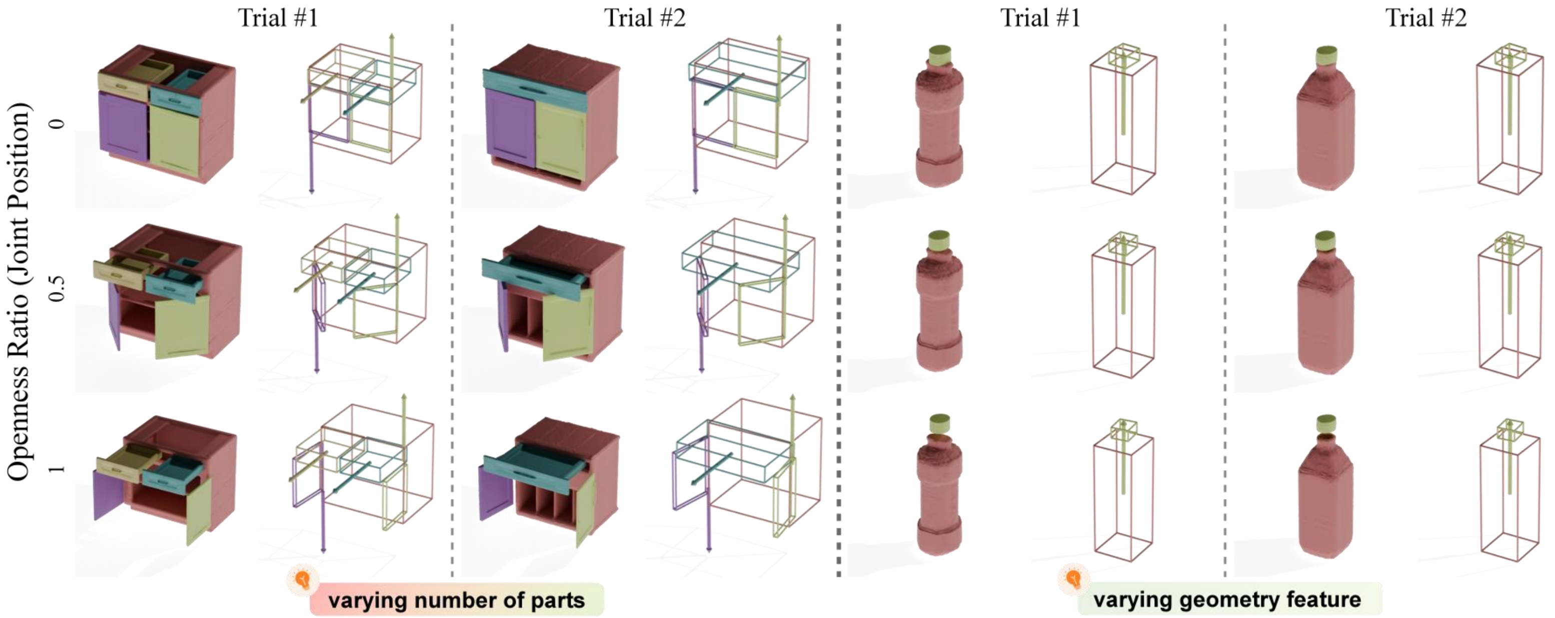
Figure 1. We present the ***Art***iculation Trans***Former***, for high-quality generation articulated objects. This figure illustrates controlled generation across random trials based on text descriptions. The openness ratio is defined within the generated joint limits and sampled for visualization. Notably, it can generate a diverse range of objects with varying numbers of sub-parts and different geometry features.

Jiayi Su
Xiamen University Malaysia, School of CST

Introduction

A storage furniture with:
- Upper drawers: Slide out horizontally.
- Lower doors: Swing open horizontally on hinges.

The bottle consists of two main parts: the body and the cap. The body has a base that gradually tapers into a narrower neck.



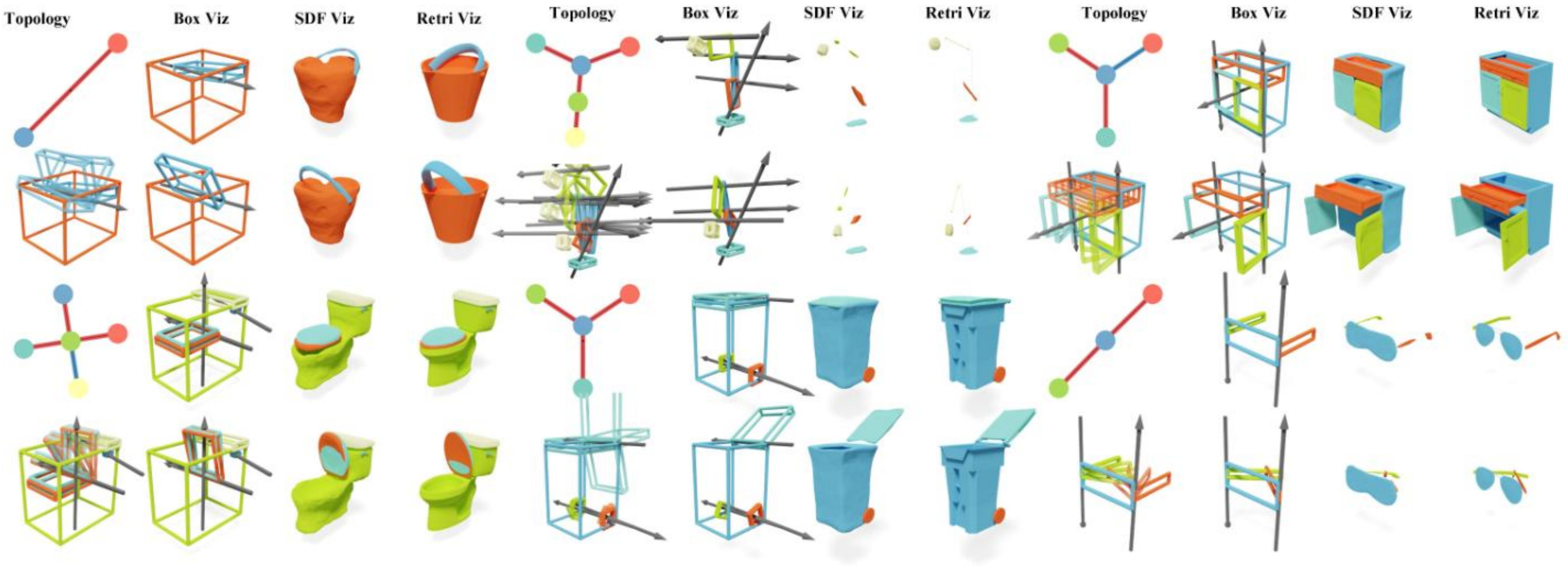
Task of our work - generate the articulated object including:

- Geometry of each subpart (Mesh)
- Articulated feature between each subpart.

Given condition with either:

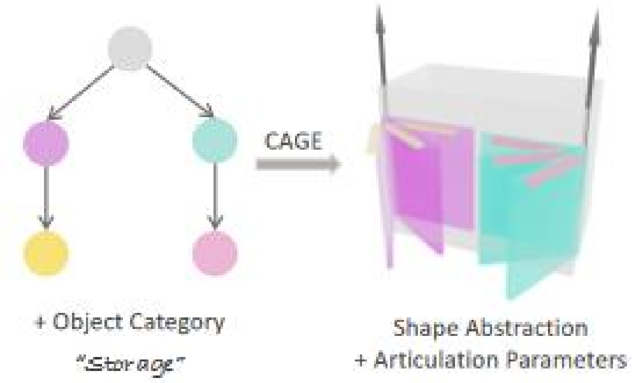
- Text description.
- Image of the object.

Previous Work (NAP)



- Diffusion Method.
- Weak Geometry Quality.
- Weak Articulated Info as well.
- Unconditional Generation.

Previous Work **(CAGE)**



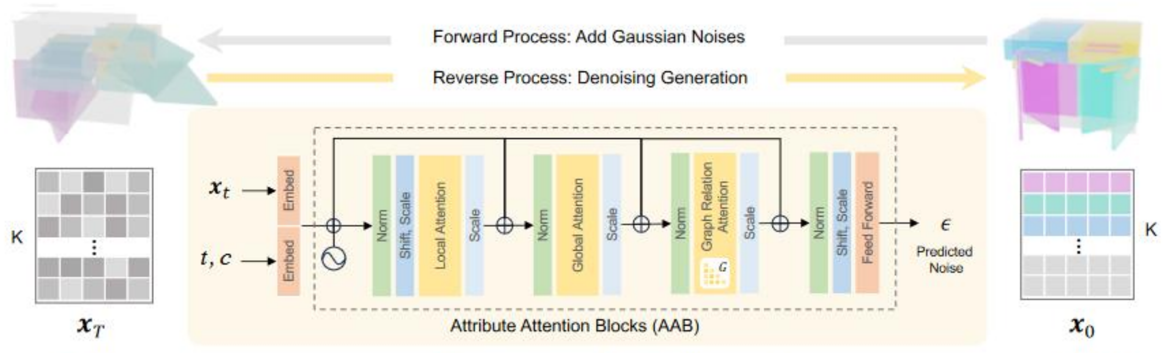


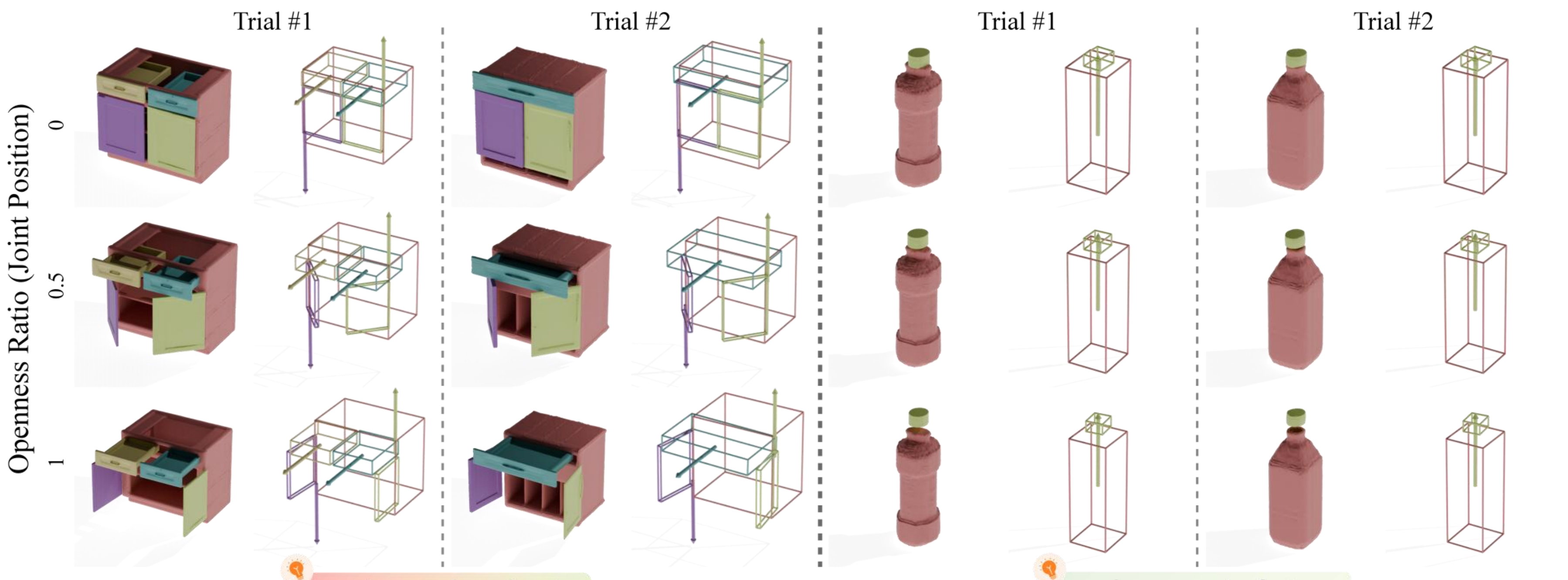
Figure 2. Method overview. Our generative model is based on DDPM [8]. In the forward pass, Gaussian noise is iteratively added to corrupt the data from $\mathbf{x}_0$ to random noise $\mathbf{x}_T$. During the reverse process, our denoiser (in yellow highlight) predicts the residual noise to be subtracted from the input data $\mathbf{x}_t$ at timestep t conditioned on the category label c and a graph adjacency G as an attention mask injected in the Graph Relation Attention module. All the timesteps share the same denoiser that is built on layers of our Attribute Attention Blocks.

- Diffusion Method.
- Does not generate geometry, use *part retrieval.*
- Weak Generational Generation (Category + Tree Structure)

Our Work

A storage furniture with:
- Upper drawers: Slide out horizontally.
- Lower doors: Swing open horizontally on hinges.

The bottle consists of two main parts: the body and the cap. The body has a base that gradually tapers into a narrower neck.

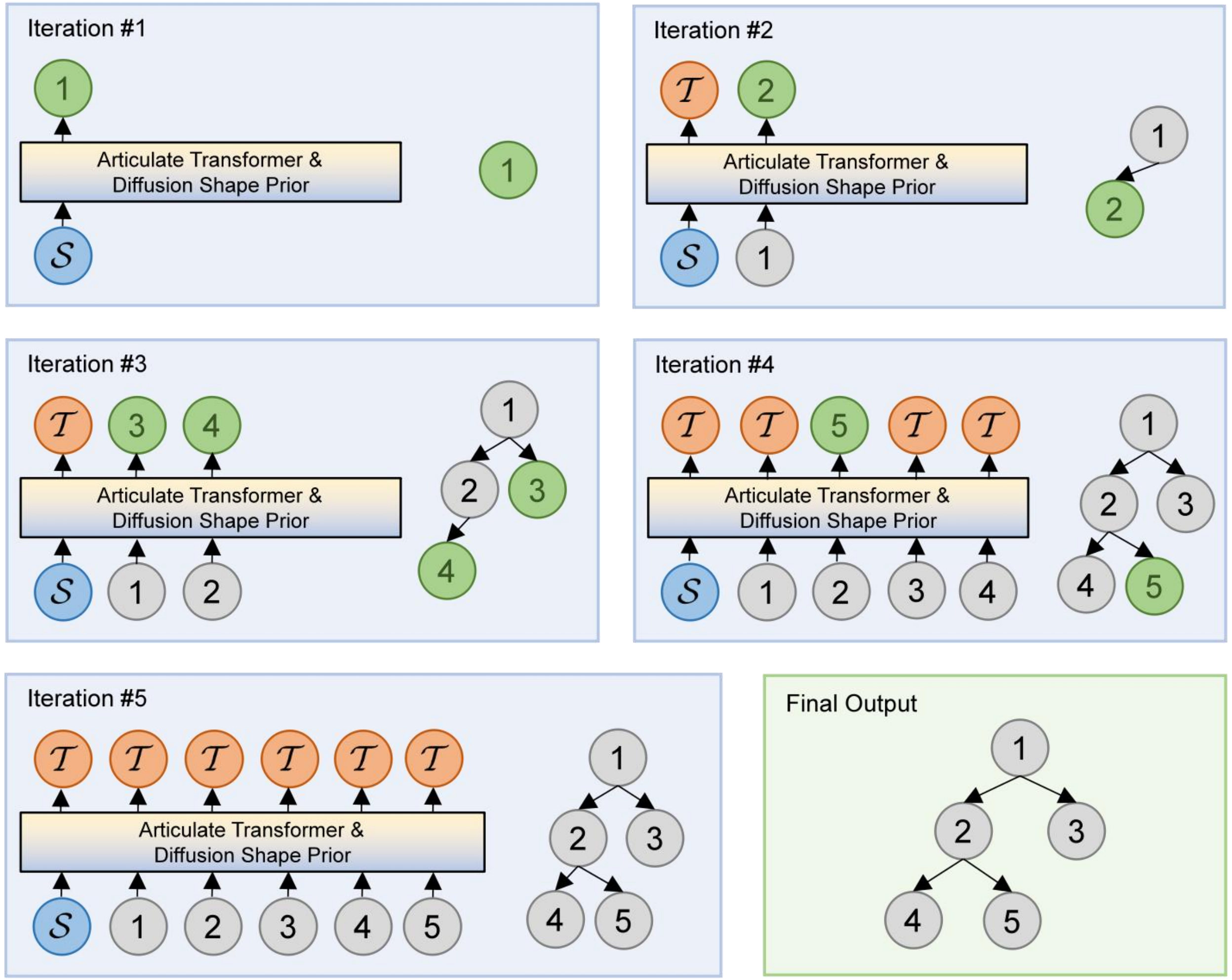


varying number of parts

varying geometry feature

Method

- **Transformer Autogressively Generate**
- Overall + Tree Articulation Parameterization (Bounding box, Geometry latent code, Joint axis, Limit)
- Tree Position Embedding.
- Shape Prior
- Gumbel Softmax.



Iteration #1

Articulate Transformer & Diffusion Shape Prior

Iteration #2

Articulate Transformer & Diffusion Shape Prior

Iteration #3

Articulate Transformer & Diffusion Shape Prior

Iteration #4

Articulate Transformer & Diffusion Shape Prior

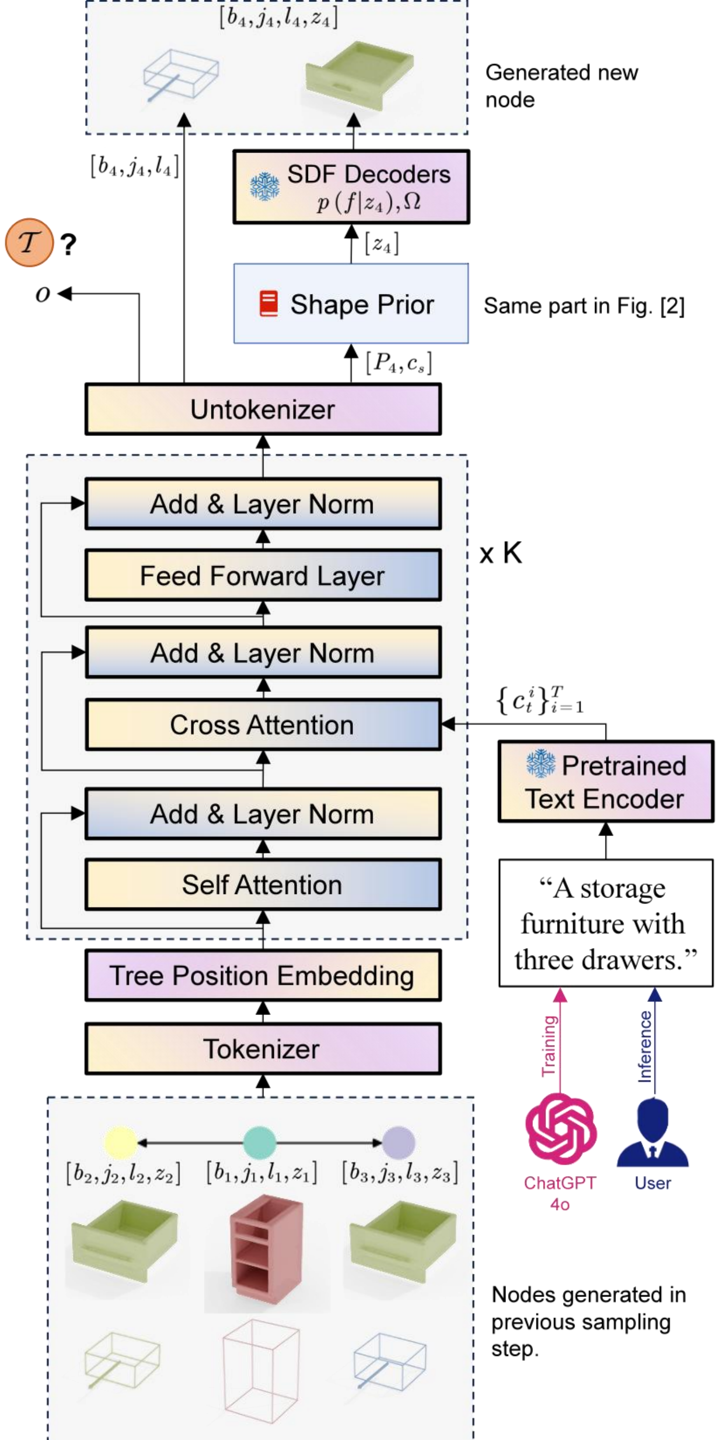
Iteration #5

Articulate Transformer & Diffusion Shape Prior

Final Output

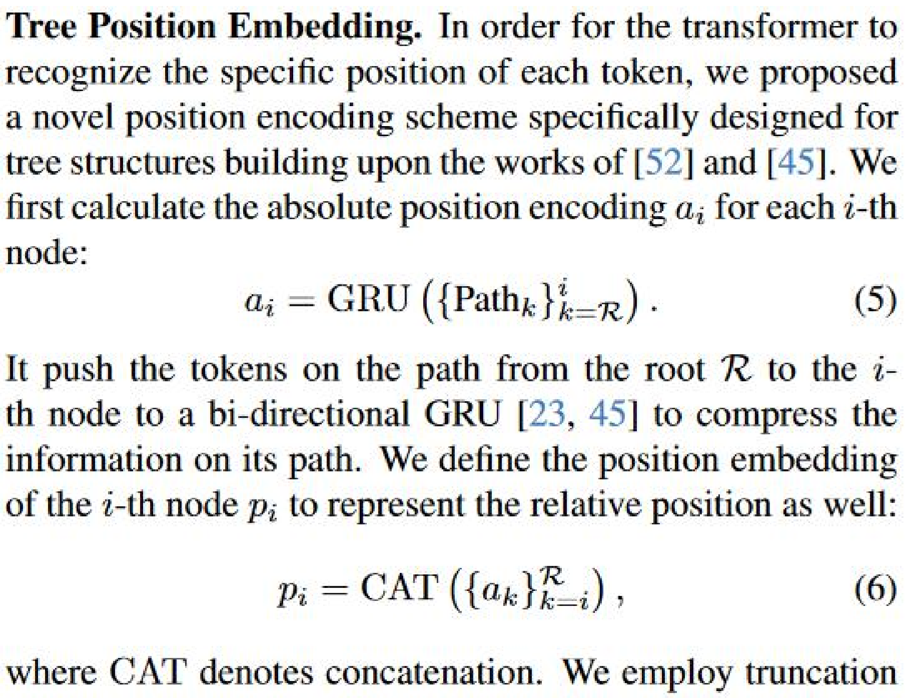
Method

- Transformer Autogressively Generate
- **Overall + Tree Articulation Parameterization (Bounding box, Geometry latent code, Joint axis, Limit)**
- Tree Position Embedding.
- Shape Prior
- Gumbel Softmax.



Method

- Transformer Autogressively Generate
- Overall + Tree Articulation Parameterization (Bounding box, Geometry latent code, Joint axis, Limit)
- **Tree Position Embedding.**
- Shape Prior
- Gumbel Softmax.



Tree Position Embedding. In order for the transformer to recognize the specific position of each token, we proposed a novel position encoding scheme specifically designed for tree structures building upon the works of [52] and [45]. We first calculate the absolute position encoding a_i for each i-th node:

$$a_i = \mathrm{GRU}\left(\{\mathrm{Path}_k\}_{k=\mathcal{R}}^{i}\right). \tag{5}$$

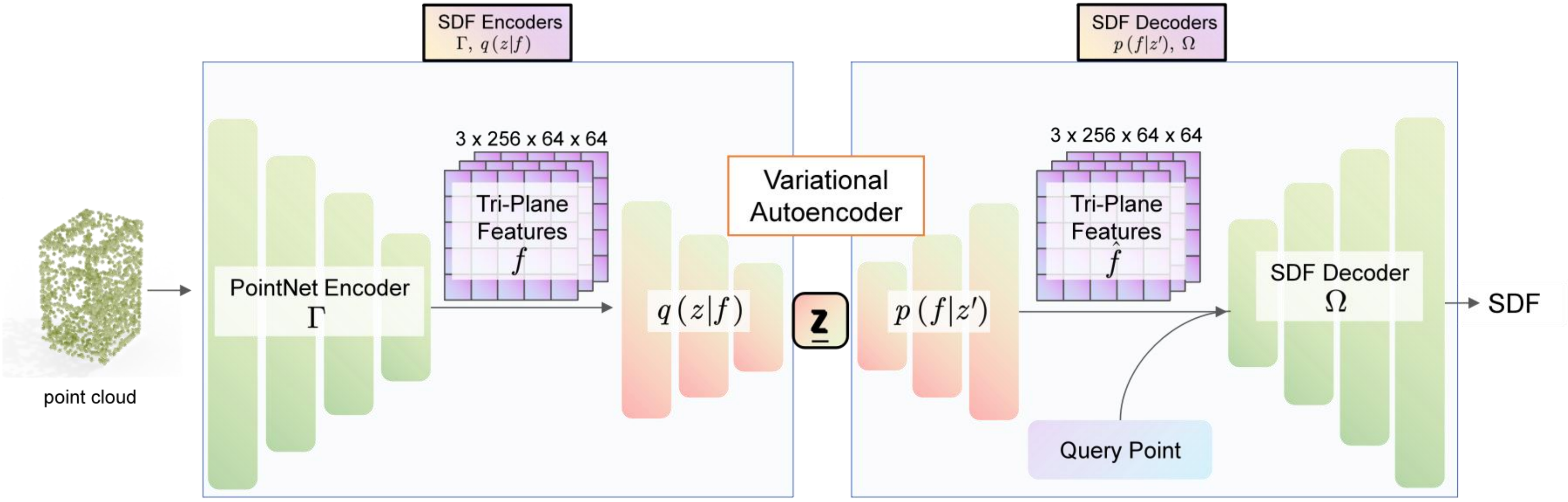
It push the tokens on the path from the root $\mathcal{R}$ to the i-th node to a bi-directional GRU [23, 45] to compress the information on its path. We define the position embedding of the i-th node p_i to represent the relative position as well:

$$p_i = \mathrm{CAT}\left(\{a_k\}_{k=i}^{\mathcal{R}}\right), \tag{6}$$

where CAT denotes concatenation. We employ truncation

Method

- Transformer Autogressively Generate
- Overall + Tree Articulation Parameterization (Bounding box, Geometry latent code, Joint axis, Limit)
- Tree Position Embedding.
- **Shape Prior**
- Gumbel Softmax.



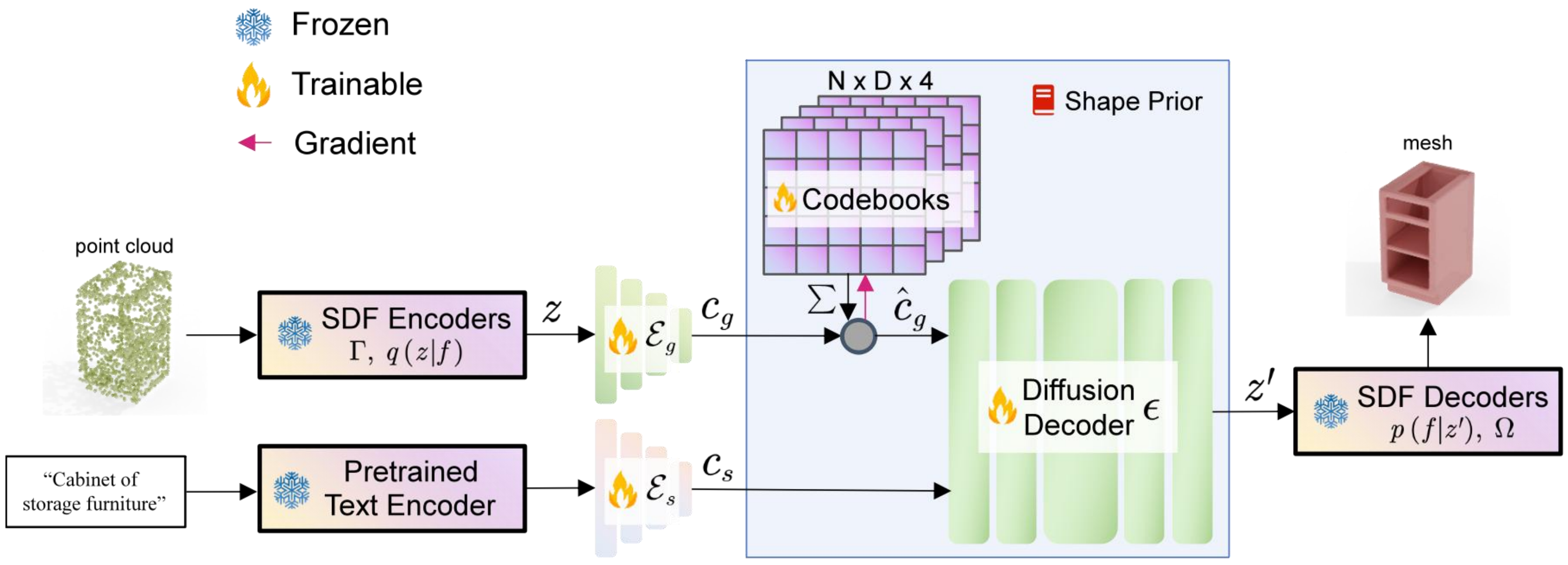
Diffusion Model: $\epsilon(z|\hat{c}_g, c_s)$

During Training:

$$\mathcal{E}_g \to c_g, \text{Codebook}(c_g) \to \hat{c}_g$$

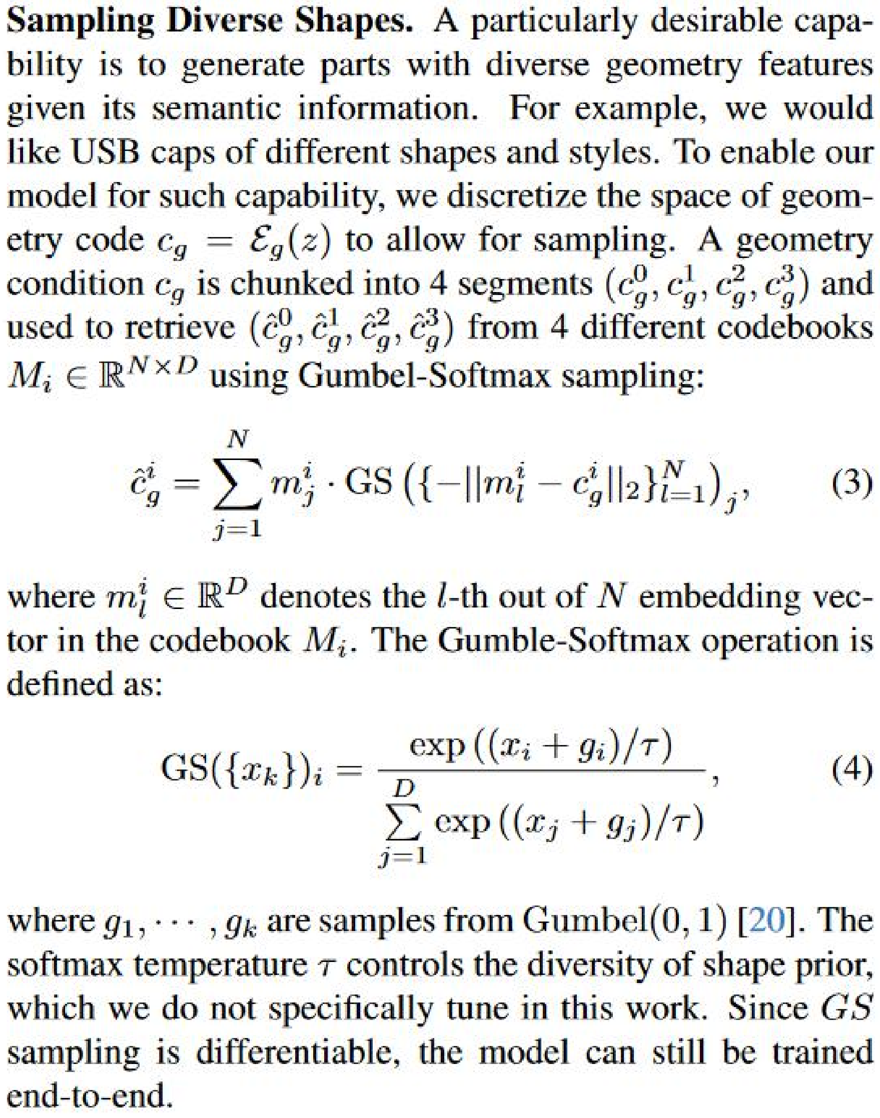
During Inference:

$$\text{Codebook}(P) \to \hat{c}_g$$



Method

- Transformer Autogressively Generate
- Overall + Tree Articulation Parameterization (Bounding box, Geometry latent code, Joint axis, Limit)
- Tree Position Embedding.
- Shape Prior
- **Gumbel Softmax.**



Sampling Diverse Shapes. A particularly desirable capability is to generate parts with diverse geometry features given its semantic information. For example, we would like USB caps of different shapes and styles. To enable our model for such capability, we discretize the space of geometry code $c_g = \mathcal{E}_g(z)$ to allow for sampling. A geometry condition c_g is chunked into 4 segments $(c_g^0, c_g^1, c_g^2, c_g^3)$ and used to retrieve $(\hat{c}_g^0, \hat{c}_g^1, \hat{c}_g^2, \hat{c}_g^3)$ from 4 different codebooks $M_i \in \mathbb{R}^{N \times D}$ using Gumbel-Softmax sampling:

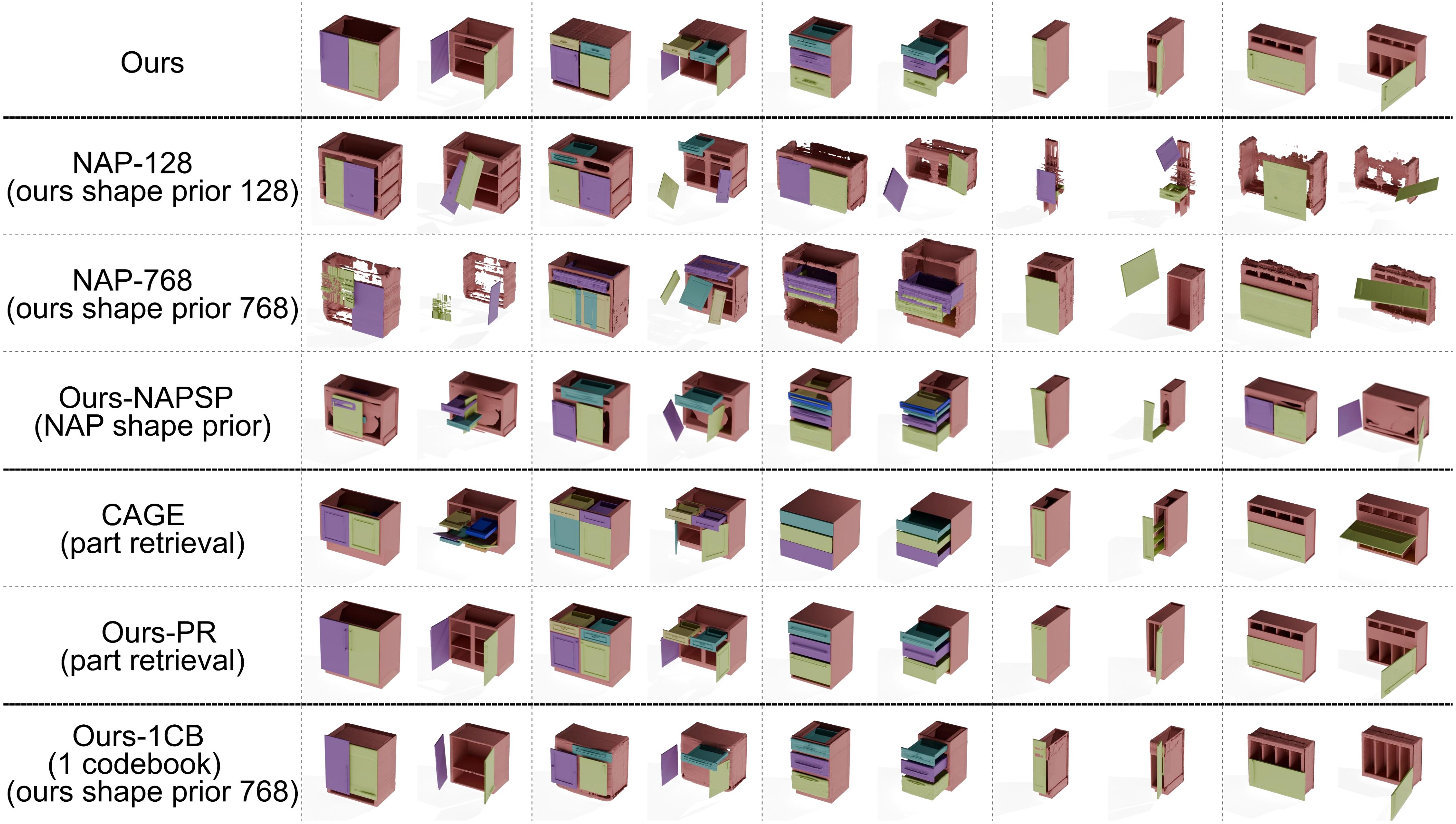
$$\hat{c}_g^i = \sum_{j=1}^{N} m_j^i \cdot \mathrm{GS}\left(\{-||m_l^i - c_g^i||_2\}_{l=1}^{N}\right)_j, \tag{3}$$

where $m_l^i \in \mathbb{R}^D$ denotes the l-th out of N embedding vector in the codebook M_i. The Gumble-Softmax operation is defined as:

$$\mathrm{GS}(\{x_k\})_i = \frac{\exp\left((x_i + g_i)/\tau\right)}{\sum\limits_{j=1}^{D} \exp\left((x_j + g_j)/\tau\right)}, \tag{4}$$

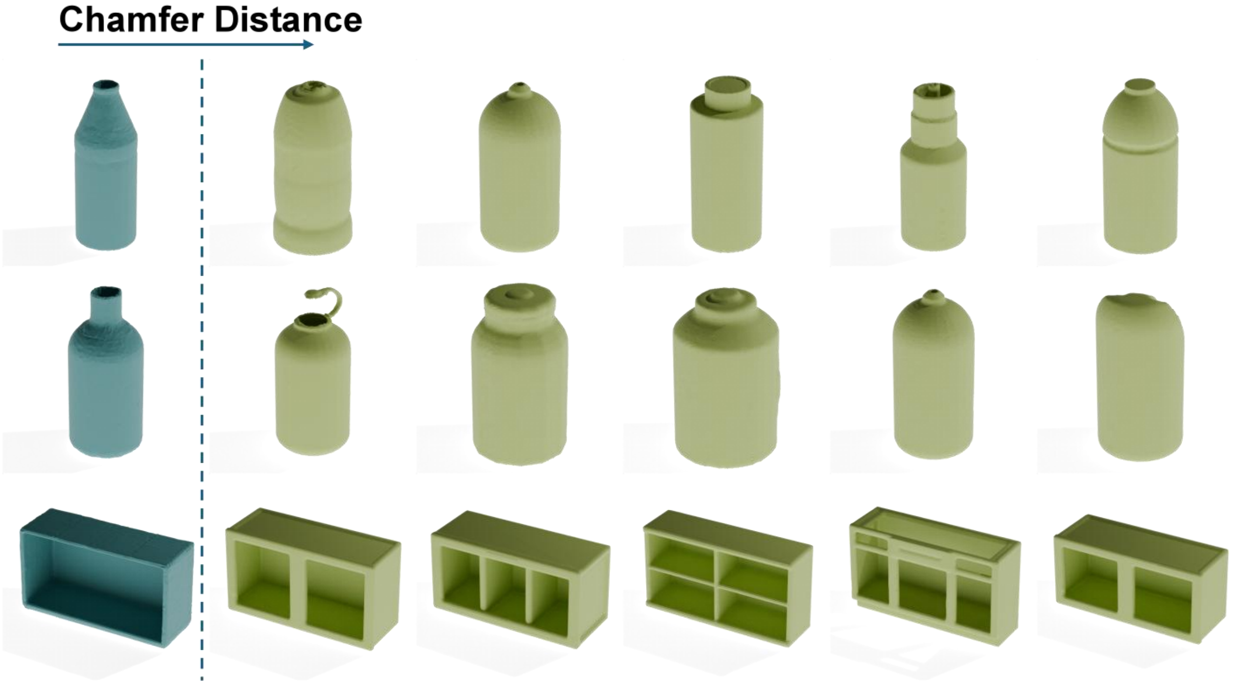
where $g_1, \cdots, g_k$ are samples from Gumbel$(0, 1)$ [20]. The softmax temperature τ controls the diversity of shape prior, which we do not specifically tune in this work. Since GS sampling is differentiable, the model can still be trained end-to-end.

Comparison


Ours
NAP-128
(ours shape prior 128)
NAP-768
(ours shape prior 768)
Ours-NAPSP
(NAP shape prior)
CAGE
(part retrieval)
Ours-PR
(part retrieval)
Ours-1CB
(1 codebook)
(ours shape prior 768)

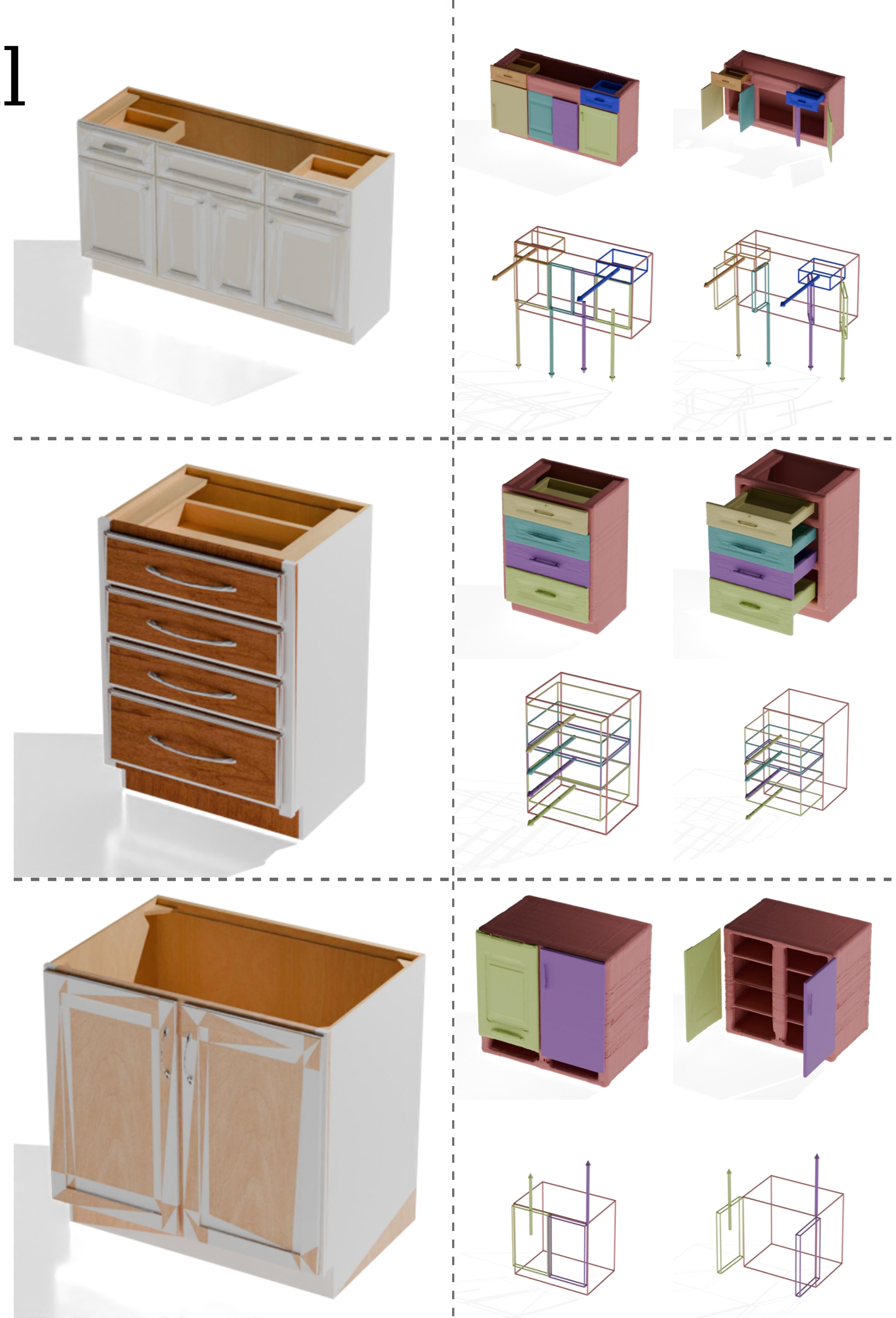
Additional Material

- **Generate new geometry**
- Image condition
- Text attention weight
- Edit Articulated Object



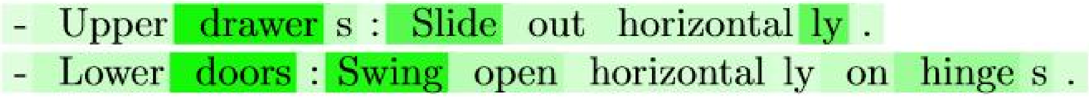
Additional Material

- Generate new geometry
- **Image condition**
- Text attention weight
- Edit Articulated Object



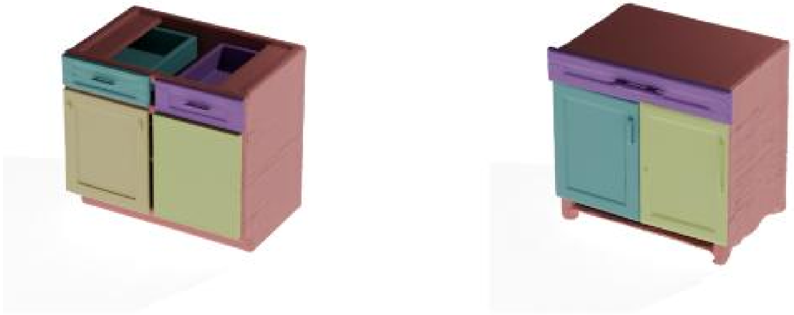
Additional Material

- Generate new geometry
- Image condition
- **Text attention weight**
- Edit Articulated Object



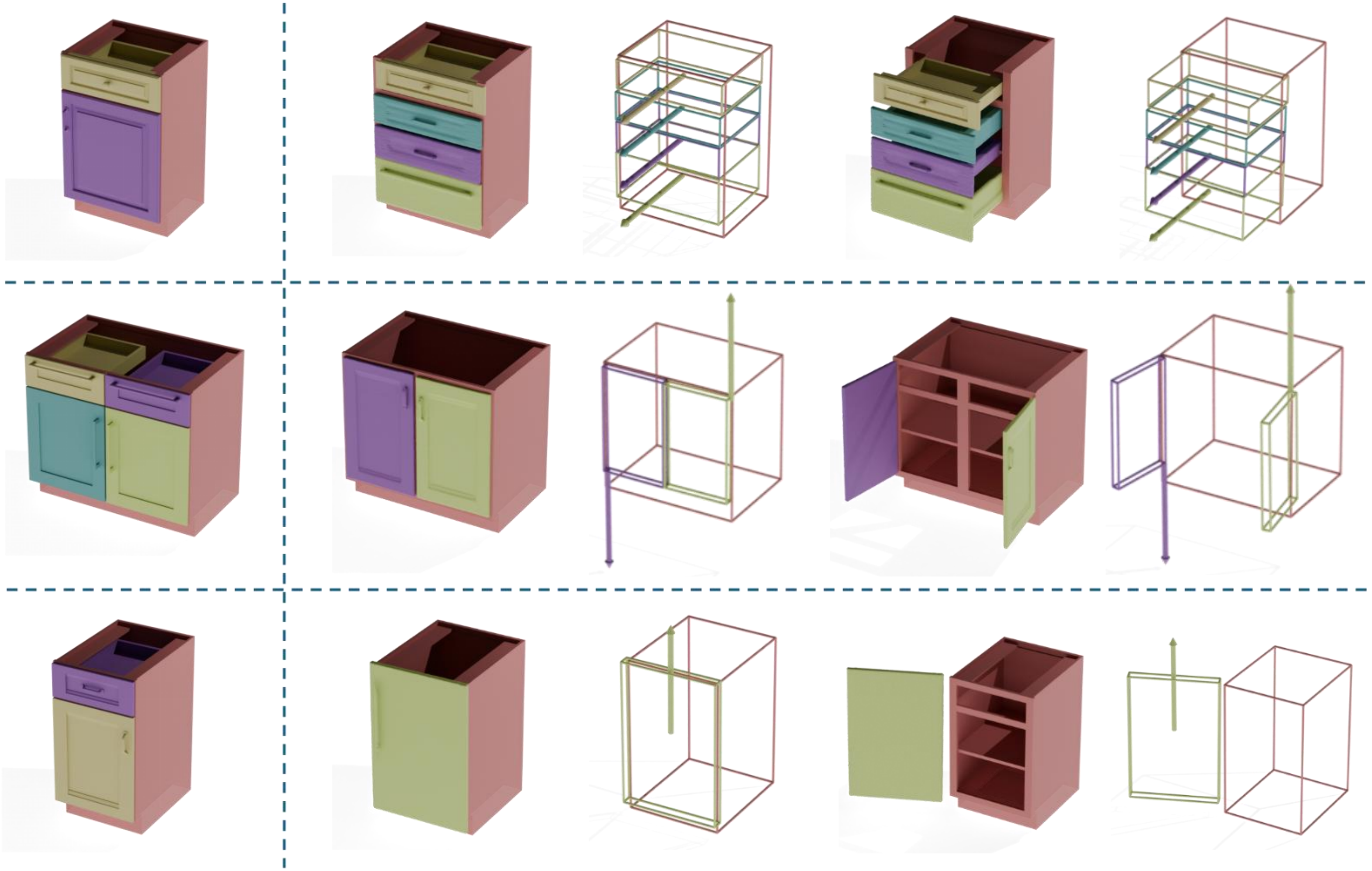
- Upper drawer s : Slide out horizontal ly .

- Lower doors : Swing open horizontal ly on hinge s .



Additional Material

- Generate new geometry
- Image condition
- Text attention weight
- **Edit Articulated Object**



1. *This storage furniture consists of a rectangular frame with multiple horizontally aligned drawers that slide in and out on tracks.*

2. *This storage furniture consists of a rectangular base with two front panels that pivot on vertical hinges to open outward.*

3. *Rectangular frame: stationary base. Front panel: hinged door, pivots outward.*

Thank You!