# Supplemental Material
# PT43D: A Probabilistic Transformer for Generating 3D Shapes from Single Highly-Ambiguous RGB Images

Yiheng Xiong
yiheng.xiong@tum.de
Angela Dai
angela.dai@tum.de

3D AI Lab
Technical University of Munich
Munich, Germany

In this supplementary material, we present more qualitative comparisons on both synthetic and real-world data in Sec. 1. We also provide more implementation details in Sec. 2.

## 1 Additional Qualitative Comparisons

We present more qualitative comparisons among AutoSDF, SDFusion and our method. Fig. 1, Fig. 2 and Fig. 3 contain visualizations from synthetic data. Fig. 4, Fig. 5 and Fig. 7 provides qualitative samples from real-world data. Our method generally generates higher quality and more plausible hypothesis shapes compared with other baselines.

## 2 Further Implementation Details

### 2.1 Multi-Hypothesis Data Augmentation

In our approach, we allow the input image to be mapped to potentially multiple ground-truth shapes that align with the image. We initially classify CAD models from the dataset into similar groups. When evaluating two models, if they exhibit identical part counts and semantics, and their geometric similarity surpasses a predefined threshold, we classify them as similar. Then, for each rendered view of the target model, we extract per-pixel part labels and visible points in 3D space. We consider models from the same similar group as mapping candidates. We iterate through these candidates, employing exactly the same rendering parameters as the target model, and obtain their per-pixel part labels and visible parts in 3D space. We then compare this information with that of the target model: if the overlap of per-pixel part labels and the geometric similarity of visible parts exceed predefined thresholds, we include the candidate in the ground-truth mapping for the image.

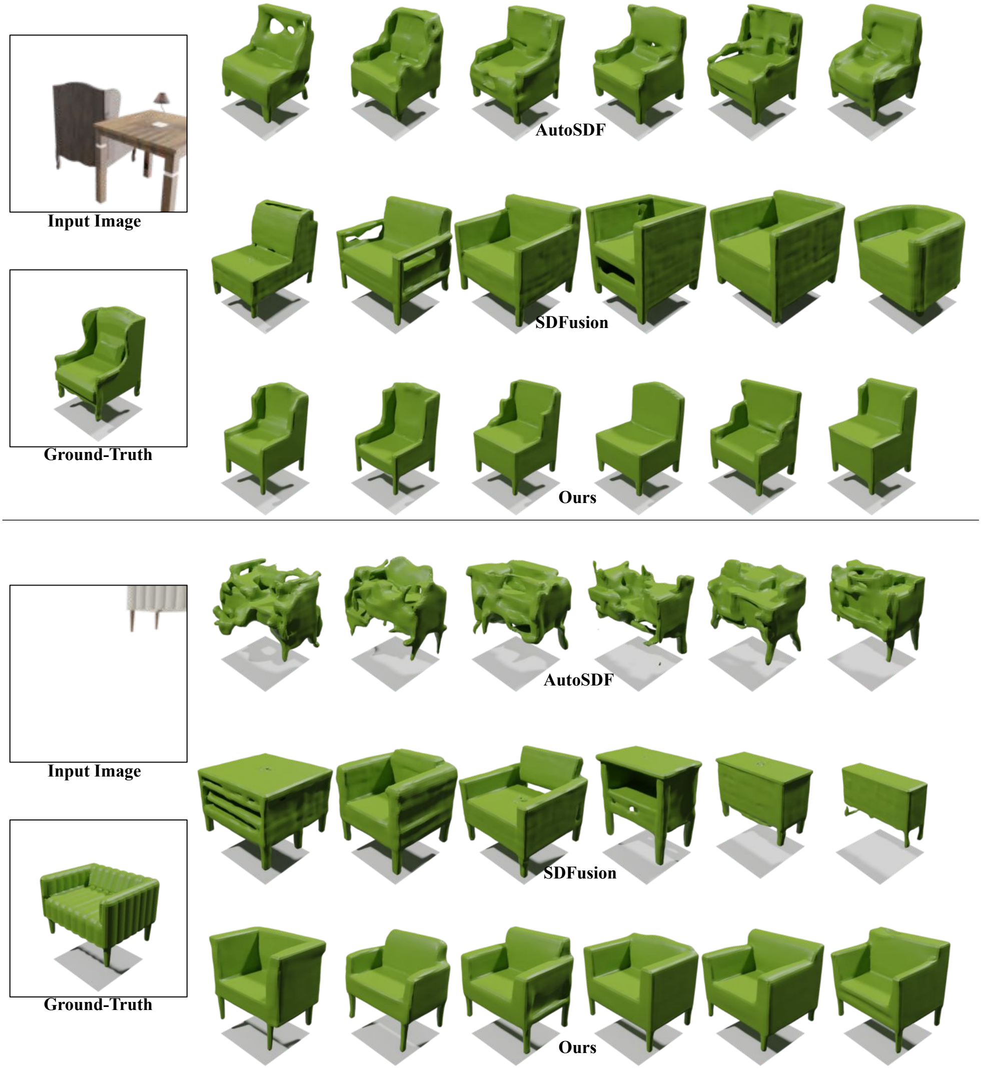

Figure 1: **More Qualitative Comparisons on Synthetic Data.** Our method generates higher quality and more plausible hypotheses compared with other baselines.

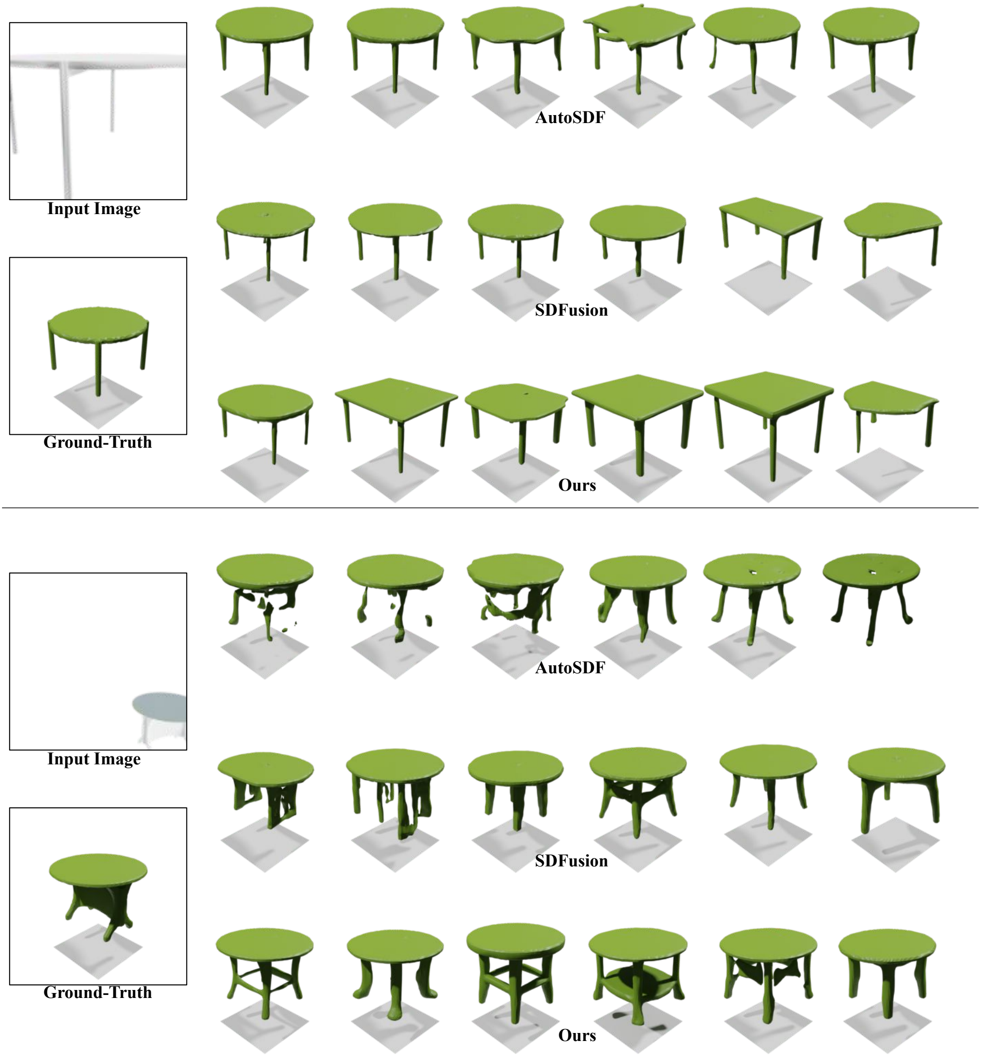

Figure 2: **More Qualitative Comparisons on Synthetic Data.** Our method generates higher quality and more plausible hypotheses compared with other baselines.

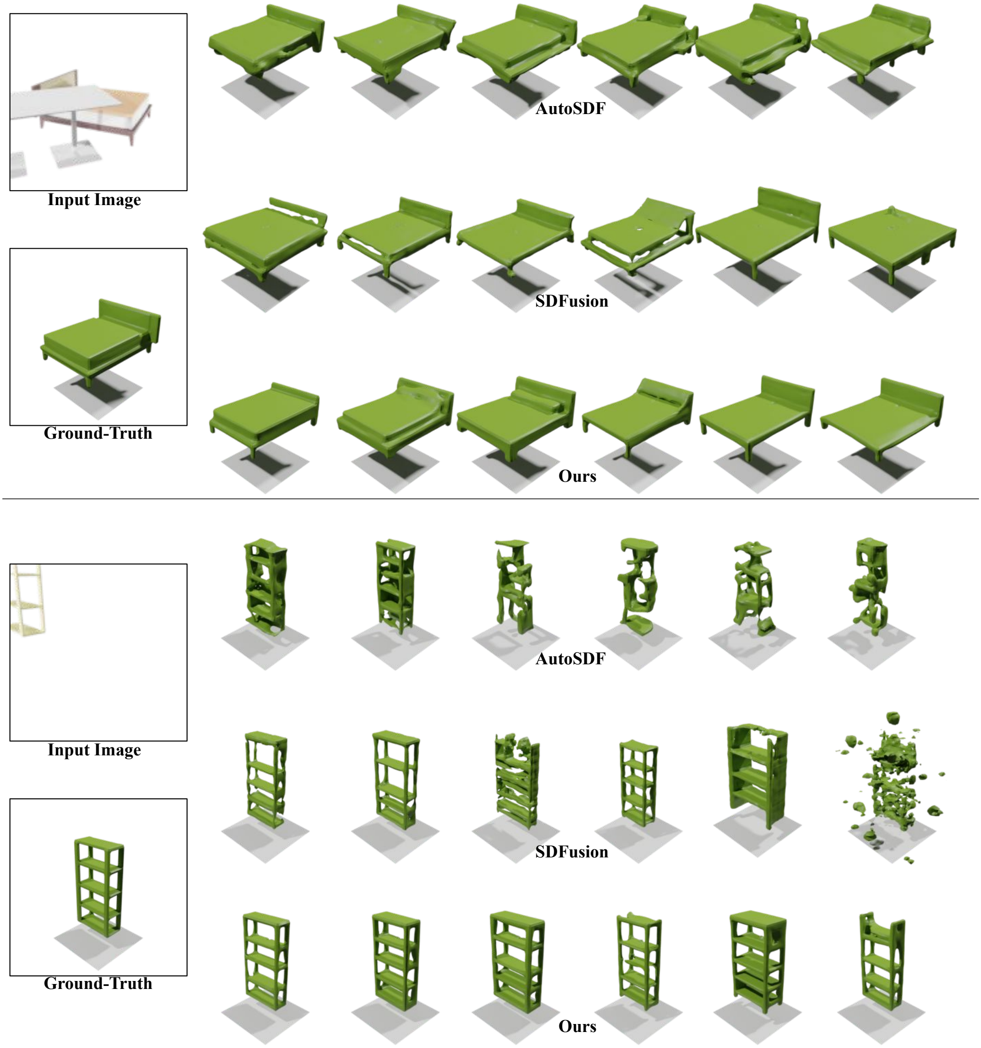

Figure 3: **More Qualitative Comparisons on Synthetic Data.** Our method generates higher quality and more plausible hypotheses compared with other baselines.

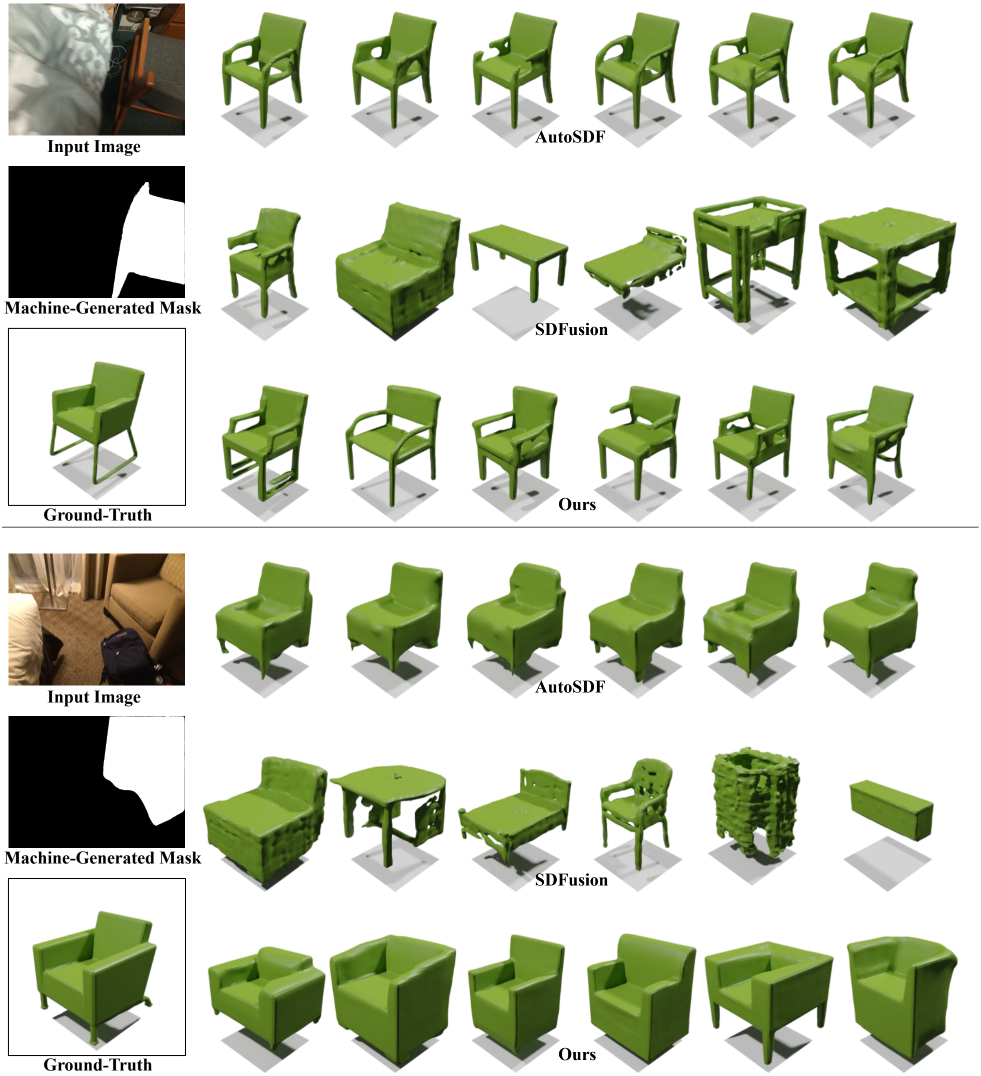

Figure 4: **More Qualitative Comparisons on Real-World Data.** Our method generates higher quality and more plausible hypotheses compared with other baselines.

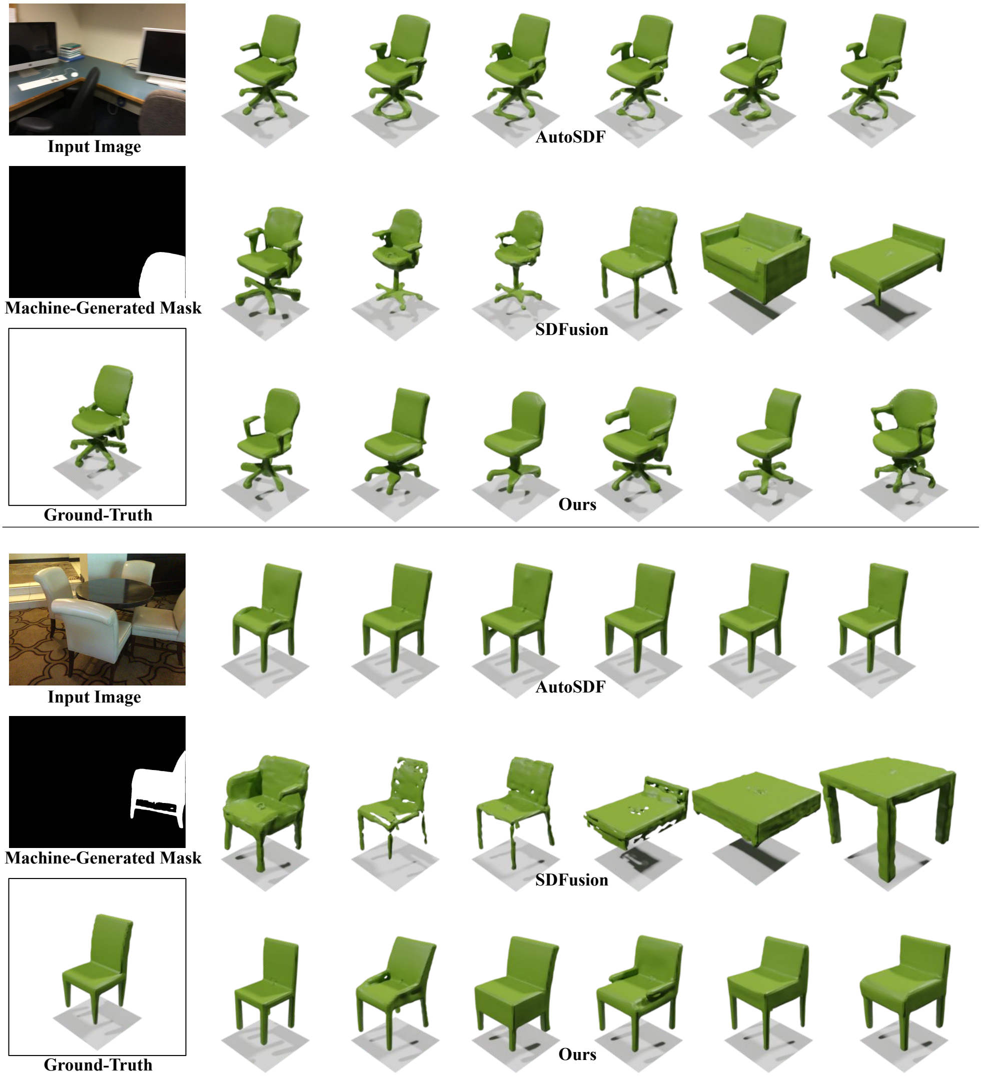

Figure 5: **More Qualitative Comparisons on Real-World Data.** Our method generates higher quality and more plausible hypotheses compared with other baselines.

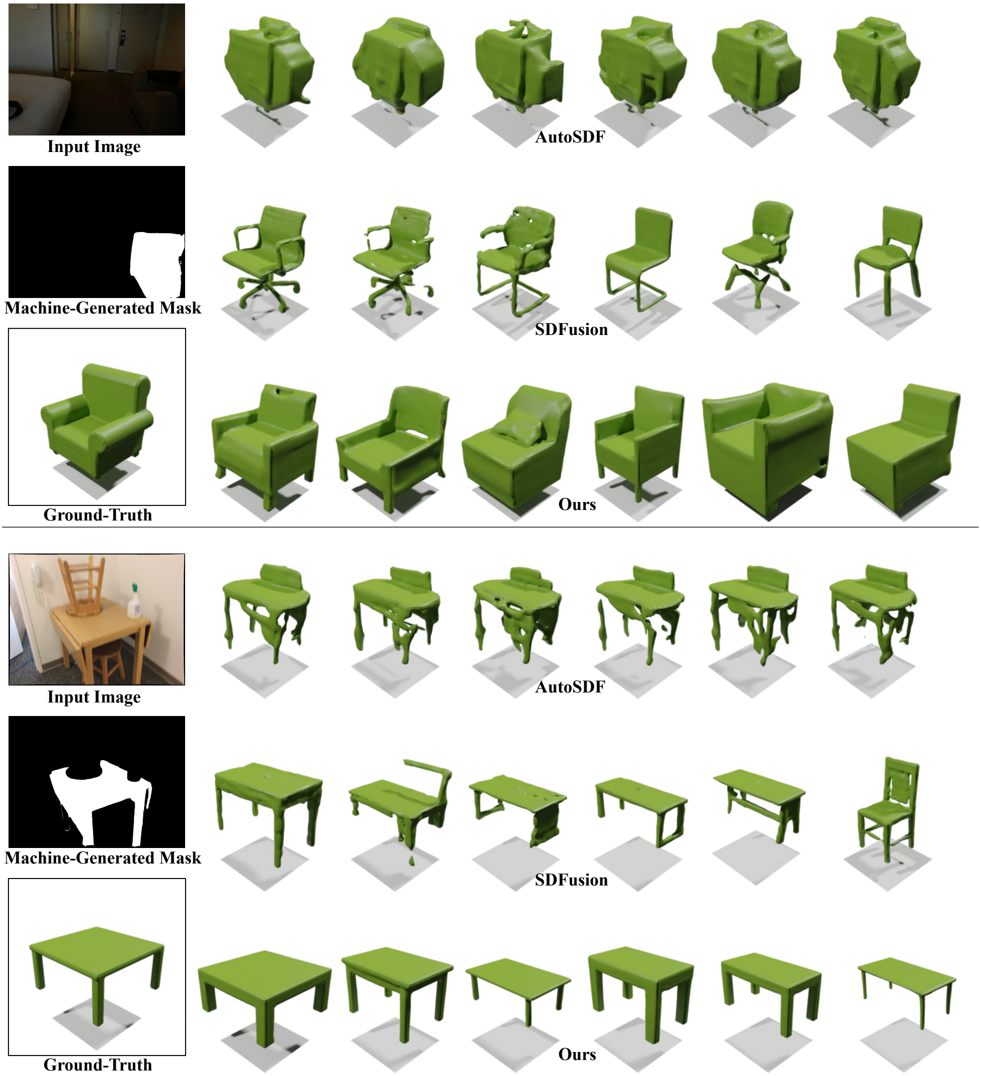

Figure 6: **More Qualitative Comparisons on Real-World Data.** Our method generates higher quality and more plausible hypotheses compared with other baselines.

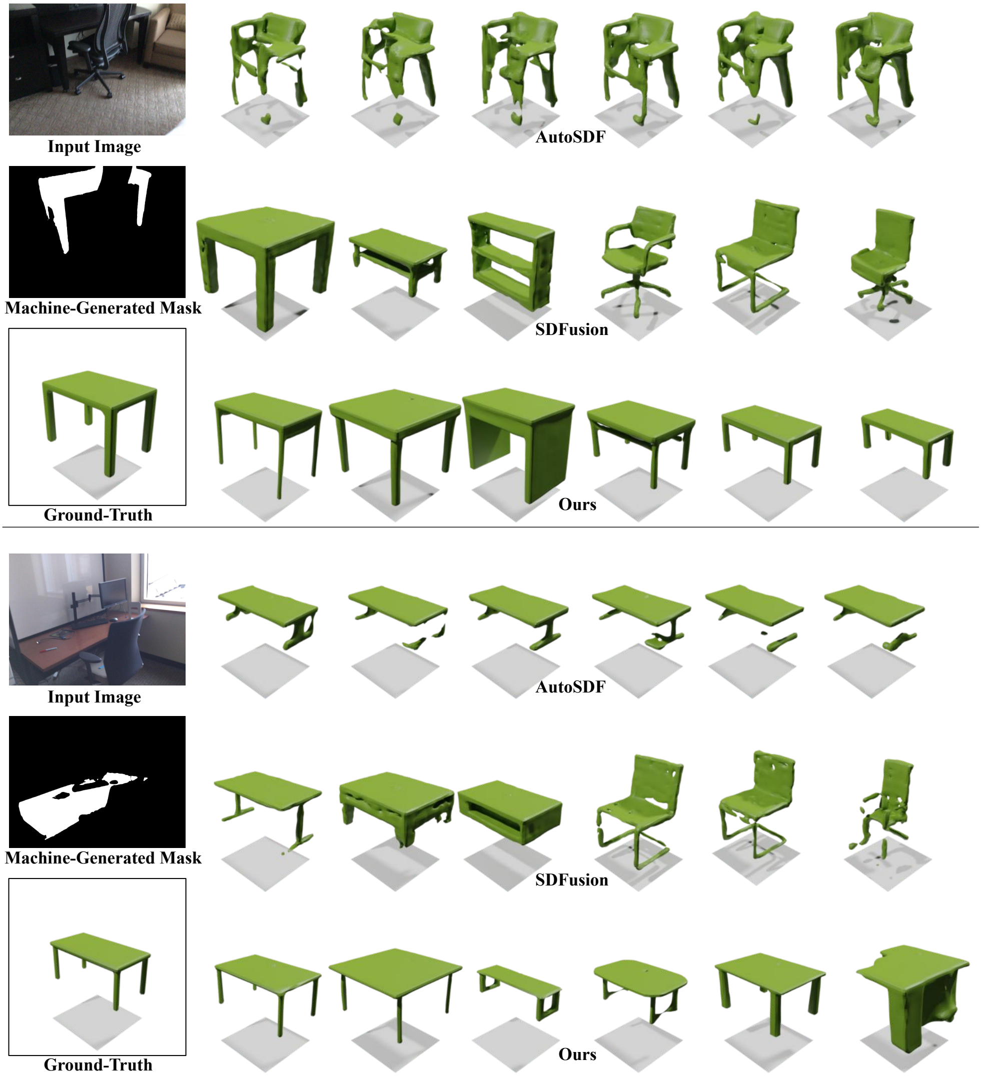

Figure 7: **More Qualitative Comparisons on Real-World Data.** Our method generates higher quality and more plausible hypotheses compared with other baselines.

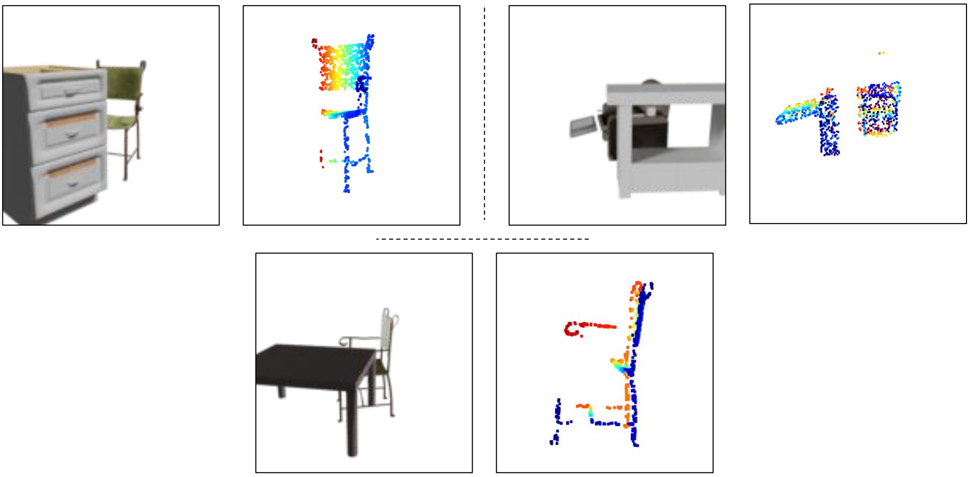

Figure 8: **Visible Points in 3D Space.** We present three sets of synthetic samples. Within each set, left is the rendering and right is the visible points of the target object in 3D space.

## 2.2 Network Architecture

**Image Encoder.** We choose the "ViT-B/32" version of CLIP[2] as our image encoder, yielding image encodings with a shape of $N' \times D'$, where $N' = 50$ represents the number of image tokens, and $D' = 768$ indicates the feature dimension.

**Conditional Cross-Attention.** The input sequence is multiplied with a weight matrix $Q$. This matrix multiplication results in a sequence of *queries* with the shape of $N'' \times d$, where $N''$ is the length of the input sequence and $d$ is a predefined hidden dimension. Likewise, the image encodings are multiplied with weight matrices $V$ and $K$ independently, generating *values* and *keys* respectively. Both are in the shape of $N' \times d$. Subsequently, each *query* performs dot product with each *key*, generating corresponding attention scores $\alpha$:

$$\alpha_{mn} = \frac{softmax(q_m \cdot k_n)}{\sqrt{d}}. \tag{1}$$

where $q_m$ is the $m$-th *query* and $k_n$ is the $n$-th key. Then for each cell $i$ of the input sequence, its embedding is replaced by the weighted sum of *values*:

$$emb_i = \sum_{j=0}^{N'-1} \alpha_{ij} \cdot v_j. \tag{2}$$

where $v_j$ is $j$-th *value*. In practice, we apply 8 multi-head attention heads and a hidden dimension of 512.

**Transformer.** The transformer[4] comprises 12 encoder layers, each with 12 multi-head attention heads and a hidden dimension of 768. Notably, it does not contain a decoder, indicating that all attention layers are self-attention. The training within the transformer is done in parallel. We feed the attention mask with upper-triangular matrix of $-\infty$, and zeros on the diagonal to make sure the information do not leak from the future elements. We use fourier features for the positional embedding for all locations $i$ following Tancik *et al.* [3].

### 2.2.1 Fine-Tuning on Real-World Data

To address the domain gap between our synthetic training pairs and real-world images, we fine-tune our pretrained model using real-world images from ScanNet[1]. To preserve the ability for generating diverse shapes, we freeze the transformer-based generation backbone and only fine-tune the CLIP encoder and the conditional cross-attention module. We fix the batch size to 10 and utilize an initial learning rate of 5e-6 for the CLIP encoder and 1e-5 for the conditional cross-attention module. We fine-tune them for around 1,000 iterations.

## References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[3] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains, 2020.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.