Supplementary Material: SDFusion: Multimodal Shape Completion, Reconstruction, and Generation

We provide the implementation details of the VQ-VAE, diffusion models, and multimodal conditional model. We also provide a **index.html** file for better visualization of the generated 3D shapes. It contains the **animated .gif files** for more shape completion, single-view reconstruction, and text-guided generation. They can be found at **supp_webpage/index.html**.

In the following we first discuss implementation details (Sec. A) before providing additional results for unconditional generation (Sec. C), shape completion (Sec. D), single-view reconstruction (Sec. E), text-guided generation (Sec. F), multi-modal conditional generation (Sec. G).

A. Implementation Details

We will release our code and learned models for reproducibility, but also describe the experiments in additional detail in the following.

A.1. VQ-VAE Training

Dataset Details. We train the VQVAE using the objects from 13 categories of the ShapeNet [1] data. These categories include [airplane, bench, cabinet, car, chair, display, lamp, speaker, rifle, sofa, table, phone, watercraft]. For BuildingNet [5], we train on all the provided shapes. To extract the signed distance function (SDF), we follow the preprocessing steps by DISN [7] and PixelTransformer [6]. We first normalize the shapes to an origin-centered cube in $[-1, 1]^3$. Their signed distance function is evaluated at locations in a uniformly sampled 64^3 grid for ShapeNet, and 128^3 for BuildingNet. To obtain the Truncated-SDF (T-SDF), we use a threshold of 0.2.

Training Details. Given an input shape $\mathbf{X}$, the encoder E_φ encodes it into a latent vector $\mathbf{z}$. We then perform the quantization step to obtain the quantized vector $\hat{\mathbf{z}} = \mathrm{VQ}(\mathbf{z})$. After the quantization, we use decoder D_τ to reconstruct the shape $\mathbf{X}'$. We then use the training objective proposed in the VQ-VAE work [4], *i.e.*,

$$\mathcal{L}_{\mathrm{VQ-VAE}} = -\log p(\mathbf{X}|\mathbf{z}) + \|\mathrm{sg}[\hat{\mathbf{z}}] - \mathbf{z}\|^2 + \|\hat{\mathbf{z}} - \mathrm{sg}[\mathbf{z}]\|^2, \tag{S1}$$

where the first term is the reconstruction loss and $\mathrm{sg}[\cdot]$ denotes the stop gradients. The second and third term in Eq. S1 is the VQ objective and the commitment loss respectively.

A.2. Multimodal Conditional Model Training

Cross attention. Given the encoded condition vector $\mathbf{c}_i$ and the latent vector $\mathbf{z}_t$, the cross-attention layer is performed as follows:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d}})V. \tag{S2}$$

where

$$Q = W_Q^{(i)} \cdot \psi_i(\mathbf{z}_t), \;\; K = W_K^{(i)} \cdot E_{\phi_i}(\mathbf{c}_i), \;\; V = W_V^{(i)} \cdot E_{\phi_i}(\mathbf{c}_i). \tag{S3}$$

and $\psi_i(\mathbf{z}_t)$ is a flattened vector produced by $i_t h$ layer of the 3D UNet.

B. Ablations, Results and Design Choices

Evaluation of unconditional generation. We compare the proposed method to GET3D [2] using their pretrained weight on ShapeNet's chair category. Following the evaluation in GET3D, in Table S1 we compare diversity with COV metric and fidelity using MMD metric. The proposed method is comparable with GET3D in terms of diversity and fidelity.

Table S1. **Quantitative evaluation of unconditional generation and the comparison with GET3D**.

	COV (%, ↑)		MMD (↓)	
Method	LFD	CD	LFD	CD
GET3D	75.70	**71.62**	3083	3.69
Ours	**78.80**	67.45	**3042**	**3.23**

Improvement over AutoSDF. To better understand the impact of each component, we train 1) SDFusion's encoder (Ours $Enc.$) with AutoSDF's transformer, and 2) AutoSDF's encoder (AutoSDF $Enc.$) with a latent diffusion model. Table S2 shows that the proposed encoder and the diffusion model improve both the fidelity and diversity, especially the diversity.

Table S2. **Ablation and comparison with AutoSDF**.

Method	UHD ↓	TMD ↑
AutoSDF	0.0567	0.0341
Ours *Enc.* + Transformer	0.0582	0.0491
AutoSDF *Enc.* + Diffusion	0.0563	0.0594
Ours	**0.0557**	**0.0885**

Importance of discrete latent space. To understand the impact of a discrete latent space, we train the proposed method on ShapeNet chairs with a plain VAE. Table S3 shows the advantage of a discrete latent space in both fidelity and diversity. The pros of a discrete latent space are two-fold: 1) stable training of the encoder due to the regularization of the latent space; 2) simplified learning of the diffusion model. The cons: 1) additional quantization operation and additional loss for the codebook; 2) the expressiveness of the model is limited by the codebook size.

Table S3. **Continuous vs. discrete latent space**.

Method	UHD ↓	TMD ↑
Continuous	0.0585	0.0620
Discrete	**0.0557**	**0.0885**

Computation and Memory Savings. In Table S4, we measure MACs (multiply–accumulate operation) and memory consumption for diffusion models using raw voxels and the encoded latent space (ours) with batch size 1 using ShapeNet. The resolution of the raw voxels is 64^3. We can see that SDFusion saves both computation and memory compared to operating on raw inputs.

Table S4. **Comparisons of computation (MACs) and memory**.

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

C. Unconditional Generation Results

We show the unconditional generation results on ShapeNet and BuildingNet at Figure S1.

D. Shape Completion Results

More Comparisons of Multimodal Shape Completion. We provide additional comparisons of our shape completion results to baselines in Figure S3. Further, we compare our generated SDF to results from AutoSDF [3] in Figure S2.

More Results of Multimodal Shape Completion. We show more results for shape completion in Figure S4.

E. Single-view Reconstruction Results

We provide more single-view reconstruction results in this section. Additional comparisons to baselines are presented in Figure S5, and more results from our proposed method are available in Figure S6.

F. Text-guided Generation Results

More comparisons of text-guided generation. We provide additional comparisons with AutoSDF in Figure S7.

More results of text-guided generation. We provide more text-guided generation results in Figure S8.

G. Multimodal Conditional Generation Results

We showcase more multimodal conditional generation results in Figure S9.

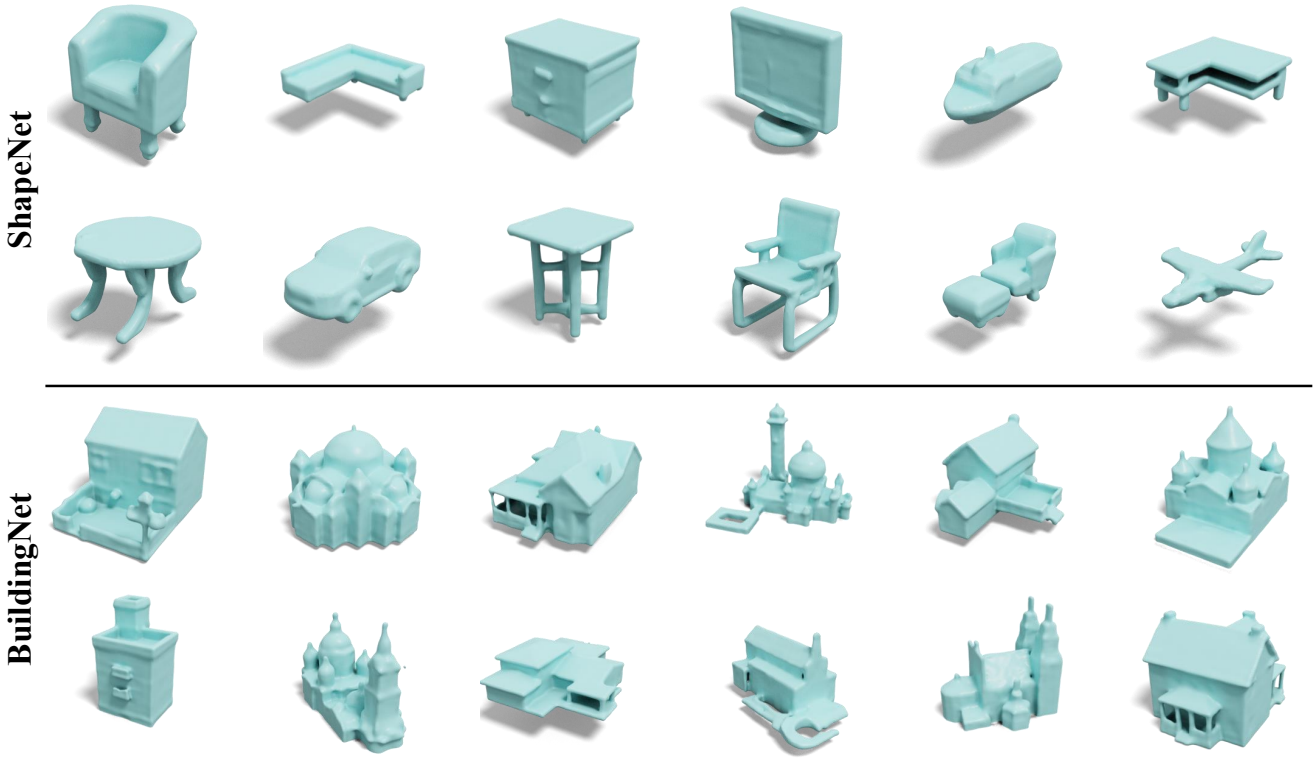

Figure S1. **Qualitative results of unconditional generation from SDFusion.**

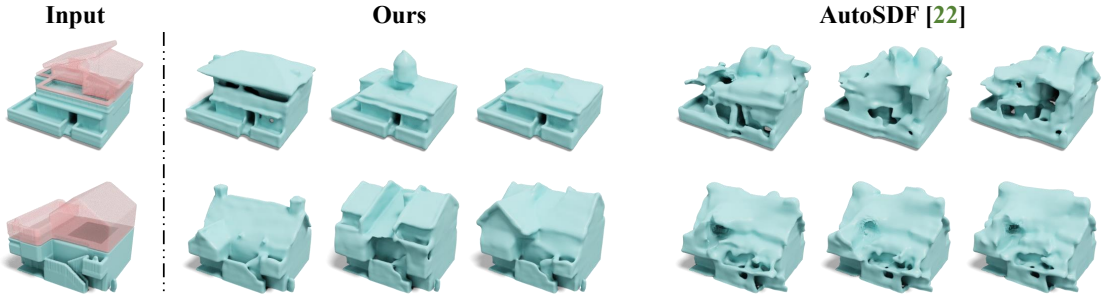

Figure S2. **Qualitative comparisons of the meshes with the AutoSDF [3].**

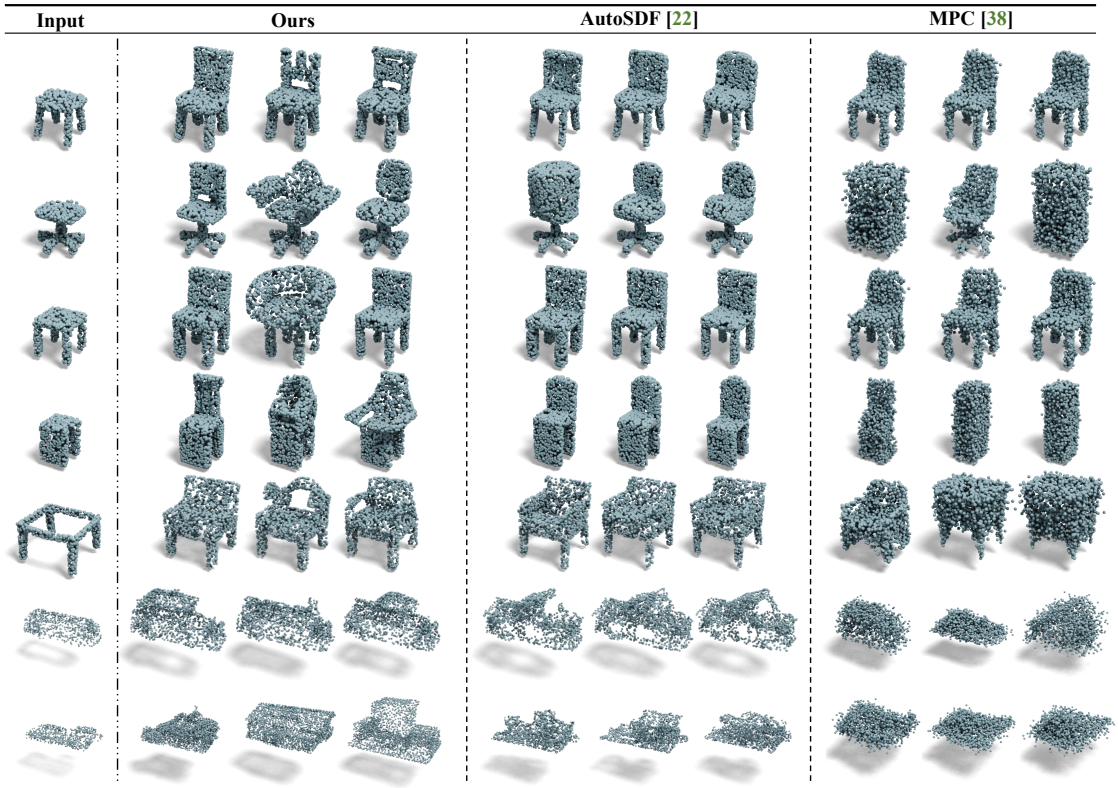

Figure S3. **Qualitative comparisons of shape completion.**

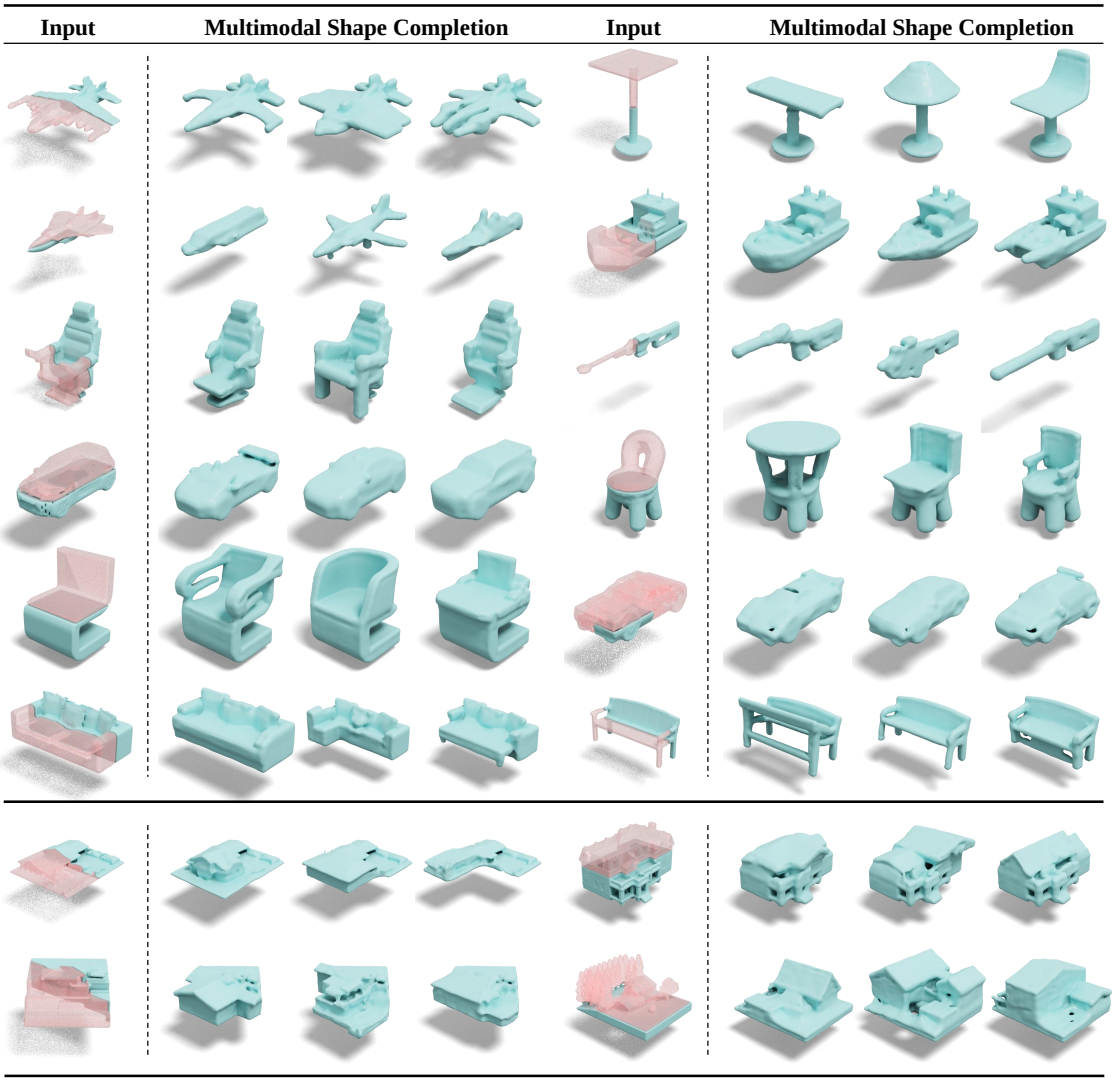

Figure S4. **Qualitative results of shape completion.**

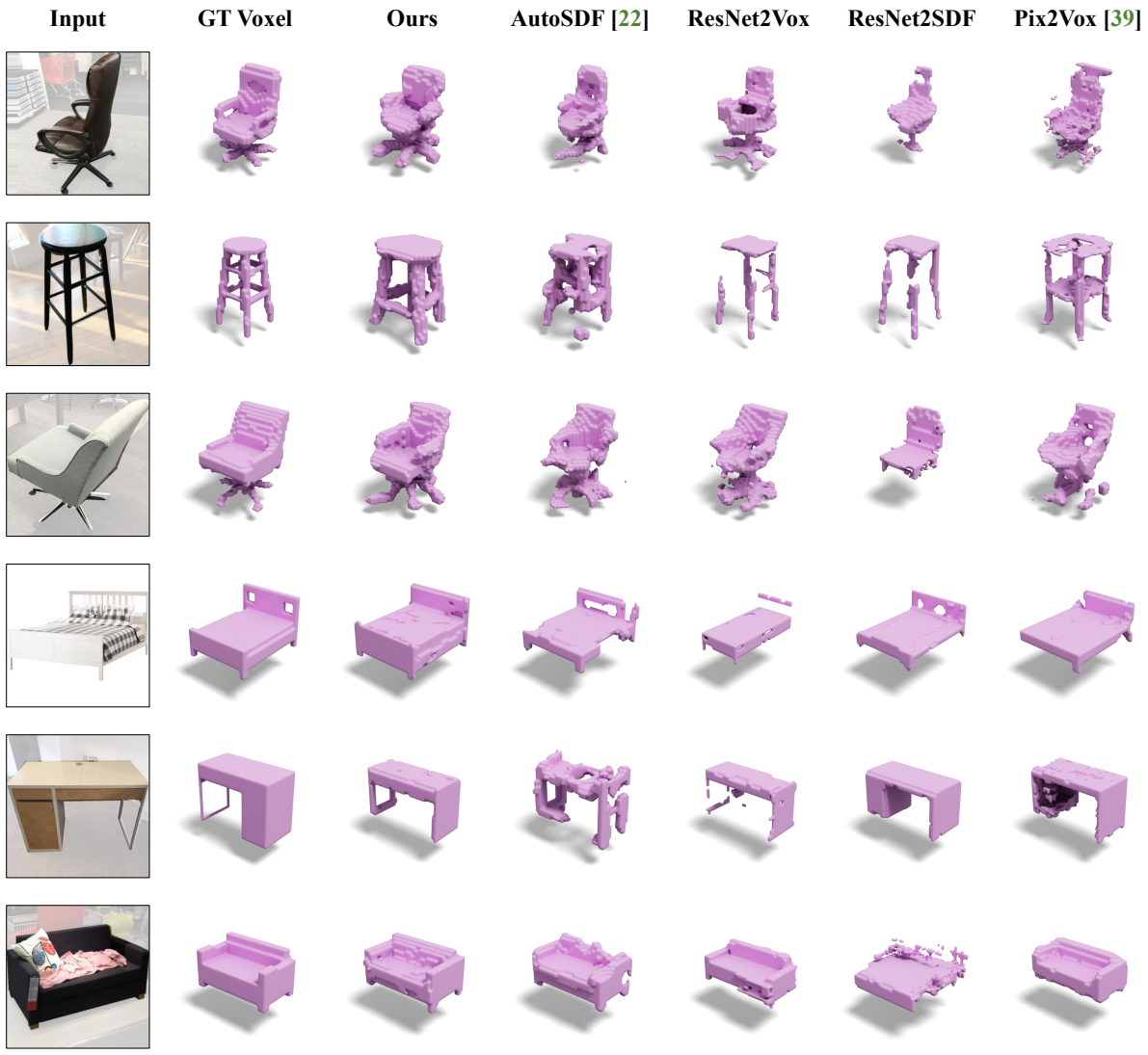

Figure S5. **Qualitative comparisons of single-view reconstruction.**

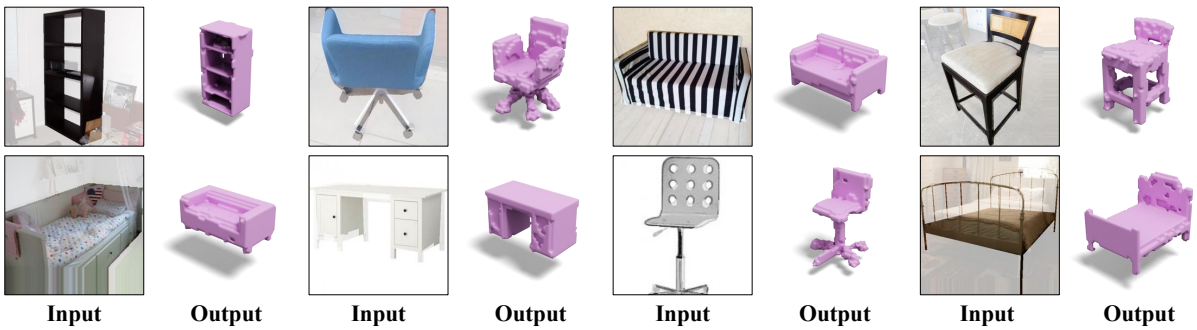

Figure S6. **Qualitative results of single-view reconstruction from SDFusion.**

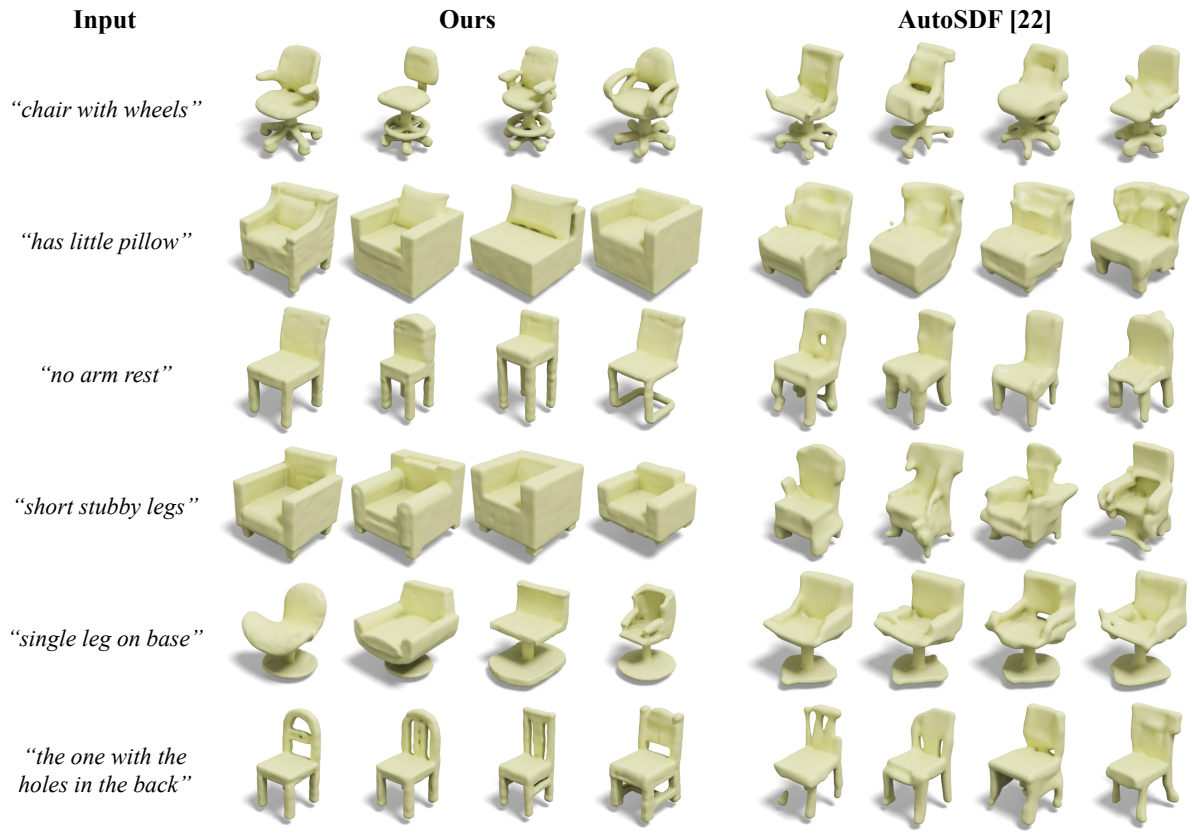

Figure S7. **Qualitative comparisons of the text-guided shape generation.**

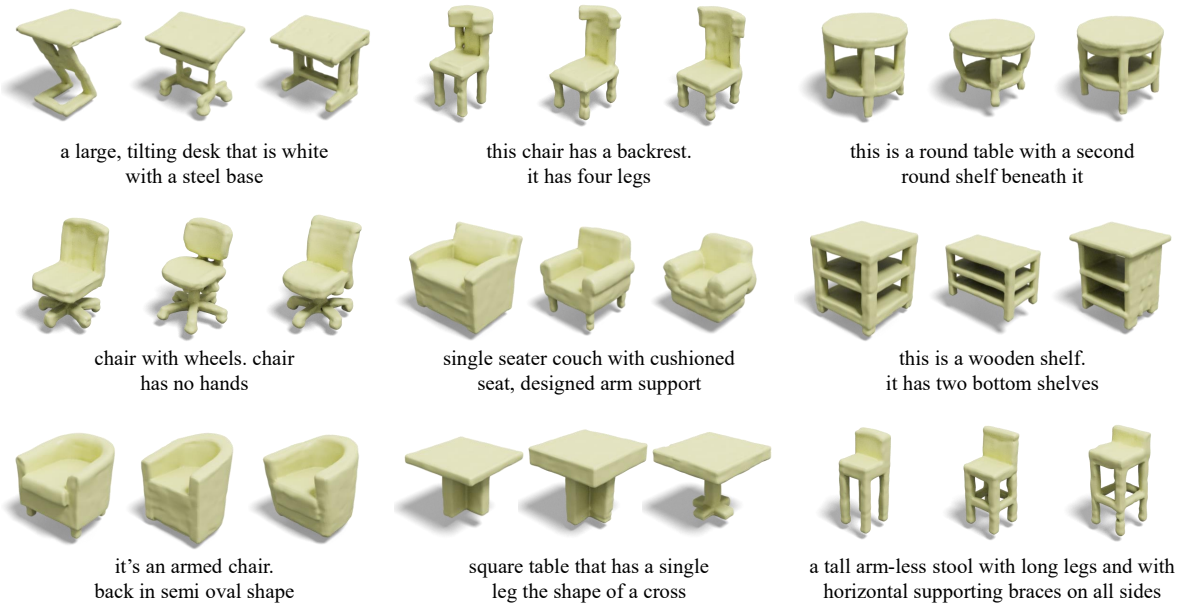

Figure S8. **Qualitative results of text-guided shape generation of SDFusion.**

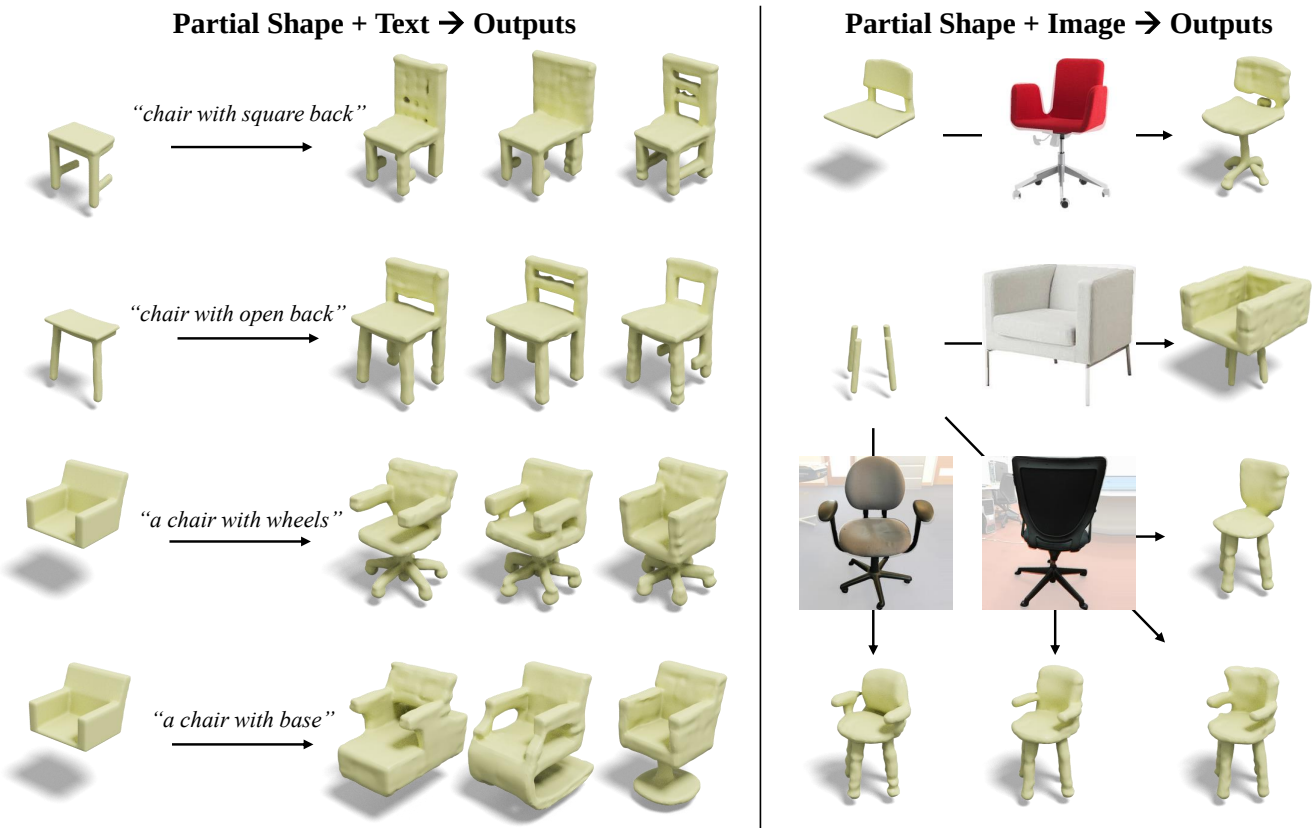

Figure S9. **Qualitative results of multimodal conditional generation results.**

References

[1] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 1

[2] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In *NeurIPS*, 2022. 1

[3] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. AutoSDF: Shape priors for 3d completion, reconstruction and generation. In *CVPR*, 2022. 2, 3

[4] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017. 1

[5] Pratheba Selvaraju, Mohamed Nabail, Marios Loizou, Maria Maslioukova, Melinos Averkiou, Andreas Andreou, Siddhartha Chaudhuri, and Evangelos Kalogerakis. Buildingnet: Learning to label 3d buildings. In *ICCV*, 2021. 1

[6] Shubham Tulsiani and Abhinav Gupta. Pixeltransformer: Sample conditioned signal generation. In *ICML*, 2021. 1

[7] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *NeurIPS*, 2019. 1