Sketch and Text Guided Diffusion Model for Colored Point Cloud Generation (Supplementary Material)

Here, we provide more details of our proposed method, including color point cloud generation from freehand sketch+text and from only freehand sketch. Section 2 presents typical attention maps for sketches and compares the instance scores before and after attention. In Section 3, we also include a more detailed comparison between DPM [2], PVD [4], point-e [3], and STPD.

We also discuss some of the interesting features of our proposed method at the end.

1. Color Point Cloud Generation

Returning to the color point cloud generation ability of our proposed model, in this section, we present additional visualizations for 3D generation of the $chair$ and $table$ categories from freehand sketch and text in point cloud generation (shown in Figure 1). Specifically, only $chair$ and $table$ categories have color descriptions to support the sketch-text based coloring. We ensure that the edge detection sketches match hand drawn ones. Our conditional method produces colored shapes that are more faithful to the input sketch-text descriptions. Notice how our method pays more attention to the shape details apparent in the sketch (e.g., see the table legs). Note that for better visualization, we up-sample each point cloud to 4096 points.

2. Typical Attention Maps Related to Sketches

We also present additional attention maps generated by our attention-capsule network. In this section, we introduce the Instance Score as the evaluation metric. For the instance score on sketch (ISS), we count the number of useful pixels and express it as a percentage of all sketch pixels. For the instance score on capsule (ISC), we compute the highest attention score of each dimension to all capsules and use their average as the evaluation metric.

All visualized sketch and attention score maps are reported in Table 2 and Figure 3. Note that, across all 24 sketches, the average number of useful pixels representing the instance is only 1.07% of the whole sketch. After our capsule-attention module, the attention weight of the instance improves to 58.68% for better diffusion.

3. More Point Cloud Generation Results

We further test our STPD model for generating object shapes and categories from sketches. In this section, we focus only on shape and do not include appearance diffusion. Figure 5 shows our results. For each generation, we input the sketch to generate the 3D shape for all four categories ($airplane$, $chair$, car, and $table$) used in our experiments. Since there is no text description associated with 3D shape to support color diffusion, we manually assigned colors for better visualization. The numerous generation results demonstrate that our STPD has a good ability to generalize and generate different 3D shapes.

4. Detailed Comparisons

We provide a more detailed visualized comparison between PVD [4], DPM [2], Point-E [3], and our STPD. PVD is trained without color. However, this approach has some drawbacks, such as not being able to capture the full range of colors present in a given image.

DPM has been extended to simultaneously capture color and geometry. However, the additional dimension for color may bias the shape and result in higher variance when simply applying the dimension extension on the sketch-text dataset.

Point-E is demonstrated here from a well-trained large dataset and performs great. However, there are two main problems with Point-E: 1) the shape generation does not strictly follow the conditions (e.g., conditions with black metallic arms but not generating the arms) and 2) the color is not accurately related to the shape parts as conditioned (e.g., showing black metallic feet but displaying gray feet).

In contrast, our method, STPD, shows faithful and fine results from the conditions.

5. Additional Discussions

In this section, we will discuss three features of our proposed method: (a) generating novel shapes, (b) the effects of slight changes in the sketch, and (c) conflicting sketch and text.

Our STPD method can learn the variance of parts from the training set. Given novel changes (such as back holes

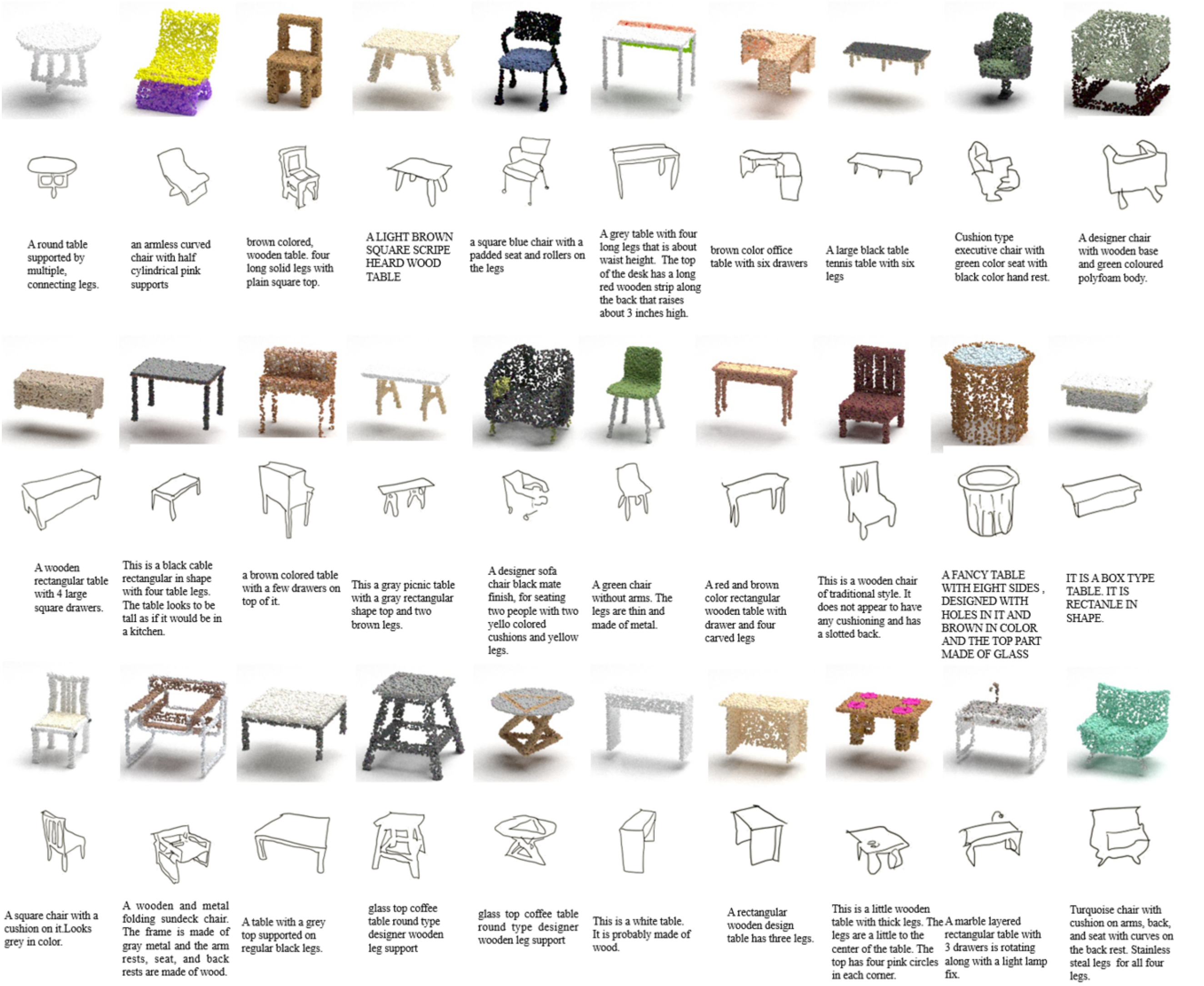
A round table supported by multiple, connecting legs.
an armless curved chair with half cylindrical pink supports
brown colored, wooden table. four long solid legs with plain square top.
A LIGHT BROWN SQUARE SCRIPE HEARD WOOD TABLE
a square blue chair with a padded seat and rollers on the legs
A grey table with four long legs that is about waist height. The top of the desk has a long red wooden strip along the back that raises about 3 inches high.
brown color office table with six drawers
A large black table tennis table with six legs
Cushion type executive chair with green color seat with black color hand rest.
A designer chair with wooden base and green coloured polyfoam body.
A wooden rectangular table with 4 large square drawers.
This is a black cable rectangular in shape with four table legs. The table looks to be tall as if it would be in a kitchen.
a brown colored table with a few drawers on top of it.
This a gray picnic table with a gray rectangular shape top and two brown legs.
A designer sofa chair black mate finish, for seating two people with two yello colored cushions and yellow legs.
A green chair without arms. The legs are thin and made of metal.
A red and brown color rectangular wooden table with drawer and four carved legs
This is a wooden chair of traditional style. It does not appear to have any cushioning and has a slotted back.
A FANCY TABLE WITH EIGHT SIDES , DESIGNED WITH HOLES IN IT AND BROWN IN COLOR AND THE TOP PART MADE OF GLASS
IT IS A BOX TYPE TABLE. IT IS RECTANLE IN SHAPE.
A square chair with a cushion on it.Looks grey in color.
A wooden and metal folding sundeck chair. The frame is made of gray metal and the arm rests, seat, and back rests are made of wood.
A table with a grey top supported on regular black legs.
glass top coffee table round type designer wooden leg support
glass top coffee table round type designer wooden leg support
This is a white table. It is probably made of wood.
A rectangular wooden design table has three legs.
This is a little wooden table with thick legs. The legs are a little to the center of the table. The top has four pink circles in each corner.
A marble layered rectangular table with 3 drawers is rotating along with a light lamp fix.
Turquoise chair with cushion on arms, back, and seat with curves on the back rest. Stainless steal legs for all four legs.

Figure 1. Visualization of the generated objects from free-hand sketch and text. Our model generates better geometry and has the additional functionality of generating colored point clouds. Being conditional, our method produces colored shapes that are more faithful to the input sketch-text descriptions.

and hollow backrest sketch) that do not exist in the training shapes, the STPD learns these changes from the whole shapes and modifies the generated shapes to make the results faithful to the condition. However, we only trained the diffusion model on some categories of ShapeNet [1], which may limit its ability to deal with totally strange objects. It would be an interesting direction to train a more generalized model across large datasets using our extending methods, like the Stable-diffusion did in the 2D area.

We also tested some slight changes in the sketch to show that the results are faithful to the conditions. However, our vanilla STPD model is trained on a small dataset with sketch-text conditions, and we did not consider conflicts during training. Thus, we tested conflicting input sketches and text (such as a sketch with no armrest and text with an armrest). During the diffusion process, the model sometimes followed the sketch and sometimes the text. We believe that the well-trained BERT mask would randomly mask the conflict information in each modality.

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[2] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 1, 4

[3] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generat-

Table 1. Instance Score Comparison before and after capsule attention module, which are denoted as Instance Score on Sketch (ISS%) and Instance Score on Capsule (ISC%).

Index	1,1	1,2	1,3	2,1	2,2	2,3	3,1	3,2	3,3	4,1	4,2	4,3	Average
ISS ISC	1.38 66.59	1.01 60.80	1.78 71.33	1.45 68.57	1.93 59.85	0.97 62.07	0.53 49.06	1.34 66.74	1.29 44.39	0.81 59.04	1.01 68.87	1.11 52.75	
Index	5,1	5,2	5,3	6,1	6,2	6,3	7,1	7,2	7,3	8,1	8,2	8,3	
ISS ISC	0.66 60.69	1.07 57.45	0.87 65.76	1.16 62.84	1.52 55.19	0.67 65.52	0.73 40.03	0.77 57.59	0.73 44.51	0.88 57.50	1.05 52.76	0.97 58.39	1.07 58.68

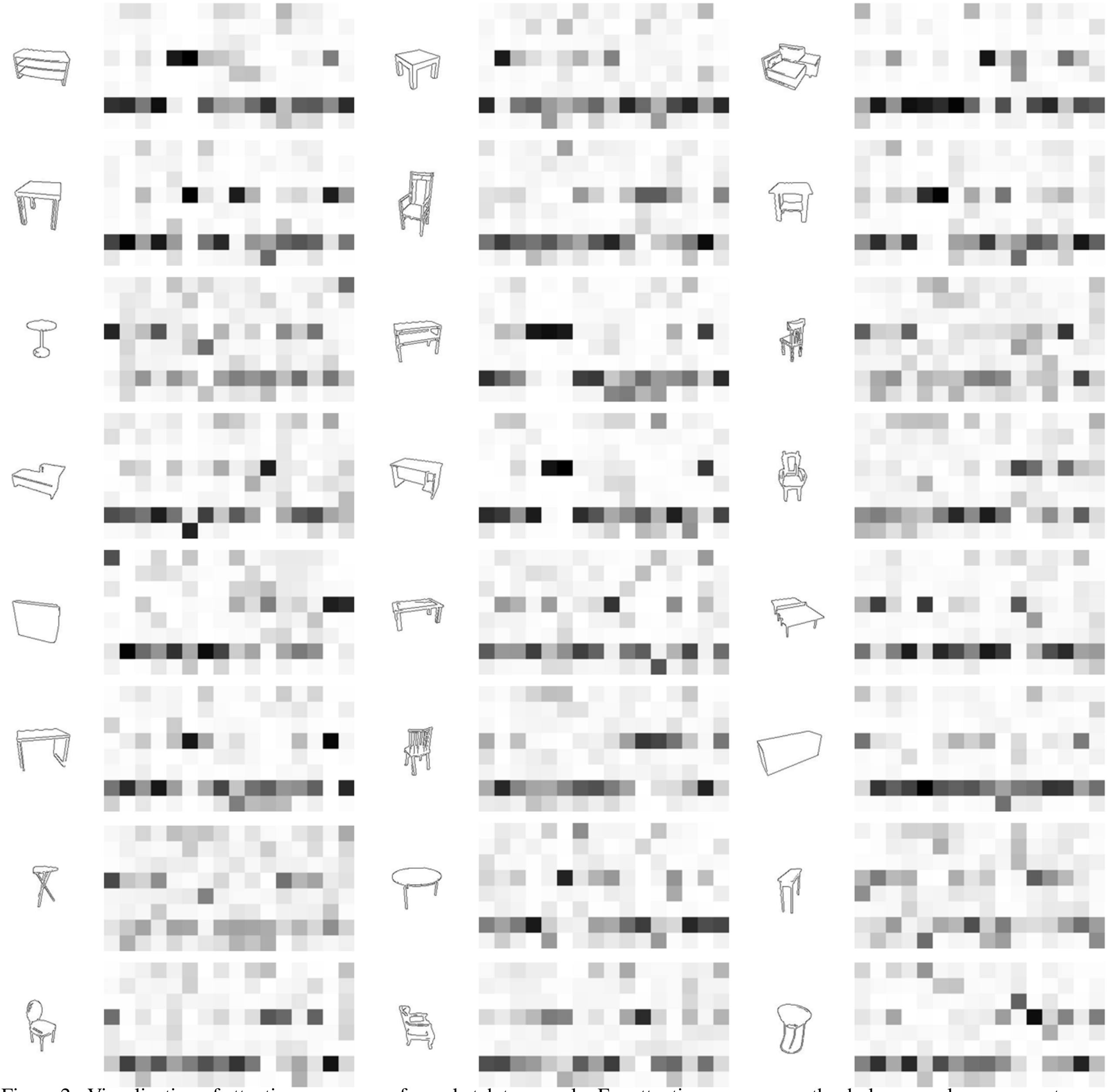

Figure 2. Visualization of attention score maps from sketch to capsule. For attention score maps, the darker capsule components pay more attention with our attention-routing module in the final feature extraction. The lighter parts denote less attention. We report the Instance score in Table 2 to show the improvement of the attention.

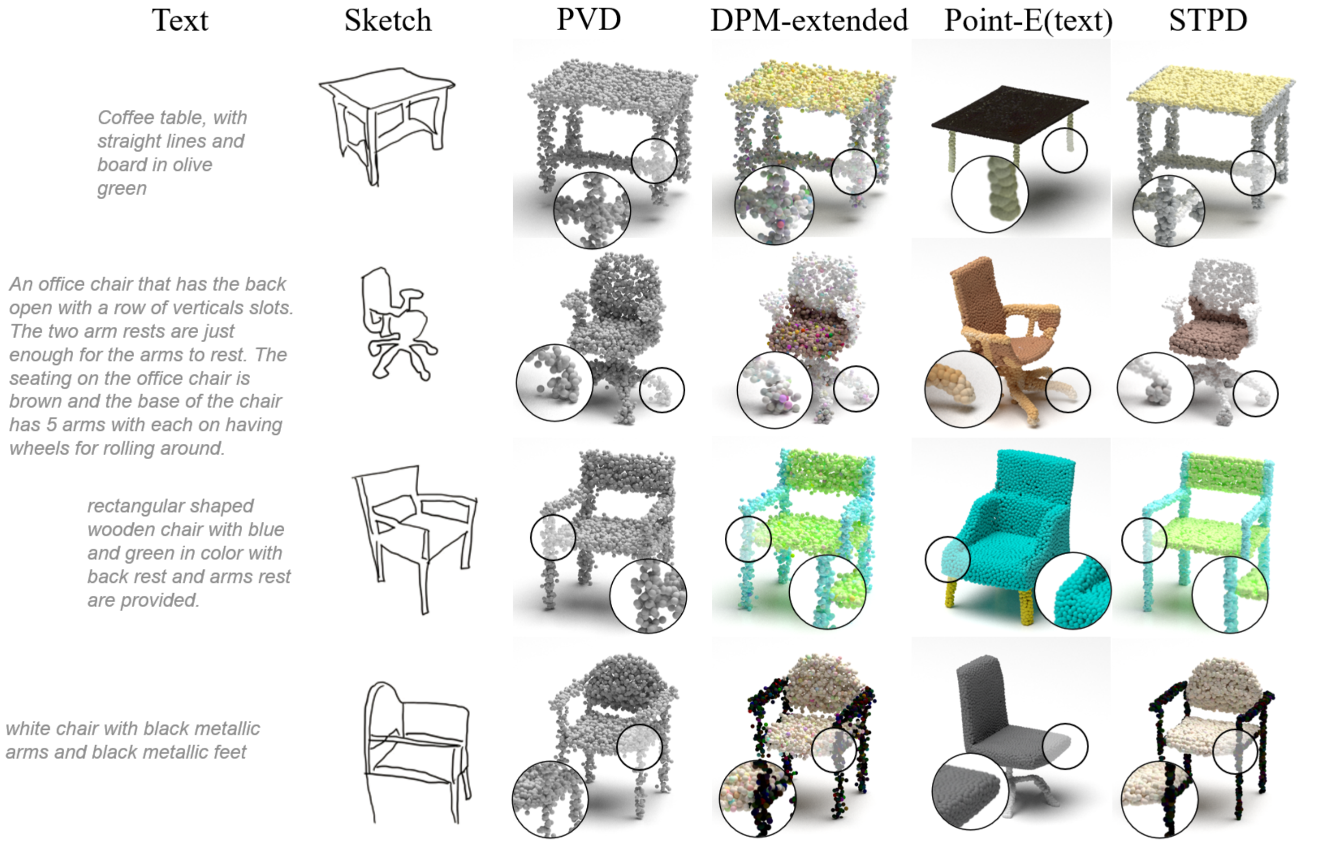
Text
Sketch
PVD
DPM-extended
Point-E(text)
STPD
Coffee table, with straight lines and board in olive green
An office chair that has the back open with a row of verticals slots. The two arm rests are just enough for the arms to rest. The seating on the office chair is brown and the base of the chair has 5 arms with each on having wheels for rolling around.
rectangular shaped wooden chair with blue and green in color with back rest and arms rest are provided.
white chair with black metallic arms and black metallic feet

Figure 3. Visualization comparison Between PVD [4], DPM [2], Point-E [3], and our STPD. We enlarge the local details to demonstrate that our STPD outperforms other baselines in terms of geometry and color generation. Our method produces faithful results from the input conditions.

ing 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 1, 4

[4] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021. 1, 4

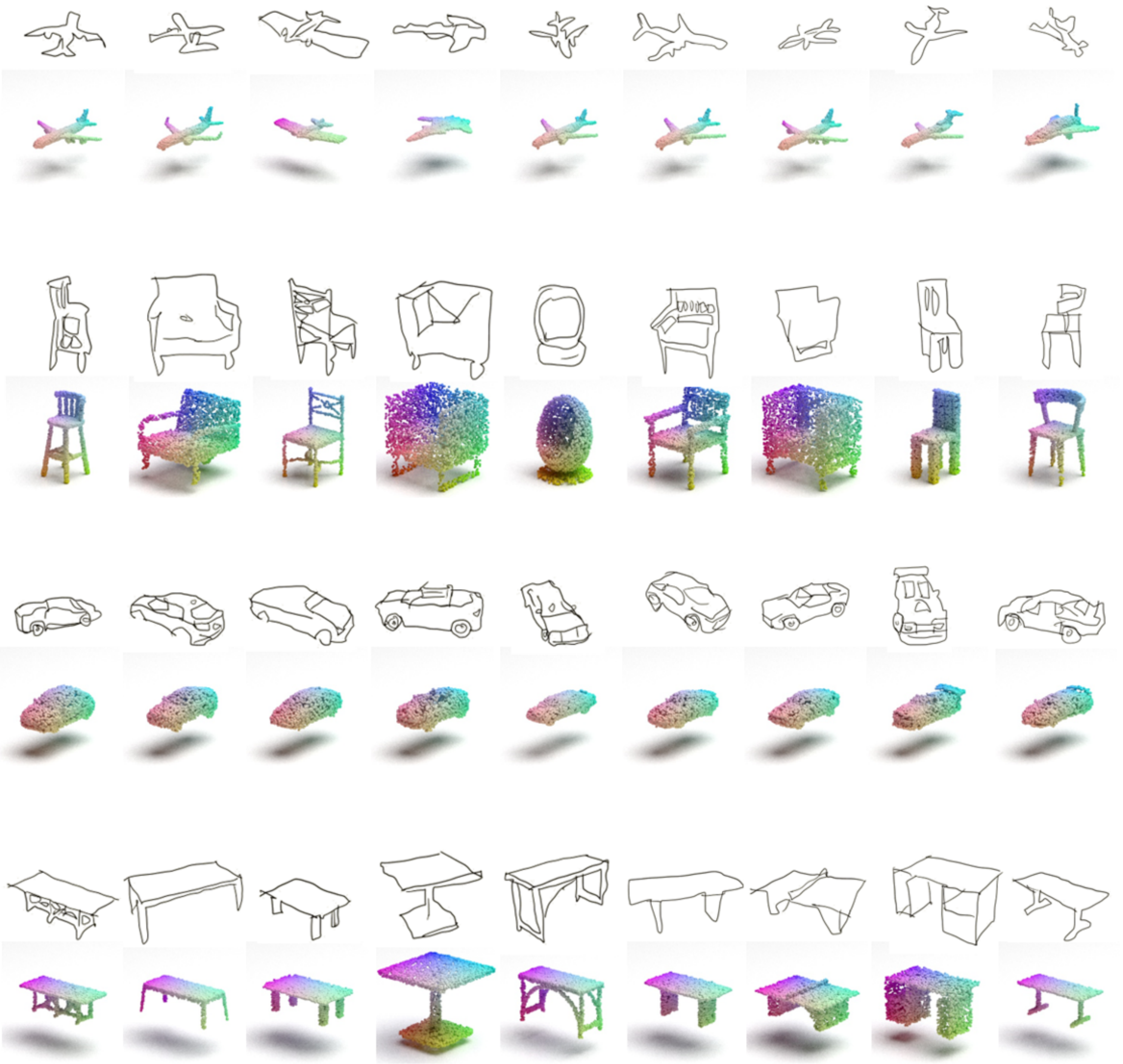

Figure 4. Visualization of point cloud generations for *airplane*, *chair*, *car*, *table* category from free-hand sketches. Our model achieves good generalization ability for generating different shapes that are faithful to the condition provided by the respective sketch.

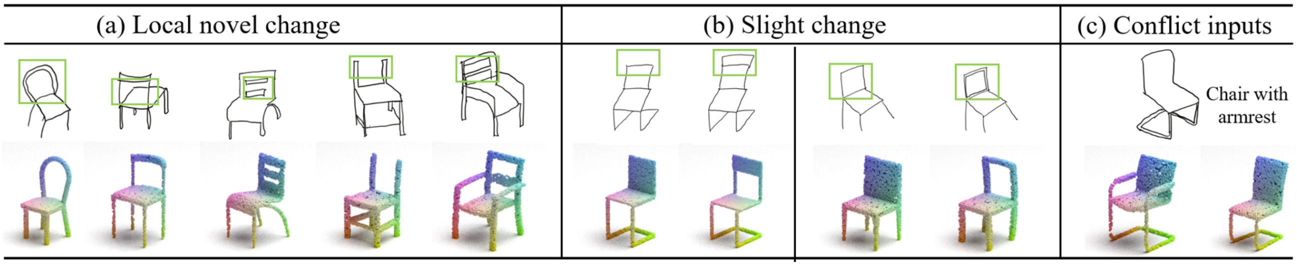

Figure 5. Feature discussions. (a) generating novel shapes, and (b) effects of slight changes in the sketch, and (c) conflicting sketch & text. The main change part of sketch is highlighted with a green box.