Supplementary Material

A. Visual Encoder Details

In our experiment, we use ViT/32, ViT/16, and RN.$\times$16 as three of the image encoders. ViT/32 is short for ViT-B/32, which represents vision transformer with 32×32 patch embeddings. ViT/16 has the same meaning. RN.×16 denotes ResNet-50 which requires 16 times more computations than the standard model.

B. More Information on the DiffCLIP Pipeline

Densify on ScanObjectNN dataset. In ModelNet datasets, only coordinates of a limited set of key points and normal vectors of faces are provided. To increase the density of our data representation, we perform densification on 2D depth maps after point sampling and projection. In contrast, the ScanobjectNN dataset provides coordinates for all points, so we use a different densification method. Specifically, we calculate the k-nearest neighbors ($k = 4$) for each point and construct triangular planes by connecting the point to all possible pairs of its neighboring points.

C. Detailed Clarification of Equations 4, 5, 6, 7

For equation(4), p_{loc} returns the diagonal entries of the matrix that represents the probabilities of the realistic images generated by text guidance j being classified into category j. p_{glo} is calculated by summing all values in the matrix that are no more than the diagonal by column, the reason for this is that the values in j_{th} column could be considered as the characteristic score of all generated with category j and the value in j_{th} row supposed to be the largest, so we ignore the larger values.

For equation (5), '*' represents the Hadamard product. we use the product of each element in p_{loc} and p_{glo} as the final probability. The function $norm(\cdot)$ could scale the values between 0 and 1.

For equation (6), p_{loc} regarded the maximum value as the local feature by each column which represent the diffusion result that is most similar to that category itself. The j_{th} element in p_{glo} is calculated by $exp\frac{1}{K}\sum_{i=1}^{K} logP_{ij}$, we considered all values in each column for the same reason as above. The purpose of fetching the log is to convert the operation of multiplication to addition, then take exp to rescale the values.

For equation (7), we directly concatenate the prefix features c_b and $[CLS_y]$ instead of the prompt embeddings designed by us. $c_b(x)$ is computed by the parameters obtained from Pre-trained Point Transformer with an original 3D point cloud input x, the process is shown in Figure3 (left).

D. Experimental Details

Construction of Pretraining Dataset In Section 3.2.3, we mentioned that we use our customized dataset, ShapeNet37, which consists of sampled data from ShapeNet-Core, to pretrain point transformer. Specifically, in order to thoroughly test the generalization ability of the DiffCLIP model, we removed all data from categories that overlapped with those in ModelNet10 and ModelNet40 datasets from the ShapeNet dataset. The remaining data belonged to 37 categories, which we named ShapeNet37.

E. Calculating the Logits of Style Transfer

To better illustrate the result of style transfer through stable diffusion and logits' calculation, we draw the bar chart (Fig. 1) of detailed logits of an example.

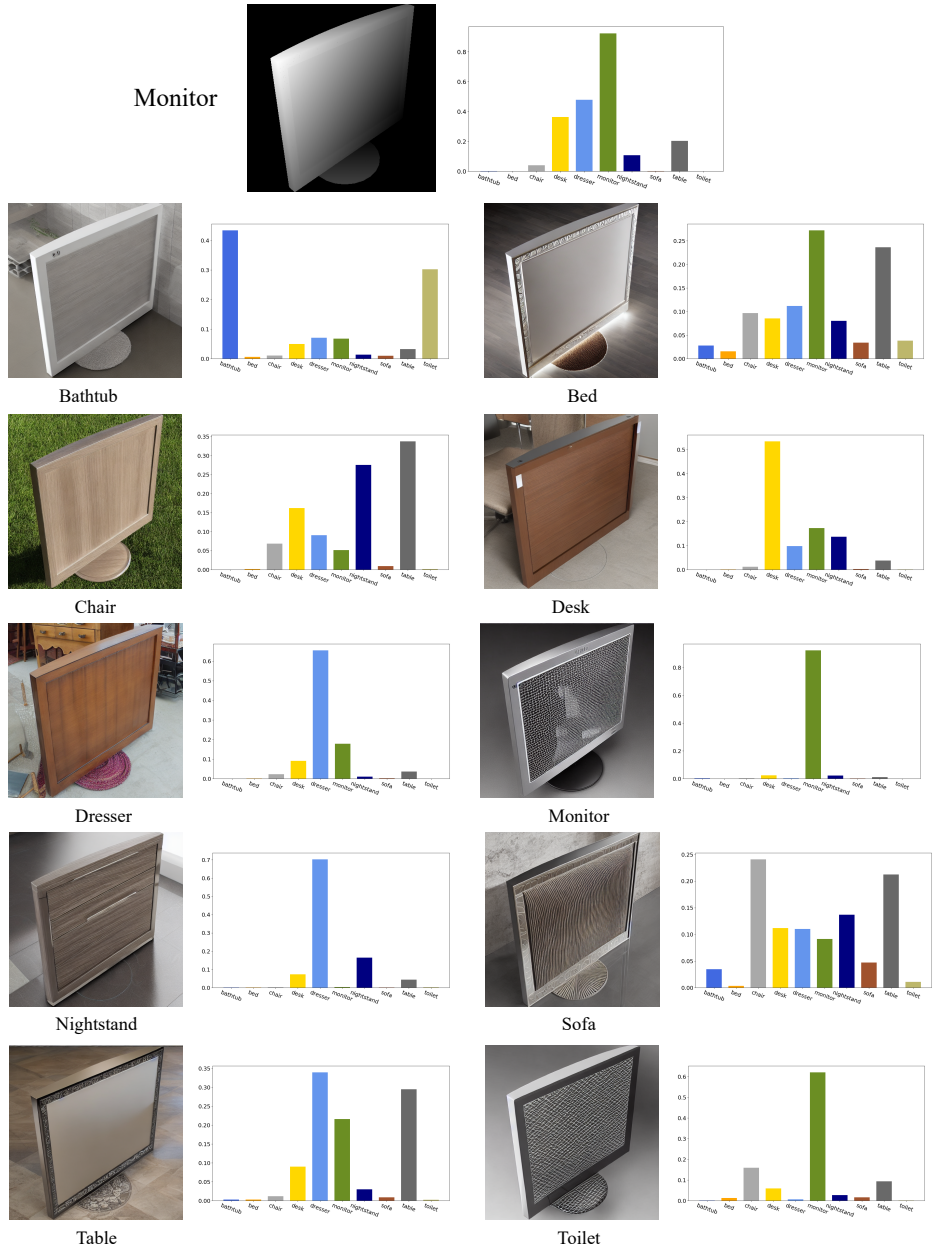

Figure 1. An example of style transfer result. Logits of ten images through stable diffusion's style transfer and the following calculation from source depth condition, the 'Monitor', are shown.