Sculpt3D: Multi-View Consistent Text-to-3D Generation with Sparse 3D Prior: Supplementary Material

A. Supplementary Materials

We have prepared supplementary materials, including a document and a video, to provide a more comprehensive understanding of our Sculpt3D. In the document, we discuss the technical details of our implementation in Sec. B. Moreover, we present additional examples and comparisons in Sec. C to demonstrate the performance of our method. Furthermore, we have prepared a video that showcases the results and comparisons of Sculpt3D.

B. Technical Details

Implementation details. This section provides more implementation details of our experiments.

- In our experiments, we observe that most objects in Objectverse [10] are aligned with the observer's frontal view. Thus we only normalized the vertices and centers of the templates without manually adjusting their poses. In addition to the 100 prompts provided by T3bench, we also use ChatGPT to generate 40 additional prompts, including common objects as well as some unusual and special items. The prompts we used with ChatGPT are as follows. *I am utilizing a text-to-3D model to generate various 3D objects. Please create 40 prompts for me, including both everyday common objects and some unusual and special items.*
- When performing appearance modulation, three adapters are utilized to correct erroneous patterns without altering the style of the generated objects. The adapters we used are a spatial color palette adapter, a structure adapter, and an image adapter. Specifically, since T2I-Adapter [27] supports combining multiple adapters to utilize complementary ability between different adapters, we use sketches extracted from the template objects as structure pattern conditions and the hue and color distribution of the generated object as spatial color conditions. Both the sketch and spatial color palette extraction model are the default models from the T2I-Adapter. This approach effectively retains the pattern information of the template while transferring it to the color distribution of the generated object. When the color distribution of the template is transferred, it is used as the image condition to modulate the diffusion process using the equation 6.

C. More Results

Loss balancing ablation In addition to the ablation studies of the shape learning provided in Sec. 4.5, we conduct another ablation study to discuss the effectiveness of sparse ray sampling, which is described in Sec. 3.2.1. We first remove sparse ray sampling and keep the value of λ in equation 3 as 0.1 to evaluate the effectiveness of sparse ray sampling.

As shown in Figure C.1, the results show that removing sparse ray sampling causes the generated objects to closely resemble the template, due to the geometric constraints being uniformly applied to all points. For example, the folds in the hat closely match those in the template, and the back cover of the water gun doesn't close. As shown in the third column of Figure C.1, by implementing sparse ray sampling our method can generate imaginative and reasonable geometry under the guidance of the reference shape.

Figure C.1. Ablation on the sparse ray sampling strategy.

For the choice of λ in equation 3, we study the effect of it by applying sparse ray sampling with λ values of 1, 0.1, and 0.01. The results are shown in Figure C.2. It's evident that even at $\lambda = 1$, our sparse sampling approach is able to provide sufficient flexibility for the model to learn new shapes. Compared to the results with $\lambda = 1$, setting λ as 0.1 can further increase the geometry freedom in the generated results. For instance, the shape of the straw hat is obviously changed. When set λ as 0.01, the model can create significantly new geometries, but it may produce undesirable outcomes. Therefore, we default λ as 0.1 in our experiments.

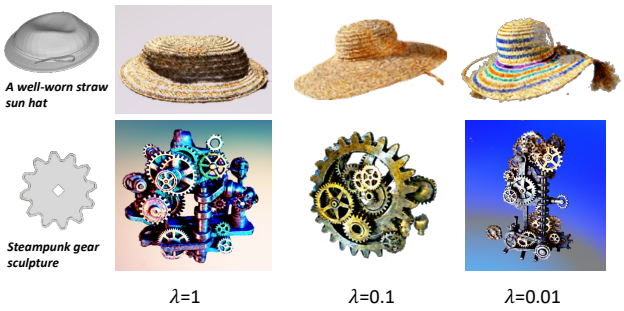

Figure C.2. Ablation on the shape co-supervision value λ in equation 3.

Details of comparison with baselines. To further validate the effectiveness of the sparse prior scheme in sculpt3d, we conduct two additional experiments. We first study the effectiveness of the sketch shape loss proposed by Latent-NeRF [25]. They propose it to allow the model to make slight changes in the template's surface, the loss is formulated as:

$$L_{\text{Sketch-Shape}} = CE(\alpha_{\text{NeRF}}(p), \alpha_{\text{GT}}(p)) \cdot (1 - e^{-\frac{d^2}{2\sigma_S}}), \quad (7)$$

where α_{NeRF} and α_{GT} are the NeRF occupancy and template shape's occupancy, respectively. The loss is applied to all points, d represents the distance of a point p from the surface, and σ_S is a hyperparameter that controls how lenient the loss is. A higher σ_S value means a more relaxed constraint to the surface of the generated object. Since their method operates with the SDS loss at a low resolution of 64 rendering, for a more comprehensive comparison, we use their code to conduct experiments in their 64 setting and combine it with the VSD loss to train at a higher resolution of 512 rendering. To fully utilize the new geometry generation capability of their method, we employ the maximum value of σ_S, 1.2, as used by them in all experiments.

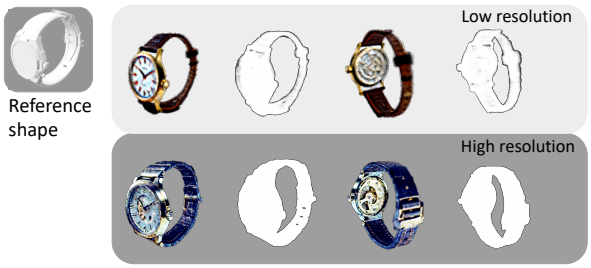

Figure C.3. Ablation of the strategy proposed by [25] in both low and high resolution rendering.

The experimental results are illustrated in Figure C.3, where we showcase the generated outcomes at two different resolutions along with their corresponding occupancy. It is observable that at lower resolutions, their method is able to alter surfaces, like thinning watch straps. However, at higher resolutions, their approach struggles to change the object's shape, which results in the generated geometries closely resembling the templates. Additionally, it is noted that their method of relaxing surface constraints often leaves residual artifacts on the surface. This is evident in the occupancy results of the watch straps in the first row, stemming from an incomplete removal of surface density.

Comparison with mesh initiation. In the main text, we mentioned that directly using the template's shape to initialize NeRF's density can not guarantee a satisfactory shape. Unlike our approach, Fantasia3D uses a mesh-based DMTet as a 3D representation, thus supporting initialization with an initial mesh. To more comprehensively verify the role of geometric initialization, we also used our template to initialize Fantasia3D. As shown in Figure C.4, the results show that simple initialization is hard to ensure the subsequent learning direction of the model. Despite the model being initialed by a reasonable shape, it still produces unsatisfactory outcomes, such as the distorted shapes of birds and books. This further underscores the necessity of employing geometry and 2D co-supervision during the learning process.

Figure C.4. Ablation on the mesh initiation strategy.

More comparisons. Here, we showcase more multi-view examples generated by our method in Figure C.5.

Furthermore, we also provide more comparisons with baselines in Figure C.6 and Figure C.7. To compare with the best results demonstrated by the baseline methods, we follow the previous works [5, 19, 41] to directly copy the figures from the corresponding papers for comparisons.

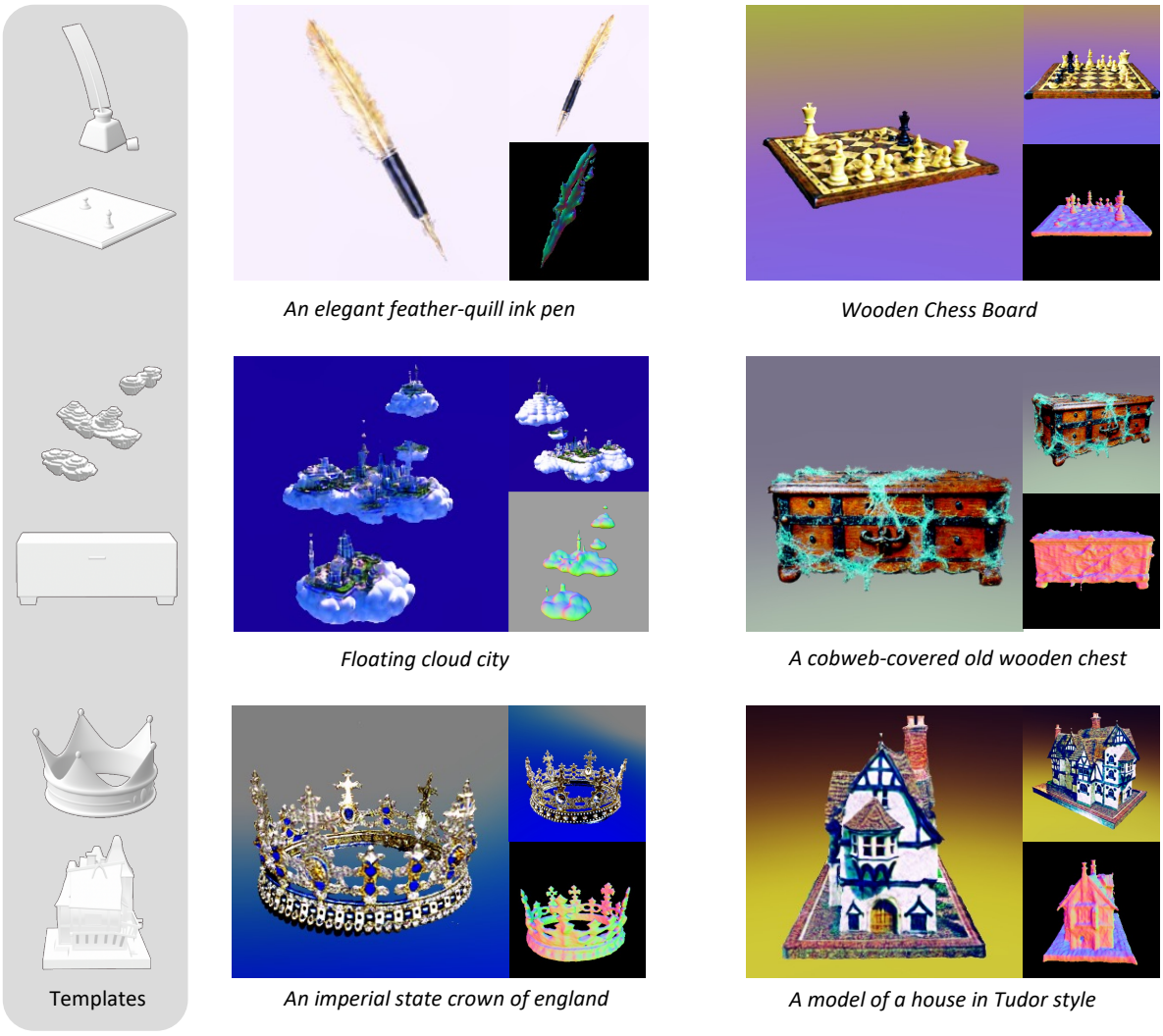

Figure C.5. More multi-view examples generated by our method, the retrieved templates are shown on the left.

A small saguaro cactus planted in a clay pot.

Ours

ProlificDreamer

Magic3D

Dreamfusion

A car made out of sushi.

Ours

ProlificDreamer

Fantasia3D

Magic3D

Figure C.6. Additional examples for qualitative comparison with baselines.

A delicious croissant.

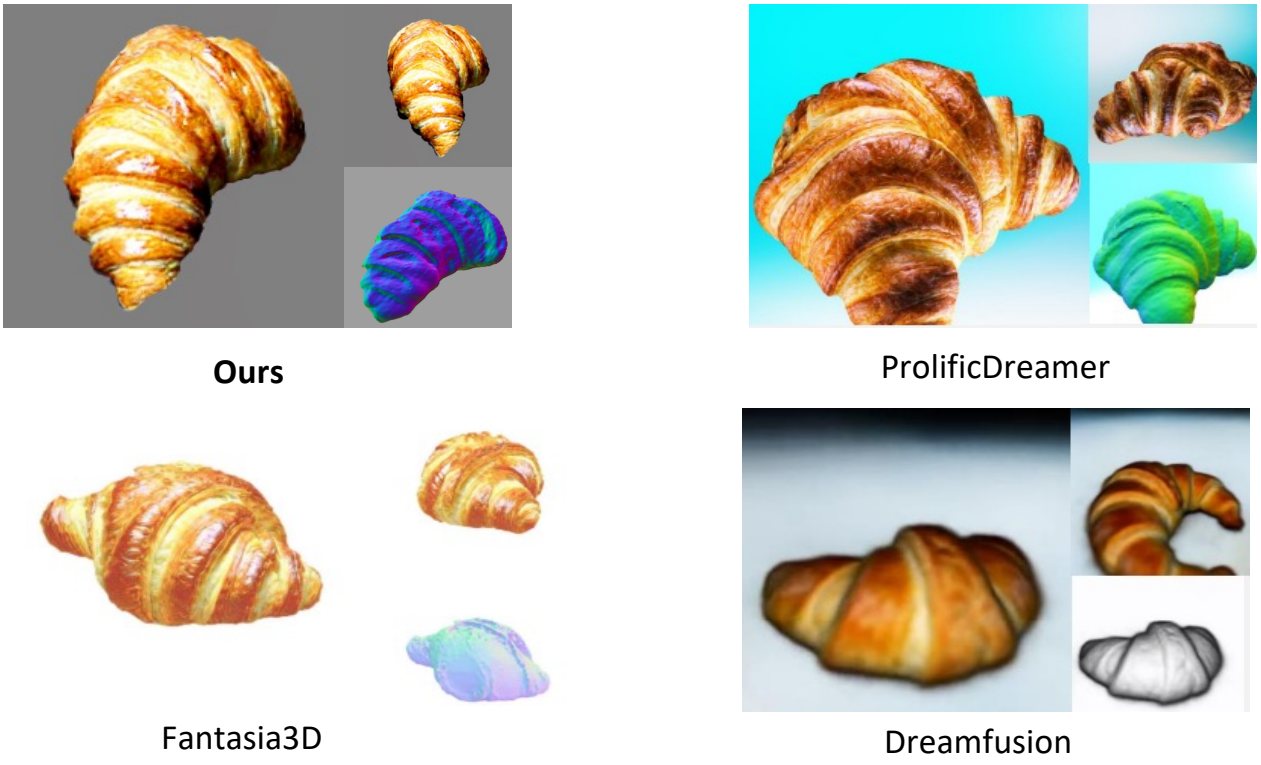

Figure C.7. Additional examples for qualitative comparison with baselines.