LDM: Large Tensorial SDF Model for Textured Mesh Generation Supplementary Materials

1. Detail for Network architecture

We use the DINOv2-ViT-B/14 [1] as our image encoder, a transformer-based model, which has 12 layers and the hidden dimension of the transformer is 768. Regarding camera features, we flatten the extrinsic parameters of each view into 16 dimensions and encode them into a 1024-dimensional vector using a 2-layer MLP. After encoding, each image yields 257 image feature tokens, including the [CLS] token. All tokens from these views are concatenated together to obtain a total of N×257 tokens, where N is the number of input images. These image feature tokens serve as condition features in the subsequent generation process.

Next, we use a transformer-based tensorial object reconstructor to predict a tensorial SDF representation from a sequence of learnable tokens with a size of (3×32×32+3×32)×1024, where 3×32×32 and 3×32 are the number token numbers align with V_k^m and $M_k^{\tilde{m}}$ represent the tensor matrix factors and vector of their corresponding spatial axes. And 1024 is the hidden dimension of the transformer decoder. After being decoded by the transformer, conducted through cross attention from image features, we obtain the same number of tensor tokens. Next, following LRM[2], we use a de-convolution layer and an MLP layer to upscale tensor matrix factors from 3×(32×32)×1024 to 3×(64×64)×40, and tensor vector from 3×32×1024 to 3×64×40.

Finally, these tensor tokens are reshaped into a tensorial SDF representation, from which we can compute the feature vector of any point with a dimension of 120. Then, for the first stage of volume rendering training, we use a 4-layer MLP with 64 hidden dimensions to decode a 1-dim SDF value from this feature. We use another 4-layer MLP with 64 hidden dimensions to decode a 6-dim color value from this feature, where the first three dimensions are considered the albedo color and the last three dimensions are considered the shading color. Finally, for the second stage of Flexcubes layer [3] training, we use a 2-layer MLP with 64 hidden dimensions to decode an 8-dim weights value from this feature. We use another 2-layer MLP with 64 hidden dimensions to decode a 1-dim deformation value from this feature.

2. Failure Cases and Limitations

As discussed in the main text, our framework can produce convincing results from a single image or text prompt input, but there are still certain instances where it fails, as shown in Figure 2. In the left side of Figure 2, we input an image of a potted plant to predict its 3D assets. The overall result meets expectations, with the plant's leaves and base being well generated, but the thin stems fail to generate properly. There are multiple reasons for this issue. On the one hand, the size of the tensorial SDF tokens produced by our model is capped at 64x64, which may result in the loss of fine geometric details. On the other hand, the multi-view images generated by the multi-view diffusion may be inconsistent, causing thin stems to misalign across different views, leading to generation failures. On the right side of Figure 2, we can see that due to misalignment in the multi-view images, the text on the predicted fire extinguisher appears blurred. The misaligned highlights also caused errors in the material prediction, with the highlights being incorrectly predicted as part of the albedo. Additionally, the prediction of the shading color for the metallic and lacquer materials on the fire extinguisher's surface is inaccurate. While part of the red metallic surface is correctly predicted in the shading color, some of the black text is mistakenly predicted as part of the shading color as well.

3. More results for 3D generation task

As shown in Figure 3, we present more 3D results generated from text prompts or single image inputs. Our method is capable of producing good results for both real-world images and unreal images.

References

[1] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[2] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 1

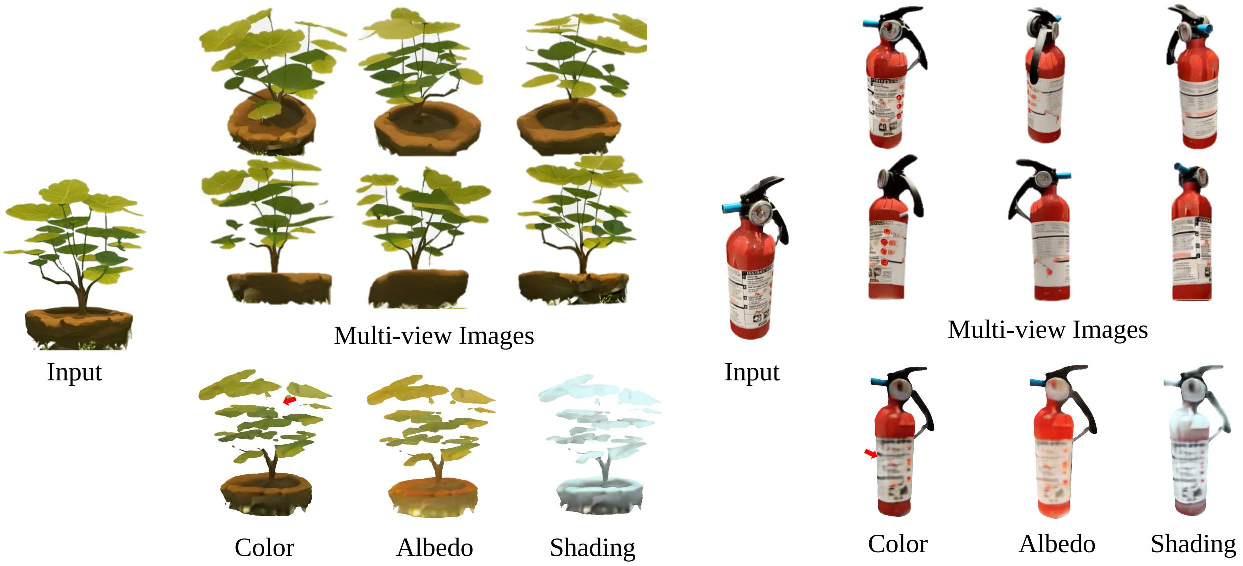

Figure 1. Some failure cases in the predictions reveal the limitations of our method.

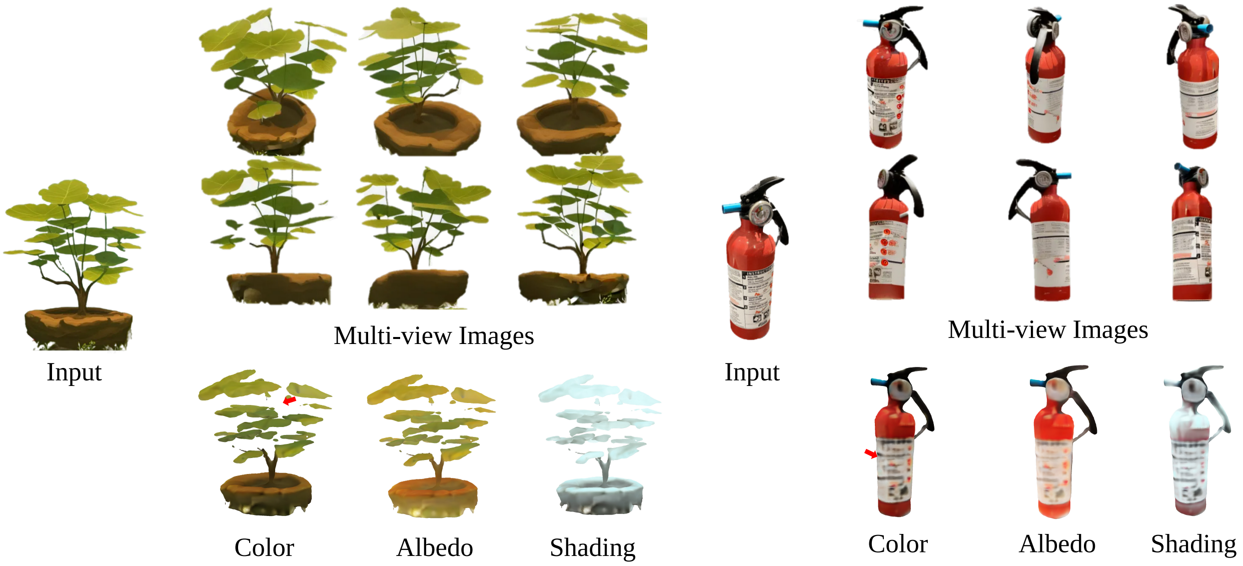

Figure 2. Some failure cases in the predictions reveal the limitations of our method.

[3] T. Shen, J. Munkberg, J. Hasselgren, K. Yin, Z. Wang, W. Chen, Z. Gojcic, S. Fidler, N. Sharp, and J. Gao. Flexible isosurface extraction for gradient-based mesh optimization. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023. 1

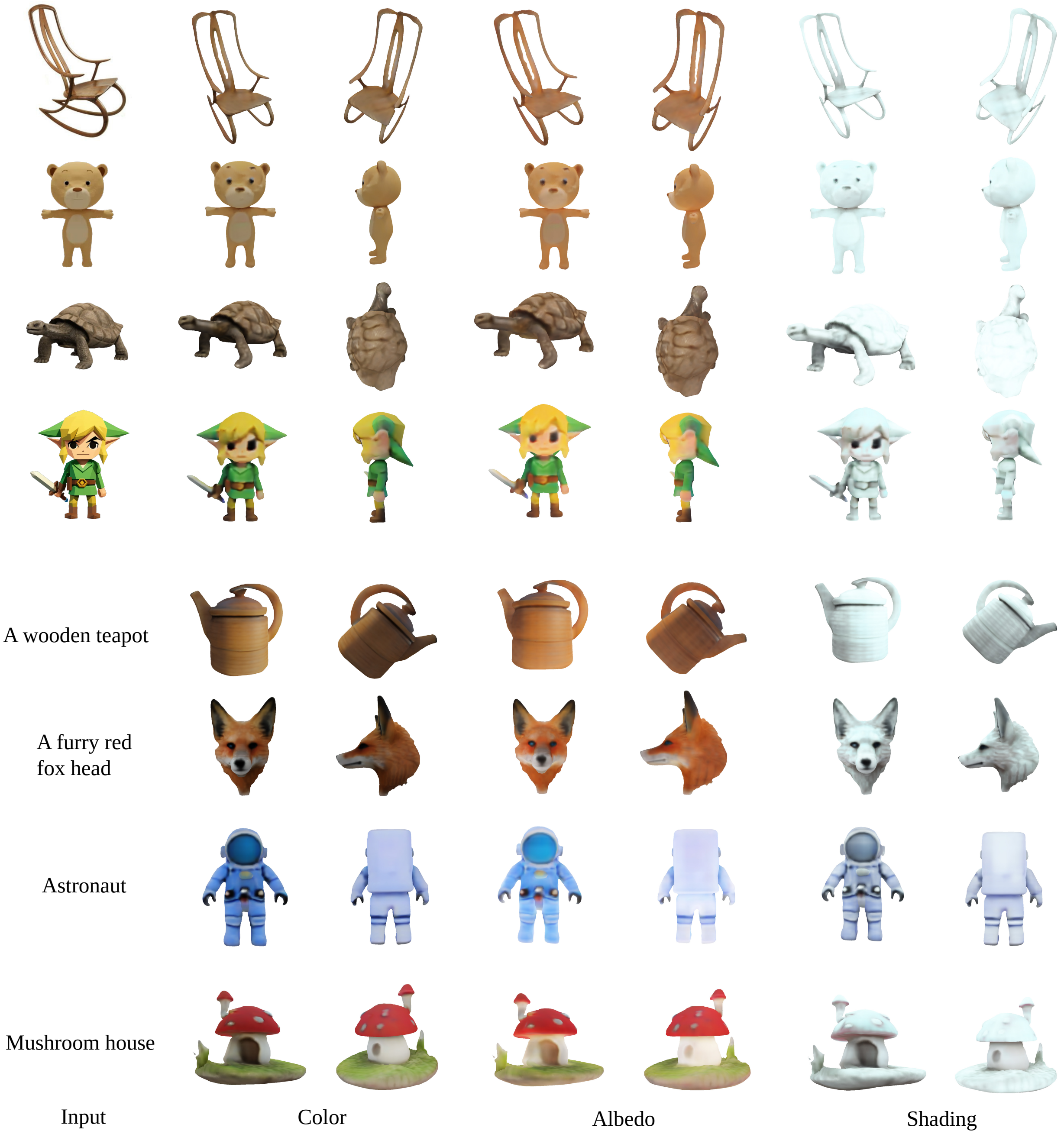

Figure 3. More predicted results for text-to-3D generation task.