ESCT3D: EFFICIENT AND SELECTIVELY CONTROLLABLE TEXT-DRIVEN 3D CONTENT GENERATION WITH GAUSSIAN SPLATTING

Author(s) Name(s)

Author Affiliation(s)

A comparison between our method, Shap-E, and DreamGaussian is presented in Fig 1. Shap-E, DreamGaussian, and our method all generate 3D content in a short amount of time. Upon observation, it is evident that Shap-E can only generate geometrically simple content, with a lack of detail in the generated content. Our method, with a time cost similar to DreamGaussian, overcomes issues such as Janus and inconsistent viewpoints present in DreamGaussian. In contrast, our approach is capable of generating more complex and richer 3D content. Overall, our method demonstrates superior performance in both generation quality and speed compared to all the other methods. It strikes a better balance between generation quality and efficiency.

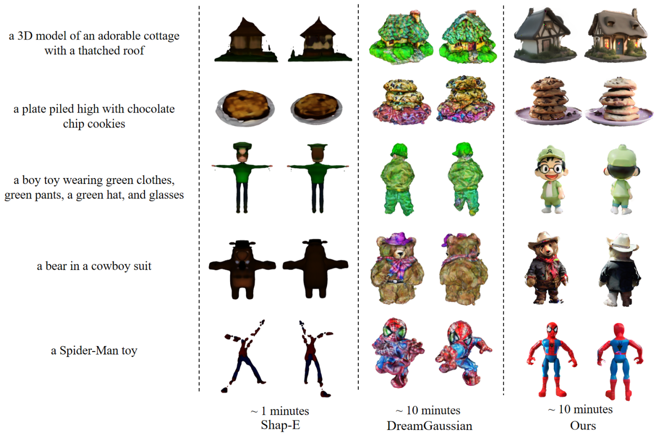
a 3D model of an adorable cottage with a thatched roof
a plate piled high with chocolate chip cookies
a boy toy wearing green clothes, green pants, a green hat, and glasses
a bear in a cowboy suit
a Spider-Man toy
∼ 1 minutes
Shap-E
∼ 10 minutes
DreamGaussian
∼ 10 minutes
Ours

Fig. 1. Comparison with Shap-E and DreamGaussian.

Fig 2 shows that when we input the same text and canny condition, the second row displays our results, where the visual effect of our multi-view prediction is significantly better than the first row (MVControl). Due to MVControl's poor performance in multi-view prediction, it directly affects the quality of its 3D generation, leading to irregular geometries. In contrast, our method ensures multi-view consistency, so the final generated content has more uniform and regular geometry.

We also demonstrate the capability of our method in artistic creation in Fig 3. Fig 3 demonstrates that our framework can effectively understand the content of images, such as the character's appearance and physical features, can also produce coherent, high-quality 3D content. Our method has a strong ability to understand both text and images, enabling the creation of rich and diverse 3D content. By providing only a very brief text and the condition image to be input, controllable content that meets the desired expectations can be generated.

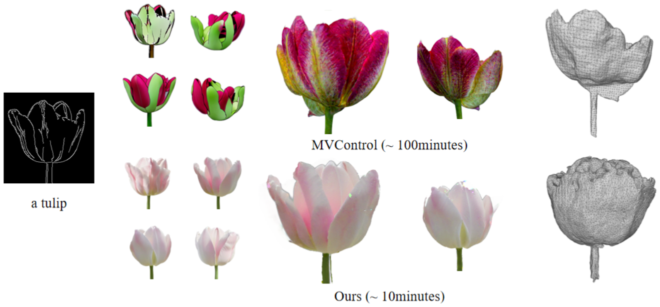

Fig. 2. Ours vs. MVControl: Text and Canny-Based.

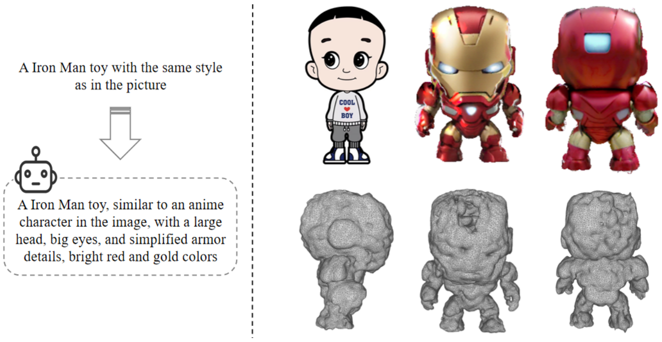
A Iron Man toy with the same style as in the picture
A Iron Man toy, similar to an anime character in the image, with a large head, big eyes, and simplified armor details, bright red and gold colors

Fig. 3. Understanding-Based Art Creation.

We further evaluate the importance of the self-optimization process. Since 2D images serve as the foundation for generating high-quality 3D content, they set the upper limit for the visual quality of 3D generation. As shown in the Fig 5, the right side displays the results generated using simple user prompts, which often result in low-quality outputs with incomplete content and various issues such as Janus problems. However, with the assistance of GPT-4V, our

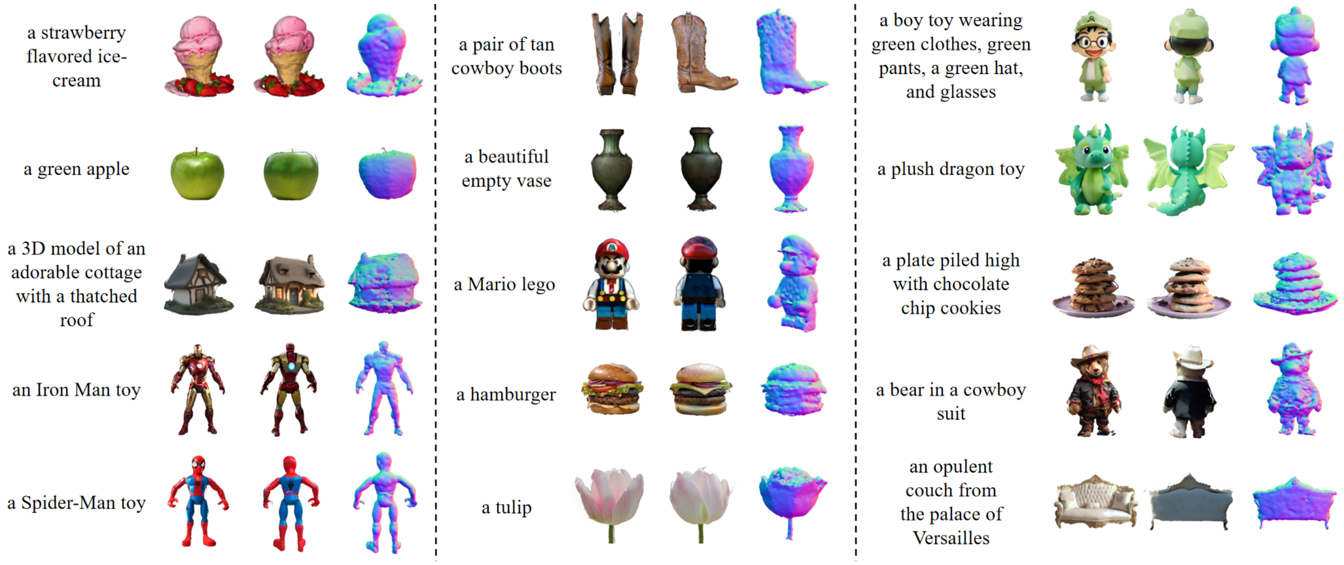
a strawberry flavored ice-cream
a green apple
a 3D model of an adorable cottage with a thatched roof
an Iron Man toy
a Spider-Man toy
a pair of tan cowboy boots
a beautiful empty vase
a Mario lego
a hamburger
a tulip
a boy toy wearing green clothes, green pants, a green hat, and glasses
a plush dragon toy
a plate piled high with chocolate chip cookies
a bear in a cowboy suit
an opulent couch from the palace of Versailles

Fig. 4. The Generation Results of ESCT3D.

self-optimization process benefits from timely revisions and quality evaluations, leading to the image on the left. We demonstrate that the self-optimization process, by improving text prompts, significantly enhances image generation quality, thereby impacting the overall quality of the generated 3D content. This capability ultimately helps in selecting more realistic and detailed images from the candidate pool, providing a solid foundation for further 3D content generation.

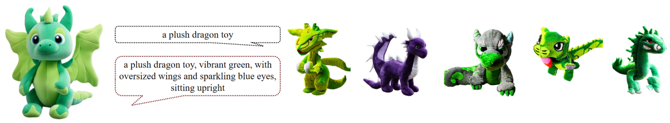
a plush dragon toy
a plush dragon toy, vibrant green, with oversized wings and sparkling blue eyes, sitting upright

Fig. 5. Ablation of Self-Optimization. It is difficult to generate high-quality results with simple text alone. However, by using the multi-modal iterative self-optimization framework, the efficiency of generating high-quality content can be significantly improved.

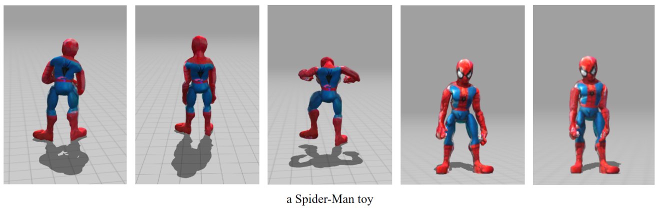

Fig. 6. Spider-Man Toy Animation Results.

Since our method can extract mesh from 3D Gaussians, these mesh can be seamlessly applied to downstream tasks, such as rigged animation. The Fig 6 below show some of our results with rigged animation.