Kiss3DGen: Repurposing Image Diffusion Models for 3D Asset Generation

Supplementary Material

The supplementary file includes a demo video showcasing the performance of our model in various tasks, including text/image-to-3D generation, and 3D enhancement/editing. Additionally, we provide further studies and detailed explanations below to offer a deeper understanding of the model and its capabilities. We will release the code upon acceptance.

1. Ablating the Initialization of Mesh

In our manuscript, we adopt the off-the-shelf LRM [2] model or a simple sphere shape to initialize the coarse mesh, then refine the mesh with ISOMER [4]. We have also experimented with different settings, such as refining the mesh from a simple, sphere-shaped initialization. As shown in Fig. 1, the results are still of excellent overall quality; however, there appear to be more geometrical errors at unseen surfaces. We also conducted quantitative evaluations, as shown in Table 1 and Table 2. The quantitative results demonstrate that the LRM initialization generally outperforms the sphere initialization across most metrics.

Table 1. Quantitative comparison of generated results for **text-to-3D** with different initializations at the reconstruction stage.

Method	CLIP↑	Quality↑	Aesthetic↑
Init-LRM	0.837	2.700	1.800
Init-Sphere	0.8012	2.559	1.566

Table 2. Quantitative comparison of generated results for **image-to-3D** with different initializations at the reconstruction stage.

Method	CD↓	FS↑	PSNR↑	SSIM↑	LPIPS↓
Init-LRM	0.149	0.769	20.348	0.902	0.116
Init-Sphere	0.173	0.719	20.122	0.902	0.117

2. Ablating the number of steps in ISOMER

In our main manuscript, we proposed using the off-the-shelf LRM [3, 5] model to initialize the coarse mesh, followed by ISOMER [4] to optimize and produce the final mesh. In the optimization step, there is a critical parameter that controls the number of geometry optimization steps. This parameter directly impacts the inference time. Specifically, when the number of steps is set to 50, the geometry optimization step takes approximately 5 seconds, while setting it to 100 increases the time to about 10 seconds. To understand the effect of this parameter, we conducted an ablation study, as shown in Fig. 2. The results indicate that increasing the number of steps leads to sharper and more refined geometry, albeit at the cost of longer computation time. It is worth

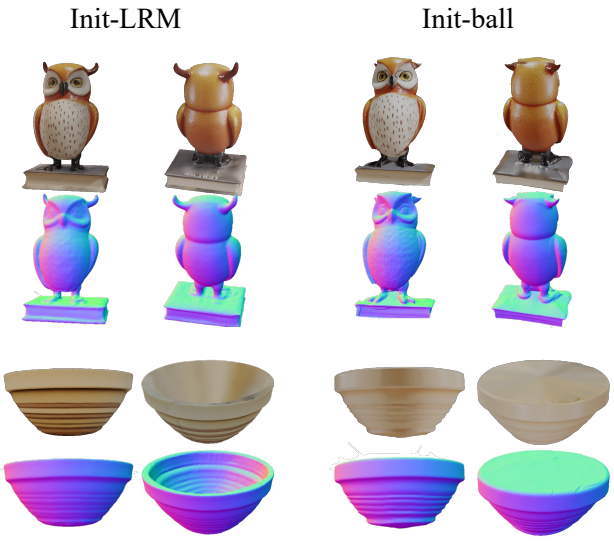


Figure 1. Qualitative comparison of 3D reconstruction results between different initializations in the reconstruction stage of our framework. The upper case (owl) shows that using LRM or sphere initialization yields similar results. The second row (bowl) shows that using sphere initialization may fail at capturing the concave geometric structure, while using LRM mitigates this problem.

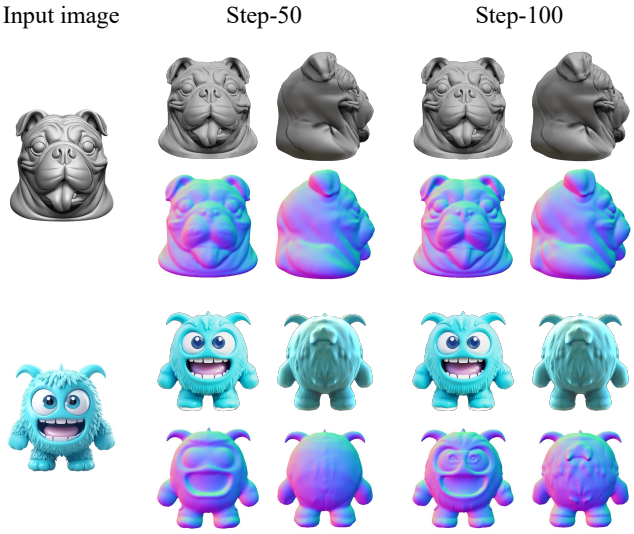


Figure 2. Qualitative comparison of 3D reconstruction results with different optimization steps with ISOMER [4]. As shown, optimizing with more steps leads to finer geometrical details.

Table 3. Comparison of inference time with other methods in different tasks. (in seconds). "–" means unapplicable.

Task	ours	MVEdit	Hunyuan3D-1.0	CraftsMan	Unique3D	3DTopia	Direct2.5
Text-to-3D	56.8	–	105.0	–	–	240.0	163.6
Image-to-3D	87.3	–	79.9	6.0	37.2	–	–
3D-to-3D	71.7	360.0	–	–	–	–	–

noting that, in our main manuscript, we used a step value of 50 for all experiments to balance experimental efficiency and result quality. This analysis highlights the trade-off between optimization time and geometry refinement, providing guidance for parameter selection based on application requirements.

3. Compatibility and extensibility of methods.

As shown in Fig. 4, our method is compatible with reconstruction techniques besides ISOMER, such as InstantNSR. Additionally, our approach retains DiT's full capabilities, enabling seamless integration with tools like IP-Adapter, ControlNet, or Flux Redux [1] (Fig. 5), highlighting its adaptability and extensibility.

4. System efficiency.

In Tab. 3, we quantitatively measure the inference time of our framework and baseline methods on an A800 GPU, our approach achieves the best performance within reasonable inference time.

5. More qualitative comparisons.

We demonstrate more comparisons against Wonder3D++ and Michelangelo for image-to-3D and LucidDreamer for text-to-3D in Fig. 3. Our method achieves better results in texture details, semantic alignment, and text-3D consistency.

6. User Study

In our manuscript, we conduct quantitative evaluations comparing our method with baseline methods, demonstrating its superior performance. We also present a user study to assess user preferences.

The user study was conducted on Amazon Mechanical Turk[2], involving 180 participants. To ensure quality, we included attention-check questions to filter out inattentive responses, resulting in 80 qualified participants whose responses were analyzed. Ultimately, we collected 2,000 valid responses covering key aspects such as geometry quality, texture quality, and overall quality. The results used in user study are generated with the default hyper-parameters without any cherry-picking.

Figure 6 shows a screenshot of the user study questionnaire. The options included GIFs displaying orbital views

[1] https://blackforestlabs.ai/flux-1-tools

[2] https://www.mturk.com

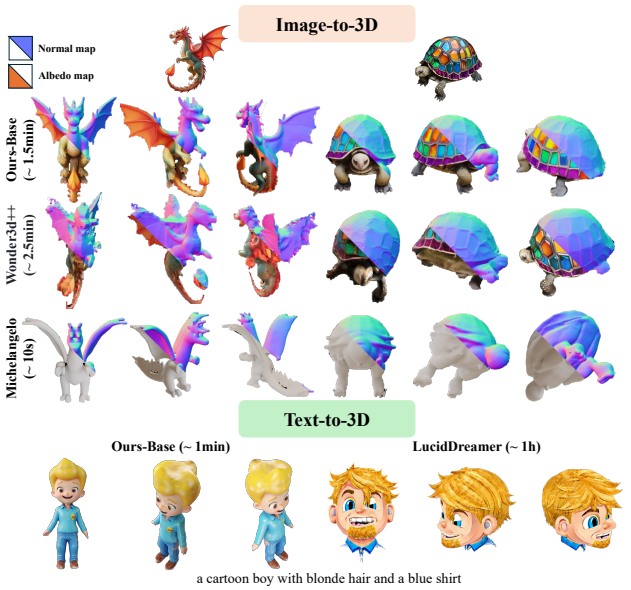


Figure 3. *Qualitative comparisons with more state-of-the-art methods for image-to-3D and text-to-3D generation.*

Figure 4. *Visual comparisons of different reconstruction methods.*

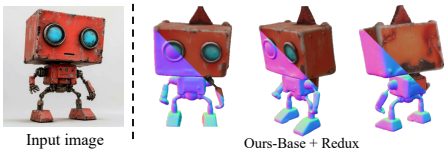


Figure 5. *Visualization of image-to-3d with redux.*

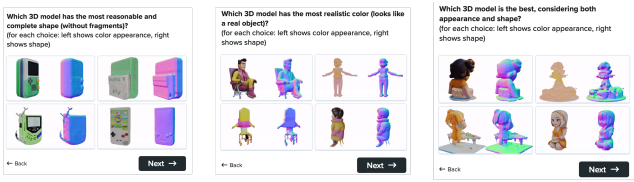


Figure 6. Screenshots of our user study questionnaire.

of the object in both color and normal space, allowing participants to better visualize the 3D structure and texture, thereby enhancing their ability to provide informed feedback.

For each case in the user study, we present a video to the users where the object rotates 360 degrees, with the left side displaying the RGB map and the right side showing the Normal map. Users are asked to select the best result based

on a series of questions. In terms of question design, we focus on several key aspects:

1. **Geometry**: "Which 3D model has the most reasonable and complete shape (without fragments)?"
2. **Texture**: "Which 3D model has the most realistic color (looks like a real object)?"
3. **Overall quality**: "Which 3D model is the best, considering both appearance and shape?"

The results of the user study are summarized in Tab. 4, where it can be observed that our method outperforms the baselines in terms of user preference for both geometry and texture quality, as well as overall impression.

Table 4. Study on user's preference on 3D generation results of ours and baseline methods.

Category	Method	Percentage
Texture	Ours	35.47%
	Hunyuan	32.37%
	Unique3D	13.13%
	3Dtopia	6.75%
	Direct2.5D	12.28%
Geometry	Ours	37.61%
	Hunyuan	36.24%
	Unique3D	10.45%
	3Dtopia	5.13%
	Direct2.5D	10.56%
Overall Quality	Ours	38.72%
	Hunyuan	32.18%
	Unique3D	15.04%
	3Dtopia	6.49%
	Direct2.5D	7.57%

7. Applications and visualization

In our main paper, we introduce various applications with our model, including text-to-3d, image-to-3d, 3D editing and enhancement. We demonstrate more results in Fig. 8, Fig 9 and Fig. 10. Also, we attach a video to this supplementary to present the 3D generation results in a dynamic approach.

7.1. Advanced image to 3D

In Fig. 7, we illustrate a 3D generation pipeline that utilizes multi-modal conditions, including both images and text. Unlike most existing image-to-3D generation methods that produce 3D assets aligned solely with the input image, our framework introduces textual control over the generation outcomes, significantly enhancing the utility of 3D content creation from images. This capability allows for more nuanced and tailored 3D outputs, catering to specific user requirements. And notably, the application of our model extends beyond the examples presented in this paper.

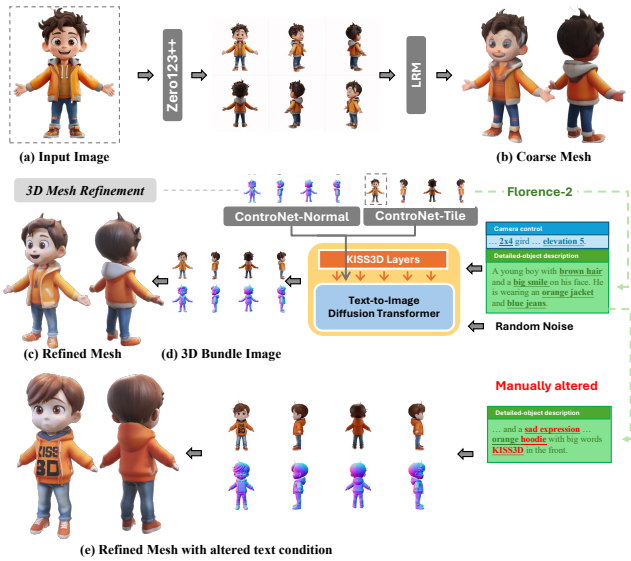


Figure 7. **Advanced image-to-3D pipeline with our framework.** In this case, we alter the text descriptions at the 3D mesh refinement stage and achieve accurate textual control on the refined result. Please zoom in for details.

8. Limitations

In this paper, we effectively adapt the pretrained 2D diffusion transformer model, specifically Flux [1], for the generation of 3D Bundle Images. To maximize the potential of the Flux model, we render our 3D dataset under varying environmental illuminations, enhancing its similarity to real-world images on which the Flux model was trained. As a result, the generated 3D Bundle Image retains lighting information, which was not disentangled from the model texture during the reconstruction phase of this work. We leave this for future study.

References

[1] BlackForestLabs. Flux.1 model family. 2024. 3

[2] Wenhang Ge, Jiantao Lin, Guibao Shen, Jiawei Feng, Tao Hu, Xinli Xu, and Ying-Cong Chen. Prm: Photometric stereo based large reconstruction model. *arXiv preprint arXiv:2412.07371*, 2024. 1

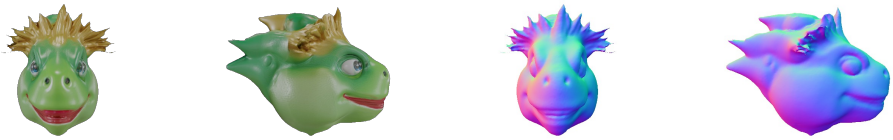
[3] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv*, 2024. 1

[4] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv*, 2024. 1

[5] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv*, 2024. 1

[6] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 7

a white sugar skull with colorful polka dots, flowers, eyes, and teeth. Cranial dome, hollow eye sockets, nasal aperture, dental arch



green lizard head with spikes symmetrical design, pronounced mane, detailed textures, elevated ridges, ornamental headpiece, sculptural form

A statue of a lion on a marble pedestal base, prominent wings, ornamental pedestal, sturdy base, beveled edges

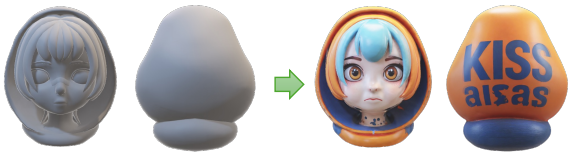
A charming owl with festive Christmas details, sitting on a simple branch. The owl wears a small, red Santa hat with fluffy white trim.

a small Chinese pagoda. elevated base, sweeping roof, overhanging eaves, multi-tiered roof, rectangular footprint

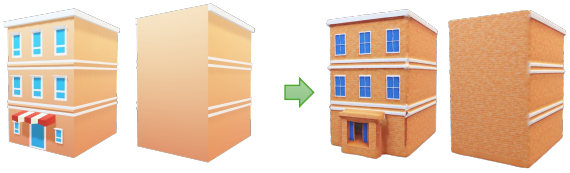
a Coca Cola monster can with arms, legs. cylindrical body, two bending arms, two bending legs, extruded circular eyes, short cylindrical snout, protruding ears.

Figure 8. More show cases of **Text-to-3D** generation with our model. Please zoom in for details.

Figure 9. More show cases of **Image-to-3D** generation with our model. Please zoom in for details.



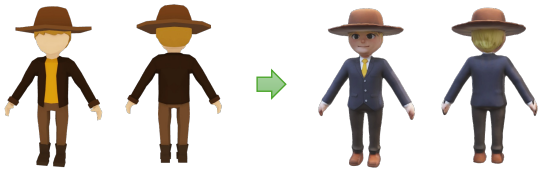
"... A girl with blue hair, she is wearing an orange hood with words KISS on the back."



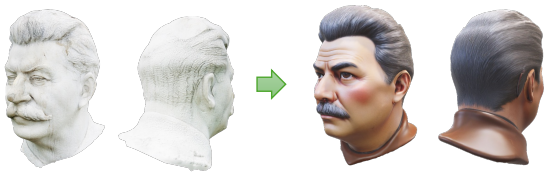
"... Orange building with white stripes, blue windows and pattern of bricks on the side."

"... A pink sedan."

"... A realistic photo of a Japanese samurai, he carries katana."

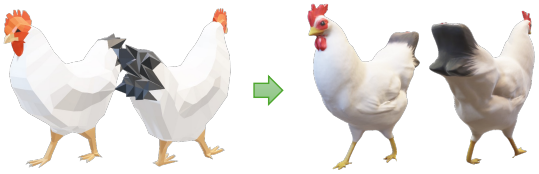


"... A cartoon-style man in black suit, and he wears a cowboy hat."



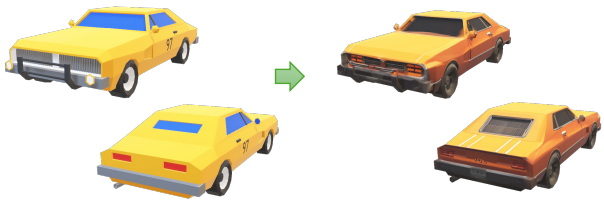
"... A portrait photo of Stalin, USSR art style."

"... A photo realistic squirrel, high-quality, rich details."



"... A chicken."

"... A character from overwatch, McCree, he is in a red cape and holding a gun."



"... 3D rendering of a classic vehicle, in orange color, super sharp texture."

Figure 10. **3D enhancement and editing results with our model.** Notably, we adopt off-the-shelf controlNets [6], e.g, Normal, Canny and Tile, with our Kiss3DGen model to align the generation results with the input 3D models. For simplicity, we denote the fixed camera control caption as "...", and the detailed-object captions are manually crafted to achieve desired results. Please zoom in for details.